

Tesis Doctoral

Generación, análisis e integración de datos moleculares para su aplicación en el control de enfermedades parasitarias: construcción de nuevas plataformas bioinformáticas

Maldonado, Lucas Luciano

2017-12-18

Este documento forma parte de la colección de tesis doctorales y de maestría de la Biblioteca Central Dr. Luis Federico Leloir, disponible en digital.bl.fcen.uba.ar. Su utilización debe ser acompañada por la cita bibliográfica con reconocimiento de la fuente.

This document is part of the doctoral theses collection of the Central Library Dr. Luis Federico Leloir, available in digital.bl.fcen.uba.ar. It should be used accompanied by the corresponding citation acknowledging the source.

Cita tipo APA:

Maldonado, Lucas Luciano. (2017-12-18). Generación, análisis e integración de datos moleculares para su aplicación en el control de enfermedades parasitarias: construcción de nuevas plataformas bioinformáticas. Facultad de Ciencias Exactas y Naturales. Universidad de Buenos Aires.

Cita tipo Chicago:

Maldonado, Lucas Luciano. "Generación, análisis e integración de datos moleculares para su aplicación en el control de enfermedades parasitarias: construcción de nuevas plataformas bioinformáticas". Facultad de Ciencias Exactas y Naturales. Universidad de Buenos Aires. 2017-12-18.

EXACTAS UBA

Facultad de Ciencias Exactas y Naturales



UBA

Universidad de Buenos Aires



Universidad de Buenos Aires
Facultad de Ciencias Exactas y Naturales

**Generación, análisis e integración de datos
moleculares para su aplicación en el control de
enfermedades parasitarias: construcción de
nuevas plataformas bioinformáticas**

Tesis presentada para optar al título de Doctor de la Universidad de Buenos
Aires en el área Ciencias Biológicas

Lic. Lucas Luciano Maldonado

Directora: Laura Kamenetzky

Consejero de estudios: Esteban Hopp

Lugar de trabajo: Instituto de Investigaciones en Microbiología y
Parasitología Médica (IMPaM, UBA-CONICET), Facultad de Medicina,
Universidad de Buenos Aires.

Ciudad Autónoma de Buenos Aires, 2017

Fecha de defensa: 18 de diciembre de 2017

Agradecimientos

- A LauKa por abrirme las puertas del laboratorio y darme la confianza y la libertad para trabajar y por siempre brindarme su apoyo. A Mara también por abrirme las puertas del laboratorio y estar siempre presente, pero principalmente a las dos por su calidad y calidez humana.
- A mis compañeros del laboratorio de BMHid, especialmente a Nati, Fede y Marce con quienes compartí y me brindaron su ayuda desde el primer día. A Gaby L, Hugo, Gaby A, Euge y Mati.
- A los chicos del 13 por su amistad, principalmente a Cris, Agus, Ema, Augusto, Nasha y Andres.
- A Guilherme por su apoyo y recibirme en su laboratorio.
- A Santiago, Yesid, Luis, Marinely, Jorge, Jose Manuel y Vicky.
- A Andre, por acompañarme, por apoyarme y estar siempre a mi lado en la carrera y en la vida.
- A mi familia, mi papá, mi mamá y mi hermano por poder contar siempre con su apoyo y estar siempre conmigo y recibirme siempre con los brazos abiertos.
- Al CONICET y a la ANPCyT por la beca de doctorado y el financiamiento otorgado.
- Al IMPaM y a la Facultad de Medicina de la UBA por brindarme el espacio físico para desarrollar mi trabajo.
- A todos quienes de una u otra manera hicieron esto posible, gracias.

Índice General

Índice General	I
Resumen	V
Abstract	VI
Publicaciones internacionales.....	VII
Abreviaturas	IX

CAPÍTULO I.....1

Introducción	1
1.1 Parásitos del género <i>Echinococcus</i>	1
1.1.1 El ciclo de vida de <i>Echinococcus</i>	1
1.1.2 Características de los estadios de <i>Echinococcus</i>	2
1.1.2.1 Adulto	2
1.1.2.2 Huevos	3
1.1.2.3 El metacestodo.....	4
1.1.2.4 El protoescoléx.....	6
1.2 Clasificación taxonómica de <i>Echinococcus</i>	6
1.3 Echinococcosis	9
1.3.1 Distribución de la echinococcosis	9
1.3.2 Echinococcosis quística en Sudamérica.....	10
1.3.3 Diagnóstico y tratamiento.....	11
1.3.4 Epidemiología molecular de <i>Echinococcus granulosus sensu lato</i>.....	12
1.4 Plelmintos	14
1.5 Características generales de los genomas de platelmintos.....	15
1.5.1 Cariotipo y número de cromosomas	17
1.5.2 Genomas de cestodos y <i>Echinococcus</i>	17
1.6 Genómica comparativa	20
1.6.1 Variantes genéticas: SNPs e Inserciones/Deleciones	21
1.6.2 Tipos de variantes de nucleótido simple	23
1.7 Estudio de características genómicas basados en métodos computacionales	25
1.7.1 Mecanismos de regulación de la expresión génica	25
1.7.1.1 Epigenética y sistema de metilación del ADN.....	26
1.7.1.2 Islas CpG y su posible participación en la regulación génica.....	28
1.7.1.3 Silenciamiento génico postranscripcional.....	28
1.7.1.3.1 Biogénesis y vía de los microARNs.....	29
1.8 Sesgo en el uso de codones.....	31
1.8.1 Relevancia del estudio del uso de codones.....	32
1.8.2 Relevancia del uso de codones en la relación Parásito-Hospedero.....	32
1.8.3 Modelos evolutivos y factores responsables del sesgo en el uso de codones ..	33
1.8.3.1 Modelo de selección y sesgo mutacional	33
1.8.3.2 Modelo de equilibrio mutación-selección-deriva.....	34

CAPÍTULO II.....	36
Hipótesis y Objetivos.....	36
2.1 Hipótesis	36
2.2 Objetivos	36
CAPÍTULO III.....	37
Materiales y Métodos.....	37
3.1 Recolección de muestras y extracción de ADN.....	37
3.1.1 Material parasitario.....	37
3.1.2 Aislamiento y purificación del ADN	38
3.2 Amplificación por PCR y secuenciación	39
3.3 Construcción de bibliotecas de ADN y secuenciación.....	41
3.4 Obtención del genoma de <i>E. canadensis</i> G7	41
3.5 Disponibilidad de los datos	42
3.6 Predicción estructural y anotación de genes	42
3.7 Anotación de genes no codificantes	43
3.8 Análisis de modelos génicos y anotación funcional de genes	44
3.9 Grupos de genes ortólogos.....	44
3.10 Blancos terapéuticos.....	45
3.11 Anotaciones de dominios y análisis estructural	45
3.12 Identificación de islas CGIs.....	47
3.13 Comparación de los genomas	47
3.14 Identificación de SNPs.....	47
3.15 Reconstrucción de la filogenia con genomas mitocondriales	48
3.16 Filogenia global de <i>Echinococcus</i> y especies modelo	49
3.17 Estudio del uso de codones.....	49
3.17.1 Métodos de medición del uso de codones	50
CAPÍTULO IV	54
Resultados.....	54
4.1 Genoma de <i>Echinococcus canadensis</i> G7	54
4.1.1 Secuenciación y ensamblado	54
4.1.1.1 Preparación de bibliotecas de ADN genómico	54
4.1.1.2 Secuenciación y análisis de calidad	55
4.1.1.3 Ensamblado	57
4.1.2 Anotación Génica.....	62
4.1.2.1 Anotación Estructural	62
4.1.2.2 Anotación Funcional.....	63
4.1.3 Conclusiones parciales	67
4.2 Genómica comparativa de tres especies de <i>Echinococcus</i>	68
4.2.1 Ortología y Organización génica de <i>Echinococcus</i>	68
4.2.1.1 Análisis de ortología con otros organismos.....	70
4.2.2 Variantes genéticas entre las especies de <i>Echinococcus</i> : estudio de SNPs.....	72
4.2.2.1 Localización y efecto de las variantes génicas en las regiones codificantes ..	75

4.2.2.2 Distribución de SNPs según vías metabólicas KEGGs	77
4.2.3 Conclusiones parciales	81
4.3 Filogenia de las especies de <i>Echinococcus</i> disponibles actualmente	81
4.3.1 Reconstrucción de la filogenia de <i>Echinococcus</i> basada en genomas mitocondriales	82
4.3.2 Reconstrucción filogenética basada en genes de copia única	83
4.3.3 Reconstrucción filogenética basada en SNPs de genomas completos	84
4.3.4 Validación de SNPs y nuevos marcadores moleculares	85
4.3.4.1 Validación de SNPs	85
4.3.4.2 Nuevos marcadores moleculares para la discriminación entre especies de <i>Echinococcus</i>	87
4.3.5 Conclusiones parciales	88
4.4 Identificación y características particulares del repertorio génico de <i>Echinococcus canadensis</i> G7	89
4.4.1 Genes como potenciales blancos de drogas o de importancia en el desarrollo del parásito	89
4.4.1.1 Antígenos	90
4.4.1.2 Péptidos antimicrobianos	91
4.4.1.3 Hormonas peptídicas	91
4.4.1.4 Transporte	92
4.4.1.5 Metabolismo	93
4.4.1.6 Proceso de transcripción	94
4.4.2 Familias de genes expandidas	95
4.4.3 Conclusiones parciales	95
4.5 Mecanismos de regulación génica en <i>Echinococcus</i>	96
4.5.1 Epigenética y sistema de metilación del ADN	96
4.5.2 Identificación y localización genómica de Islas CpG	99
4.5.3 Maquinaria de microARNs; disección de Proteínas Argonautas - Modelado por Homología	107
4.5.4 Conclusiones parciales	113
4.6 Estudio del uso de codones en <i>Echinococcus</i>	114
4.6.1 Contenido de nucleótidos en los codones en genes de <i>Echinococcus</i>	114
4.6.2 Uso de codones sinónimos relativos en las especies de <i>Echinococcus</i>	117
4.6.3 Sesgo y variabilidad en el uso de codones en <i>Echinococcus</i>	120
4.6.4 Estudio de los fenómenos responsables en el sesgo del uso de codones	121
4.6.4.1 Análisis de Neutralidad	121
4.6.4.2 Evaluación de la regla de paridad 2 de Chargaff (PR2)	122
4.6.4.3 Relación entre ENc y GC3s	123
4.6.4.4 Análisis de los principales factores responsables del sesgo en el uso de codones	124
4.6.4.5 Efecto composicional de los genes sobre el sesgo del uso de codones	130
4.6.4.6 Efecto del nivel de expresión de los genes sobre el sesgo del uso de codones	133
4.6.4.7 Efecto de la hidrofobicidad de las proteínas codificadas sobre el sesgo del uso de codones	135
4.6.4.8 Efecto del tamaño de los genes sobre el sesgo del uso de codones	136
4.6.5 Análisis global e integrado del efecto de los distintos factores sobre el sesgo en el uso de codones en las especies de <i>Echinococcus</i>	138

4.6.5.1 Tipos de genes involucrados en la variabilidad del uso de codones identificados por ACP.....	141
4.6.6 Conclusiones parciales	142
<u>CAPÍTULO V</u>	<u>144</u>
Discusión y Conclusiones	144
5.1 Genoma de <i>Echinococcus canadensis</i> G7	144
5.2 Análisis de ortología y genómica comparativa de las especies de <i>Echinococcus</i>	146
5.3 Relación filogenética entre tres especies de <i>Echinococcus</i>	147
5.4 Blancos terapéuticos.....	148
5.5 Regulación de la expresión génica.....	150
5.6 Uso de codones en <i>Echinococcus</i>	151
5.7 Conclusión general	155
Referencias.....	156

Generación, análisis e integración de datos moleculares para su aplicación en el control de enfermedades parasitarias: construcción de nuevas plataformas bioinformáticas

Resumen

El parásito *Echinococcus canadensis* G7 (filo Platyhelminthes, clase Cestoda) es uno de los agentes causantes de la echinococcosis, zoonosis crónica que afecta tanto a los seres humanos como a mamíferos domésticos y silvestres, y considerada una enfermedad prioritaria por la Organización Mundial de la Salud. A pesar de su gravedad e incidencia en el mundo aún no se dispone de datos genómicos o herramientas terapéuticas y de diagnóstico eficaces para el tratamiento de la infección.

La información presentada en esta tesis permite comprender características biológicas de esta especie y provee información que puede ser utilizada para el diseño de nuevas herramientas de control. En este trabajo se secuenció, ensambló y anotó el genoma de 115 Mb de *E. canadensis* G7. Se identificaron 11435 genes y mediante análisis de ortología se determinaron 881 grupos de genes específicos de cestodos y 581 grupos específicos del género *Echinococcus*. Estos análisis permitieron identificar genes que podrían ser nuevos blancos terapéuticos. Este es el primer trabajo de genómica comparativa entre las especies de *Echinococcus*, y mediante el cual se determinó un alto grado de variabilidad genética entre *E. canadensis* G7 y *E. granulosus* G1, a pesar de que ambas especies pertenecen al complejo *E. granulosus sensu lato* y presentan un fenotipo similar del estadio de metacestodo. Los análisis filogenéticos basados en el análisis de SNPs de genomas completos confirmaron a *E. canadensis* G7 como una especie diferente respecto a *E. granulosus* G1. Asimismo, se identificaron SNPs en genes del metabolismo de *Echinococcus* que podrían tener un impacto en su función. La validación de SNPs permitió desarrollar nuevos marcadores moleculares nucleares para diferenciar las especies de *Echinococcus* que infectan a animales y al hombre en Argentina, y complementan a los marcadores mitocondriales y ribosomales utilizados. Además, se analizaron tres características epigenéticas particulares: la distribución de islas CpG, el sistema de metilación del ADN y componentes de la vía de pequeños ARN basados en el análisis estructural proteína-ARN. Los resultados sugieren que *Echinococcus* posee mecanismos de regulación génica aún desconocidos y diferenciales entre las especies. Por otra parte, el estudio del uso de codones de *Echinococcus* reveló que la principal fuerza evolutiva que moldea su uso es la selección. Si bien se identificaron codones que son utilizados más frecuentemente no se encontraron diferencias entre las tres especies de *Echinococcus*.

Los datos generados en esta tesis contribuyeron a la construcción de la base de datos WormBase ParaSite (<https://parasite.wormbase.org/>), especializada en parásitos platelmintos y nematodos y del Nodo Bioinformático IMPaM (<http://impam2.fmed.uba.ar/>) en el que se implementaron herramientas especializadas para el análisis de cestodos. La disponibilidad de un nuevo genoma es fundamental para entender la biología de un organismo patógeno, y más aún en aquellos con limitaciones para su mantenimiento y manipulación en el laboratorio.

Palabras clave: *Echinococcus*, genoma, SNPs, parásitos helmintos, genómica

Generation, analysis and integration of molecular data for the application in the control of parasitic diseases: construction of new bioinformatics platforms

Abstract

The parasite *Echinococcus canadensis* G7 (phylum Platyhelminthes, class Cestoda) is one of the causative agents of echinococcosis, a chronic zoonosis that affects humans and domestic and wild mammals and is considered a priority disease by the World Health Organization. Despite its severity and incidence worldwide, neither genomic data, nor diagnostic and therapeutic tools for the treatment of infection are yet available.

The information presented in this thesis provides the understanding of biological features of these species and valuable information that can be used for the design of novel control tools of these parasites. In this work, the *E. canadensis* G7 genome of 115 Mb in size was sequenced, assembled and annotated. A total of 11435 genes were identified which were classified into 881 cestode-specific groups of genes and 581 *Echinococcus* genus-specific groups of genes using orthology analysis. These analyses allowed the identification of genes that could be new drug targets. This is the first work performing comparative genomics analysis among the different *Echinococcus* species. The study of genetic variability among the *Echinococcus* species revealed a high degree of variability between *E. canadensis* G7 and *E. granulosus* G1, even though both species have similarities in the phenotype of the metacestode stage and belong to the *E. granulosus sensu lato* complex. Phylogenetic analyses based on whole genome SNPs analysis confirmed *E. canadensis* G7 as a different species from *E. granulosus* G1. SNPs were also identified in *Echinococcus* metabolic pathways genes suggesting that they could have an impact on their function. The SNPs validation allowed the identification of regions that work as new molecular markers to distinguish among the *Echinococcus* species that infect animals and humans in Argentina; and complement the mitochondrial and ribosomal markers that are currently used. Furthermore; three particular epigenetic features were analysed: the distribution of CpG islands, the DNA methylation system and the small RNA pathway components based on structural protein-RNA analysis. The results suggested that *Echinococcus* species have unknown gene regulation mechanisms with some differences among the species. In addition, the study of codons usage bias of *Echinococcus* revealed that the main evolutionary force shaping the codon usage is selection. Although there are codons that are more frequently used than others, no differences among the three *Echinococcus* species were found.

The data reported in this thesis contributed to the construction of the WormBase ParaSite database (<https://parasite.wormbase.org/>), specialized in Platyhelminthes and nematodes parasites and the IMPaM bioinformatics node (<http://impam2.fmed.uba.ar/>) in which specialized tools were implemented for the analysis of cestodes genomes. The availability of a new genome is essential to understand the biology of any pathogenic organism, mainly in those having limitations for the maintenance and manipulation in the laboratory.

Keywords: *Echinococcus*, genome, SNPs, helminths parasites, genomics.

Publicaciones internacionales

1. **Lucas L. Maldonado**, Natalia Macchiaroli, Zarowiecki M, Cucher M, Gismondi MI, Kamenetzky L, Rosenzvit MC. "Genome-wide identification of microRNA targets in the neglected disease pathogens of the genus *Echinococcus*", Mol Biochem Parasitol. 2017 Apr 3. pii: S0166-6851(17)30048-8. doi: 10.1016/j.molbiopara.2017.04.001
2. **Lucas L Maldonado**; Juliana Assis; Flávio Gomes Araújo; Natalia Macchiaroli; Marcela Cucher; Federico Camicia, Adolfo Fox, Mara Rosenzvit; Guilherme Oliveira and Laura Kamenetzky. "The *Echinococcus canadensis* (G7) genome: a key knowledge of parasitic plathyhelminth human diseases", BMC Genomics. 2017 Feb 27;18(1):204. doi: 10.1186/s12864-017-3574-0.
3. Georgina Stegmayer; **Lucas Maldonado**; Natalia Macchiaroli; Cristian Yones; Diego H Milone. MicroRNA discovery in the human parasite *Echinococcus multilocularis* from genome-wide data, Genomics. 2016 Jun;107(6):274-80. doi: 10.1016/j.ygeno.2016.04.002. Epub 2016 Apr 21.
4. Cucher MA, Macchiaroli N, Baldi G, Camicia F, Prada L, **Maldonado L**, Avila HG, Fox A, Gutiérrez A, Negro P, López R, Jensen O, Rosenzvit M, Kamenetzky L. Cystic echinococcosis in South America: systematic review of species and genotypes of *Echinococcus granulosus sensu lato* in humans and natural domestic hosts. Trop Med Int Health. 2015 Nov 26. doi: 10.1111/tmi.12647.
5. Cucher M, Macchiaroli N, Kamenetzky L, **Maldonado L**, Brehm K, Rosenzvit MC. High-throughput characterization of *Echinococcus* spp. Metacestode miRNomes. Int J Parasitol. 2015 ;45 (4) :253-67.
6. Macchiaroli N, Cucher M, Zarowiecki M, **Maldonado L**, Kamenetzky L, Rosenzvit M. microRNA profiling in the zoonotic parasite *Echinococcus canadensis* using a high-throughput approach. Parasit Vectors. 2015 6;8(1):83.
7. Federico Camicia; Ana María Celentano; Malcolm Johns; Chan John; **Lucas Maldonado**; Hugo Vaca; Nicolás Di Siervi; Laura Kamenetzky; Ana Gamo; Silvia Ortega-Gutiérrez; Mar Martin-Fontecha; Carlos Davio; Jonathan S Marchant; Mara Cecilia Rosenzvit. "Unique pharmacological properties of

serotonergic G-protein coupled receptors from cestodes"; PLOS Neglected Tropical Diseases; Manuscrito en revisión.

8. Basika T, Araujo FMd, Salim AC, Pais F, **Lucas L. Maldonado**, Macchiaroli N, Rosenzvit M, Oliveira G, Kamenetzky L, Ferreira HB. "Transcriptomic profile of two developmental stages of the cestode parasite *Mesocostoides corti*"; Genes, Elsevier, Manuscrito en revisión.
9. **Lucas L. Maldonado**, Mara Rosenzvit, Laura Kamenetzky. "Codon usage of *Echinococcus* species and host association studies". Manuscrito en preparación.

Abreviaturas

°C: grados Celsius

µg: microgramos

µl: microlitro

µM: micromolar

3'UTR: región tres prima no traducida, del inglés *three prime untranslated region*

3p: brazo tres prima del precursor de miARN

5p: brazo cinco prima del precursor de miARN

ACP: análisis de componentes principales

ADN: ácido desoxirribonucleico

AED: Distancia de edición de anotación, puntaje de la métrica que determina la confiabilidad de un modelo génico, del inglés “*Annotation Edit Distance*”

Ago: proteína Argonauta

ARN: ácido ribonucleico

ARNm: ARN mensajero

ARNr: ARN ribosómico

ARNt: ARN de transferencia

Arom: aromaticidad

ATAC-seq: ensayo de cromatina transposasa accesible mediante secuenciación

BLAST: herramienta de alineamiento local, del inglés “*Basic Local Alignment Search Tool*”

CAI: índice de adaptación de codones, del inglés “*codon adaptation index*”

CBI: índice del sesgo de codones, del inglés “*codón bias index*”

Cc: componente celular

CDS: secuencia codificante, del inglés *coding DNA sequence*

CGIs: islas ricas en contenido CG

CNVs: variaciones en el número de copias, del inglés “*Copy number variants*”

COX1: citocromo oxidasa subunidad 1

COA: análisis de correspondencia

Contig (contiguos): conjunto de lecturas de secuenciador que se han podido ordenar para formar un segmento continuo de secuencia en base al solapamiento de sus extremos y que juntos representan una región consenso de ADN

CP: componente principal

CpG: par de nucleótidos de C y G adyacentes unidos por enlace fosfodiéster

ChIP-seq: inmunoprecipitación de cromatina seguida de secuenciación

DALY: años de vida ajustados en función de la discapacidad, del inglés “*Disability-Adjusted Life Year*”

DNMT: ADN metil transferasa, del inglés “*DNA methyltransferase*”

DPM: del inglés “*dolicol fosfato manosil transferasa*”

EMBOSS: del inglés “*European Molecular Biology Open Software Suite*”

ENA: del inglés “*europaen nucleotide archive*”

ENc: número efectivo de codones, del inglés “*effective number of codons*”

ER: retículo endoplasmático

ERAD: retículo endoplasmático asociado a la degradación de proteínas, del inglés “*Endoplasmic-reticulum-associated protein degradation*”

EST: marcador de secuencia expresada, del inglés “*Expressed sequence tag*”

e-valor: valor esperado

Fm: Función molecular

Fop: frecuencia óptima de codones

FTP: Protocolo de Transferencia de Archivos, del inglés de “*File Transfer Protocol*”

g: gramos

Gap: (hueco); lugar en la secuencia de ADN donde no se cuenta con información de las bases que hay en esa posición.

GO: ontología génica, del inglés “*gene ontology*”

GPCRs: receptor acoplado a proteína G, del inglés “*G protein coupled receptor*”

GPI: glicosilfosfatidilinositol

Hf: hidrofobicidad

Hsp: proteínas de golpe térmico, del inglés “*Heat shock protein*”

InDel: del inglés “*insertions/deletions*”

JTT: modelo de Jones-Taylor-Thornton

kb: kilobases

log: logaritmo

LTR: del inglés “*Long Terminal Repeat*”

M: molar

Mb: megabases

MBD: del inglés “*Methyl binding domain*”

MCL: del inglés “*Markov Cluster algorithm*”

MeCP2: del inglés “*methyl-CpG binding protein 2*”

Metilseq: secuenciación del ADN metilado

microARN: microARN

min: minutos

miRNA: microARN maduro

miRISC: complejo silenciador inducido por miARN, del inglés “*miARN-induced silencing complex*”

mm: milímetros

NCBI: del inglés “*National Center for Biotechnology Information*”

ng: nanogramos

NGS: Secuenciación de última generación, del inglés “*Next Generation Sequencing*”

nt: nucleótidos

P: probabilidad de error

P1: contenido de GC en la primera posición de codones sinónimos (GC1s)

P12: contenido de GC en la primera y segunda posición de codones sinónimos (GC12s)

P2: contenido de GC en la segunda posición de codones sinónimos (GC2s)

P3: contenido de GC en la tercera posición de codones sinónimos (GC3s)

PAIR: punción, aspiración del contenido, inyección agente escolicida, reaspiración del contenido de un quiste hidatídico

pb: pares de bases

Pb: Proceso biológico

PBS: solución amortiguadora fosfato salino, del inglés “*phosphate-buffered saline*”

PCR: reacción en cadena de la polimerasa, del inglés “*polymerase chain reaction*”

PS: protoescolices

pH: potencial hidrógeno

Phred: escala del valor de calidad de lecturas de secuenciación. Se define como:

Phred(Q) = - 10 log (probabilidad de error).

piARN: ARNs asociados a Piwi

PR2: del inglés “*parity rule 2*”

Pre-miARN: precursor de microARN

Pri-miARN: microARN primario

PTGS: de inglés “*PosTranscriptional Gene Silencing*”

p-valor: valor de probabilidad

qPCR: reacción en cadena de la polimerasa cuantitativa, del inglés “*quantitative polymerase chain reaction*”

RBH: mejor hit recíproco, del inglés “*Reciprocal best hit*”

RNAseq: secuenciación de alto rendimiento del ARN, del inglés “*RNA sequencing*”

RSCU: frecuencia relativa del uso de codones, del inglés “*Relative Synonymous Codon Usage*”

RT: retrotranscripción

SAM: S-adenosil metionina

Scaffold: conjunto de contigs ordenados y orientados en base a información obtenida de lecturas apareadas. Contiene gaps o segmentos de secuencia sin determinar

SINES: elementos nucleares cortos interespaciados, del inglés “*Short interspersed nuclear elements*”

SNPs: polimorfismo de nucleótido simple, del inglés “*single nucleotide polymorphism*”

SNV: variante de nucleótido simple, del inglés “*single nucleotide variant*”

STR: repeticiones cortas en tándem, del inglés “*Short tandem repeats*”

TRD: dominio represor transcripcional

UBC: Ubiquitina C

UHRF: del inglés “*ubiquitin-like, containing PHD and RING finger domain*”

VNTR: número variable de repeticiones en tándem, del inglés “*Variable number tandem repeat*”

WHO: organización mundial de la salud, del inglés “*World Health Organisation*”

INTRODUCCIÓN



Capítulo I

Introducción

1.1 Parásitos del género *Echinococcus*

1.1.1 El ciclo de vida de *Echinococcus*

Los parásitos del género *Echinococcus* pertenecen al Filo Platelminotos, Clase Cestoda, Orden Cyclophyllidea, Familia Taeniidae. El género *Echinococcus* incluye a las especies: *E. granulosus sensu stricto (s.l)*, *E. equinus*, *E. ortleppi*, *E. canadensis*, *E. felidis* (todos incluidos en el complejo de especies: *E. granulosus sensu lato*), *E. multilocularis*, *E. shiquicus*, *E. oligarthrus* y *E. vogeli*.

Como sucede con todos parásitos miembros del Filo platelmintos los parásitos del género *Echinococcus* poseen ciclos de vida complejos que requieren de dos hospederos mamíferos para completar su ciclo [1]. El ciclo de vida consiste en una fase de larva (metacestodo) que se desarrolla en el hospedero intermediario herbívoro, como roedores y ungulados, y accidentalmente humanos y una fase de gusano adulto que habita en el intestino delgado de cánidos y félidos como hospederos definitivos. Durante su estadio adulto las especies del género *Echinococcus* se albergan en el intestino delgado y presentan ciertas particularidades que los diferencian del resto de los miembros de esta familia Taeniidae. El individuo adulto mide entre 2 mm y 7 mm de longitud y usualmente posee 3 o 4 proglótides, mientras que en otras especies el adulto de la misma familia puede alcanzar varios metros de longitud y poseen miles de proglótides [2]. En el interior del intestino delgado del hospedero definitivo el gusano adulto estrobilado se desarrolla sexualmente y produce huevos con oncósferas infectivas que son liberados al ambiente con las heces del hospedero. Los hospederos intermediarios se infectan al ingerir los huevos. Las especies reportadas como hospederos intermediarios incluyen ovinos, caprinos, bovinos, porcinos, camélidos, especies de herbívoros silvestres y al hombre como hospedero accidental [3]. Los huevos ingeridos liberan sus oncósferas por acción de las enzimas digestivas del estómago, permitiendo la migración de las mismas hacia el intestino donde atraviesan la pared intestinal y se diseminan por circulación sistémica a los diferentes órganos, principalmente al hígado y a los pulmones. Una vez alcanzado el órgano el parásito se desarrolla a su estadio larvario de metacestodo generando un quiste hidatídico y produce numerosos protoescolices por reproducción asexual, cada uno con la capacidad de desarrollarse en un individuo adulto si logra infectar a un hospedero definitivo. El

hospedero definitivo se infecta al ingerir vísceras de animales conteniendo quistes con protoescólices. Ocasionalmente los huevos infectivos liberados por el hospedero definitivo pueden ser ingeridos por el hombre convirtiéndose en un hospedero accidental ([Figura I.1](#)).

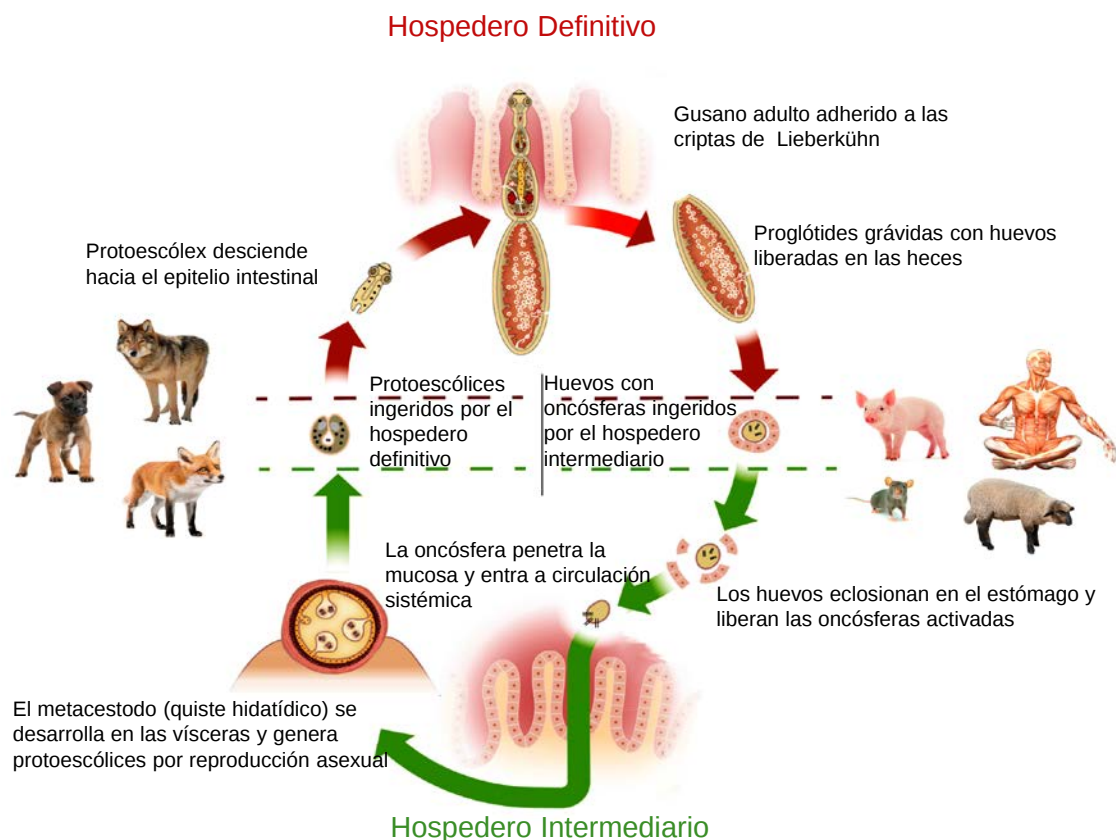


Figura I.1: Ciclo de vida esquemático de *Echinococcus*. Extraído y modificado de Thompson 2017 [1]. El metacestodo corresponde a la especie *E. granulosus* s. l.

1.1.2 Características de los estadios de *Echinococcus*

1.1.2.1 Adulto

El hospedero definitivo adquiere la infección al ingerir protoscolices viables. En el hospedero definitivo, el tiempo requerido para la activación de los protoscolices es muy variable llegando hasta 3 días. Inicialmente deben ser muy activos ya que rápidamente deben localizarse y adherirse a la superficie mucosa de las criptas de Lieberkuhn para evitar ser barridos fuera del intestino delgado [1]. Los gusanos en desarrollo se adhieren principalmente al tejido epitelial con sus ventosas y los ganchos sólo penetran superficialmente en el epitelio mucoso, pero su forma asegura que actúen como anclajes para ayudar a evitar ser expulsados y se mantienen hasta su madurez. La localización y distribución del adulto en el intestino parece ser igual para todas las especies de *Echinococcus*. Las criptas de Lieberkuhn servirían de sitio nutricional para

que el adulto crezca y madure y posteriormente libere proglótides conteniendo huevos fértiles, proceso que demora desde 30 a 45 días dependiendo de la especie. El adulto maduro es hermafrodita y capaz de realizar autofecundación y fecundación cruzada, aunque generalmente se desarrolla por autofecundación. La producción de huevos es un proceso continuo que dura un período de 90 días. El número de huevos producidos es incierto, pero se cree que varían entre 100 y 1500 por proglótide. Sin embargo, no se sabe con exactitud la frecuencia con que las especies de *Echinococcus* producen proglótides grávidas. El esquema de la estructura del gusano adulto puede observarse en la [figura I.2](#)

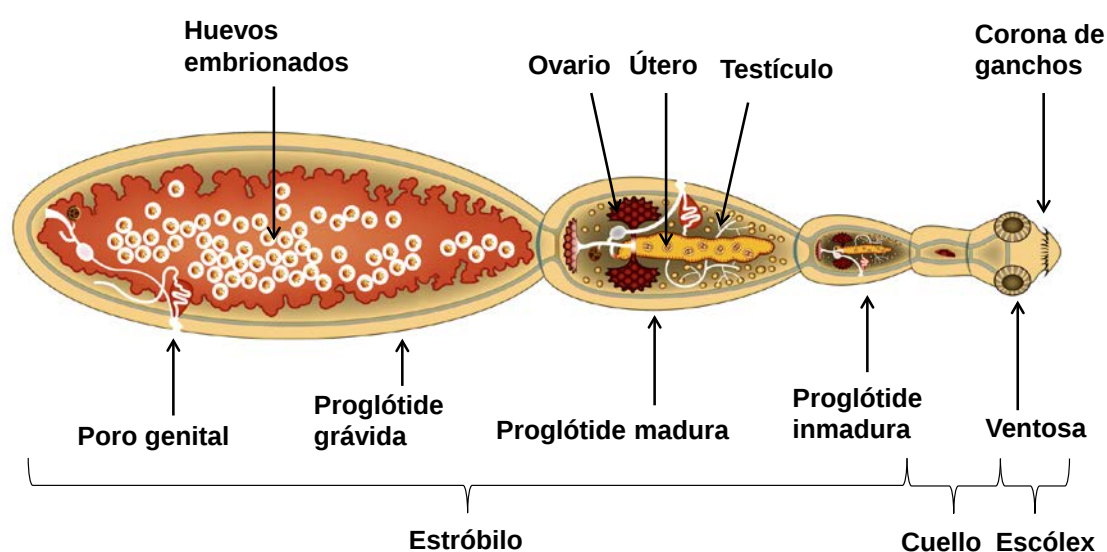


Figura I.2: Esquema de la morfología del estadio adulto de *Echinococcus*. Extraído y adaptado de Thompson 2017 [1].

1.1.2.2 Huevos

Los huevos de *Echinococcus* son de forma esférica a elipsoide y suelen tener un tamaño de 30 a 50 μm y de 22 a 44 μm en sus dos diámetros. Poseen estructuras compuestas por varias capas y membranas que son similares en todas las especies de *Echinococcus* e indistinguibles de los huevos del género *Taenia* haciéndolo imposible de diferenciar por técnicas coproparasitológicas clásicas. Los huevos de *Echinococcus* son extremadamente resistentes, lo que les permite soportar un amplio rango de temperaturas ambientales durante muchos meses. Cuando son liberados del hospedero definitivo, se presume que el huevo de *Echinococcus* está completamente embrionado y es infeccioso para un hospedero intermediario adecuado. El embrióforo es la principal capa que proporciona protección física al embrión u oncósfera ya que la capa vitelina ("cáscara de huevo" o envoltorio exterior) se extrae pasivamente del huevo antes de su liberación. El embrióforo es relativamente grueso e impermeable formado por bloques poligonales compuestos de una proteína inerte parecida a la queratina, que se mantienen unidos por una sustancia adherente [1] ([Figura I.3](#)).

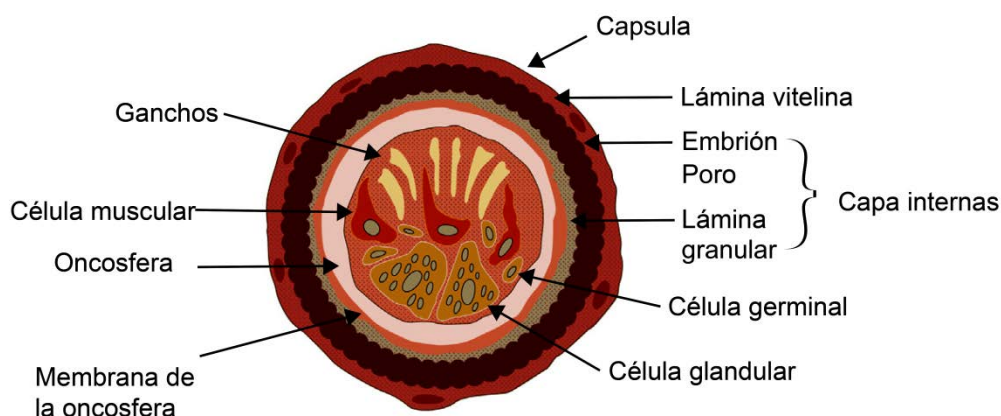


Figura I.3: Diagrama de un huevo de *Echinococcus*. Extraído de Thompson 2017 [1].

1.1.2.3 El metacestodo

Cuando los huevos del parásito son ingeridos por un hospedero intermediario adecuado, los huevos viables de *Echinococcus* eclosionan en el estómago y el intestino delgado. La eclosión es un proceso de dos etapas que involucra: (1) la desagregación pasiva de los bloques embriofóricos en el estómago y en el intestino y (2) la activación de la oncósfera y su liberación de la membrana oncosférica, proceso que es ayudado por la acción de las enzimas digestivas en el estómago de hospedero intermediario. Luego migran a través de las vellosidades epiteliales, alcanzando la lámina propia y penetrando la pared intestinal para luego alojarse en un tejido final donde se dará lugar a la formación del metacestodo [1,3].

La mayor parte del conocimiento actual sobre la biología del desarrollo de *Echinococcus* y la relación parásito-hospedero se ha obtenido a partir de estudios del metacestodo. Esto se debe a la posibilidad de mantener los estadios larvales en cultivo axénico, y la disponibilidad de modelos de roedores para mantener infecciones quísticas después de una infección primaria o secundaria [1]. A su vez, el metacestodo es el estadio de relevancia sanitaria causante de la echinococcosis en animales y humanos. Este estadio larvario de los parásitos cestodos del género *Echinococcus* es el que más diferencias fenotípicas presenta entre sus especies.

De acuerdo al tipo de quiste desarrollado por el metacestodo en el hospedero intermediario, la echinococcosis se puede clasificar en: “unilocular o quística”, causada por parásitos del complejo *E. granulosus* s.l.; “alveolar”, causada por *E. multilocularis* y *E. shiquicus* y “neotropical o poliquística”, causada por *E. oligarthrus* y *E. vogeli*. El metacestodo de *E. granulosus* s.l. consiste en una vesícula llena de fluido en su interior (líquido hidatídico) que está compuesto por una cámara única (unilocular) y su crecimiento se produce de forma concéntrica. Dicha vesícula se encuentra delimitada por una pared compuesta por dos capas: una capa celular interna fina denominada capa germinal y una capa acelular externa de mayor espesor denominada capa laminar. Además la rodea una capa más externa, denominada capa adventicia que se genera como resultado de la respuesta inflamatoria del hospedero inducida por la presencia del parásito. La capa germinal se compone de un tegumento sincicial con un citoplasma

distal del cual se proyectan microtrichias hacia la capa laminar. Las conexiones citoplasmáticas mantienen la continuidad entre las dos capas. En su cara basal que da a la cavidad de la vesícula se encuentran los cuerpos celulares y núcleos que conforman la capa celular, germinal o proliferativa, la cual contiene además otros tipos de células tales como células musculares, células nerviosas, células germinales indiferenciadas y de almacenamiento de glucógeno y lípidos. Las células indiferenciadas de la capa germinal son proliferativas y responsables de la formación de pequeñas masas nucleares, o brotes, que se desarrollan hacia la cavidad quística dando lugar a las vesículas prolíferas en cuyo interior se generan los protoescólices mediante reproducción asexual ([Figura I.4](#)).

Por otra parte, la capa laminar es una matriz extracelular especializada que es secretada por la capa germinal y es una característica propia del género *Echinococcus*. Está compuesta por mucinas que son glicoproteínas con alto nivel de glicosilación y, en el caso de *E. granulosus* s. l. se compone por depósitos de inositol hexaquifosfato de calcio. Su función es mantener la integridad física del metacestodo, mantener su turgencia y protegerlo de la respuesta inmune celular del hospedero [4]. Asimismo, debe permitir la absorción de nutrientes y la eliminación de productos de desecho, ambos requisitos necesarios para el crecimiento del parásito. El metacestodo de *E. multilocularis* es más complejo y se desarrolla de forma muy diferente a la de *E. granulosus* s. l. Se trata de una estructura multivesicular y penetrante, no se encuentra delimitada de la barrera tisular del hospedero y está compuesta por numerosas vesículas pequeñas incrustadas en un estroma denso de tejido conectivo. La masa larvaria generalmente contiene una matriz semisólida en lugar de fluida y la proliferación ocurre tanto endógena como exógenamente por las células indiferenciadas de la capa germinal. Por lo tanto, el metacestodo consiste en una red de protuberancias celulares sólidas y filamentosas de la capa germinal que son responsables del crecimiento invasor tipo tumoral que se transforma en estructuras tubulares y quísticas [1].

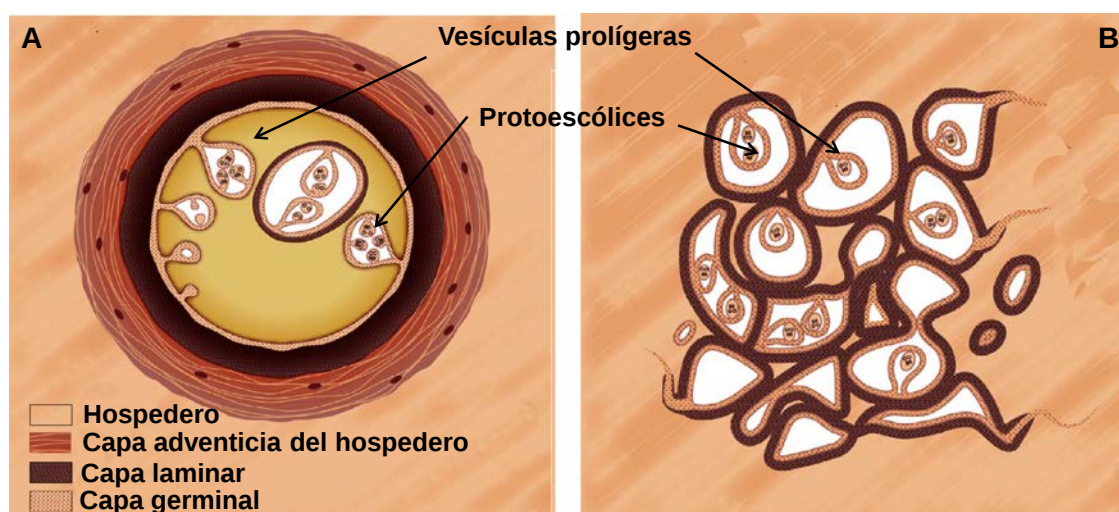


Figura I.4: Esquema comparativo de la estructura de los metacestodos de **A)** *E. granulosus* s.l. **B)** *E. multilocularis*. Extraído y adaptado de Thompson 2017 [1].

1.1.2.4 El protoescólex

Los protoescólices son el estadio larval que antecede al adulto y se originan en el interior de las vesículas prolíferas producidas por la capa germinal del metacestodo. Su proceso de formación es asincrónico por lo que es posible encontrar protoescólices con distinto grado de desarrollo y cada uno de ellos puede dar lugar al gusano adulto. En los protoescólices se distinguen 4 regiones morfológicas: el rostelo con ganchos, el cuello, las ventosas y el cuerpo [5], el cual está recubierto por un tegumento similar al del parásito adulto y al de la capa germinal ([Figura I.5](#)). Los protoescólices poseen además un alto grado de plasticidad fenotípica; no solo tienen la capacidad de convertirse en un gusano adulto en el hospedero definitivo, sino que también pueden diferenciarse en quistes hidatídicos secundarios en caso que se produzca la ruptura y liberación del contenido del quiste dentro del hospedero intermediario [6]. Debido a la dificultad que presenta la recolección de muestras de otros estadios del parásito, el metacestodo representa la mayor fuente de material parasitario en la mayoría de los experimentos, ya sea que provengan de cultivos o de material proveniente de muestras naturales. En esta tesis el estadio de metacestodo es el utilizado como material de partida para todos los ensayos.

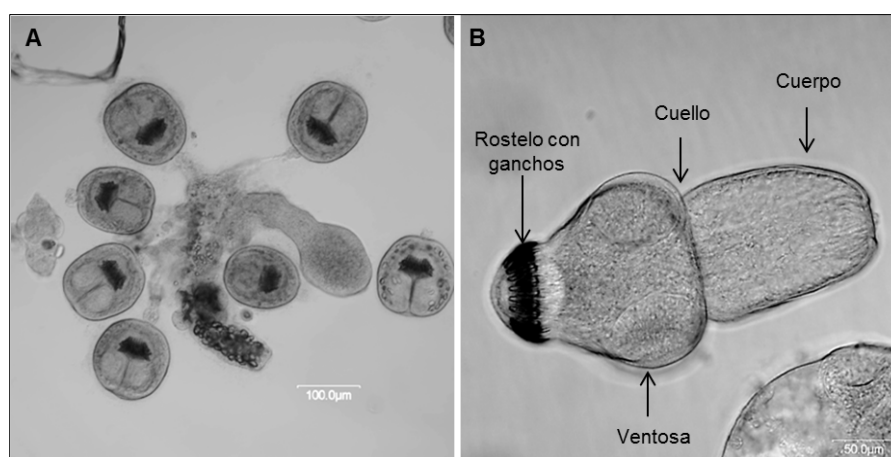


Figura I.5: A) Protoescólices invaginados unidos a la cápsula prolífera. B) Protoescólex evaginado.

1.2 Clasificación taxonómica de *Echinococcus*

Tradicionalmente las especies de *Echinococcus* han sido clasificadas en cuatro grupos basados en la observación de características fenotípicas: *E. granulosus*, *E. multilocularis*, *E. vogeli* y *E. oligarthrus*. Entre las características que diferencian a las especies de *Echinococcus* se encuentran: morfología del metacestodo y del adulto, tamaño de los ganchos, especificidad de hospedero, antigenicidad, perfil proteico, requerimientos metabólicos, desarrollo en el modelo murino y características de relevancia epidemiológica tales como período prepatente (tiempo de demora desde la

ingestión de quistes por parte del hospedero definitivo hasta la liberación de huevos infectivos en las heces) y patogenicidad en humanos entre otras [7].

Debido a las variaciones fenotípicas observadas en distintos aislamientos de *E. granulosus*, obtenidos a partir de especies de hospederos intermediarios diferentes, se definieron cepas o genotipos según su eficiencia en la producción de metacestodos fértiles (conteniendo protoescolices). Inicialmente estas diferencias fueron determinadas por características morfológicas o bioquímicas. Posteriormente, con el advenimiento de las técnicas de biología molecular como por ejemplo la reacción en cadena de la polimerasa (PCR) y la secuenciación del ADN, junto con estudios epidemiológicos moleculares en zonas endémicas, se determinó de forma precisa la clasificación de los diferentes genotipos [1]. Se demostró que algunas especies conforman complejos de especies y algunos genotipos son, en realidad, especies diferentes. Las técnicas basadas en la secuenciación de genes mitocondriales tales como COX1 y NADH, revelaron polimorfismos en genes de las distintas cepas de *E. granulosus* [8,9] y se asignaron diferentes genotipos a las cepas anteriormente definidas: genotipo G1 (cepa oveja), genotipo G2 (cepa oveja de Tasmania), G3 (cepa búfalo), G4 (cepa caballo), G5 (cepa vaca), G6 (cepa camello), G7 (cepa cerdo), G8 (cepa cérvido) y G10 (cepa cérvidos de Escandinavia). En los últimos años, en base a los análisis filogenéticos a partir de genomas mitocondriales completos se determinó que los genotipos de *E. granulosus* no conforman un grupo monofilético [10] sino que son más divergentes de lo que se creía. Actualmente se considera a *E. granulosus* como un complejo de especies denominado *E. granulosus* s. l. Este complejo está formado por diferentes especies de *Echinococcus*: *E. granulosus sensu stricto* (s. s.) que contiene a los genotipos G1/G2/G3, *E. equinus* (genotipo G4), *E. ortleppi* (genotipo G5) y *E. canadensis* que comprende a los genotipos G6/G7/G8/G10 [11]. La cepa león pasó a considerarse una nueva especie a la que se denominó *E. felidis* y se descubrió una más relacionada a *E. multilocularis*, denominada *E. shiquicus* que fue identificada en Tibet, China [12]. El estado actual de la taxonomía de las especies de *Echinococcus* y sus hospederos característicos se resumen en la [Tabla I.1](#).

El estado taxonómico de *Echinococcus* sigue siendo controvertido a pesar de los numerosos intentos de revisar su clasificación. El análisis de las secuencias de ADN del genoma mitocondrial es el utilizado con mayor frecuencia para reconstruir la filogenia del género, sin embargo los resultados derivados de estos análisis han dado lugar a incoherencias significativas respecto a las clasificaciones obtenidas previamente basadas tanto en análisis fenotípicos como en análisis de secuencias nucleares [13–15]. En esta tesis, se aborda por primera vez el estudio filogenético de *Echinococcus* haciendo uso del nuevo genoma obtenido junto con los genomas completos disponibles hasta la fecha.

Tabla I.1: Taxonomía actual del género *Echinococcus*.

Espece	Cepa/Genotipo	Hospedero intermediario *	Hospedero definitivo *	Distribución	Metacestodo	Infección en humanos
<i>Echinococcus granulosus</i> s. s.	Oveja/G1	Oveja, cabra y vaca	Perro	Mundial	Unilocular	Sí
	Oveja de Tasmania/G2					
	Búfalo/G3					
<i>Echinococcus equinus</i>	Caballo/G4	Caballo	Perro	Mundial	Unilocular	Incierto
<i>Echinococcus ortleppi</i>	Vaca/G5	Vaca	Perro	Mundial	Unilocular	Si
<i>Echinococcus canadensis</i>	Camello/G6	Cerdo, camello, vaca, cabra	Perro	Mundial	Unilocular	Sí
	Cerdo/G7					
	Cérvido/G8	Alce y ciervo canadiense	Lobo	Ártico Norte y boreal	Unilocular	Si
	Cérvido de Escandinavia/G10	Alce, reno y ciervo canadiense	Lobo y perro	Ártico Norte y boreal	Unilocular	Si
<i>Echinococcus multilocularis</i>	Variación entre aislamientos	Roedores arvicolinos	Zorro rojo, perro lobo, coyote	Holártica	Alveolar	Sí
<i>Echinococcus vogeli</i>		Paca	Perro grullero	Neotropical	Poliquística	Sí
<i>Echinococcus oligarthra</i>		Agutí	Felinos silvestres	Neotropical	Poliquística	Sí
<i>Echinococcus shiquicus</i>		Pica	Zorro tibetano	Meseta tibetana	Unilocular	Incierto
<i>Echinococcus felidis</i>	León/-	Desconocido	León	Africa	Desconocido	Incierto

* Solo se mencionan algunos hospederos más relevantes. Extraído de Nakao y col, 2013; Thompson & Jenkins 2014 y McManus y col, 2013 [13,16,17].

1.3 Echinococcosis

La echinococcosis es una enfermedad zoonótica ocasionada por las larvas de los cestodos del género *Echinococcus* spp. Los hospederos definitivos son carnívoros domésticos y silvestres portadores de los céstodos adultos y los principales factores de contagio para el ser humano. La echinococcosis quística o unilocular (hidatidosis) causada por *E. granulosus* s.l. y la echinococcosis alveolar causada por *E. multilocularis* son las dos formas más importantes de la enfermedad [18] y es considerada una zoonosis mundial de gran importancia en la salud pública que ha sido reconocida por la WHO (*World Health Organization*) como una enfermedad desatendida prioritaria. Ambas se caracterizan por tener un periodo de incubación asintomático que puede durar muchos años debido al lento desarrollo del metacestodo. Sin embargo, el crecimiento de las larvas da lugar a la formación de quistes en órganos vitales como el hígado y los pulmones que pueden causar la muerte de los individuos afectados.

La echinococcosis quística es la forma más común de la enfermedad en personas y animales domésticos. Dado que se producen quistes únicos independientes, es la forma menos grave y más tratable. No obstante, los quistes grandes o múltiples pueden producir un daño irreversible a los órganos, y la ruptura o perforación del quiste causar reacciones anafilácticas. O bien, si un quiste fértil sufre la ruptura puede producirse la siembra de los protoescólices liberados con la subsiguiente generación de nuevos quistes, conocida como hidatidosis o echinococcosis secundaria. Por otra parte, la echinococcosis alveolar es menos común que la echinococcosis quística, pero es muy grave y difícil de tratar. Las larvas de este organismo crecen como quistes embrionados múltiples que pueden penetrar todos los órganos en forma metastásica y diseminarse hacia otros órganos. Tanto la echinococcosis quística como la alveolar representan una carga de morbilidad considerable con más de 1 millón de personas que viven con esta enfermedad en algún momento de su vida. Muchas de estas personas experimentarán síndromes clínicos severos que serán potencialmente mortales si no se las trata. Incluso con tratamiento, las personas a menudo enfrentan una calidad de vida reducida. Para la echinococcosis quística, la tasa media de mortalidad postoperatoria de los pacientes quirúrgicos es del 2,2% y la de recaída después de una intervención es del 6,5%, requiriendo un tiempo de recuperación prolongado. Se estima que la echinococcosis es la causa de 19300 muertes y de alrededor de 871000 años de vida ajustados en función de la discapacidad (DALY) en todo el mundo cada año. Los costos anuales asociados al tratamiento de los casos de echinococcosis quística y las pérdidas que genera en la industria ganadera se estiman en aproximadamente 3 billones de dólares [19,20].

1.3.1 Distribución de la echinococcosis

La echinococcosis quística está distribuida en todo el mundo y en todos los continentes, con la excepción de la Antártida ([Figura 1.6](#)), mientras que la

echinococcosis alveolar solo se encuentra en el hemisferio norte y en particular en algunas zonas de China, de la Federación Rusa y de los países de Europa continental y América del Norte.

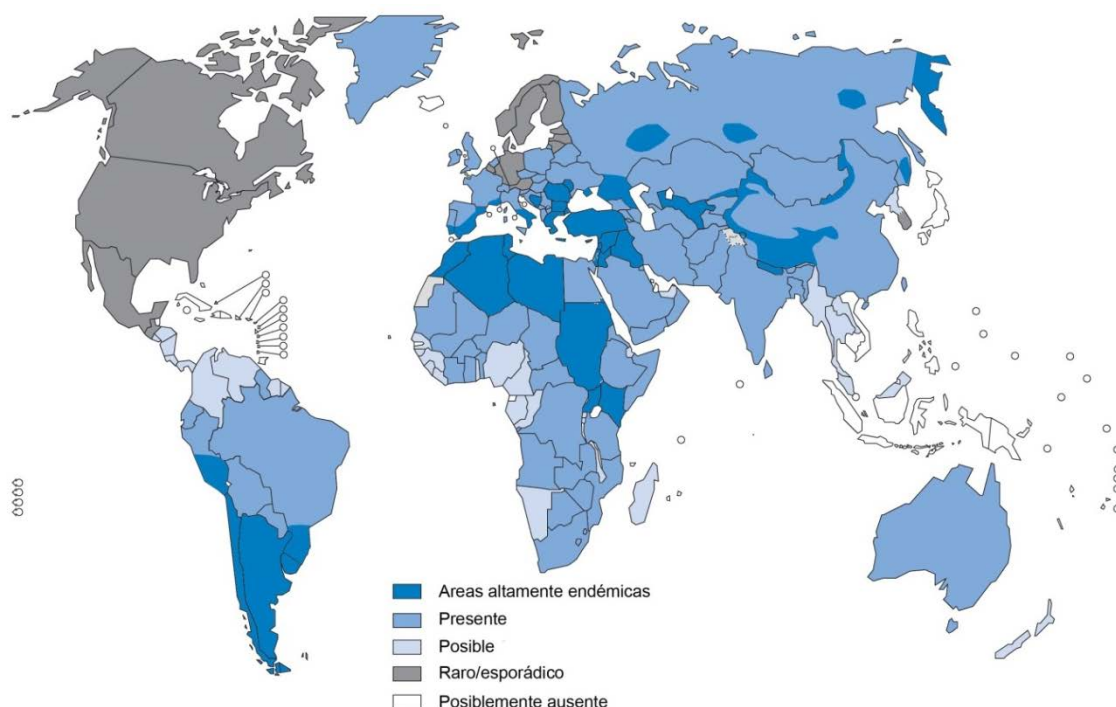


Figura I.6: Distribución global de *E. granulosus* s. l. Figura modificada de la WHO, <http://www.who.int/echinococcosis/en>.

En las regiones endémicas, la tasa de incidencia de la echinococcosis quística en humanos puede alcanzar más de 50 por cada 100000 personas al año. En algunas regiones de la Argentina, Perú, África oriental, Asia central y China los niveles de prevalencia pueden llegar a valores de entre el 5% y 10%. En el ganado, la prevalencia de echinococcosis quística encontrada en mataderos en zonas hiperendémicas de Sudamérica es de entre el 20% y el 95% de los animales sacrificados. La mayor prevalencia se encuentra en las zonas rurales donde se sacrifican los animales de mayor edad. Dependiendo de la especie del animal infectado, las pérdidas de producción ganadera atribuibles a la echinococcosis quística se basan en la declaración del hígado como no apto para el consumo, la reducción del peso, la disminución del valor de la piel, la disminución de la producción de leche y la reducción de la fertilidad (WHO, 2017).

1.3.2 Echinococcosis quística en Sudamérica

La echinococcosis quística es una enfermedad endémica o hiperendémica en Argentina, sur de Brasil, Uruguay, Chile y regiones montañosas de Perú. La distribución de la echinococcosis es heterogénea dentro de los países y se concentra en las áreas de producción ganadera, y zonas rurales y frecuentemente asociada a factores sociales tales como comunidades pobres, alta ruralidad, pueblos originarios y

economías de subsistencia entre otros. En el periodo Enero 2009-Diciembre 2014, los cinco países reportaron 29556 casos humanos. De ellos Perú aparece como el país con el mayor número de casos humanos (20785 casos), aun cuando en dicho país no es obligatoria la notificación de la enfermedad. Asimismo, se estima que la notificación de casos está considerablemente subestimada en toda la región. Se calcula que la echinococcosis causó la muerte de más de 820 personas en el periodo 2009-2014 lo que resulta en una letalidad media de 2,9% y fue responsable de más de 300000 días de hospitalización teniendo en cuenta los datos de Argentina, Brasil, Chile y Uruguay. En la Argentina la echinococcosis quística es una enfermedad de notificación obligatoria y está presente en todas las provincias presentando una elevada tasa de transmisión principalmente en tres regiones: la región patagónica (provincias de Tierra del Fuego, Neuquén, Santa Cruz, Chubut y Río Negro), la provincia de Buenos Aires y la región del litoral al este del país (Corrientes y Entre Ríos). Los registros oficiales han informado que la echinococcosis es la zoonosis de mayor prevalencia y la que mayor número de muertes genera cada año (<http://www.paho.org/arg/>).

1.3.3 Diagnóstico y tratamiento

Hasta el momento no existe un método de diagnóstico suficientemente eficaz para la detección de la enfermedad que pueda emplearse ante la infección producida por todas las especies de *Echinococcus*. La ecografía es la técnica de imagen de elección para el diagnóstico de la echinococcosis quística y la echinococcosis alveolar, sin embargo muchas veces los quistes o metacestodos son descubiertos por casualidad mediante radiografía y son detectables solo si han alcanzado un tamaño considerable. Hay diferentes pruebas serológicas que detectan anticuerpos específicos contra antígenos del parásito, como por ejemplo Antígeno B [21] y pueden complementar el diagnóstico por imágenes; sin embargo las limitaciones de estas pruebas utilizadas para diagnosticar la echinococcosis han contribuido a la subestimación de casos humanos. La elección del tratamiento se basa principalmente en las imágenes ecográficas del quiste, dependiendo de la fase específica en que se encuentre, de la ubicación, el tamaño del quiste, así como de la infraestructura médica y los recursos humanos disponibles. Entre las opciones para el tratamiento de la echinococcosis quística se encuentran: el tratamiento percutáneo con la técnica PAIR (punción, aspiración del contenido, inyección agente esclerizante, reaspiración del contenido), la cirugía, el tratamiento farmacológico con benzimidazoles (albendazol y mebendazol) y la conducta expectante [21]. En el caso de la echinococcosis alveolar, el tratamiento es la cirugía radical (similar a la extirpación de tumores), seguida de profilaxis anti infecciosa con albendazol [22]. En muchos pacientes la enfermedad se diagnostica en una fase avanzada y debido al desarrollo metastásico que tiene esta especie presenta altas tasas de mortalidad. Por lo tanto es muy importante el diagnóstico temprano de la infección.

1.3.4 Epidemiología molecular de *Echinococcus granulosus sensu lato*

El genotipo de mayor relevancia epidemiológica a nivel mundial es *E. granulosus* G1 seguido por *E. canadensis* G6 y G7 ([Figura I.7](#)).

La situación epidemiológica molecular en América del Sur, recientemente revisada por nuestro grupo de investigación [7] reveló que en toda la región circulan seis genotipos pertenecientes al complejo *E. granulosus* s.l. (G1, G2, G3, G5, G6 y G7), los cuales residen en diferentes hospederos intermediarios tales como ovejas, vacas, cerdos, cabras y alpacas, siendo los perros hospederos definitivos de todos los genotipos descriptos y demostrando que el ciclo completo de cada genotipo se encuentra en la Argentina. Al igual que lo que sucede a nivel mundial, la especie con mayor prevalencia y relevancia epidemiológica en Argentina es *E. granulosus* G1. Esto podría deberse a la gran variedad de hospederos intermediarios en los que puede desarrollarse y/o a la mayor infectividad en humanos. La segunda especie de mayor importancia es *E. canadensis* G6/G7 [7].

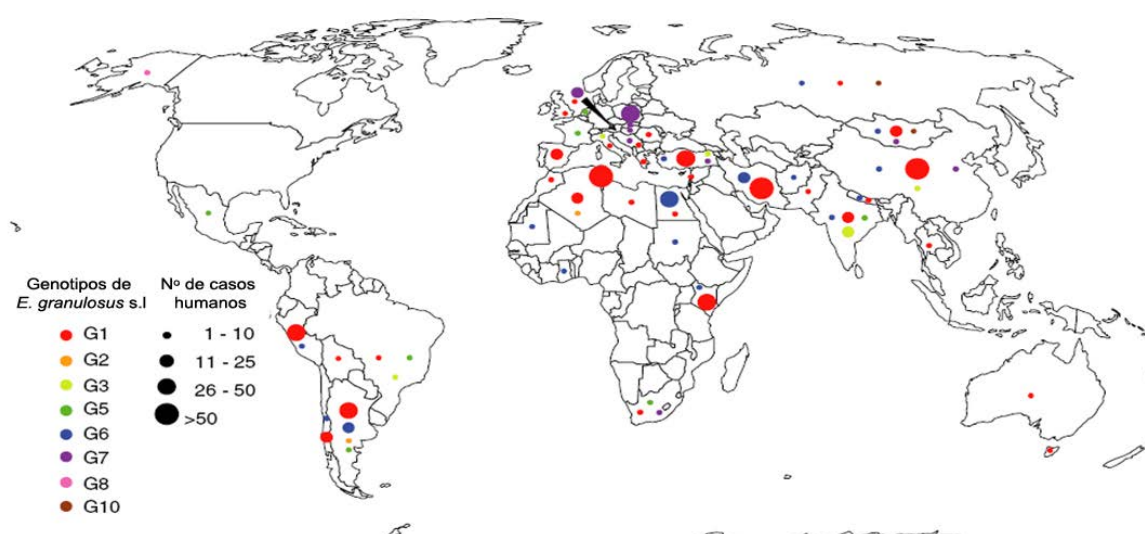


Figura I.7: Distribución mundial de casos humanos correspondientes a genotipos de *Echinococcus granulosus sensu lato*. Extraído y modificado de Cucher y col 2016 [7].

Estudios llevados a cabo en nuestro laboratorio han revelado que en Argentina, circulan un total de seis genotipos: *E. granulosus* G1 en ovejas, cabras, vacas y cerdos; *E. granulosus* G2 en ovejas y vacas; *E. granulosus* G3 en ovejas; *E. ortleppi* G5 en vacas; *E. canadensis* G6 en cabras y en vacas; *E. canadensis* G7 en cerdos y *E. oligarthrus* en agutíes [7,23–25]. Más aún, en relación a las infecciones en humanos, se han descripto hasta el momento casos correspondientes a *E. granulosus* (G1 y G2); *E. ortleppi* (G5) y *E. canadensis* (G6 y G7) [7,26] ([Figura I.8](#)).

aborda el estudio de las especies de *Echinococcus* cuyos genomas se encuentran disponibles y el de otros platelmintos, basado en técnicas de genómica comparativa.

1.4 Platelminotos

Los platelmintos son animales pequeños multicelulares con cuerpos alargados que tienen extremos anterior y posterior claramente definidos y son bilateralmente simétricos. Por lo general tienen una cabeza reconocible, que alberga los órganos receptivos y carecen de sistema circulatorio y respiratorio. A veces presentan una cavidad digestiva, con una sola abertura hacia el exterior que sirve de boca y de ano. En las formas más simples está ausente o no ramificado, pero en las formas superiores se ramifica a todas las partes del cuerpo. Los órganos de los sentidos principales, cuando están presentes, se concentran en la cabeza, o parte delantera. Aunque algunas de las formas más simples poseen una red nerviosa primitiva, otras tienen varias cuerdas nerviosas que se extienden desde un cerebro a lo largo del cuerpo. Este último patrón de organización se mantiene en el sistema nervioso de los invertebrados superiores, específicamente anélidos y artrópodos. Sus cuerpos aplanados se componen de tres capas: el ectodermo, el endodermo y el mesodermo. Pueden estar cubiertos por una cutícula protectora o por pelos microscópicos, llamados cilios [37]. Todos los platelmintos, excepto los más simples, tienen los túbulos nefridiales (protonefridios), distribuidos generalmente a través del cuerpo. Dichas estructuras consisten en una abertura externa y un túbulo que se ramifica internamente, terminando en una serie de estructuras ciegas, en forma de bulbo llamadas “bulbos de llama o células flamígeras”, que probablemente funcionan como órganos excretores y osmorreguladores [38]. El sistema reproductor de los platelmintos es característicamente hermafrodita (es decir, cada individuo produce óvulos y esperma), y pueden reproducirse tanto por autofecundación como por fecundación cruzada. Existen alrededor de unas 29000 especies descritas en el filo “Platelminotos” [39] los cuales comprenden especies de vida libre y parasitarias. El filo se encuentra formado por dos clados monofiléticos, Rhabditophora y Catenulida. El clado Rhabditophora agrupa a todas las especies de platelmintos menos a las 100 especies que pertenecen al clado Catenulida [37,40]. La mayoría de las especies de platelmintos son parásitos obligados y la mayoría de las especies ([Figura I.9](#)) se agrupan en un clado monofilético conocido como Neodermata que incluye tres clases: Trematoda, Monogenea y Cestoda. Muchos de estos parásitos planos tienen ciclos de vida complejos que implican múltiples hospederos. A diferencia de los otros grupos parasitarios, los monogeneos son parásitos externos que infectan a los animales acuáticos y sus larvas sufren una metamorfosis hacia la forma adulta después de unirse a un hospedador adecuado. Por otra parte el clado Turbellaria es parafilético e incluye a los miembros tanto de Catenulida como a algunos de los miembros de Rhabditophora. Los linajes de Turbellaria incluyen especies de vida libre, simbióticas y algunas de ellas son parasitarias [37].

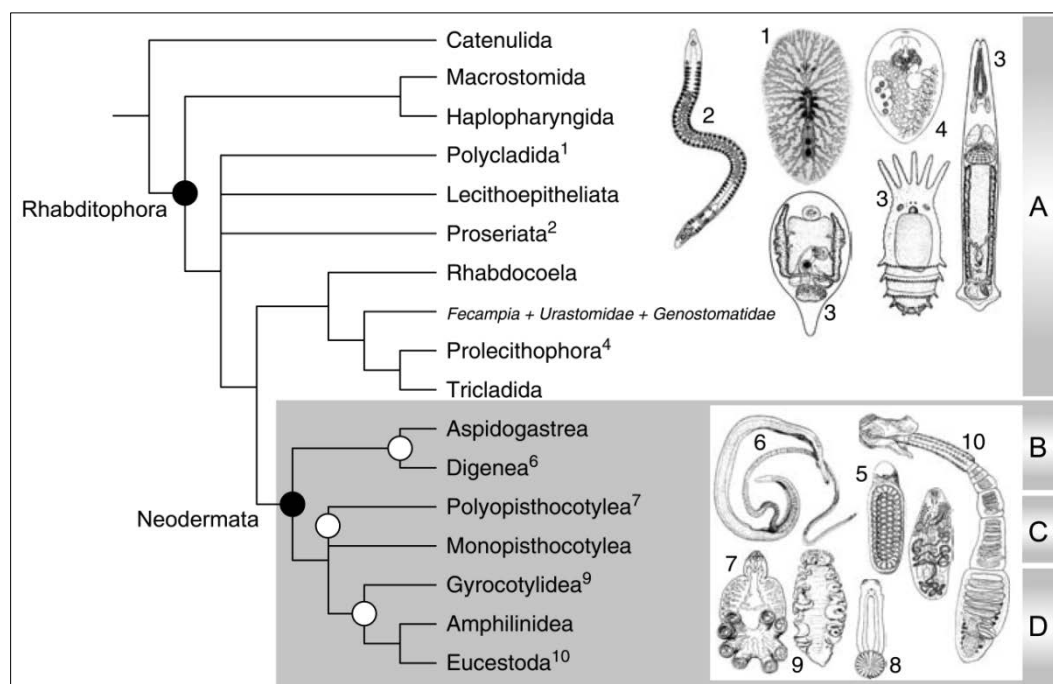


Figura 1.9. Relaciones filogenéticas entre los principales grupos de platelmintos parásitos. A) Turbellaria, B) Trematoda, C) Monogenea, D) Cestoda. Extraído y modificado de Maule [37].

1.5 Características generales de los genomas de platelmintos

Durante los últimos 15 años se han realizado enormes avances en materia de secuenciación del ADN, y en las tecnologías de computación científica para el análisis de datos genómicos y transcriptómicos. Las tecnologías de última generación (NGS) tales como 454 Roche, Ion Torrent e Illumina son las más usadas para la generación de este tipo de datos, siendo esta última la que posee una aplicación más amplia: secuenciación de genomas completos o secuenciación exómica; aplicaciones en epigenómica, como ChIP-seq (inmunoprecipitación de cromatina seguida de secuenciación), ATAC-seq (ensayo de cromatina transposasa accesible mediante secuenciación) o secuenciación del ADN metilado (metilseq) y transcriptómica [41]. Durante el mismo período, el costo de adquirir y explotar estas tecnologías también ha disminuido significativamente y en consecuencia ha favorecido la generación de gran cantidad de información genómica para muchos organismos en poco tiempo. Los parásitos platelmintos están entre los sistemas experimentales menos manipulables con los cuales trabajar en comparación con los organismos modelo tradicionales como *Drosophila* o *Caenorhabditis*. Su ciclo de vida completo puede ser difícil o imposible de mantener en el laboratorio, ya sea *in vitro* o *in vivo*, ya que poseen ciclos de desarrollo complejos y prolongados que no logran llegar a los estadios de desarrollo completos (madurez sexual) y producen cantidades limitadas de material biológico. En raras ocasiones se logran establecer como poblaciones clonales o cepas definidas con algunos estadios bien caracterizados en cuanto a su morfología, bioquímica,

comportamiento y desarrollo. Algunos ejemplos de ellos incluyen a los cultivos axénicos del metacestodo de *E. multilocularis* que ha podido mantenerse en cultivos primarios o los cultivos *in vitro* de *T. crassiceps*. Es por ello que la secuenciación de los genomas de estos organismos es fundamental para poder abordar su estudio. Actualmente, los proyectos más avanzados involucran a *Schistosoma* spp [40,42] y recientemente a *E. multilocularis* [43] y sobre ellos se han basado la mayoría de los estudios que describen sus características. Hasta hace 10 años atrás solo algunas características genómicas de platelmintos habían sido descritas basadas en técnicas clásicas como estudios de cariotipo y bandeo cromosómico y observaciones macroscópicas de la estructura del ADN o por secuenciación de unas pocas regiones del genoma de algunos organismos [37,44]. El tamaño del genoma se ha calculado para una variedad de especies de platelmintos. Para Turbellaria, la medición directa del contenido de ADN de los núcleos teñidos con Feulgen por densitometría o análisis de imágenes han revelado una enorme variación en el tamaño estimado del genoma haploide, desde 50 Mb en *Stenostomum arevaloi* hasta 18390 Mb en *Otomesostoma auditivum* [37]. También se ha observado una variación en el tamaño del genoma dentro de ciertos géneros 'turbelarios' (por ejemplo, *Dugesia*, *Phaenocora* y *Mesostoma*). Para Trematoda y Cestoda, se han utilizado análisis de reasociación cinética del ADN (C_{0t}) y medición directa del contenido de ADN por tinción de núcleos para estimar el tamaño del genoma observándose tamaños que van desde 150 hasta 1600 Mb para cestodos (*Echinococcus granulosus*, *Hymenolepis diminuta* y *Spirometra mansonioides*) y desde 270 hasta 1120 Mb para trematodos (*Diplostomum pseudospathaceum*, *Schistosoma mansoni* y *Trichobilharzia* spp) [37]. La composición nucleotídica de los genomas de platelmintos es variable según la especie. Más aún, entre los miembros de una misma clase el porcentaje del contenido GC es diferente para las distintas especies. Del mismo modo se ha observado que el porcentaje de GC es diferente según se trate de regiones codificantes, intergénicas o intrónicas. En las regiones codificantes de *S. mansoni* se han observado un sesgo hacia un mayor contenido de A y T en la última base de los codones más frecuentes [45,46]. Por el contrario en especies de *Echinococcus* se ha descrito un sesgo hacia altos contenido de GC en la última base de los codones de genes altamente expresados [47]. Los estudios de la composición genómica global basada en análisis de cinética de reasociación [48–50] y el análisis de secuencias de clones de cromosomas bacterianos artificiales (BAC) del proyecto del genoma de *S. mansoni* han demostrado que los genomas de platelmintos contienen secuencias altamente y moderadamente repetidas [37]. Gran parte del ADN repetido comprende dos clases de elementos móviles integrados; Clase I, que incluyen retrotransposones y retrovirus de repetición terminal larga (LTR), retrotransposones no LTR y elementos nucleares intercalados cortos (SINES) y que se transponen a través de un intermedio de ARN, y elementos de clase II (transposones) que se transponen como ADN [51]. Hasta el momento no se conoce exactamente la función de estos elementos móviles, sin embargo prácticamente todos los elementos móviles integrados de platelmintos descritos hasta la fecha parecen transcribirse hasta cierto punto. En *S. mansoni*, aproximadamente el 14% de los transcritos de las cercarias codifican para secuencias

que poseen alta similitud con la transcriptasa reversa. También se han observado elementos de bajo número de copias con actividad transcripcional alta y viceversa [52].

1.5.1 Cariotipo y número de cromosomas

Los cromosomas de los platelmintos son cromosomas típicos de los metazoos que se condensan durante la mitosis y su tamaño genómico es tal que los cromosomas son visibles bajo microscopio óptico. Esto ha permitido el análisis del cariotipo en una variedad muy amplia de especies parasitarias y de vida libre. En general, los platelmintos parecen poseer un complemento cromosómico relativamente pequeño, desde $n = 2$ para *Polycelis auriculata* [53] hasta $n = 28$ en *Clonorchis sinensis* [37], con $n < 10$ como una regla general [54] y un $n = 9$ que es más común entre los cestodos [55]. Los platelmintos son generalmente diploides ($2n$) a lo largo de su ciclo de vida a excepción de las gametas en los casos en que exista reproducción sexual. Sin embargo, se ha informado de una amplia gama de complementos cromosómicos "no estándar", tales como taxones poliploides, cromosomas súper numerarios y aneuploidías. La triploidía se ha observado en varias especies en distintos miembros de trematodos (*Fasciola* spp y *Paragonimus westermani*), turbelarios (*Dugesia* spp, *Polycelis nigra* y *Crenobia alpina*) y cestodos (*Spirometra erinaceieuropaei*) [37]. Las poliploidías y cromosomas supernumerarios han sido informadas para *P. nigra* [56,57] y el estado poliploide más extremo que se ha descrito hasta la fecha se encuentra en *Phagocata ullala* de la clase Turbellaria con una serie poliploide de $4n = 68$, $8n = 136$, $10n = 170$ y $14n = 238$ [58,59]. Las aneuploidías se han descrito en *Macrostomum* spp [60]. La pérdida de cromosomas también se ha descrito en cestodos, particularmente en *Taenia crassiceps* ($2n = 16$) cuya cepa denominada "ORF" presenta un cariotipo de $2n = 14$ por pérdida de las dos copias del cromosoma 2 en su estadio de metacestodo, preservando la viabilidad pero perdiendo su capacidad infectiva [37].

1.5.2 Genomas de cestodos y *Echinococcus*

Hasta 2014, año en el cual se comenzó el proyecto de secuenciación del genoma de *E. canadensis* G7, solo se contaban con 28 genomas de platelmintos secuenciados, entre los cuales 14 pertenecían a la clase Cestoda con 2 miembros del género *Echinococcus* (Tabla I.2). La secuenciación de estos genomas forma parte del proyecto "The 50 Helminth Genomes Initiative" que tiene como objetivo el estudio de los genomas de nematodos y platelmintos que son de importancia médica y/o veterinaria, incluyendo a especies que son usadas como modelos de estos parásitos. A pesar de su importancia zoonótica, su impacto en la salud humana y su mayor disponibilidad de secuencias genómicas, las características importantes de los genomas de platelmintos están lejos de haber sido elucidados en su totalidad. Hasta la fecha sólo hay dos genomas de platelmintos que han sido secuenciados y ensamblados en sus cromosomas dando lugar a genomas relativamente completos, los genomas de *S. mansoni* y *E. multilocularis* [40,43].

Tabla I.2: Genomas disponibles de platelmintos generados por diferentes consorcios y disponibles actualmente en la base de datos WormBase Parasite (parasite.wormbase.org).

Especies	Proveedor	BioProject ID	Clado	N50 (kb)	Tamaño de genoma (Mb)	Nº de genes	%GC
<i>Diphyllobothrium latum</i>	Wellcome Trust Sanger Institute	PRJEB1206	Cestoda	6,7	531,4	19966	43,2
<i>Echinococcus granulosus</i>	Wellcome Trust Sanger Institute	PRJEB121	Cestoda	5228,7	114,5	10245	41,9
<i>Echinococcus granulosus</i>	Chinese National Human Genome Center	PRJNA182977	Cestoda	682,6	110,8	11319	41,8
<i>Echinococcus multilocularis</i>	Wellcome Trust Sanger Institute	PRJEB122	Cestoda	13762,5	115,0	10663	42,2
<i>Hydatigera taeniaeformis</i>	Wellcome Trust Sanger Institute	PRJEB534	Cestoda	12,4	103,7	11649	43,0
<i>Hymenolepis diminuta</i>	Wellcome Trust Sanger Institute	PRJEB507	Cestoda	49,8	165,9	11271	35,2
<i>Hymenolepis microstoma</i>	Wellcome Trust Sanger Institute	PRJEB124	Cestoda	7673,8	182,1	12368	36,0
<i>Hymenolepis nana</i>	Wellcome Trust SangerInstitute	PRJEB508	Cestoda	19,2	162,9	13777	36,7
<i>Mesocetoides corti</i>	Wellcome Trust Sanger Institute	PRJEB510	Cestoda	65,8	117,2	10614	43,0
<i>Schistocephalus solidus</i>	Wellcome Trust Sanger Institute	PRJEB527	Cestoda	31,5	539,4	20228	43,0
<i>Spirometra erinaceieuropaei</i>	Wellcome Trust Sanger Institute	PRJEB1202	Cestoda	4,6	1258,7	39557	44,9
<i>Taenia asiatica</i>	Wellcome Trust Sanger Institute	PRJEB532	Cestoda	51,9	136,0	10331	42,9
<i>Taenia asiatica</i>	Chinese Academy of Agricultural Sciences	PRJNA299871	Cestoda	34,2	168,7	13322	43,1
<i>Taenia solium</i>	National University of Mexico	PRJNA170813	Cestoda	67,8	122,4	12481	37,7
<i>Protopolystoma xenopodis</i>	Wellcome Trust Sanger Institute	PRJEB1201	Monogenea	2,9	617,3	37906	37,7

<i>Schmidtea mediterranea</i>	Genome Institute at Washington University	PRJNA12585	Rhabditophora	4,1	901,6	29850	29,9
<i>Clonorchis sinensis</i>	Sun Yat-sen University	PRJDA72781	Trematoda	415,8	547,3	13634	44,0
<i>Echinostoma caproni</i>	Wellcome Trust Sanger Institute	PRJEB1207	Trematoda	26,9	834,6	18607	42,5
<i>Fasciola hepatica</i>	CGR University of Liverpool	PRJEB6687	Trematoda	203,7	1275,1	22676	44,1
<i>Fasciola hepatica</i>	Genome Institute at Washington University	PRJNA179522	Trematoda	16,0	1139,2	15739	44,1
<i>Schistosoma curassoni</i>	Wellcome Trust Sanger Institute	PRJEB519	Trematoda	13,8	344,2	23546	34,2
<i>Schistosoma haematobium</i>	University of Melbourne	PRJNA78265	Trematoda	304,8	385,1	13073	34,3
<i>Schistosoma japonicum</i>	Chinese National Human Genome Center	PRJEA34885	Trematoda	174,8	402,7	12738	34,1
<i>Schistosoma mansoni</i>	Wellcome Trust Sanger Institute	PRJEA36577	Trematoda	32115,4	364,5	10772	35,2
<i>Schistosoma margrebowiei</i>	Wellcome Trust Sanger Institute	PRJEB522	Trematoda	35,2	367,4	26189	34,3
<i>Schistosoma mattheei</i>	Wellcome Trust Sanger Institute	PRJEB523	Trematoda	12,3	340,8	22997	34,1
<i>Schistosoma rodhaini</i>	Wellcome Trust Sanger Institute	PRJEB526	Trematoda	18,6	343,3	24089	34,4
<i>Trichobilharzia regenti</i>	Wellcome Trust Sanger Institute	PRJEB4662	Trematoda	7,7	701,8	22185	37,4

La obtención del genoma de *E. multilocularis* fue realizada mediante una combinación de tecnologías de secuenciación basadas en 454, Illumina y secuenciación capilar generando fragmentos de hasta 4 kb con lecturas de 100 pb utilizándose un ensamblado “*de novo*” y el ensamblado final del genoma se terminó manualmente mediante el clonado de fragmentos de hasta 80 kb (40-80 kb) de ADN genómico en fósmidos e integrando el análisis con mapeo óptico. De esta forma se logró obtener el 89% de la secuencia genómica contenida en 9 cromosomas enumerados de 1-9 de mayor a menor según su tamaño. Uno de sus cromosomas está completo de telómero a telómero (cromosoma 5), y 13 de los 18 telómeros esperados están presentes en los cromosomas reconstruidos. Por otra parte, el genoma de *E. multilocularis* reveló tetraploidía en protoescolices en cultivo, y una trisomía del cromosoma 9 (el cromosoma más pequeño y posiblemente el único para el que se tolera una trisomía) tanto en protoescolices como en el metacestodo en cultivo. Además mediante la secuenciación de transcriptomas provenientes de ARN total obtenido en diferentes etapas de su ciclo de vida se identificaron 10345 genes [43].

1.6 Genómica comparativa

La genómica comparativa es un campo de investigación biológica en el que se comparan las características genómicas de diferentes organismos. Dichas características genómicas pueden ser la propia secuencia de ADN, los genes, el orden genético en el genoma, las secuencias regulatorias u otros marcadores genéticos. En este contexto regiones genómicas o la totalidad del genoma son comparados para estudiar similitudes o diferencias biológicas básicas entre los genomas en estudio y las relaciones evolutivas entre los organismos [61]. El mayor principio en la genómica comparativa radica en que las características que son comunes a los organismos son generalmente codificadas por regiones de ADN que están evolutivamente conservadas entre ellos [62]. La genómica comparativa ha revelado altos niveles de similitud entre organismos estrechamente relacionados, tales como humanos y chimpancés o incluso entre organismos aparentemente muy distantes como humanos y levaduras y ha permitido demostrar la existencia de diversidad en la composición genética en diferentes linajes celulares. Una de las técnicas más usadas en muchos proyectos genomas es la búsqueda de ortólogos y construcción de grupos de ortología dado que es fundamental para muchos objetivos, tales como la anotación de genomas y comparación de vías metabólicas entre diferentes organismos ya que se basan en que dichos genes ortólogos compartirían las mismas características funcionales en diferentes especies [63,64].

Con el aumento del número de proyectos genomas debido a los grandes avances en materia de tecnología de secuenciación, particularmente métodos de NGS, el campo del estudio genómico ha mejorado haciendo posible la manipulación de muchos genomas a la vez para la realización de un estudio en particular [62]. El gran paso de la era genómica ha sido el primer genoma secuenciado de un organismo autorreplicante, el de la bacteria *Haemophilus influenzae* [65]. Desde entonces, las secuencias de miles de genomas se han puesto a disposición de la comunidad científica. El genoma de

Saccharomyces cerevisiae fue el primer eucariota en publicarse en el año 1996 [66] y desde ese momento, las especies elegidas para los proyectos de secuenciación se seleccionaban basándose principalmente en su tamaño del genoma, su papel como organismo modelo y su relevancia para los seres humanos [67]. La liberación del primer genoma completo de una Archaea, *Methanococcus jannaschii* [68], representó un gran avance en el campo, ya que brindó la primera oportunidad de explorar los tres dominios de la vida a nivel de genomas completos [66]. El siguiente gran paso de la era de la genómica fue, sin duda, el anuncio del proyecto genoma humano en 2001 financiado con fondos públicos [69] y privados [70]. Desde entonces, los investigadores comenzaron a comparar sistemáticamente genomas de organismos modelo con *Homo sapiens* para obtener información sobre la evolución y las enfermedades humanas, entre otras características [71–74]. Es por ello, que gracias a la generación de un alto número de datos de secuencias disponibles, el campo de la genómica comparativa ha crecido rápidamente y se ha convertido en una parte importante en la investigación biológica. Un estudio pionero fue sin duda el análisis de genómica comparativa de *Drosophila melanogaster*, *Caenorhabditis elegans* y *Saccharomyces cerevisiae* [75]. Por primera vez se obtuvieron las secuencias genómicas completas de tres organismos eucariotas modelo y pudieron ser analizadas de manera comparativa. Se determinó que el tamaño del proteoma de la mosca es sólo el doble que el de la levadura unicelular, siendo la mosca un metazoo con mayor complejidad y que, a su vez, comparte el tamaño de varias familias de proteínas con el nematode, a pesar de las diferencias en cuanto a su desarrollo y morfología. Este trabajo fundacional develó, por lo tanto, que la complejidad de un organismo no correlaciona con un aumento en el número total de genes [75].

1.6.1 Variantes genéticas: SNPs e Inserciones/Deleciones

La variación genética ocurre debido a cambios en la secuencia de bases de ADN. Dicha variabilidad genética juega un rol muy importante en la supervivencia y adaptación de una especie a su entorno. Cuando un hábitat se modifica, los organismos de la población que viven en él deben adaptarse al cambio del medio ambiente. Las variaciones en el acervo génico de una población de determinada especie le otorgan características fenotípicas que le permiten adaptarse a los cambios en el medio ambiente mediante el fenómeno de selección natural. Así la selección natural lleva al cambio evolutivo cuando individuos con ciertas características genéticamente heredables, poseen una tasa de supervivencia o reproducción más alta que otros individuos de la población que no portan dichas características. Cuando esto sucede, las características que otorgan mayor supervivencia son más probables de ser transmitidas a la siguiente generación.

Inicialmente, previo al desarrollo de gran cantidad de técnicas de secuenciación, las diferencias que se observaban en la composición genética de los organismos estaban asociadas con cambios en la cantidad y estructura de los cromosomas y se limitaban a cambios detectables mediante microscopía tales como aneuploidias, rearreglos,

heteromorfismos y sitios frágiles generalmente asociados con una enfermedad [76]. Posteriormente los avances en las técnicas de biología molecular y una mayor disponibilidad de las tecnologías de secuenciación permitieron observar en el ADN pequeñas alteraciones no detectadas anteriormente y más abundantes como: SNPs, secuencias cortas de ADN altamente repetitivas (micro y mini satélites) y secuencias menores a 1 kb involucradas en inserciones, deleciones, inversiones y duplicaciones. Se estima que el conjunto de estas variaciones constituyen la mayor proporción de variabilidad genética. En la población humana hay por lo menos 10 millones de SNPs, promediando 1 cada 300 nucleótidos entre los ~3 mil millones de pares de bases de nucleótidos que constituyen el genoma de un individuo [76].

Existen diferentes tipos de variantes genéticas que básicamente pueden agruparse en dos categorías: las que no afectan el contenido del ADN debido a que el número de nucleótidos permanece inalterado y las que causan cambios en el contenido nucleotídico por ganancia o pérdida de bases de ADN. La primer categoría involucra a variantes o polimorfismos de nucleótido simple (SNV o SNP) en las que una base en una posición dada cambia por otra o debido a la sustitución de un pequeño número de ellas (denominadas en ocasiones mutaciones puntuales). Esta categoría también agrupa a translocaciones de uno o de un número pequeño de nucleótidos a otra región de un mismo cromosoma y si la región que sufre translocación es de un tamaño considerable este es catalogado dentro de un subgrupo denominado variaciones estructurales. Esta categoría donde no hay cambio en la composición nucleotídica también incluye a las inversiones, otra clase de variación estructural en el cual un segmento de ADN invierte su orientación en un cromosoma de principio a fin. [77]

Las variaciones estructurales son cambios en la estructura de los cromosomas de un organismo. Son rearrreglos cromosómicos que comprenden secuencias generalmente mayores a 50 pb y suelen clasificarse en dos grupos (balanceadas y no balanceadas) dependiendo si hay o no cambio en la composición nucleotídica del ADN. En las variaciones estructurales balanceadas las variantes de ADN tienen el mismo contenido nucleotídico pero algunas secuencias de ADN están localizadas en posiciones diferentes en el genoma pero sin pérdida o ganancia de bases. Este grupo incluye a las ya mencionadas translocaciones e inversiones [76,77]. En la variación estructural no balanceada las variantes de ADN difieren en su contenido nucleotídico por ganancia o por pérdida de ADN. En ella se encuentran comprendidas las inserciones y las deleciones (InDels) y las variaciones en número de copia (CNVs). Los InDels son inserciones o pérdidas de un segmento de ADN a nivel submicroscópico (<1 kb); aunque, la convención moderna reserva el término InDel para describir las deleciones o inserciones cuyos tamaños varían entre 1 y 50 nucleótidos (elegidos arbitrariamente ya que los métodos de secuenciación masiva de ADN producen secuencias muy cortas y a menudo no es posible detectar deleciones o inserciones de más de 50 nucleótidos). Las CNVs, a diferencia de las anteriores, involucran segmentos mayores a 1 kb (o mayores a 100 pb en la convención moderna) pero que la mayoría de las veces son del orden de Mb y pueden presentarse en un número variable de copias [76]. Otras variaciones que aportan a la denominada variación estructural son las copias en tándem o repeticiones múltiples en tándem (STR) de regiones de ADN que son denominadas según su tamaño

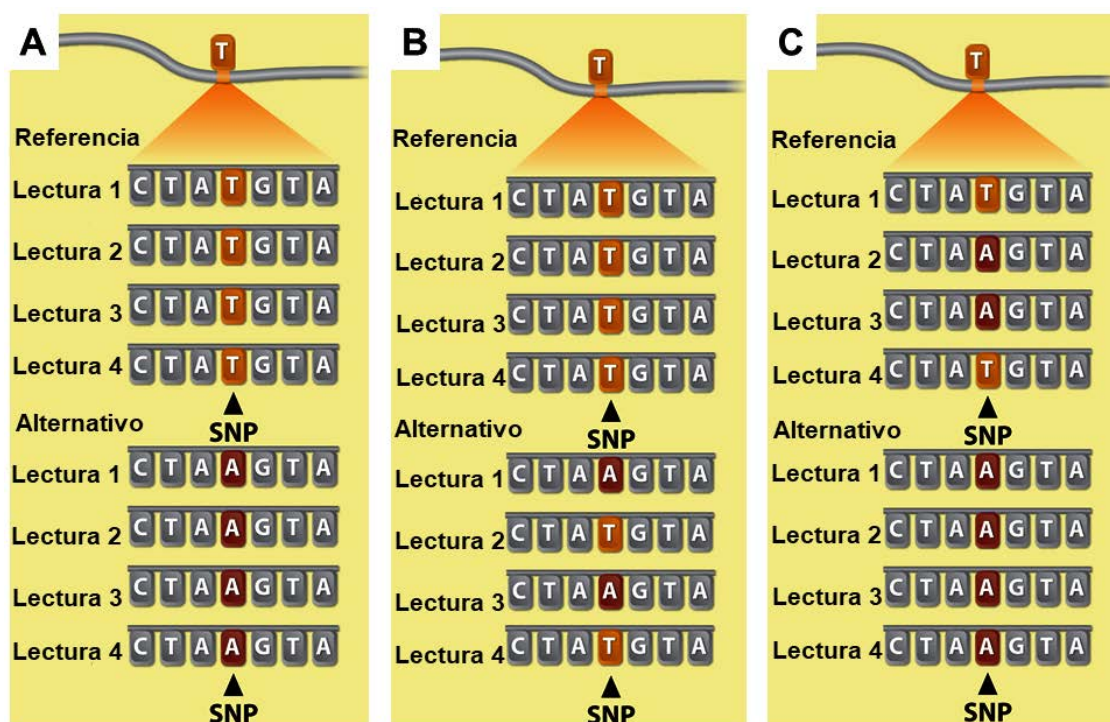
en satélite, minisatélite o microsatélite de ADN. La inestabilidad del ADN conteniendo STRs genera variantes de ADN que difieren en el número de repeticiones en tándem denominadas polimorfismos de número variable de repeticiones en tándem (VNTRs) [77].

1.6.2 Tipos de variantes de nucleótido simple

Cuando los SNPs o SNVs están presentes en regiones génicas, estos pueden clasificarse según el efecto que genera la sustitución del nucleótido en la proteína para la cual codifica dicho gen. Los SNPs que se localizan dentro de una secuencia codificante pueden modificar o no la secuencia de aminoácidos. Cuando una sustitución de nucleótido genera un cambio de aminoácido en la secuencia de esa proteína se denomina sustitución o SNP no sinónimo. Si por el contrario, la sustitución nucleotídica no se traduce en un cambio de aminoácido se denomina sustitución sinónima. La presencia de un SNP en la región codificante también puede provocar la aparición de un codón de terminación prematuro que resulta en una proteína de menor tamaño a la original, en ese caso se lo denomina SNP sin sentido. Los SNPs que se encuentren en regiones no codificantes también pueden tener efectos en la traducción, particularmente si se localizan en regiones genómicas de unión de factores de transcripción, en regiones donde se transcriben los ARNs no codificantes o en las regiones intrónicas en los sitios de *splicing* de los ARNm, alterando las señales normales del procesamiento de los ARNm y resultando en un *splicing* anormal [77]. En cualquier caso, los SNPs que alteren de algún modo la expresión génica reciben el nombre de SNPs de expresión y también pueden encontrarse tanto río arriba como río abajo de la secuencia codificante.

Otra clasificación utilizada en esta tesis está relacionada con la heterocigocis del sitio polimórfico. A diferencia de los términos usados en genética clásica donde las palabras homocigota, heterocigota y hemicigota son usados para describir el genotipo de un organismo diploide para un sitio en particular en el ADN, en la comparación de los genomas de dos organismos a nivel de su secuencia nucleotídica realizada mediante el mapeo de lecturas contra un genoma de referencia en búsqueda de SNPs, estos términos son utilizados de forma diferente y su uso se describe a continuación. En este caso si las lecturas provienen de un organismo diploide, se podría suponer que las mismas provienen estocásticamente de cualquiera de los dos alelos, por lo tanto el porcentaje de lecturas que exhiban una sustitución de nucleótido en un sitio en particular con respecto a la referencia determina si ese polimorfismo es homocigota o heterocigota para ese sitio. De este modo se define que los SNPs son homocigotas si para un sitio o *locus* en particular los dos cromosomas del genoma en estudio poseen el mismo alelo y este es diferente al del genoma de referencia. Es decir que si todas las lecturas que mapean contra la referencia poseen la misma sustitución de nucleótido en un determinado sitio y diferente al que posee el alelo de la referencia, entonces se lo denomina SNP homocigota ([Figura I.10 A](#)).

Asimismo, si para un sitio en particular uno de los alelos es igual al alelo de la referencia y el otro alelo es diferente se dice que los SNPs son heterocigotas. Es decir que si un cierto porcentaje de las lecturas que mapean contra la referencia exhiben una sustitución de nucleótido en un determinado sitio y el porcentaje restante de lecturas para ese mismo sitio presenta el mismo nucleótido que la referencia (sin cambio) entonces esos sitios se clasifican como SNPs heterocigotas ([Figura I.10 B](#)). El mismo principio puede utilizarse al mapear las lecturas del genoma de referencia contra sí mismo y de esta forma identificarse los SNPs homocigotas y heterocigotas tanto para la especie que se usa de referencia como para la especie con el alelo alternativo ([Figura I.10 C, D, E y F](#)). La posibilidad de encontrar sitios heterocigotas trialelicos es menos frecuente y en este último caso ninguno de los dos alelos alternativos son iguales al alelo de la referencia y a su vez son diferentes entre sí, mientras que el sitio de la referencia puede ser homocigota o heterocigota para ese sitio o no tener polimorfismo. ([Figura I.10 E y F](#)). En la práctica utilizar solamente los porcentajes de las lecturas mapeadas para la identificación de sitios polimórficos no es suficiente, por lo cual se utilizan además otros criterios con el fin de evitar errores en la detección de SNPs. Los criterios generales a tener en cuenta son los siguientes: frecuencia de variación $> 40\%$ con al menos 20 lecturas que cubran los sitios polimórficos, una calidad de base (Q) tanto para el sitio de referencia como para el sitio polimórfico $Q > 30$ en la escala Phred y localizarse a una distancia no menor a 5 pb de un *gap* o InDel. Adicionalmente, para la clasificación de SNPs homocigotas o heterocigotas, al menos el 40% de las lecturas deben estar apoyando dicho cambio y en proporciones similares de la hebra directa y la reversa (Véase capítulo materiales métodos, sección 3.14 para mayor detalle).



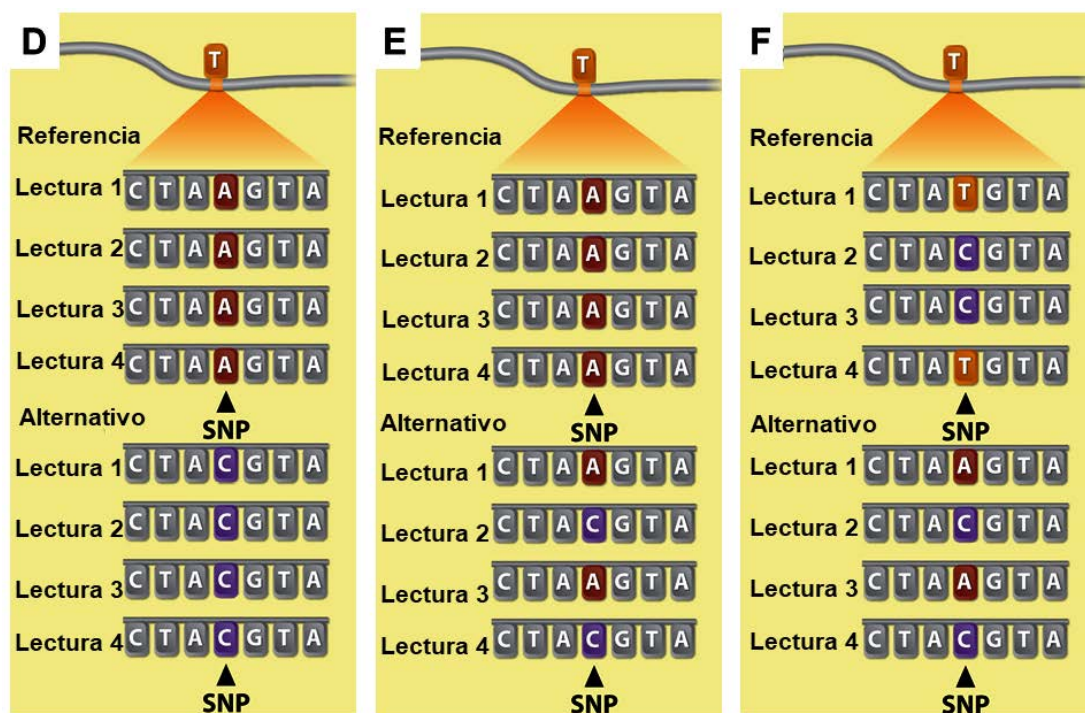


Figura I.10: Sitios polimórficos homocigotas y heterocigotas según la secuencia de referencia y la alternativa. A). SNPs homocigotas: sitio no polimórfico para la referencia (sin SNP) y polimórfico homocigotas para el alternativo. B). SNPs heterocigotas: sitio no polimórfico para la referencia (sin SNP) y polimórfico heterocigotas para el alternativo. C). Sitio polimórfico heterocigota para la referencia y polimórfico homocigota para el alternativo. D). Sitio polimórfico homocigota para la referencia y polimórfico homocigota para el alternativo. E). Sitio polimórfico homocigota para la referencia y polimórfico heterocigota para el alternativo. F). Sitio polimórfico heterocigota para la referencia y polimórfico heterocigota para el alternativo.

1.7 Estudio de características genómicas basados en métodos computacionales

1.7.1 Mecanismos de regulación de la expresión génica

Los platelmintos parásitos son un grupo diverso de metazoos adaptados a vivir en diversos ambientes. Sus ciclos de vida complejos requieren contar con los niveles y repertorios apropiados de ARNm y proteínas que les permitan desarrollarse, sobrevivir y adaptarse a sus hospederos, como así también responder de manera rápida a los cambios ambientales a los que se enfrentan [43,78–81]. Los niveles de expresión de los diferentes ARNm y proteínas observados en cada estadio del ciclo de vida de *Echinococcus* se encuentran modulados por diferentes mecanismos de regulación de la expresión génica como por ejemplo *trans-splicing* y metilación del ADN [82,83] o bien mediante modificaciones post-traduccionales de histonas como se ha encontrado en otros platelmintos parásitos [84]. Otro mecanismo de regulación de la expresión génica, que ha sido descrito por primera vez en los parásitos del género *Echinococcus* por nuestro grupo es el de los pequeños ARNs no codificantes denominados microARNs.

Sin embargo, el conocimiento sobre los mecanismos de regulación de la expresión en *Echinococcus*, y en particular el que involucra a los microARNs aún es limitado [85]. Asimismo, las adaptaciones morfológicas y metabólicas de *E. multilocularis* y *E. granulosus* G1 reveladas por el análisis de los genomas recientemente secuenciados [43,78] junto a otras características biológicas observadas, tales como la especificidad de hospedero, el hecho de que estén adaptados a sobrevivir por largo tiempo en sus hospedero y sus particularidades del desarrollo como la plasticidad fenotípica y estrobilización del parásito adulto ponen de manifiesto la importancia de profundizar en el estudio de los mecanismos moleculares subyacentes con el objetivo de avanzar en el conocimiento de la compleja biología que poseen estos parásitos.

En esta tesis se aborda el estudio de la regulación de la expresión génica mediante la identificación de posibles regiones promotoras o regulatorias, por marcaje epigenético y regulación post-transcripcional mediada por microARNs por lo cual a continuación se describen algunas características conocidas sobre los factores recién mencionados.

1.7.1.1 Epigenética y sistema de metilación del ADN

No fue sino hasta 1980 que se demostró que la metilación del ADN estaba involucrada en la regulación génica y la diferenciación celular. La metilación del ADN es uno de los principales factores epigenéticos de control sobre la actividad de los genes. La metilación de ADN es catalizada por una familia de ADN metiltransferasas (DNMTs) que transfieren un grupo metilo desde S-adenosil metionina (SAM) al carbono 5 de una citosina. La familia de DNMTs está formada por las DNMT1, DNMT2, DNMT3L, DNMT3a y DNMT3b que comparten similitud estructural con un dominio extenso N-terminal y un dominio C-terminal con actividad catalítica, sin embargo poseen diferentes funciones y patrones de expresión. Se cree que la DNMT3a y la DNMT3b son las metiltransferasas *de novo* que pueden establecer un nuevo patrón de metilación en un ADN no modificado y que actúan en la etapa de desarrollo temprano. DNMT3L es una proteína que es homóloga a los demás DNMT3s pero no tiene ninguna actividad catalítica. En su lugar, DNMT3L se asocia a las metiltransferasas *de novo* provocando un aumento de su capacidad de unión al ADN y así estimular su actividad catalítica [86]. Por otra parte, la DNMT1 se encuentra activa durante la replicación del ADN y copia el patrón de metilación desde la hebra de ADN parental a la nueva hebra sintetizada [86,87]. Finalmente, DNMT2 ha sido identificada como una proteína homóloga a la ADN metiltransferasa, que contiene los 10 motivos de secuencia comunes a todas las metiltransferasas de ADN. Sin embargo, hasta el momento no se ha descrito metilación de ADN vía DNMT2 sino que se la describe involucrada en la metilación de la citosina-38 en el bucle anti codón del ARN de transferencia del ácido aspártico [88]. Las grandes distancias genéticas entre los miembros de la familia DNMT2 probablemente se atribuyen a sus presuntas especificidades duales de sustrato (ADN y ARN) [89]. El mayor porcentaje de la metilación del ADN ocurre en las citosinas que preceden a una guanina denominados sitios CpG. Los genomas de mamíferos contienen un gran número de sitios CpG que

son propensos a sufrir una reacción de desaminación y por consiguiente producir la mutación de la citosina del CpG a una timina. Los sitios CpG están altamente distribuidos en todo el genoma y se encuentran potencialmente metilados, a excepción de las islas CpG (CGIs) (ver en la siguiente sección). [90].

Un segundo tipo de proteína que interviene en este mecanismo de regulación de la expresión génica son las proteínas de unión a ADN metilado que impiden la unión de los factores de transcripción al ADN. El ADN metilado es reconocido por proteínas de tres familias diferentes: proteínas MBD (*methyl binding domain proteins*), proteínas UHRF (*ubiquitin-like, containing PHD and RING finger domain*) y proteínas con dedos de zinc. Las proteínas MBD poseen un dominio MBD que les confiere gran afinidad por los sitios CpG metilados [91]. Esta familia incluye a las MeCP2 (*methyl-CpG binding protein 2*), MBD1, MBD2, MBD3 y MBD4 [92–94]. La proteína MBD3 es incapaz de unirse al ADN debido a una mutación en su dominio MBD, en cambio la proteína MBD4 se une normalmente al ADN y reconoce preferentemente un apareamiento incorrecto entre guanina y timina y forma parte del sistema de reparación del ADN. Los otros miembros de las MBDs se unen directamente al ADN metilado mediante su dominio MBD, y poseen un dominio extra TRD (*transcriptional repression domain*) que permite la asociación a varios complejos represores de la transcripción [92,95,96].

La metilación del ADN es una característica epigenética que se encuentra conservada en las cuatro clases de platelmintos (Cestoda, Monogenea, Trematoda y Turbellaria), habiéndose observado particularmente la metilación de citosinas del ADN genómico en individuos adultos de las especies de *S. mansoni*, *Protopolystoma xenopodis*, *S. haematobium*, *S. japonicum*, *P. nigra* y en protoescolices de *E. multilocularis*. Asimismo, se ha identificado que estos organismos poseen un gen que codifica para la proteína DNMT2, pero carecen de genes codificantes para DNMT1 o DNMT3 [83,97]. Estudios de expresión y silenciamiento de las DNMT2 y de cuantificación e inhibición de la metilación del ADN realizados en *S. mansoni* han revelado que las marcas epigenéticas en su ADN son directamente dependientes del grado de metilación de las citosinas del genoma y de las proteínas DNMT2 enzimáticamente activas [83,97]. Estos estudios mostraron que un menor grado de metilación inhibe la producción de huevos *in vitro*, sugiriendo que la modificación epigenética en estas especies de gusanos, y particularmente el grado de metilación de la citosina está ligada al proceso de ovoposición de los platelmintos, maduración de los huevos y desarrollo del sistema de órganos sexuales [97]. Además de las DNMTs, también se han observado en platelmintos proteínas MBDs que aparentemente serían miembros de la familia MBD2/3 [83,98]. Las MBDs de platelmintos exhiben regiones conservadas de los dominios de unión a CpG los cuales consisten en 70 aminoácidos en el extremo N-terminal (PF01429) y 97 aminoácidos del extremo carboxilo terminal (PF14048). La conservación de estos dos dominios, la ausencia de los motivos dedos de zinc característicos (CxxCZn) y la ausencia de los dominios de reparación de ADN glicosilasas sitúa a estas MBDs en la familia ancestral MBD2/3 [98–100]. En las especies de platelmintos estudiadas hasta el momento, solo se encontró un gen que codifica para MDB. Esta proteína se ha descrito en trematodos (*S. japonicum*, *F. hepatica*, *P. westermani*, *C. sinensis* y *O. viverrini*), turbelarios

(*Schmidtea mediterranea* y *M. lignano*) y cestodos (*E. multilocularis*, *E. granulosus*, *T. solium* y *H. microstoma*) y pareciera estar ausente en la clase Monogenea [83].

1.7.1.2 Islas CpG y su posible participación en la regulación génica

Las islas CpG (CGIs) son segmentos de ADN de 1000 pb o más que contienen una mayor densidad de CpG que el resto del genoma, con un contenido de GC mayor al 55% y se encuentran no metilados [86]. Los dinucleótidos CpG de regiones ricas en GC pueden tener un rol activo en la regulación de la expresión génica ya que la mayoría de los promotores de genes, aproximadamente el 70 %, están embebidos en regiones que contienen CGIs [101]. También los elementos potenciadores (*enhancers*) pueden localizarse entre las CGIs o estar contenidos en ellas. De este modo las CGIs pueden poseer ambas propiedades, las características promotoras y el efecto potenciador para la expresión génica [102]. Algunos estudios han demostrado que las CGIs pueden tener un rol en la regulación de la expresión génica tejido-específico y que pueden sufrir metilación principalmente en los estadios embrionarios en regiones intragénicas y en las regiones llamadas “*gene body*” (región del gen más allá el primer exón). La metilación del primer exón, así como la metilación de la región promotora, conduce al silenciamiento de los genes [103]. Las CGIs asociadas a promotores se encuentran frecuentemente sin metilar, y por lo general son encontradas río arriba de genes constitutivos [104–108]. Las CGIs, particularmente aquellas que están asociadas con promotores, están altamente conservadas entre ratones y humanos. La localización y conservación de CGIs a través de la evolución implicaría una importancia funcional de dichas regiones. Aparentemente, las CGIs se han conservado evolutivamente debido a que promueven la expresión génica mediante la regulación de la estructura de la cromatina y favorecen la unión de factores de transcripción. Una de las características de las CGIs es que contienen menos nucleosomas que otras regiones de ADN, haciéndolas más permisivas para la unión de los factores de transcripción. Más aún, los pocos nucleosomas que se encuentran unidos a las CGIs usualmente contienen histonas modificadas que favorecen la potenciación de la expresión génica [109–111].

1.7.1.3 Silenciamiento génico postranscripcional

El silenciamiento génico es un mecanismo que está presente en la mayoría de los organismos eucariotas y que tiene como función controlar la expresión de los genes ya sean endógenos o exógenos. Este mecanismo desempeña un papel decisivo en la diferenciación celular, en la regulación del desarrollo y en la defensa contra genes foráneos y elementos transponibles [112]. La señal para la activación del mecanismo la constituyen moléculas de ARN (mayormente moléculas de doble cadena). Tras la activación de este mecanismo se producen pequeños fragmentos de ARN que actúan de señal y median la inactivación del ARNm blanco por medio de la complementariedad de bases impidiendo que sea traducido y en muchos casos conducen a la degradación del ARNm [113]. Este mecanismo recibe el nombre de silenciamiento génico

postranscripcional (PTGS, de las siglas en inglés “*PosTranscriptional Gene Silencing*”). Uno de los principales efectores en la vía de PTGS involucra a los microARNs. Los microARNs son pequeños ARN no codificantes que se generan a partir de precursores específicos codificados en el genoma, y que al transcribirse se pliegan en horquillas (*hairpins*) con segmentos de complementariedad imperfecta dando lugar a los llamados ARN interferentes. Su tamaño es de aproximadamente 22 nucleótidos de longitud y tienen un rol fundamental en diversos procesos biológicos tales como el desarrollo, la proliferación celular, la diferenciación celular y la apoptosis [114,115], encontrándose involucrados en el control de casi todas las vías celulares desde el desarrollo hasta oncogénesis [116,117]. A su vez, los microARNs poseen patrones de expresión temporal y espacial que son específicos de un tipo celular, tejido o de la etapa del desarrollo en la que el organismo se encuentra [118]. Los microARNs fueron identificados por primera vez en *Caenorhabditis elegans* [119] y han sido identificados en diversos organismos, incluyendo humanos, plantas y parásitos helmintos. El número total de microARN es diferente entre las diversas especies de metazoos [120]. Nuestro grupo ha identificado el repertorio de genes de microARNs y su perfil de expresión en varias especies de platelmintos, *E. granulosus* s. s. G1, *E. canadensis* G7, *E. multilocularis*, *Mesocostoides corti*, *T. crassiceps* y *T. solium* [85,121,122]. Particularmente en *Echinococcus* se han identificado 38 genes de microARNs, algunos conservados en metazoos y otros específicos de platelmintos, agrupados en 30 familias diferentes. A su vez, se determinó la expresión diferencial para determinados microARNs en diferentes etapas del ciclo de vida sugiriendo un rol en el mantenimiento de las características de cada etapa del desarrollo en platelmintos.

1.7.1.3.1 Biogénesis y vía de los microARNs

Los genes de los microARNs codificados en el genoma de un organismo se transcriben por acción de la ARN polimerasa II generando microARNs primarios largos (pri-microARNs). Los genes de los microARNs pueden encontrarse como estructura génicas normales en el genoma o bien localizarse en intrones de genes que codifican proteínas [123]. En el núcleo, los pri-microARNs son procesados por la enzima con actividad ARNasa III Drosha/DGCR8 para producir un precursor de microARN de aproximadamente 70 nucleótidos de longitud con estructura de tallo y rulo (pre-microARN o *hairpin*). Luego, el pre-microARN es transportado desde el núcleo hacia el citoplasma por Ran-GTP y Exportin-5 donde es procesado por otra enzima con actividad ARNasa III Dicer/TRBP, generando un dúplex microARN-microARN*. La cadena, conocida como microARN estrella (microARN*), o hebra pasajera generalmente es degradada [124,125]; mientras que la otra hebra, el microARN maduro, es cargado en el complejo efector (miRISC, *microARN-induced silencing complex*). El componente activo del complejo RISC es una proteína de la familia de las endonucleasas denominada Argonauta (Ago), que según las características de sus dominios y familia a la que pertenecen cortan la hebra de ARNm blanco complementaria al microARN cargado en el complejo RISC o bloquean su traducción. El componente fundamental de este complejo es una secuencia de 6-7 nucleótidos

localizado en el extremo 5' de la secuencia del microARN maduro cargado en la proteína Ago y conocida como la "región de siembra o semilla" (*Seed*) que sirve de guía a la proteína hacia el ARNm blanco. Aunque el emparejamiento de bases entre el microARN y su ARNm blanco no coincidan perfectamente, la secuencia "*Seed*" tiene que ser perfectamente complementaria para que sea capaz de reprimir la expresión de la proteína ([Figura I.11](#)). Los microARNs se diferencian de otras clases de ARNs interferentes principalmente debido a que muchos de ellos están filogenéticamente conservados, particularmente en la secuencia *Seed* de la secuencia madura del microARN y en la estructura de "*hairpin*" generada durante su biogénesis [115].

En esta tesis se aborda el estudio de este mecanismo enfocado en la función de las proteínas Ago codificadas en el genoma de *Echinococcus* mediante el modelado por homología de las proteínas y el análisis de los dominios, motivos y residuos esenciales involucrados en su rol funcional.

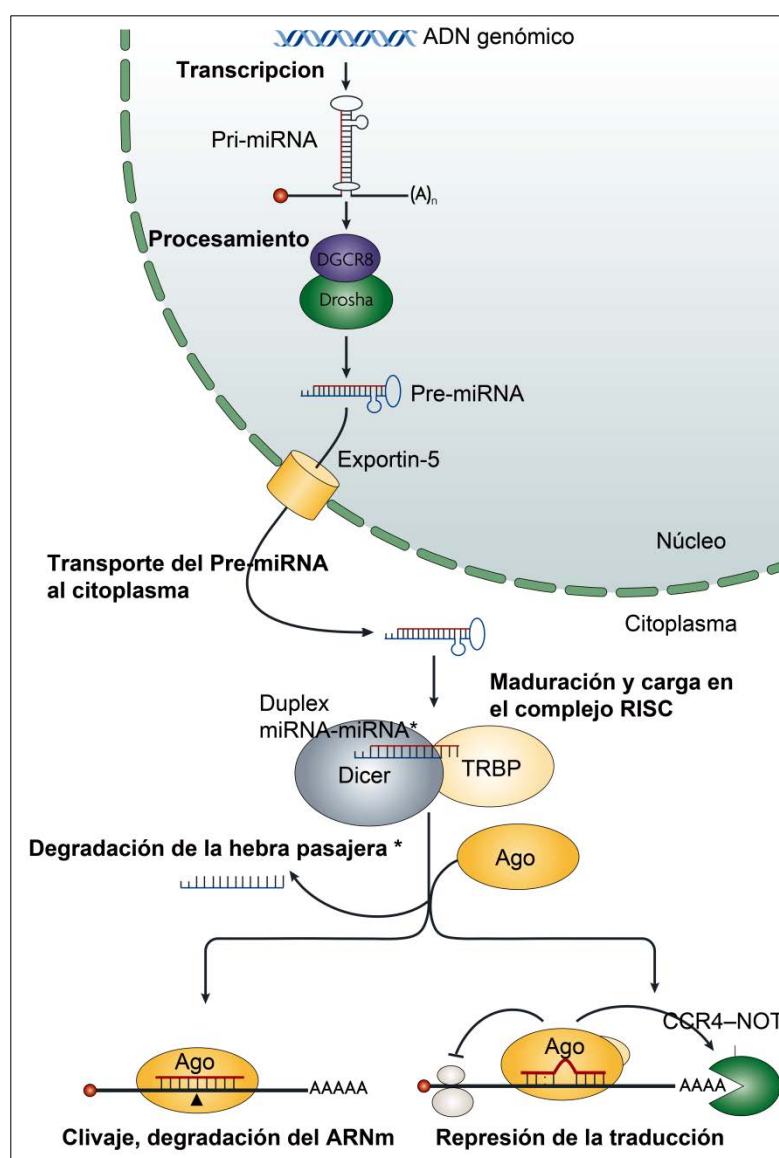


Figura I.11. Esquema de la vía de microARNs desde su transcripción en el núcleo hasta su carga en el complejo RISC-Ago y unión al ARN mensajero. Extraído y modificado de Filipowicz 2008 [123].

1.8 Sesgo en el uso de codones

En el código genético universal cada familia contiene entre 1 y 6 codones dependiendo del aminoácido para el cual codifican. Cuando los aminoácidos de las proteínas de un organismo están codificados por dos o más codones alternativos se los denominan codones sinónimos. Aunque la elección entre los codones sinónimos no altera la estructura primaria de una proteína, se sabe desde hace más de 20 años que el uso de codones sinónimos alternativos es un proceso no aleatorio [126–128]. Se ha observado que durante la codificación de genes a proteínas, algunos codones son utilizados con mayor frecuencia comparados con otros codones sinónimos. Esto en sí mismo no es sorprendente, ya que cabe esperar que el uso de codones se vea influenciado, al menos, por un sesgo mutacional [129]. Esta característica del uso preferencial de determinados codones en el proceso de traducción se conoce como "sesgo de uso de codones". El grado o patrón de sesgo del uso de codones puede variar dentro y entre las distintas especies de organismos [127]. Se dice que el patrón del uso de codones de un gen está altamente sesgado, si este utiliza un codón sinónimo determinado la mayor parte del tiempo en la incorporación de un aminoácido específico durante la traducción de la proteína. Este tipo de codones suelen denominarse codones óptimos o codones frecuentes. La naturaleza de la optimización entre el uso de codones y los ARNt ha sido ampliamente debatida por la comunidad científica, sin embargo aún no se conoce con exactitud si el uso de codones conduce a la evolución de los ARNt o viceversa. No obstante, está aceptado que el patrón del uso de codones depende del balance entre el sesgo mutacional y la selección natural para la optimización de la traducción y que este fenómeno está muy extendido en todas las especies contribuyendo a la evolución del genoma de manera significativa [129–131]. Asimismo, se han determinado algunos patrones generales en el uso de codones basados en el estudio de diversos organismos, desde bacterias levaduras, gusanos, moscas y hasta humanos [127,132,133]. Se ha sugerido que la optimización del uso de codones en organismos de crecimiento rápido está determinada principalmente por la abundancia de los ARNt en los genomas y se cree que los codones óptimos contribuyen a la eficiencia y precisión de la traducción. Se denomina selección traduccional a la eficiencia y precisión de la traducción que depende de la abundancia de los ARNt isoaceptores codificados en el genomas de un organismo. Como resultado de este fenómeno se espera que la selección de la traducción ocurra más frecuentemente en genes con alto nivel de expresión. Por otra parte, en los organismos cuya tasa de crecimiento no es alta o presentan genomas más pequeños la optimización del uso de codones está presente solo en algunos genes y la preferencia en el uso de codones suele estar determinada principalmente por un sesgo mutacional característico [127]. La variación de la optimización de los codones en un subconjunto de genes proporciona una eficiencia diferencial y una mayor precisión para la traducción de los dichos genes.

1.8.1 Relevancia del estudio del uso de codones

Dado que las secuencias codificantes de un genoma son las que proporcionan información sobre la función génica y la evolución de un organismo las investigaciones genómicas de los patrones de sesgo en el uso de codones, sus causas y consecuencias, y la identificación de fuerzas de selección que conforman su evolución son relevantes en los estudios genómicos para la comprensión de la biología del mismo. Por otra parte, el análisis del sesgo del uso de codones tiene importantes aspectos aplicados. La expresión génica heteróloga es una metodología muy importante utilizada en los procesos biotecnológicos mediante la cual se generan productos recombinantes como insulina, antibióticos y vacunas. El sesgo en el uso de codones del organismo hospedero en el que se expresan los genes foráneos puede ser un factor limitante que determina el rendimiento de los productos deseados [134,135]. Por lo tanto, es necesaria la optimización de la producción del gen foráneo basado en los patrones de uso de codones del organismo hospedero para que la traducción del gen recombinante no se vea afectada [136,137]. Asimismo, también se ha demostrado que los pares de codones óptimos pueden ser reemplazados por codones sinónimos “no óptimos” en genomas virales para el desarrollo de vacunas virales atenuadas [138–140].

1.8.2 Relevancia del uso de codones en la relación Parásito-Hospedero

Aunque los mecanismos utilizados por los organismos para la traducción de las proteínas permanecen altamente conservados a través de los taxones, existe una gran variación a nivel del reconocimiento entre el codón y el anti codón, tanto dentro como entre especies. Esta variación se refleja en el uso de codones, número de copia de ARNt y redundancia y degeneración de anti codones. En estudios evolutivos se ha explorado la degeneración y redundancia de la maquinaria traduccional del código genético en relación con la infección con parásitos que utilizan vías de síntesis de proteínas del hospedero. Estos estudios han sugerido que las características de la maquinaria traduccional podrían ser adaptaciones de los organismos parásitos para reducir los costos energéticos de la infección. Para ello, a lo largo de la evolución, los ARNt isoaceptores de parásitos fluctúan con el tiempo a medida que los parásitos rastrean el uso del codón del hospedero, y los hospederos al mismo tiempo modifican su uso para escapar de la infección y de este modo los mecanismos de traducción del hospedero imponen una presión de selección sobre el uso de codones de los parásitos [141]. Con el advenimiento de las técnicas de secuenciación de genomas enteros de numerosas especies, se han realizado estudios de patrones de sesgo del uso de codones en diferentes organismos tanto procariotas como eucariotas [142] y las investigaciones se dirigen ahora a analizar el patrón del sesgo de uso de codones de una manera holística a lo largo del genoma para comprender mejor los patrones de sesgo del codón dentro y

entre los genomas, en lugar de considerar solamente genes específicos o conjuntos de genes particulares [128].

1.8.3 Modelos evolutivos y factores responsables del sesgo en el uso de codones

El sesgo de uso de codones es un fenómeno complejo que puede involucrar muchos factores dando como resultado un patrón definible y característico de cada grupo de organismos. Dentro de un mismo genoma determinados genes individuales muestran una variación en el sesgo de uso de los codones, cuyo grado depende de factores tales como el nivel de expresión, la longitud del gen [143], el sesgo en la composición de nucleótidos (dado por un sesgo en el porcentaje de GC en distintas regiones) entre otros [144,145]. Incluso dentro de un mismo gen, los patrones del uso de los codones pueden variar dependiendo de la posición del codón en el gen [146]. En esta tesis, dichos factores son evaluados para determinar su implicancia en la formación del patrón de uso de codones en *Echinococcus*, como así también se aborda el estudio de las fuerzas evolutivas que moldean el uso de codones. A continuación se describen los dos modelos evolutivos más importantes que se han desarrollado para explicar la variación en el sesgo del uso de codones.

1.8.3.1 Modelo de selección y sesgo mutacional

Aunque el mecanismo de selección del sesgo en el uso de codones sigue siendo controvertido, las posibles explicaciones de este sesgo se dividen en dos categorías generales. Una de ellas se basa en la teoría de selección, en la que el sesgo en el uso de codones contribuye a la eficiencia y/o a la precisión de la expresión de proteínas y por lo tanto, se somete a una selección positiva. El modelo seleccionista también explica por qué los codones más frecuentes son reconocidos por las moléculas de ARNt más abundantes, como así también la correlación entre los codones preferidos, los niveles de ARNt y los números de copias génicas. Aunque se ha demostrado que la incorporación de aminoácidos en los codones más frecuentes se produce a una tasa mucho mayor que la de los codones raros, no se ha demostrado que la velocidad de traducción se vea directamente afectada y, por lo tanto, el sesgo hacia codones más frecuentes puede no ser directamente ventajoso. Sin embargo, el aumento en la velocidad de elongación de la traducción puede ser indirectamente ventajoso al provocar el aumento de la concentración celular de ribosomas libres y de esta forma favorecer el inicio de la traducción de los ARNm [126]. Asimismo, dado que la velocidad de la traducción del ARNm influye en el plegamiento de la proteína y que además el uso de codones está ligado a la velocidad de la elongación en la traducción, el uso de un codón u otro influiría sobre el plegado y correcto funcionamiento de las proteínas traducidas. En efecto, se ha observado que la presencia de mutaciones sinónimas tienen un efecto importante sobre el plegamiento de algunas proteínas llegando incluso a cambiar la

especificidad de su sustrato [147]. Además el tipo de codones reflejado en la composición nucleotídica del ARNm influye en la estructura secundaria de dicho ARNm y puede ser uno de los factores que determina la preferencia en el uso de codones en el que interviene la selección [148]. La segunda explicación para el uso de codones es el “sesgo de mutación”, esta teoría postula que el sesgo de codones existe debido a la no aleatoriedad en los patrones de mutación. En otras palabras, algunos codones pueden sufrir más cambios que otros y por lo tanto resultar en frecuencias de equilibrio más bajas, también conocidas como codones "raros". Diferentes organismos también exhiben diferentes sesgos mutacionales, y existe evidencia creciente de que el nivel de contenido de GC total de un genoma es el parámetro más significativo para explicar las diferencias en el sesgo del uso de codones entre los organismos. Sin embargo, este modelo por sí solo no puede explicar completamente por qué los codones preferidos son reconocidos por ARNts más abundantes [126].

1.8.3.2 Modelo de equilibrio mutación-selección-deriva

El sesgo en el uso de codones también puede explicarse mediante el modelo de equilibrio de mutación, selección y deriva que contempla a los fenómenos anteriormente mencionados [149–154]. Esta hipótesis postula que la selección favorece a los codones más frecuentes sobre los menos frecuentes, pero estos últimos pueden persistir debido a la presión de mutación y a la deriva genética. También sugiere que la selección es generalmente débil, pero que la intensidad de la selección se adapta a una mayor expresión y a las limitaciones funcionales de las secuencias codificantes [155]. De este modo, las mutaciones de un codón no preferido que dan lugar a un codón preferido deberían ser beneficiosas, mientras que las mutaciones en la dirección opuesta deberían ser perjudiciales.

Los genes altamente expresados generalmente muestran un mayor sesgo en el uso de codones [143,156–160]. La correlación positiva entre el sesgo del uso de codones y los niveles de expresión génica se atribuyen generalmente a la selección con el fin de priorizar la eficiencia de la traducción de aquellos genes que son fundamentales para la supervivencia de un organismo [130,143,157,161,162]. En este modelo, el sesgo del uso de codones se explica mediante una fuerza de selección negativa frente a los codones raramente utilizados y una fuerza de selección positiva en favor de los codones usados con mayor frecuencia, y de este modo los codones se seleccionan para genes que están altamente expresados. Por lo tanto, los codones óptimos se definen generalmente como codones que están presentes con mayor frecuencia en genes altamente expresados en relación a la baja frecuencia con la que aparecen en genes con bajo nivel de expresión [131]. La selección de la traducción se explica principalmente sobre la base de los genes de los ARNt en las especies donde se sugiere que la co evolución desempeña un papel importante entre el uso de codones y la abundancia de los ARNt isoaceptores codificados en el genoma [131,154,163,164]. La hipótesis de co evolución entre codón-anticodon, establece que el sesgo del uso de codones es un factor determinante que ejerce una presión de selección sobre la abundancia de los ARNt, y la abundancia de los ARNt debe co evolucionar con el uso

de codones para coincidir con el codón más abundante de una familia de codones sinónimos. Esto explica por qué hay abundancias desiguales de ARNt isoaceptores que conducen a un uso sesgado de codones entre un grupo de codones reconocidos por dichos ARNt [130,153,165]. Sin embargo, la causa o consecuencia de esta asociación aún no está clara. Aunque la selección traduccional podría ser un factor importante que determina el sesgo del uso de codones dentro de las especies, tal selección puede no estar directamente correlacionada con el patrón del uso de codones entre las especies debido a que gran parte de la variabilidad del patrón de uso de codones entre las especies parece depender de los diferentes patrones de mutación [127].

HIPÓTESIS Y OBJETIVOS



Capítulo II

Hipótesis y Objetivos

2.1 Hipótesis

El desarrollo de las nuevas tecnologías de secuenciación ha permitido abordar el estudio de organismos que no se encuentran dentro de la escala experimental rutinaria de los laboratorios. La obtención de datos genómicos, su análisis y estudio mediante genómica comparativa permite extraer una cantidad de información relevante de organismos que poseen ciclos de vida complejos y metodologías de experimentación limitadas. El análisis exhaustivo de este tipo de datos permite explorar los genomas de un modo más eficiente y de esta manera identificar estructuras genómicas y genes que son esenciales para la sobrevivencia del individuo en estudio y particularmente para el estudio de los patógenos que son de gran importancia en la salud pública.

2.2 Objetivos

Dada la complejidad del estudio de parásitos helmintos del género *Echinococcus* y su relevancia a nivel sanitario y epidemiológico se propusieron los siguientes objetivos principales:

- I.1. Generar y analizar nuevos datos genómicos y transcriptómicos de uno de los principales causantes de la echinococcosis, *Echinococcus canadensis* G7.
- II.1. Realizar análisis de genómica comparativa entre las especies de *Echinococcus* que permitan explicar diferencias biológicas.
- III.1. Estudiar la relación filogenética de las especies de *Echinococcus* mediante el análisis de SNPs en todo el genoma.
- IV.1. Identificar y analizar genes específicos del género *Echinococcus* como posibles blancos de intervención para el control de la echinococcosis y sus posibles mecanismos de regulación de la expresión génica.
- V.1. Transferir los datos genómicos obtenidos como bases de datos de consulta y herramientas bioinformáticas aplicables a las líneas de investigación y estudio de otros parásitos helmintos.

MATERIALES Y MÉTODOS



Capítulo III

Materiales y Métodos

3.1 Recolección de muestras y extracción de ADN

3.1.1 Material parasitario

El material parasitario utilizado para la secuenciación del ADN de *E. canadensis* G7 fue recolectado en Buenos Aires, Argentina. Se obtuvieron quistes hidatídicos fértiles a partir de hígados de cerdos naturalmente infectados suministrados por mataderos de Buenos Aires, Argentina. Los animales involucrados en este estudio no fueron sometidos a ningún procedimiento experimental y todas las muestras utilizadas se recogieron *post mortem* en mataderos comerciales. Para la secuenciación del genoma, el líquido hidatídico fue asépticamente aspirado con una jeringa a partir de un quiste unilocular de un hígado de cerdo y se recuperaron los protoscolices (PS) del fluido aspirado. Los mismos se lavaron en una solución salina tamponada con fosfato (PBS) para eliminar los PS muertos y los residuos de la pared del quiste. Una fracción de los PS de cada quiste se usó para determinar su viabilidad mediante la prueba de exclusión de eosina. Las muestras que mostraron más del 90% de viabilidad se seleccionaron para el aislamiento del ADN. La especie y el genotipo se determinaron mediante la amplificación seguida de secuenciación de un fragmento de la subunidad 1 mitocondrial de la citocromo c oxidasa (COX1) [27,85]. El esquema del diagrama de trabajo se muestra en la [figura III.1](#).

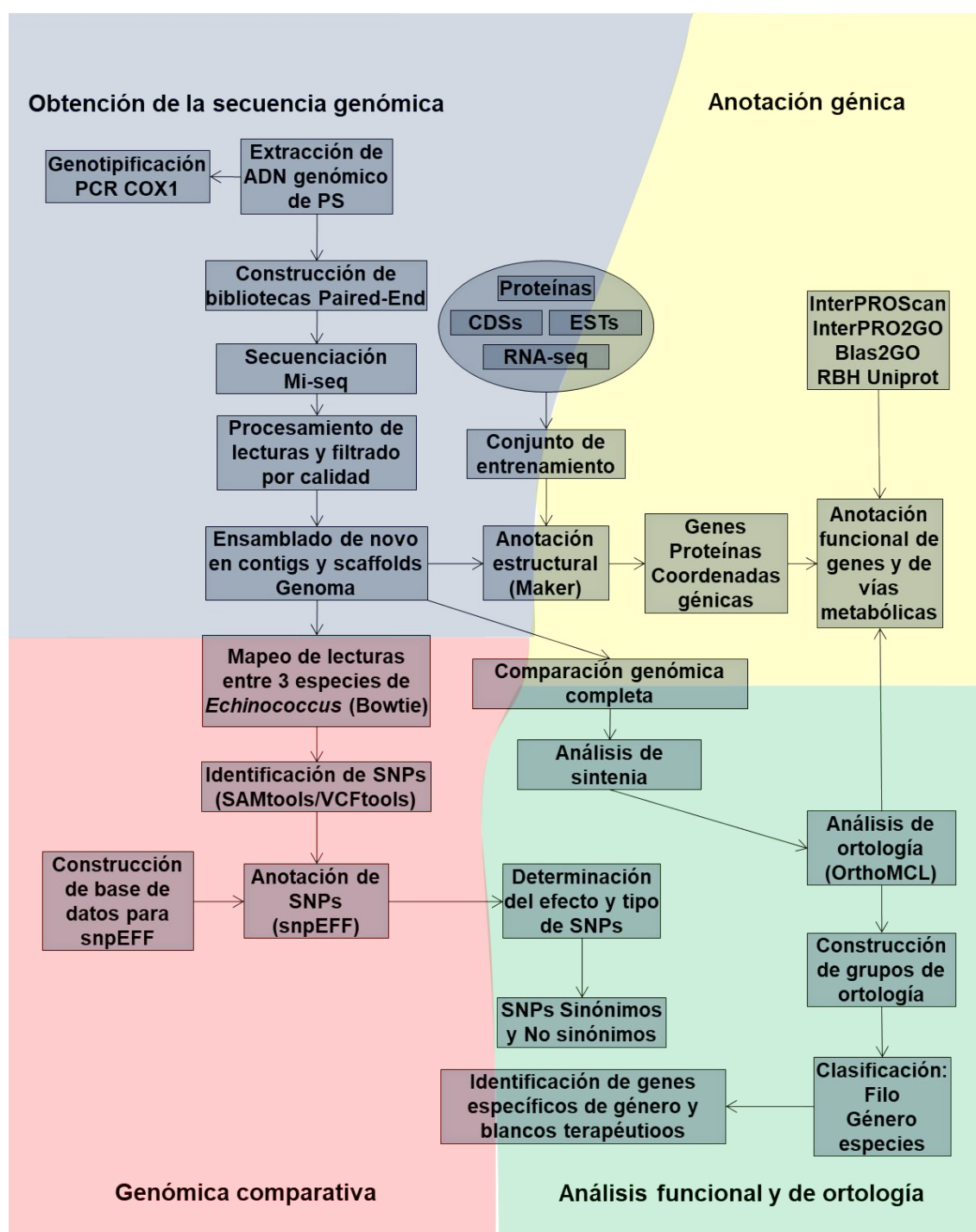


Figura III.1: Esquema del diagrama de trabajo bioinformático implementado en esta tesis para la obtención, anotación génica, análisis funcional y genómica comparativa del genoma de *Echinococcus*.

3.1.2 Aislamiento y purificación del ADN

El aislamiento de ADN genómico de alta calidad se realizó mediante la técnica de fenol/cloroformo. Para ello, a un volumen de 100 µl de PS se le adicionó un volumen 1,2 ml de solución tampón de lisis Tris-HCl de PH = 8 (100 mM NaCl, 50 mM EDTA,

0,5% SDS, 20 ug/ml ARNasa A y 0,1 mg/ml de proteinasa K) y se incubó a 50°C durante 24 Hs. Las muestras que contenían restos visibles de tejido del hospedero se las trató con pepsina con el fin de eliminarlo. Posteriormente se agregó 1 volumen de fenol/cloroformo/alcohol isoamílico con una relación 25:24:1 (SIGMA-Life science) y se centrifugó a 5000 g durante 15 minutos. Se extrajo cuidadosamente el sobrenadante y se repitió una vez más la extracción con fenol/cloroformo, seguido de una extracción con cloroformo y centrifugando a 5000 g durante 15 minutos luego de cada extracción. Posteriormente se extrajo el sobrenadante y se adicionó 0,1 volúmenes de acetato de sodio (NaAc) de PH = 5,8 y 2 volúmenes de etanol 100%. Inmediatamente el precipitado de ADN se extrajo mediante un capilar sellado y se trasvasó a un tubo seco, Luego se lavó dos veces con etanol 70%, se secó el ADN por evaporación y se lo resuspendió en 20 ul de agua. Las muestras se cuantificaron utilizando un fluorómetro Qubit (Invitrogen) específico para ADN de doble cadena y la calidad del ADN se evaluó mediante la relación de DO 260/280 y DO 260/230 usando un NanoDrop (ThermoFisher Scientific).

3.2 Amplificación por PCR y secuenciación

La reacción de PCR se realizó en un volumen final de 50 µl conteniendo el ADN de la muestra (5-20 ng), 200 µM de cada dNTP (Invitrogen), 1,5 mM de MgCl₂, 0,2 µM de cebadores directos y reversos (ver tabla III.1; cebadores y temperaturas de unión) y 2 U de ADN taq polimerasa. Las condiciones de ciclado fueron las siguientes: una desnaturalización inicial a 95 °C durante 5 minutos, 35 ciclos a 95 °C durante 45 segundos, seguido de 50 segundos a la temperatura de unión óptima (ver [Tabla III.1](#)) y a 72 °C durante 60 segundos, con una reacción de terminación de 5 minutos a 72 °C. Los productos de amplificación se verificaron en geles de agarosa al 1,5% teñidos con gelRed para verificar la presencia de un único producto de amplificación. La secuenciación se realizó en un secuenciador capilar Applied Biosystems 3130 usando el kit “Big Dye Terminator v3.1 Cycle Sequencing Kit”, de acuerdo con las instrucciones del fabricante y se secuenciaron ambas cadenas del producto de PCR. Para la verificación del genotipo se compararon las secuencias COX1 obtenidas con todas las secuencias de *Echinococcus* conocidas según Kamenetzky y col [166]. Para la validación de la presencia de SNPs en los genomas completos se utilizaron 3 aislamientos representativos de *E. canadensis* G7, 3 de *E. granulosus* G1 y 4 de *E. multilocularis*, 2 de *E. orteppi*, 2 de *E. canadensis* G6 y 2 de *E. oligarthrus* [24]. La detección de bases del cromatograma, el ensamblado de secuencias, la identificación de polimorfismos y la inspección manual del ensamblado de secuencias y polimorfismos se realizó mediante una combinación de softwares Phred - (version phred-dist-071220.c), Phrap (version 1.090518) y Consed (version 0.29). La detección de bases de los cromatogramas fue realizado por Phred. Las secuencias fueron ensambladas por Phrap-Consed y los SNPs candidatos fueron identificados y sometidos a una inspección manual.

Tabla III.1: Cebadores utilizados en esta tesis para genotipificación y validación de SNPs.

Genes	Orientación	Secuencia del cebador	Tamaño de amplicon (pb)	Tm (°C)	Estructura genómica molde	Referencia
cox1	F	5'-tttgggcatcctgaggtt-3'	410	55	Mitochondrial (exon)	[166]
	R	5'-taaagaaagaacataatgaaaatg-3'				
Proteína Argonauta-Piwi	F	5'-tcatcaccgggtaacatgac-3'	196	57	Nuclear (exon)	Esta tesis
	R	5'-ggagcaatagtttggcatcg-3'				
Proteína Kinasa	F	5'-tgaacgtgagaagaacaaatcc-3'	181	57	Nuclear (exon)	Esta tesis
	R	5'-aaatgatttgcgctccagat-3'				
Factor de elongación (EFG) con dominio de unión a GTP	F	5'-gcacatgagatgtggtgac-3'	205	57	Nuclear (exon)	Esta tesis
	R	5'-ctcaacgtccgtccttgg-3'				
Subunidad alfa/beta de proteosoma	F	5'-gggattgccttaacgactt-3'	568	57	Nuclear (exon)	Esta tesis
	R	5'-acaaatcgggtagggtttca-3'				
Núcleotidodisulfuro de piridina oxidoductasa; Glutaredoxina	F	5'-tgccatactcttggctcct-3'	100	57	Nuclear (exon)	Esta tesis
	R	5'-aaacagggcatagtaaagaacatca-3'				
Núcleotidodisulfuro de piridina oxidoductasa; Glutaredoxina	F	5'-acgatcgaagtcctcctcc-3'	107	57	Nuclear (exon)	Esta tesis
	R	5'-acaaccgcgtggtaggatt-3'				
Proteína MBD	F	5'-agctaacgcatccagtccac-3'	332	57	Nuclear (exon)	Esta tesis
	R	5'-ccagtccttcaaccaattcc-3'				
Proteína MBD	F	5'-cggacctctattctccaaa-3'	220	57	Nuclear (exon)	Esta tesis
	R	5'-aaatcacgacgctgttga-3'				

3.3 Construcción de bibliotecas de ADN y secuenciación

La preparación de las bibliotecas y la secuenciación del ADN se realizaron en la plataforma de secuenciación MiSeq (Illumina) con colaboración del Grupo de Genómica y Biología Computacional del Centro de Pesquisa, René Rachou (CPqRR), Fundación Oswaldo Cruz, Minas Gerais, Brasil. Para la preparación de cada biblioteca, se usaron 50 ng de ADN genómico que fueron sometidos a una reacción de fragmentación donde el ADN se fragmentó y simultáneamente se unió a adaptadores específicos usando el Kit Nextera® XT DNA Sample Preparation Kit de acuerdo con las instrucciones del fabricante. A continuación, los fragmentos generados se purificaron y se sometieron a una reacción de amplificación utilizando cebadores complementarios a los adaptadores y los productos fueron cuantificados por qPCR utilizando el kit KAPA™ SYBR® FAST qPCR. Se prepararon dos bibliotecas: L1 con fragmentos de 1050 pb y L2 con fragmentos de 1200 pb. Luego las bibliotecas se diluyeron en solución de Tris-HCl + Tween 0,1%, se sembraron en un micro chip y se secuenciaron con 500 ciclos 2x250bp (L1) o 600 ciclos 2x300pb (L2) usando el Kit MiSeq v2 Reagent Kit. Las imágenes fueron procesadas y analizadas con el software suministrado por el fabricante. La calidad de las lecturas de Illumina se evaluó con FastQC v0.10.1, y se procesaron de acuerdo a los parámetros de calidad Phred ($Q > 33$) utilizando el software Trimmomatic [167] y se removieron secuencias remanentes de adaptadores. Phred es una medida de la calidad con la que se detecta un nucleótido en la secuenciación del ADN y se define como la probabilidad de que una base nucleotídica sea leída con cierta precisión.

$$Q = -10 \times \log_{10} P$$

donde Q es el valor de calidad Phred y P la probabilidad de que un nucleótido haya sido leído incorrectamente durante la secuenciación. Un valor de $Q > 33$ significa una precisión mayor al 99,99%.

3.4 Obtención del genoma de *E. canadensis* G7

El genoma de *E. canadensis* G7 se obtuvo por medio un ensamblado *de novo* de las lecturas en formato fastq obtenidas por la secuenciación de dos bibliotecas genómicas en la plataforma MiSeq de Illumina. El ensamblado de las lecturas dio lugar a secuencias largas denominadas *contigs*. Los contigs obtenidos fueron comparados contra el genoma del hospedero y aquellos con resultados de búsqueda positivos fueron removidos. Los contigs son segmentos de secuencias de ADN obtenidos por superposición de secuencias más cortas obtenidas del secuenciador que reconstruyen las secuencias originales de ADN que conforman los cromosomas de un organismo. El

grupo de contigs ordenados y orientados correctamente que reconstruyen la secuencia original de ADN y separados por *gaps* se denomina “*scaffolds*”.

Con el fin de optimizar el ensamblado del genoma de *E. canadensis* G7, se probaron diferentes ensambladores de *nov*o y se evaluaron parámetros estadísticos de calidad para definir el mejor ensamblado. Los softwares usados en las distintas pruebas de ensamblado fueron Velvet1.2.07 [168], SOAPdenovo2 [169] y SPAdes 3.6 [170,171]. Las lecturas y los contigs obtenidos se reutilizaron para generar los genomas mejorados usando PAGIT, IMAGE [172]; ICORN [173]; ABACAS [174]. Posteriormente las métricas de la calidad de los diferentes ensamblados obtenidos se calcularon con QUAST [175] se seleccionó el mejor ensamblado en base a los valores de N50, el número total de contigs y la longitud total del ensamblado. El N50 es una estadística básica para describir la contigüidad de un ensamblado de un genoma. Cuanto mayor sea el N50, mejor será el ensamblado. El cálculo de N50 se realiza ordenando los contigs por tamaño de mayor a menor. A continuación se suman los tamaños los contigs, hasta que la suma sea igual a la mitad del número de bases total los contigs en el conjunto y el tamaño del contig más corto de dicha lista es el que determina el valor del N50 del ensamblado. El proceso de ensamblado dio lugar a la versión v.1 del genoma de *E. canadensis* G7 y los contigs menores a 1 kb de longitud fueron descartados. El espacio génico total fue validado usando CEGMA 2.4 [176] y la comparación de la calidad de los genomas de las especies de *Echinococcus* fue realizada mediante el mapeo de contigs contra el genoma de *E. multilocularis* utilizando ABACAS.1.3.2 [174]. La cobertura y la profundidad de la cobertura fueron calculadas con scripts personalizados. La profundidad de la cobertura se refiere al número de veces que fue leída cada base en una posición determinada del ensamblado final. Y la cobertura se refiere al porcentaje del tamaño total del genoma de referencia que está representado por el genoma ensamblado.

3.5 Disponibilidad de los datos

Las secuencias ensambladas del genoma de *E. canadensis* (G7) se depositaron en ENA (BioProject PRJEB8992, <https://www.ebi.ac.uk/ena/PRJEB8992>) y Wormbase Parasite ([http://parasite.wormbase.org/Echinococcus_canadensis_prjeb8992 / Info / Index /](http://parasite.wormbase.org/Echinococcus_canadensis_prjeb8992/Info/Index/)). Los datos sobre los grupos ortólogos se pueden descargar de la página web del proyecto FlatDB (<http://www.bmhid.org.ar/flatdb/>).

3.6 Predicción estructural y anotación de genes

Para la anotación génica estructural se realizó la búsqueda de genes *ab initio*, utilizando el software automatizado MAKER 2.31 [177]. Para la predicción de los modelos génicos, previamente se seleccionó y descargó de diferentes bases de datos (<https://www.ebi.ac.uk/ena>; <https://www.ncbi.nlm.nih.gov/genbank/> y <ftp://ftp.sanger.ac.uk/>), un grupo confiable de genes y proteínas representativas de las

especies relacionadas del género *Echinococcus* para generar así un conjunto de datos fiables que sirvan como conjunto de entrenamiento para los softwares Augustus [178] y SNAP [179]. La elección de los genes y/o proteínas que conformaron estos grupos se realizó acorde a si poseían evidencia transcripcional o estudios asociados que den evidencia de ello y que cuya secuencia se encuentre completa. El conjunto de entrenamiento consistió en 300 ESTs de *Echinococcus*, 4321 proteínas de *E. multilocularis* y 2531 proteínas de *E. granulosus* G1.

Los datos de proteínas se incorporaron a través del software Exonerate v.2.2.0 [180]. La predicción génica además fue acompañada y sustentada por la incorporación de datos de transcriptómica de RNA-seq de *E. canadensis* G7 generados por el Wellcome Trust Sanger Institute y Tsai y col. [43]. Las lecturas de RNA-Seq se mapearon contra el genoma de *E. canadensis* G7 usando Tophat v2.0.12 [181,182] y el resultado de este mapeo se utilizó para construir modelos ocultos de Markov mediante GeneMark-ET [183] para luego ser utilizados por GeneMark [184] incorporado a MAKER2 para la predicción génica. La anotación estructural génica mediante MAKER 2.31 [177] se realizó de la siguiente manera: en primer lugar se identificaron regiones de ADN repetitivas utilizando RepeatModeler [185] las cuales fueron enmascaradas durante la predicción génica usando RepeatMasker [186]. Los modelos génicos identificados por CEGMA 2.3 [176] fueron también incorporados a MAKER 2.31 como evidencia génica. El protocolo de predicción consistió en tres rondas de predicción con MAKER2: La primer ronda se ejecutó utilizando la base de datos de proteínas personalizadas como evidencia física y con GeneMark [184], SNAP [179] y Augustus [178] activados para las predicciones *ab initio*. Posteriormente, se realizó una segunda ronda de MAKER2 utilizando la evidencia física de transcriptos del archivo GFF3 obtenido a partir del mapeo de las lecturas de RNA-Seq obtenido con GeneMark-ET [183] y con todos los predictores *ab initio* desactivados. El paso final de MAKER2 se realizó utilizando la evidencia física de transcriptos del archivo GFF3 obtenido a partir del mapeo de las lecturas de RNA-Seq obtenido con GeneMark-ET [183] y con todos los predictores de genes activados, y se incorporaron además los modelos de genes obtenidos en la ronda anterior de corrida través de MAKER2 como EST con una puntuación umbral AED < 0,5. Así, MAKER2 recopiló la evidencia física basada en la homología de los resultados de los modelos génicos predichos *ab initio* por cada uno de los programas integrados en MAKER2 para generar la anotación final.

3.7 Anotación de genes no codificantes

Los ARNs de transferencia fueron buscados por tRNAscan-SE v.1.23 [187] y cada *locus* de ARNt fue confirmado comparandolos contra la bases de datos “RNA central database” (<http://rnacentral.org/>). Dos genes adyacentes de ARNt se definieron como clúster si su distancia era inferior a 1000 nucleótidos, de acuerdo con Bermudez y col. [133]. Los genes de ARN ribosómicos (ARNr) fueron identificados por el software RNAmmer-1.2 (www.cbs.dtu.dk/services/RNAmmer/) y se los corroboró mediante BLAST contra las bases de datos del NCBI y contra la base de datos

“5S ARNr (<http://combio.pl/rRNA/>)”. Los microRNAs específicos de *Echinococcus* se identificaron como se describe anteriormente por Macchiaroli y col. [122]. Los microARNs fueron mapeados contra el genoma de *E. canadensis* G7 utilizando BLAST [188]. Todos los datos de anotación relativos a la ubicación en el genoma de *E. canadensis* G7 fueron agregados como coordenadas al archivo GFF de *E. canadensis* G7 utilizando *scripts* personalizados.

3.8 Análisis de modelos génicos y anotación funcional de genes

La estructura génica obtenida se evaluó utilizando Eval [189]. Los modelos génicos, tamaño y número de intrones y exones y sitios de *splicing* fueron verificados por la implementación de este método. La anotación funcional de los genes se realizó mediante una combinación de tres métodos. En primer lugar todos los modelos génicos fueron analizados en búsqueda de dominios conocidos utilizando InterProScan-5.7-48.0 [190] que incluye a las bases de datos PROSITE [191], PRINTS [192], SMART [193], PFAM [194] y SUPERFAMILY [195]. Por otra parte los genes fueron mapeados mediante BLAST [188] contra la base de datos UniProt/Swiss-Prot (<http://www.uniprot.org/>) y todos los datos obtenidos fueron integrados y asociados con términos GO mediante InterPro2GO y Blast2GO [196] utilizando la base de datos de “Gene Ontology” [197]. La asociación de los términos GO a los distintos niveles de jerarquía se realizó con R utilizando el paquete “GO.db” (<https://bioconductor.org/packages/release/data/annotation/html/GO.db.html>) y fueron procesados y analizados usando *scripts* personalizados en BASH. Los modelos génicos que no pudieron ser anotados por medio del procedimiento anterior fueron anotados mediante la asignación de la anotación de grupos de ortología creados para organismos parásitos usando OrhtoMCL v2.0.9 [198,199]. Y por último se utilizó BLAST [188] contra la base de datos UniProt/Swiss-Prot para definir las anotaciones finales de los genes que no pudieron ser anotados por ninguno de los métodos anteriores. Posteriormente, los genes se asociaron a vías metabólicas. Para ello los proteomas fueron mapeados contra las bases de datos KEGGs utilizando la BlastKOALA [200] y Blast2GO [196]. Con el propósito de profundizar en el estudio del metabolismo se seleccionaron manualmente algunas vías reguladoras importantes y otros módulos funcionales, que se utilizaron en “KEGG Mapper Reconstruct Pathway” para la visualización y el análisis de dichas vías metabólicas.

3.9 Grupos de genes ortólogos

Para llevar a cabo la búsqueda de ortólogos, primeramente se buscaron y descargaron los proteomas de 14 taxones de metazoos ([Tabla III.2](#)) y sus secuencias se compararon por mejor hit recíproco (RBH) utilizando BLAST [188]. Los resultados se

utilizaron posteriormente para identificar los grupos de genes ortólogos mediante el algoritmo OrthoMCL v2.0.9 [198,199]. Para establecer la relación de ortología se estableció un valor de corte de 1,5 y un subconjunto de proteínas se agrupó según las siguientes categorías: platelmintos, platelmintos parásitos, platelmintos de vida libre, cestodos y *Echinococcus*. Para detectar las familias de proteínas expandidas, se utilizaron grupos ortólogos construidos con OrthoMCL y se identificaron conjuntos que contenían más de tres proteínas por especie en todas las especies de metazoos y en el género *Echinococcus*.

3.10 Blancos terapéuticos

Con el fin de identificar genes específicos de cestodos que están ausentes en humanos se seleccionaron los genes ortólogos de *E. canadensis* G7 a partir de los grupos ortología que son exclusivos de cestodos. Se utilizó un valor de corte estricto de 0,8 para llevar a cabo el agrupamiento. Los nuevos genes candidatos a blancos terapéuticos se seleccionaron teniendo en cuenta los siguientes criterios: estar presente en todas las especies cestodos, tener alta conservación de la secuencia entre los cestodos (MCL puntuación <0,8) y estar ausente o tener un alto grado de divergencia en los seres humanos. Para todos los grupos de ortología seleccionados se realizaron análisis filogenéticos mediante alineamientos múltiples con ClustalX, se calcularon las distancias con el modelo JTT y se construyeron árboles con el algoritmo de Máxima verosimilitud implementado en el software MEGA5. Los valores de cada nodo representan los porcentajes de *bootstrap* mayores al 50%, obtenidos sobre 500 réplicas.

3.11 Anotaciones de dominios y análisis estructural

Las proteínas estudiadas en este trabajo fueron buscadas utilizando BLAST [188] contra las bases de datos UniProt/Swiss-Prot. Los dominios de proteínas se identificaron en las bases de datos PFAM y Prosite utilizando PFAM_scan [201] o HMMscan 3.0. Para un subset de proteínas de interés se realizaron los siguientes análisis (ver sección Resultados 4.4): identificación de péptidos señal mediante Phobius (<http://www.ebi.ac.uk/Tools/pfa/phobius/>) y SignalP 4.0 (<http://www.cbs.dtu.dk/services/SignalP/>), obtención de estructura secundaria mediante PSIPRED (<http://bioinf.cs.ucl.ac.uk/psipred/>) y Jpred 4 (<http://www.compbio.dundee.ac.uk/jpred/>), modelado de la estructura de proteínas mediante combinación de los softwares PHYRE2 [202] y SWISS-MODEL [203–206], búsqueda por homología de estructuras mediante la base de datos PDB (<http://www.rcsb.org/pdb>, [207]).

Tabla III.2: Proteomas utilizados en la búsqueda de genes ortólogos entre distintas especies de metazoos.

Especies	Filo	Clase	Nº de proteínas*	Parasitismo	Hospedero	Fuente
<i>H. sapiens</i>	Cordados	Mammalia	30529	no		ftp://ftp.ncbi.nlm.nih.gov/pub/CCDS/current_human/CCDS_protein.current.faa.gz
<i>M. musculus</i>	Cordados	Mammalia	23874	no		ftp://ftp.ncbi.nlm.nih.gov/pub/CCDS/current_mouse/CCDS_protein.current.faa.gz
<i>D. rerio</i>	Cordados	Actinopterygii	41677	no		ftp://ftp.ensembl.org/pub/release-65/fasta/danio_rerio/pep/Danio_rerio.Zv9.65.pep.all.faa.gz
<i>B. floridae</i>	Cordados	Leptocardii	50817	no		ftp://ftp.jgi-psf.org/pub/JGI_data/metazome/v3.0/Bfloridae/annotation/Bfloridae_17_peptide.faa.gz
<i>D. melanogaster</i>	Artropodos	Insecta	24132	no		ftp://ftp.flybase.net/genomes/Drosophila_melanogaster/dmel_r5.41_FB2011_09/fasta/dmel-all-translation-r5.41.fasta.gz
<i>C.elegans</i>	Nematodos	Chromadorea	27367	no		ftp://ftp.wormbase.org/pub/wormbase/species/c_elegans/sequence/protein/c_elegans.PRJNA13758.current.protein.faa.gz
<i>E. canadensis</i>	Platelmintos	Cestoda	11437	si	vertebrados	ftp://ftp.ebi.ac.uk/pub/databases/wormbase/parasite/releases/WBPS3/species/Echinococcus_canadensis/PRJEB8992/Echinococcus_canadensis.PRJEB8992.WBPS3.genomic.faa.gz
<i>E. granulosus</i>	Platelmintos	Cestoda	10274	si	vertebrados	ftp://ftp.sanger.ac.uk/pub/pathogens/Echinococcus/granulosus/genome/Egranulosus_v3_2014_05_27.aa
<i>E. multilocularis</i>	Platelmintos	Cestoda	10551	si	vertebrados	ftp://ftp.sanger.ac.uk/pub/pathogens/Echinococcus/multilocularis/genome/Emultilocularis_v3_2014_04_27.aa
<i>H. microstoma</i>	Platelmintos	Cestoda	10145	si	vertebrados/ invertebrados	ftp://ftp.sanger.ac.uk/pub/pathogens/Hymenolepis/microstoma/Hmicrostoma_v1_2014_05_27.aa
<i>T. solium</i>	Platelmintos	Cestoda	12353	si	vertebrados	ftp://bioinformatica.biomedicas.unam.mx/TsM1_13.12.11/TsM1_13.12.11.prot.faa
<i>G. salaris</i>	Platelmintos	Monogenea	15488	si	Invertebrado/ Ectoparasito	http://invitro.titan.uio.no/gyrodactylus/ftp/Gsa-gene-models.fasta.gz
<i>S. mansoni</i>	Platelmintos	Trematoda	11825	si	vertebrado/ invertebrado	ftp://ftp.sanger.ac.uk/pub/pathogens/Schistosoma/mansoni/genome/Gene_models/Smansoni_v5_0_proteins_21032013.fasta
<i>S. mediterranea</i>	Platelmintos	Turbellaria	29649	no	Vida libre	ftp://ftp.ebi.ac.uk/pub/databases/wormbase/parasite/releases/WBPS1/species/schmidtea_mediterranea/PRJNA12585/schmidtea_mediterranea.PRJNA12585.WBPS1.proteins.faa.gz

3.12 Identificación de islas CGIs

Con el objetivo de identificar las islas CGIs, se usó el software “CgiHunterLight software tool v1.0” (<http://cgihunter.bioinf.mpi-inf.mpg.de/>), aplicando el algoritmo de Takai y Jones [208], el cual sugiere los siguientes parámetros de búsqueda: contenido de GC $\geq 55\%$, Obs CpG / Exp CpG ≥ 0.65 y longitud de CGIs ≥ 500 pb. Este algoritmo identifica y excluye a los falsos positivos de las secuencias repetitivas. Las CGIs identificadas fueron anotadas en los genomas de las tres especies de *Echinococcus* analizadas y su localización genómica y densidad fue evaluada en cada caso. Posteriormente se estudió de manera comparativa la presencia de CGIs en las tres especies de *Echinococcus* utilizando *scripts* personalizados que permitieron discriminar las diferentes regiones: genomas completos, regiones río arriba de los genes (posibles secuencias regulatorias) y las regiones codificantes.

3.13 Comparación de los genomas

La comparación de genomas completos de las especies de *Echinococcus* se realizó mediante el mapeo de los *contigs* sobre los cromosomas de *E. multilocularis* utilizando ABACAS.1.3.2 [174]. Una vez dispuestos los contigs en el orden y en la orientación correcta se reanotaron los genes de *E. canadensis* G7 y de *E. granulosus* G1. Los resultados de los análisis de ortología se utilizaron para definir las regiones de ortología génica y sintenia en un contexto cromosómico entre las especies de *Echinococcus*. La sintenia de los genes ortólogos se representó gráficamente con CIRCOS plot [209].

3.14 Identificación de SNPs

Para la determinación de la variabilidad genética las bibliotecas de Illumina de *E. canadensis* G7, *E. granulosus* G1 y *E. multilocularis* se filtraron por calidad y luego se mapearon contra los genomas de las especies de *Echinococcus* usando bowtie2 [210]. Las estadísticas de mapeo se calcularon con Bamtools (<https://github.com/pezmaster31/bamtools>), y las duplicaciones fueron marcadas y descartadas usando picard-tools v.1.129 (<http://broadinstitute.github.io/picard/>). Los SNPs fueron reconocidos utilizando Samtools [211], con los siguientes parámetros: frecuencia de variación $> 40\%$ con al menos 20 lecturas que cubran sitios polimórficos, y una calidad de base tanto para el sitio de referencia como para el sitio polimórfico > 30 . Las inserciones y deleciones (InDels) fueron filtradas con VCFtools [212]. Para disminuir el número de falsos positivos durante la identificación de SNPs, aquellos ubicados a una distancia menor a 10 pb de los InDels fueron removidos del análisis. Para todos los análisis de SNPs las lecturas fueron primero mapeadas contra su correspondiente genoma de referencia y, a continuación, contra el

genoma de cada una de las diferentes especies analizadas. Los sitios polimórficos heterocigotas (SNPs heterocigotas) fueron anotados como tales solo si ambas lecturas de la biblioteca “*paired end*” (izquierda y derecha) mapeadas contra la referencia contenían tanto al alelo alternativo como al alelo de la referencia y en ambas hebras de la referencia (directa y reversa) y con más de 20 lecturas totales sustentando el tipo de nucleótido en esa posición. Los sitios polimórficos homocigotas (SNPs homocigotas) fueron anotados como tales si ambas lecturas (derecha e izquierda) mapearon sobre el alelo de referencia en ambas hebras (directa y reversa) y con más de 20 lecturas totales apoyando el tipo de nucleótido en esa posición y ninguna lectura conteniendo el tipo de nucleótido presente en la secuencia de referencia. Este procedimiento se utilizó para registrar todos los sitios con SNPs homocigotas y heterocigotas de las tres especies de *Echinococcus*. La relación de sustituciones transición/transversión se calculó usando VCFtools [212] y la anotación y clasificación de SNPs basados en el efecto que generan en las regiones codificantes fueron llevadas a cabo con SnpEff v4.0 [213]. Las comparaciones entre las tres especies de *Echinococcus* se realizaron con *scripts* personalizados. Los análisis estadísticos se realizaron implementando el test de ANOVA, con un nivel de confianza del 95% y se graficaron mediante el software R implementando el paquete “car”. (<https://www.r-project.org/>).

3.15 Reconstrucción de la filogenia con genomas mitocondriales

Los contigs mitocondriales fueron identificados por BLAST [188] contra una base de datos de ADN mitocondrial personalizada de especies de *Echinococcus* y se concatenaron sobre el genoma mitocondrial de referencia de *E. canadensis* G7 (AB235847) para la obtención del genoma mitocondrial final completo. La anotación de sus genes se realizó por transferencia de anotación mediante RATT [214]. Las secuencias nucleotídicas de 13 genes que codifican para proteínas de los genomas mitocondriales de *Echinococcus* se tradujeron a las secuencias de aminoácidos utilizando el código genético mitocondrial de platelmintos [215]. Las secuencias de aminoácidos fueron alineadas usando ClustalX [216]. Los análisis filogenéticos usando los genomas mitocondriales completos se realizaron utilizando el método de Máxima verosimilitud con matrices de distancia JTT. El árbol consenso se obtuvo por el método de bootstrap inferido a partir de 500 réplicas. Las ramas mostradas son las obtenidas en al menos un 50% de las réplicas de bootstrap. El análisis incluyó 13 genomas mitocondriales con 12 proteínas cada uno (3276 aminoácidos). Los valores de cada nodo representan los porcentajes de bootstrap.

3.16 Filogenia global de *Echinococcus* y especies modelo

Para construir la filogenia de genomas completos se realizaron tres análisis: filogenia a partir de secuencias de aminoácidos derivados de genes de copia única, filogenia a partir de secuencias de nucleótidos derivados de los CDS de copia única y a partir de secuencias nucleotídicas artificiales generadas a partir la concatenación de sitios polimórficos de genomas completos. Para el análisis de genes de copia única se seleccionaron grupos ortólogos conteniendo una única copia de genes de cada especie y se concatenaron las secuencias de las proteínas y de CDS por especies de organismos. Las secuencias de proteínas se alinearon utilizando ClustalX. El alineamiento múltiple fue editado manualmente eliminando todas las posiciones con menos del 60% de cobertura y se calcularon las matrices de distancia con el modelo JTT y se reconstruyó la filogenia mediante el método de máxima verosimilitud. El análisis involucró 14 secuencias de aminoácidos de 29 genes de copia única extraídos de los grupos de ortología. Los bloques analizados en el conjunto de datos final incluyeron un total de 7001 posiciones. Las secuencias de nucleótidos de los CDS de copia única se alinearon con el software transAlign (EMBOSS), y el análisis filogenético se llevó a cabo por el método de Máxima verosimilitud basado en el modelo de Tamura-Nei. Se eliminaron todas las posiciones que contenían gaps y con datos faltantes y el conjunto de datos final incluyó 14364 posiciones. Para la reconstrucción filogenética a partir de genomas completos se seleccionaron los SNPs homocigotas de los genomas de *Echinococcus* y se integraron todos los sitios polimórficos de las tres especies de *Echinococcus* usando GATK [217]. Después de la eliminación de SNPs heterocigotas, los sitios de SNPs homocigotas de los genomas de las tres especies fueron concatenados formando secuencias de nucleótidos artificiales y el alineamiento resultante se utilizó para crear un árbol filogenético utilizando FastTree [218], se calcularon las matrices de distancia con el modelo GTR y se implementó el método de máxima verosimilitud para la construcción de los árboles. El árbol consenso se obtuvo por el método de bootstrap inferido a partir de 1000 réplicas. El conjunto final de datos incluyó 562759 posiciones. Los sitios polimórficos compartidos entre las especies se identificaron y cuantificaron con scripts personalizados.

3.17 Estudio del uso de codones

Para el estudio del uso de codones de las especies de *Echinococcus* se seleccionaron 2754, 2417 y 2442 CDSs de *E. canadensis* G7, *E. multilocularis* y *E. granulosus* G1 respectivamente. Los mismos fueron seleccionados habiéndose filtrado estrictamente los que contaban con el codón de inicio ATG y con los codones de terminación de la traducción TAA, TAG o TGA, y además ser ortólogos entre las tres especies de *Echinococcus*. Los grupos de genes de cada especie fueron analizados mediante CodonW [219], CodonO [220,221] y el contenido de nucleótidos de cada posición de los codones y la ocurrencia de dinucleótidos GC fue calculado con scripts

personalizados escritos en lenguaje Perl. Los parámetros calculados mediante CodonW y el contenido de nucleótidos de las posiciones en los codones fueron utilizados para realizar análisis de componentes principales (ACP). Previamente los datos fueron normalizados según Peden [219] para definir la adaptabilidad relativa de cada codón. Y luego se aplicó el ACP basado en la adaptabilidad relativa para identificar las principales tendencias de la variación génica en el uso de codones sinónimos entre los genes [222]. Se utilizó el análisis de correspondencia (COA) para el análisis estadístico multivariante de patrones de uso de codones. Dado que hay un total de 59 codones sinónimos (los cual incluye 61 codones, menos los codones Met y Trp únicos), los grados de libertad se reducen a 40 al eliminar las variaciones causadas por el uso desigual de los aminoácidos al generar un análisis de correspondencia de RSCU [219]. Todos los gráficos que relacionan los distintos índices y/o parámetros fueron realizados con R (www.r-project.org) y los cálculos estadísticos se realizaron con el paquete Hmisc (<https://cran.r-project.org/web/packages/Hmisc/>).

3.17.1 Métodos de medición del uso de codones

Los datos obtenidos y medidos con CodonW fueron: uso relativo de codones sinónimos (RSCU), número efectivo de codones (ENc), índice de adaptación de codones (CAI) y contenido de GC en la primera, segunda y tercera posiciones de codones (GC1, GC2 y GC3), la frecuencia de G o C en la tercera posición de codones sinónimos (GC3s), y la media de GC total, CBI, ENc, Fop, Hidrofobicidad (Hf) y aromaticidad (Arom). Para el cálculo de los índices CAI, CBI y Fop se seleccionaron previamente un grupo de genes con alta y baja expresión como referencia. El significado de cada índice se explica a continuación:

Uso relativo de codones sinónimos (RSCU): es la relación de las frecuencias observadas y esperadas de los codones bajo un uso de codones sinónimo uniforme,

$$RSCU_i = \frac{X_i}{\frac{1}{n} \sum_{i=1}^n X_i}$$

con n = número de codones sinónimos cuyos valores ($1 \leq n \leq 6$) según el aminoácido para el que codifica y X_i = número de veces que aparece el codón i . Si los codones sinónimos de un aminoácido se utilizan con frecuencias iguales, sus valores de RSCU serán iguales a 1. Los valores de RSCU mayores que 1 indican un uso de codón superior al esperado [223].

Numero efectivo de codones (ENc) Es una medida general del sesgo en el uso de codones para genes individuales. Su valor representa el número total de codones utilizados que generaran el mismo sesgo de uso de codones que el observado.

$$ENc = 2 + \frac{9}{F_2} + \frac{1}{F_3} + \frac{5}{F_4} + \frac{3}{F_6}$$

donde F_i ($i = 2, 3, 4$ o 6) es la probabilidad media de que dos codones elegidos al azar para un aminoácido codificado por i codones sinónimos sean idénticos.

En un rango de valores de 20 a 61, los valores más bajos indican un mayor sesgo en el uso de codones. En términos generales, los valores de ENc inferiores a 36 indican un fuerte sesgo en el uso de codones [224].

Índice de adaptación de codones (CAI): Los valores de CAI miden el grado de predisposición hacia el uso de codones presentes en genes altamente expresados o de codones preferidos/óptimos. Primero se elabora una tabla de valores RSCU para genes altamente expresados. A partir de dicha tabla se identifican los codones que se utilizan con mayor frecuencia para cada aminoácido y se calcula la adaptabilidad relativa de los codones (w_i) mediante:

$$w_i = \frac{RSCU_i}{RSCU_{max}}$$

donde $RSCU_{max}$ = valor RSCU para el codón utilizado mas frecuentemente para un aminoácido determinado. Luego; el valor CAI de un gen se calcula como la media geométrica de los valores de w_i para todos los codones utilizados en ese gen.

$$CAI = \left(\prod_{i=1}^L w_i \right)^{\frac{1}{L}}$$

donde L = Número de codones, excluyendo a los codones AUG y UGG y w_i es el valor w para el codón i en el gen. En un rango de valores entre 0 y 1.0, los valores CAI más altos indican una mayor expresión de genes [223,225]. El conjunto de secuencias utilizadas para calcular los valores CAI en este estudio fueron los genes que con altos RFPKMs calculadas por Tsai y col. [43].

Índice del sesgo de codones (CBI) es una medida de sesgo de codones direccional hacia un subconjunto de codones óptimos. Mide el grado con el que un gen usa un subgrupo de codones óptimos [160].

$$CBI = \frac{N_{opt} - N_{esp}}{N_{tot} - N_{esp}}$$

donde N_{opt} = número de codones óptimos; N_{tot} = número de codones sinónimos; N_{esp} = número esperado de codones óptimos si los codones se asignaron al azar. Si el

uso de codones es totalmente aleatorio, el valor de CBI es 0. Si Nopt es menor que Nesp, el índice será negativo.

Frecuencia de codones óptimos (Fop): los codones óptimos se definen como los codones que son reconocidos por los ARNt isoaceptores más abundantes en genes altamente expresados, comparados con genes de baja expresión [131].

$$Fop = \frac{N_{\text{codones}_{\text{óptimos}}}}{N_{\text{codones}_{\text{sin óptimos}}}}$$

La significancia se evalúa mediante una prueba de contingencia de chi cuadrado con $p\text{-val} < 0.01$.

Gráfico de número efectivo de codones (ENc-plot). En un ENc-plot los valores de ENc son relacionados contra los valores de GC3s y se utiliza para analizar la influencia de la composición de bases en el uso de codones en un genoma [226]. Se genera una curva estándar para mostrar la relación funcional entre los valores ENc y GC3 bajo influenciados por un sesgo mutacional en lugar de la presión de selección.

En genes donde la presión de selección está totalmente ausente los valores de ENc dependen del contenido de GC3 según la ecuación siguiente:

$$ENc = 2 + s + \frac{29}{s^2 + (1 + s^2)}$$

donde s representa el porcentaje de G + C en la tercer posición de los codones.

Aromaticidad (Arom); es la suma de la frecuencia relativa de la aromaticidad de cada aminoácido.

$$Arom = \sum_{i=1}^{20} \delta_i f_i$$

donde f_i es la frecuencia relativa de los aminoácidos de tipo i de la proteína y $\delta_i = 1$ cuando el aminoácido es aromático (Fenil-alanina, tirosina y triptófano) y $\delta_i = 0$ cuando es no aromático [227].

Hidrofobicidad (Hf): es la estimación de la hidrofobicidad total de la proteína, cuanto mayor es el valor de Hf mayor es su carácter hidrofóbico. El valor de Hf es una combinación lineal de frecuencias relativas de los aminoácidos:

$$Hf = \sum_{i=1}^{20} \alpha_i f_i$$

donde f_i es la frecuencia relativa de aminoácido del tipo i de la proteína y α_i es el índice de hidrofobicidad de dicho aminoácido [223].

Gráfico de sesgo PR2: El análisis de diagrama de la regla de paridad 2 (PR2) se utilizó para determinar el grado mutacional y de selección en el uso de codones [228]. Este análisis se basa en la relación de sesgos de AT ($[A3/ (A3 + T3)]$) y de sesgos GC ($[G3/ (G3 + C3)]$) en la tercera posición de los codones que codifican para dichos aminoácidos. Los aminoácidos de cuatro codones son alanina, arginina (CGA, CGT, CGG, CGC), glicina, leucina (CTA, CTT, CTG, CTC), prolina, serina (TCA, TCT, TCG, TCC), treonina y valina [229].

Gráficos de neutralidad: Los gráficos de neutralidad (P12-P3); se calcularon en base al contenido de GC en la primera y segunda posición (GC12) y en la tercera posición de codones sinónimos GC3s. P12-P3 fueron utilizados para evaluar las relaciones entre las tres posiciones de los codones de *Echinococcus*. Se realiza una regresión lineal entre las tres posiciones y luego se analiza la pendiente. En el análisis de regresión lineal, una pendiente nula indica una ausencia de presión de mutación direccional (fuerzas selectivas completas), mientras que una pendiente de 1 indica neutralidad completa [223,228,230,231].

RESULTADOS



Capítulo IV

Resultados

4.1 Genoma de *Echinococcus canadensis* G7

La disponibilidad de las nuevas técnicas de secuenciación aplicadas a la obtención de datos genómicos y transcriptómicos ha permitido abordar el estudio de enfermedades desatendidas desde una perspectiva diferente. La disponibilidad de los genomas de parásitos y las anotaciones génicas de los mismos nos ha permitido realizar estudios que no habían podido llevarse a cabo mediante las técnicas clásicas. Sin embargo, debido al gran volumen de datos que estas tecnologías generan es relevante su cuidadosa manipulación y más aún la correcta interpretación de los resultados pertinentes. La obtención del genoma de *E. canadensis* G7 y los análisis y resultados obtenidos en la presente tesis son pioneros en el área de genómica asociada al estudio de parásitos platelmintos.

4.1.1 Secuenciación y ensamblado

La extracción del ADN genómico de protoescolices de *E. canadensis* G7 aislados de los quistes de hígado de cerdo (*Sus scrofa*) fueron genotipificados por PCR del gen mitocondrial citocromo oxidasa 1 (COX1) seguido de secuenciación. La secuencia obtenida fue 100% idéntica a la reportada para *E. canadensis* G7 (nro de acceso GenBank AB235847).

4.1.1.1 Preparación de bibliotecas de ADN genómico

Para la secuenciación del genoma de *E. canadensis* G7 se prepararon dos bibliotecas de ADN genómico *paired end*. Para ello, el ADN fue sometido a fragmentación por transposición y adición de adaptadores en cada extremo del ADN doble cadena (*tagmentation*) obteniéndose fragmentos de 1050 y 1200 pb para la primera y segunda biblioteca respectivamente. Posteriormente se secuenciaron 250 pb y 300 pb de cada extremo de los fragmentos ([Figura IV.1](#)).

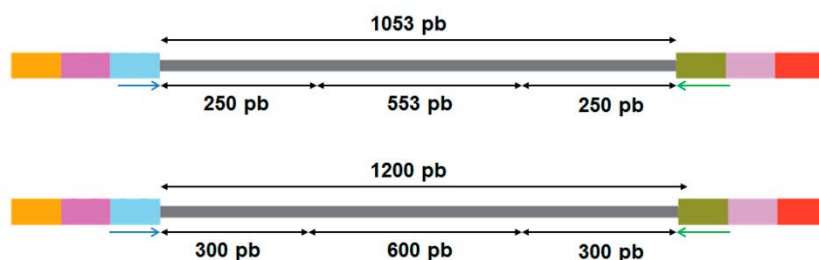
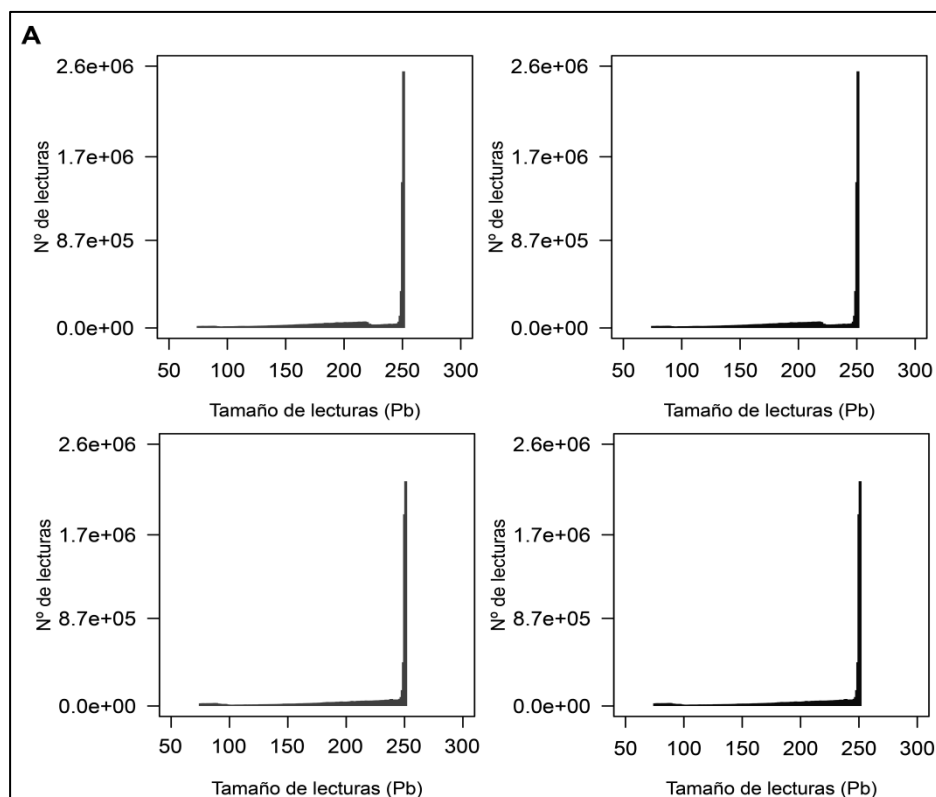
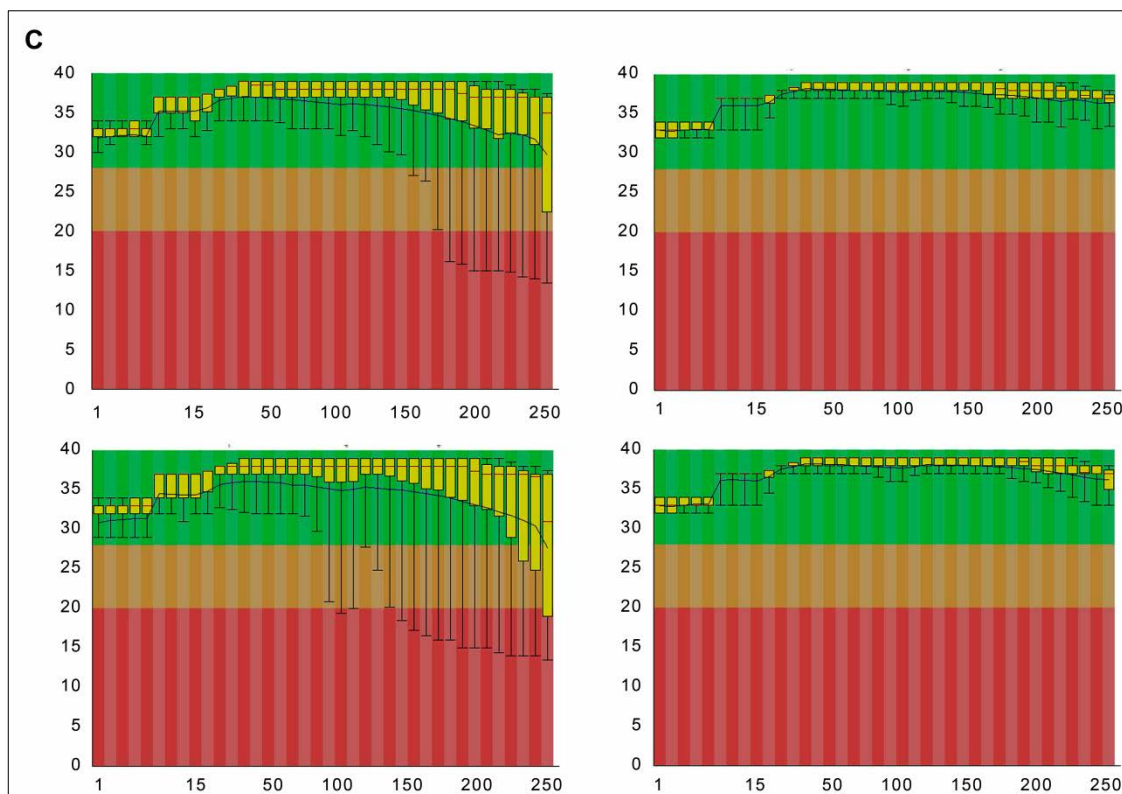
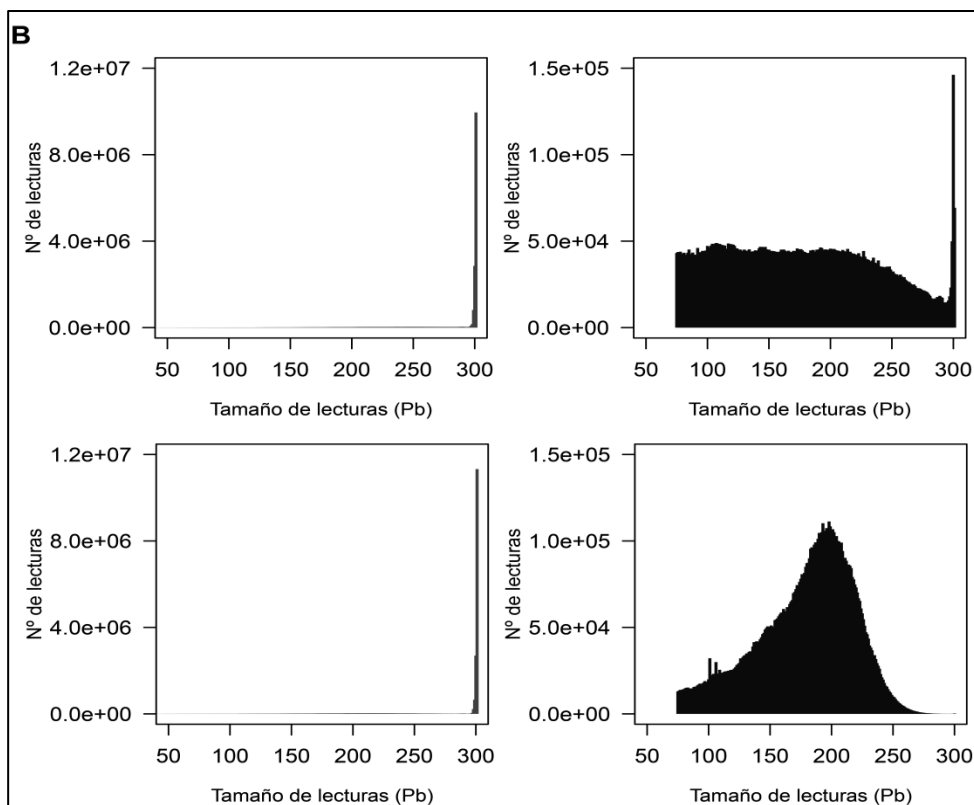


Figura IV.1: Esquema de los fragmentos obtenidos en la construcción de las dos bibliotecas *paired end* para la secuenciación. Las flechas azules y verdes representan a los cebadores para la secuenciación seguido de las flechas negras que representan el tamaño de lectura que se secuenció en cada extremo.

4.1.1.2 Secuenciación y análisis de calidad

En la secuenciación de las bibliotecas genómicas se produjeron lecturas de secuencia codificadas en formato *fastq*. Para la primera biblioteca se obtuvieron 2x16644529 lecturas con tamaños de entre 35 y 251 pb, y luego del proceso de filtrado por calidad quedaron 2x10549520 lecturas con tamaños entre 75 y 251 pb. Para la segunda biblioteca se obtuvieron 2x18836422 lecturas que luego de filtrarlas por calidad se obtuvieron 2x9015899 lecturas de tamaños entre 75 y 300 pb. El rendimiento de la secuenciación fue del 63,4% y 47,9% para la primera y segunda biblioteca genómica respectivamente ([Figura IV.2](#)).





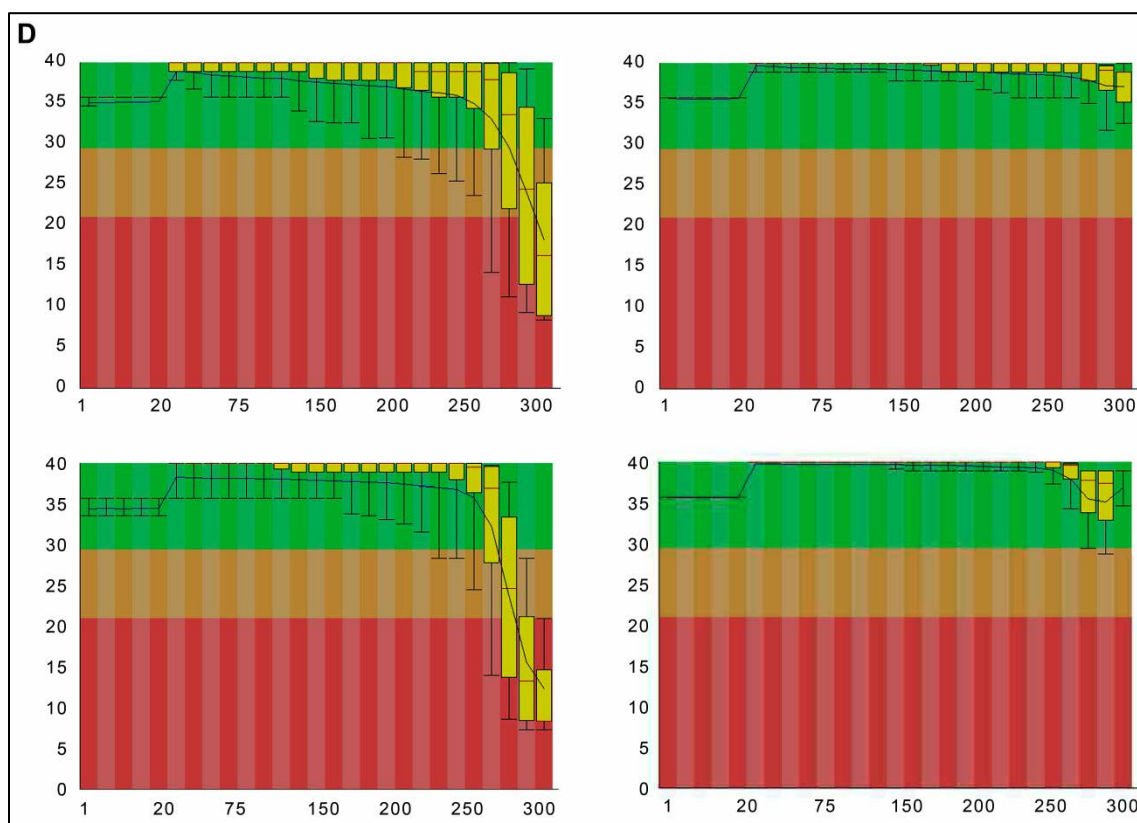


Figura IV.2: Distribución de tamaño de las lecturas: A). Número de lecturas vs. tamaño de las lecturas de la primera biblioteca para el total de secuencias antes del procesamiento y filtrado por calidad (gris) y después del procesamiento y filtrado por calidad (negro). B). Número de lecturas vs tamaño de las lecturas de la segunda biblioteca para el total de secuencias antes del procesamiento y filtrado por calidad (gris) y después del procesamiento por calidad (negro). C). Calidad de bases en escala *Phred* por posición en las lecturas para la primera biblioteca antes del procesamiento y filtrado por calidad (izquierda) y después del procesamiento y filtrado por calidad (derecha). D). Calidad de bases en escala *Phred* por posición en las lecturas para la segunda biblioteca antes del procesamiento y filtrado por calidad (izquierda) y después del procesamiento y filtrado por calidad (derecha).

4.1.1.3 Ensamblado

Para obtener el genoma de *E. canadensis* G7 se plantearon dos alternativas; el mapeo de lecturas utilizando el genoma de *E. multilocularis* como genoma de referencia o mediante la realización de un ensamblado *de novo*. En primer lugar se realizó el ensamblado utilizando el genoma de referencia y se obtuvo un porcentaje de lecturas mapeadas del 76,1%. Debido a los *gaps* que presenta el genoma de *E. multilocularis* y ante la posibilidad de perder información genómica específica de *E. canadensis* G7 dado el sesgo que representa el mapeo de lecturas sobre un genoma de referencia existente y recientemente generado se decidió realizar un ensamblado *de novo*. Para obtener el ensamblado de mejor calidad se ensayaron 34 estrategias diferentes usando cinco softwares de ensamblado. El ensamblado óptimo se eligió según las medidas estadísticas del genoma tales como el N50, la profundidad, el número de *contigs*, el %GC, la cobertura del genoma de referencia y la evaluación del espacio génico (Tabla IV.1). El mapeo de las lecturas sobre el ensamblado *de novo* del genoma de *E. canadensis* G7 mostró un 92,1% de lecturas mapeadas (ver sección 4.2.2).

Tabla IV.1: Optimización del ensamblado del genoma de *E. canadensis* G7. Resumen de métricas para los ensamblados obtenidos por diferentes softwares utilizados en la obtención del genoma. Sólo se consideran *contigs* > 1000 pb. DBG - Gráfos de Bruijn; PDBG- Gráfos de Bruijn paralelos iterados.

Ensamblador	K-mer	Nº de contigs	Tamaño total [Mb]	Mayor contig [kb]	N50 [pb]	Gaps N [kb]	Observaciones
SOAPDeNovo	23	257453	114,06	119,65	10667	11056	*
SOAPDeNovo	23	4408970	408,26	3,34	147	3212	*
SOAPDeNovo	31	4108901	484,80	3,50	173	416	*
SOAPDeNovo	31	145612	113,50	181,72	24061	6120	*
SOAPDeNovo	37	4108901	484,80	3,50	173	416	*
SOAPDeNovo	37	145612	113,50	181,72	24061	6120	*
SOAPDeNovo	41	4108901	484,80	3,50	173	416	*
SOAPDeNovo	41	145612	113,50	181,72	24061	6120	*
SOAPDeNovo	127	848909	140,16	95053,29	4471	0	*
SOAPDeNovo	127	4408893	408,26	9208,98	147	0	*
SOAPDeNovo	137	4108447	484,78	8481,15	173	0	*
SOAPDeNovo	137	594270	139,80	100168,33	9207	0	*
SOAPDeNovo	149	3977274	529,62	8019,60	188	0	*
SOAPDeNovo	149	504547	140,12	101437,01	11428	0	*
Velvet	23	3397665	304,63	2,53	94	0	*
Velvet	31	75736	110,64	180,18	23120	0	*
Velvet	37	2144947	295,29	2,77	141	0	*
Velvet	41	1908241	289,92	3,59	111	0	*
Velvet	31	57418	111,15	436,06	49829	363	*
Velvet	37	57440	111,14	436,06	50099	360	*
Velvet	31	74637	110,58	202,27	23416	0	*

Velvet	31	74596	110,58	202,27	23416	0	*
Velvet	31	75003	110,87	202,27	24043	0	*
Velvet	31	74637	110,58	202,27	23416	0	*
Velvet	31	74596	110,58	202,27	23416	0	*
Velvet	31	75053	110,87	202,27	24043	0	*
Velvet	31	73403	110,92	202,27	24061	0	*
Velvet	31	75003	110,87	202,27	24043	0	*
Velvet + IMAGE	31	15367	104,40	202,27	26296	0	**
Velvet + IMAGE	31	15367	104,40	202,27	26296	0	***
Velvet + IMAGE	31	10277	96,54	202,27	28388	0	***
SPAdes	29,31,37,41,43,47,53	41665	128,51	573,99	60762	0,007	
SPAdes + IMAGES	29,31,37,41,43,47,53	34984	127,31	573,99	62416	0	
SPAdes + IMAGES	29,31,37,41,43,47,53	9333	115,02	573,99	74555	0	****
SPAdes +IMAGES	29,31,37,41,43,47,53	9326	115,02	573,99	74555	0	*****

*Distintos parámetros de software; **Con ABACAS; ***Sin ABACAS; ****Con *contigs* mitocondriales; *****Genoma mitocondrial removido

El mejor ensamblado se logró utilizando el software SPAdes [170]. Luego de remover los *contigs* que potencialmente no corresponden a *E. canadensis* G7 y los menores a 1000 pb, el genoma quedó compuesto por 9326 *contigs* cuyos parámetros de calidad del ensamblado fueron los siguientes: N50 74,6 kb, profundidad 55X y una distribución y longitud de *contigs* similar a la de *E. multilocularis* y *E. granulosus* G1. El ensamblado del genoma de *E. canadensis* G7 exhibió un tamaño de 115 Mb con 41,86% de contenido GC y un tamaño acumulativo de 9000 *contigs* de *E. canadensis* G7 que alcanzaron el tamaño total del genoma de *E. multilocularis* ([Figura IV.3](#)).

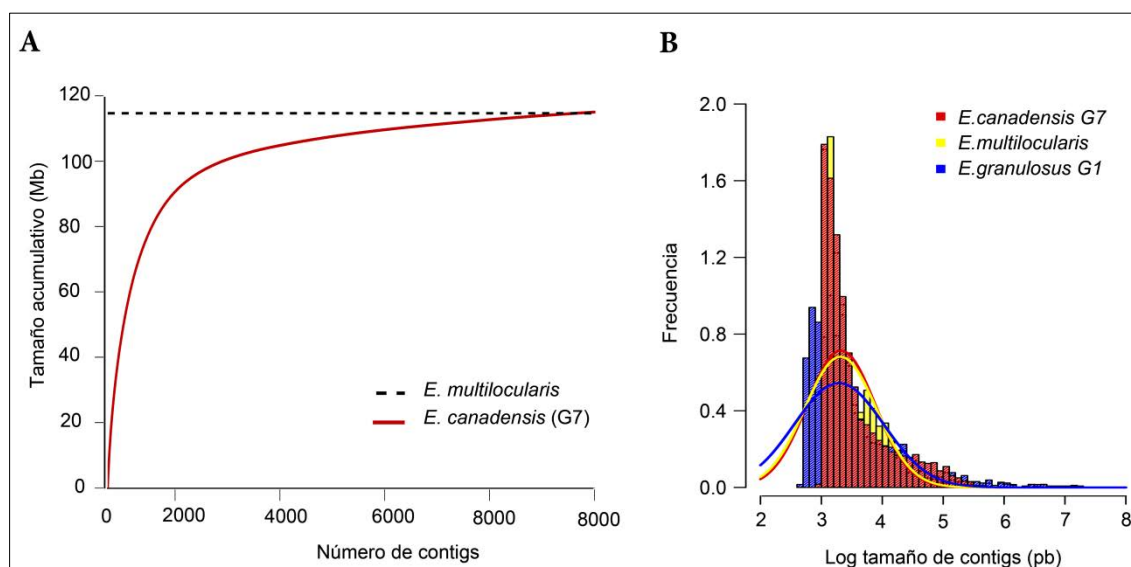


Figura IV.3: Estadísticas métricas de la calidad del ensamblado del genoma de *E. canadensis* G7: A). Tamaño acumulativo de los *contigs* (Mb) por el número de *contigs* usando al genoma ensamblado de *E. multilocularis* como referencia, B). Distribución de tamaño de *contigs* de *E. canadensis* G7. Los histogramas representan la frecuencia de *contigs* en función del logaritmo del tamaño de los *contigs* (pb) para cada genoma. Las curvas representan la distribución normal del logaritmo de tamaño de los *contigs*.

Posteriormente se evaluó el genoma mediante el análisis del espacio génico que debería contener un todo organismo eucariota. Este análisis se realizó mediante la búsqueda de 248 genes centrales (CEGs) que forman parte de todos los organismos pertenecientes a este grupo utilizando el software “Core Eukaryotic Genes Mapping Approach” (CEGMA) [176]. De este modo se estimó que el espacio génico fue del 85.08% y el porcentaje de cobertura fue del 95%. Adicionalmente, se ensambló el genoma mitocondrial se identificaron y anotaron el 95% de sus genes. El ensamblado consistió en un *contig* de 13719 pb con 85% de cobertura y 99,9% de identidad respecto al genoma mitocondrial de referencia de *E. canadensis* G7 (Nro. de acceso del GenBank AB235847 ([Figura IV.4](#) y [Tabla IV.2](#)).

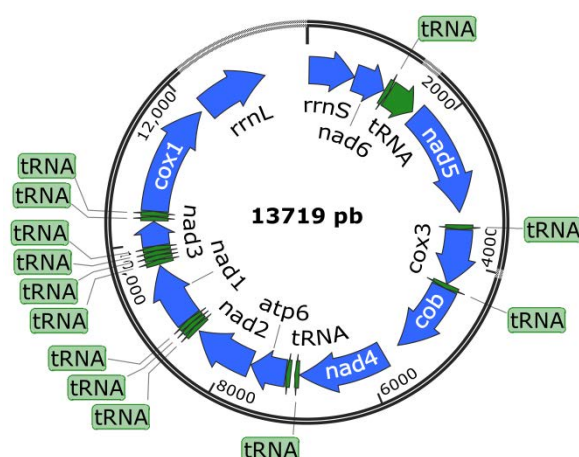


Figura IV.4: Esquema del ensamblado del genoma mitocondrial de *E. canadensis* G7.

Tabla IV.2: Genes mitocondriales del genoma de *E. canadensis* G7.

Contig_ID	Clase de gen	Inicio	Final	Anotación
E.canG7_Mit.Genome	gen	33	692	rrnS
E.canG7_Mit.Genome	gen	34	609	cox2
E.canG7_Mit.Genome	ARNt	625	691	ARNt-Glu
E.canG7_Mit.Genome	gen	695	1150	nad6
E.canG7_Mit.Genome	ARNt	1163	1228	ARNt-Tyr
E.canG7_Mit.Genome	ARNt	1235	1661	ARNt-Leu
E.canG7_Mit.Genome	ARNt	1318	1759	ARNt-Ser
E.canG7_Mit.Genome	ARNt	1580	1644	ARNt-Leu
E.canG7_Mit.Genome	ARNt	1656	1714	ARNt-Arg
E.canG7_Mit.Genome	gen	1716	3287	nad5
E.canG7_Mit.Genome	ARNt	3466	3531	ARNt-Gly
E.canG7_Mit.Genome	gen	3536	4336	cox3
E.canG7_Mit.Genome	ARNt	4338	4401	ARNt-His
E.canG7_Mit.Genome	gen	4405	5472	cob
E.canG7_Mit.Genome	gen	5488	5748	nad4L
E.canG7_Mit.Genome	gen	5709	6968	nad4
E.canG7_Mit.Genome	ARNt	6972	7033	ARNt-Gln
E.canG7_Mit.Genome	ARNt	7033	7095	ARNt-Phe
E.canG7_Mit.Genome	ARNt	7092	7157	ARNt-Met
E.canG7_Mit.Genome	gen	7163	7675	atp6
E.canG7_Mit.Genome	gen	7684	8565	nad2
E.canG7_Mit.Genome	ARNt	8593	8655	ARNt-Val
E.canG7_Mit.Genome	ARNt	8661	8725	ARNt-Ala
E.canG7_Mit.Genome	ARNt	8726	8790	ARNt-Asp
E.canG7_Mit.Genome	gen	8794	9668	nad1
E.canG7_Mit.Genome	ARNt	9687	9752	ARNt-Asn
E.canG7_Mit.Genome	ARNt	9759	9821	ARNt-Pro
E.canG7_Mit.Genome	ARNt	9822	9883	ARNt-Ile
E.canG7_Mit.Genome	ARNt	9889	9952	ARNt-Lys
E.canG7_Mit.Genome	gen	9955	10302	nad3
E.canG7_Mit.Genome	ARNt	10314	10372	ARNt-Ser
E.canG7_Mit.Genome	ARNt	10381	10447	ARNt-Trp
E.canG7_Mit.Genome	gen	10451	12058	cox1
E.canG7_Mit.Genome	ARNt	12053	12136	ARNt-Thr
E.canG7_Mit.Genome	gen	12137	13114	rrnL

4.1.2 Anotación Génica

El proceso de anotación génica puede agruparse en dos componentes: anotación estructural y anotación funcional [232]. Estructuralmente los genes deben ser asignados a regiones específicas dentro del genoma dispuesto en *contigs*, *scaffolds* o cromosomas y adicionalmente describir la estructura de cada gen, su marco de lectura incluyendo los intrones, exones, sitios de inicio y de terminación de la traducción como así también las regiones no traducidas que se encuentran al principio y al final de cada unidad transcripcional (5'UTR y 3'UTR). En esta tesis también se realizó un análisis adicional y poco frecuente en los trabajos de rutina de anotación de genomas que es la identificación de otros elementos no codificantes tales como elementos repetidos, microARNs, entre otros y un esfuerzo adicional en la identificación de regiones regulatorias que permitan proporcionar información sobre la regulación de la expresión de los genes anotados. Por otra parte, la anotación funcional de los genes consiste en que, una vez definida la estructura y los límites de los genes, es posible evaluar dicha región, compararla con bases de datos de genes y realizar búsqueda de dominios que permitan asignar posibles funciones a cada gen. Encontrar los genes completos, detallar su ubicación, asignarle una posible función y entender su relación con otros genes es el punto de partida para el estudio funcional de los genes de interés.

4.1.2.1 Anotación Estructural

La anotación estructural génica de un genoma es un procedimiento delicado en el que deben tenerse en cuenta varios factores para la correcta definición de los genes ya que de ello depende la subsiguiente anotación funcional. Una anotación automática puede resultar defectuosa si no se cuenta previamente con un set de entrenamiento confiable y bien definido que aporte evidencia sustentable a los algoritmos de predicción de los modelos génicos, y si bien no es excluyente es sumamente determinante que se cuente con datos de transcripción para una correcta definición del espacio génico. La anotación estructural génica del genoma de *E. canadensis* G7 se llevó a cabo en una primera instancia utilizando como modelos a *C. elegans* y *S. mansoni* lo que condujo a la anotación de aproximadamente unos 500 y 1000 genes respectivamente. Esto evidenció la necesidad de no contemplar estas anotaciones y generar un set de entrenamiento personalizado para la anotación génica de *E. canadensis* G7. Los modelos génicos finales fueron predichos utilizando el software MAKER2 [177] con algoritmos entrenados a partir de genes de *E. multilocularis* y *E. granulosus* G1 (CDSs, ESTs y proteínas) y datos transcriptómicos de *E. canadensis* G7 (véase capítulo de Materiales y Métodos). Luego de tres rondas de anotación y un filtrado por tamaño de proteínas mayores o iguales a 10 aminoácidos se logró anotar un total de 11435 genes que codifican para proteínas con un puntaje de AED < 0,5 y se obtuvo las coordenadas de los mismos en los *contigs*. Se evaluó la región codificante y los elementos que lo componen y se observó una densidad de 13 genes por Mb y un contenido de GC de la región codificante del 48% ([Tabla IV.3](#)).

Tabla IV.3: Predicción génica del genoma de *E. canadensis* G7.

Parámetros estadísticos del genoma	
Tamaño del genoma (Mb)	115
Contenido de GC (%)	42
Número de <i>contigs</i>	9261
N50 (kb)	75
Mayor <i>contig</i> (kb)	574
Profundidad	55x
Número de genes	11435
Densidad génica por Mb	13
Contenido de GC codificante (%)	48%
Tamaño del proteoma (aminoácidos)	4915068
Tamaño máximo de proteína (aminoácidos)	7886
Tamaño promedio de las proteínas (aminoácidos)	440
Tamaño promedio de los exones (pb)	219
Mediana del tamaño de los exones (pb)	159
Promedio de exones por transcripto	6
Mediana de exones por transcripto	4
Total del tamaño de los intrones (kb)	40117
Tamaño promedio de los intrones (pb)	714
Mediana del tamaño de los exones (pb)	273

4.1.2.2 Anotación Funcional

Del total de 11435 proteínas predichas por técnicas bioinformáticas se obtuvo la anotación funcional para 6876 de ellas, lo cual representó el 60,1% de las proteínas totales. Para lograr este rendimiento en la anotación de los genes fue necesario recurrir a diferentes estrategias. Principalmente se utilizó InterProScan mediante el cual se asignaron 6205 anotaciones, 5986 de ellas asociadas a términos GO (Gene Ontology terms) mediante la combinación de InterPro2GO y Blast2GO. Adicionalmente, 493 genes pudieron ser rescatados mediante la construcción de grupos de ortología conformados por grupos taxonómicos definidos de parásitos y grupos específicos de cestodos (véase sección 3.9) y se les asignó la anotación correspondiente a dichos grupos. Finalmente, se pudieron rescatar 144 anotaciones génicas adicionales usando la técnica de mejor hit recíproco (RBH) por BLAST contra la base de datos UniProt. Por otra parte, las 4559 proteínas restantes fueron anotadas como proteínas hipotéticas y presentaron también evidencia transcripcional dada por el mapeo de lecturas de RNA-seq. A partir del total de las anotaciones se analizaron las frecuencias de los genes con términos GO agrupados en las categorías “Función Molecular” (Mf), “Proceso biológico” (Pb) y “Componente celular” (Cc). Dichas categorías describen la actividad molecular de los productos génicos, el rol biológico en el que intervienen y la localización de la proteína respectivamente. Mediante este estudio se analizó el primer nivel ancestral de cada categoría y se calcularon los porcentajes de términos GO

asociados a los genes. Cada gen anotado tiene asociado al menos un término GO y el porcentaje de términos GO de cada categoría se calculó sumando el total de genes agrupados en un determinado término GO relativo al número total de términos GO de dicha categoría. Se observó que el 77% de los términos GO de la categoría Fm se distribuyeron en dos subcategorías, “Unión” (GO:0005488) (4913 genes) y “Actividad catalítica” (GO:0003824) (2476 genes) ([Figura IV.5 A](#)). Las siguientes subcategorías más abundantes fueron "actividad transportadora" (GO:0005215) (523 genes) y "regulador de función molecular" (GO:0098772) (402 genes). En cuanto a la categoría Pb, el 51,2% del total de términos GO se agrupó en cinco subcategorías "proceso celular" (GO:0009987) (5106 genes), "proceso metabólico" (GO:0008152) (3998 genes), "regulación biológica" (GO:0065007) (3552 genes), "regulación del proceso biológico" (GO:0050789) (3383 genes) y “respuesta a estímulo” (GO:0050896) (2621 genes) ([Figura IV.5 B](#)). Por último, en la categoría Cc se observó que el 73% de los términos GO se agruparon en cuatro subcategorías "célula" (GO:0005623) (5314 genes), "organela" (GO:0043226) (4673 genes), “complejo macromolecular” (GO:0032991) (2528 genes) y “membrana” (GO:0016020) (2213 genes); ([Figura IV.5 C](#)).”. Estos datos concuerdan con los observados en las anotaciones de otros parásitos platelmintos [233].

Para entender mejor la variabilidad génica agrupada en las dos categorías más abundantes de Fm se estudió la distribución de los genes al nivel de la segunda jerarquía de estos términos GO (“Unión” y “Actividad catalítica”). Asociados al término “Unión” se encontraron cuatro subcategorías principales que representaron el 72% del total de los términos GO. Estas categorías fueron "unión de compuestos cíclicos orgánicos" (GO:0097159) (2194 genes), "enlace de compuesto heterocíclico" (GO:1901363) (2173 genes), "unión a proteínas" (GO:0005515) (2044 genes) y "unión a iones" (GO:0043167) (1826 genes). El número de genes totales en estas cuatro categorías fue de 3625, dado que muchos de estos genes presentan función molecular asociada a dos o más términos GO ([Figura IV.5 D](#)). Por otra parte; dentro de la categoría “Actividad catalítica” solo dos subcategorías conformaron el 83% de los términos GO; "actividad hidrolasa" (GO:0016787) (1089 genes) y "actividad transferasa" (GO:0016740) (1054 genes). En este caso, el número total de genes fue de 2039 indicando que la mayoría de los genes posee un solo tipo de actividades catalítica ([Figura IV.5 E](#)).

Al analizar la combinación de estas dos subcategorías pertenecientes a la categoría “Actividad catalítica” con las cuatro subcategorías de la categoría “Unión” se observó que aproximadamente al 25% de los genes contenidos en cada subcategoría perteneciente a “Unión” le corresponde además los términos “Actividad hidrolasa” o “Actividad transferasa”, mientras que el 50% restante no posee términos de actividad catalítica o posee otros términos asociados. Asimismo se observó que los genes agrupados con función “Actividad hidrolasa” y “Actividad transferasa” poseen asociados al menos dos terminos de la subcategoría “Unión”. Estos análisis confirmaron la correcta asignación de términos GO a los genes dada la presencia necesaria de términos GO de ambas categorías (“Unión” y “Actividad catalítica”).

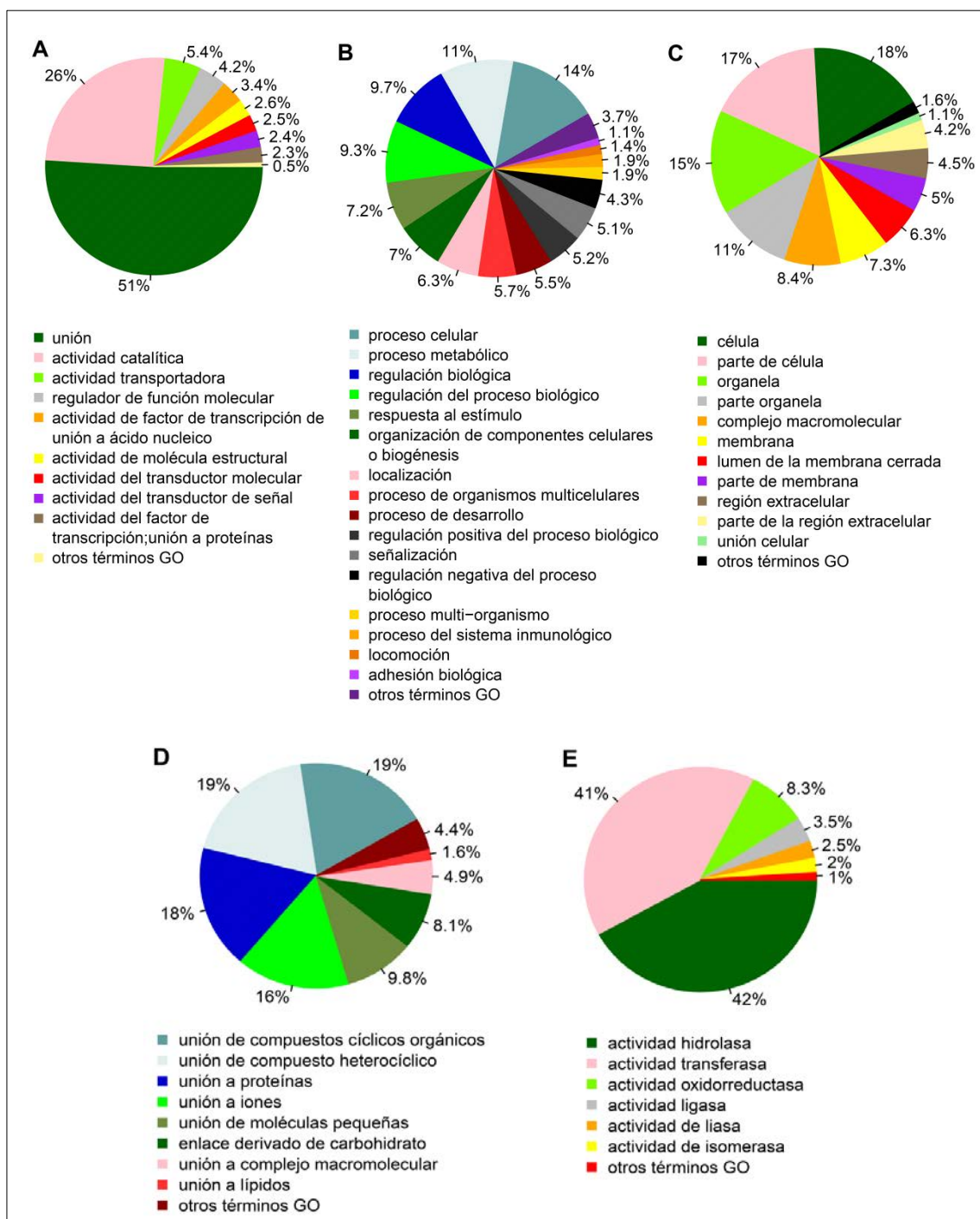


Figura IV.5: Anotación funcional de los genes de *E. canadensis* G7. Porcentaje de términos GO asociados a la categoría A). "Función Molecular", B). "Proceso biológico"; C). Componente celular. D). Porcentaje de términos GO de la categoría "Unión"; E). Porcentaje de términos GO de la categoría "Actividad catalítica". Las categorías menores a 1% fueron sumadas y se muestran como "otros términos GO".

Adicionalmente un total de 4202 proteínas de *E. canadensis* G7 fueron asignadas a una o más identificadores KO (KEGG (*Kyoto Encyclopedia of Genes and Genomes Orthology*) clasificados en 31 categorías de las principales 5 vías metabólicas KEGGs. A partir de ello se encontraron 656 términos KO asociados con metabolismo agrupados en 11 vías metabólicas. La distribución de genes en las distintas vías metabólicas KEGGs puede observarse en la [tabla IV.4](#).

Tabla IV.4: Distribución de genes de *E. canadensis* G7 en las principales vías metabólicas KEGGs.

Vías metabólicas	<i>E. canadensis</i> (G7)		
	Nº de KO único	Nº de KO	Nº de Genes
Proceso celular	488	698	1047
Desarrollo y muerte celular	134	253	360
Motilidad celular	47	47	78
Comunidad celular	132	176	325
Transporte y Catabolismo	175	189	284
Procesamiento de la información ambiental	466	1044	1610
Transporte de membrana	12	11	21
Transducción de señales	407	944	1520
Moléculas de señalización e interacción	47	47	69
Sistemas Organísmicos	737	1467	2354
Sistema circulatorio	56	75	153
Desarrollo y muerte celular	62	69	123
Sistema digestivo	55	93	199
Sistema endocrino	174	382	706
Adaptación ambiental	44	44	76
Sistema excretorio	56	65	113
Sistema inmune	130	262	444
Sistema nervioso	121	253	419
Sistema sensorial	39	52	121
Metabolismo	656	1838	1268
Metabolismo de amino acidos	70	109	130
Biosíntesis de otros metabolitos secundarios	9	13	19
Metabolismo de carbohidratos	116	199	312
Metabolismo de energía	99	101	126
Biosíntesis de Glicano y metabolismo	76	105	159
Metabolismo de lípidos	76	103	158
Metabolismo de cofactores y vitaminas	54	54	71
Metabolismo de otros amino acidos	31	33	46
Metabolismo de terpenoides y policétidos	19	18	19
Metabolismo de ácidos nucleicos	84	124	178
Biodegradación de xenobioticos y metabolismo	22	26	50
Procesamiento de la información genética	709	809	927
Plegado, estructuración y degradación	219	233	279
Replicación y reparación	77	116	137
Transcripción	142	141	162
Traducción	271	289	349
Total	3616	7666	8590

En *E. canadensis* G7 se encontraron aproximadamente el 80% de las vías metabólicas KEGGs esperadas para el género *Echinococcus* [43]. Vías tales como la síntesis *de novo* de bases nucleotídicas, amino ácidos y lípidos parecen estar ausentes lo cual es acorde con los resultados obtenidos en trabajos anteriores descritos por Zhang y col. y Tsai y col [43,78] reforzando la hipótesis de la necesidad importar metabolitos del hospedador para la supervivencia de los parásitos del género *Echinococcus*. La asignación de los genes a las vías metabólicas KEGGs permitió llevar a cabo el estudio de algunas vías particulares y asociarlas a los análisis de variabilidad genética en las siguientes secciones. Por otra parte, en relación a los genes que no codifican proteínas de *E. canadensis* G7 se identificaron 39 genes de microARNs previamente descritos en nuestros trabajos [85,121,122]. También se identificaron 10 genes que codifican para ARN ribosómico (ARNr): cinco 18S, tres 5.8S y dos 28S. Adicionalmente se identificaron 124 genes que codifican para ARNs de transferencia (ARNt), con porcentajes de identidad que van desde el 76,8% al 100% respecto a otras secuencias de ARNt descritas en bases de datos [234]. Se observó además que la disposición de dichos genes consiste en 21 clústers de genes de ARNt con hasta 6 genes por clúster. Este tipo de organización también se ha descrito en genomas de nematodos y mamíferos [133].

4.1.3 Conclusiones parciales

En este trabajo se secuenció y se ensambló el genoma de *E. canadensis* G7, uno de los principales responsables de la echinococcosis en el mundo. El genoma se obtuvo a partir de ADN genómico de protoescolices obtenidos de un único quiste hídático de cerdo. El tamaño de su genoma fue de 115 Mb, y se encuentra en el orden del tamaño de los genomas de otros cestodos. La calidad del genoma basado en las métricas estadísticas y en el espacio génico posicionan a este genoma a un nivel de calidad por encima de la media de muchos genomas de parásitos secuenciados hasta el momento.

La anotación estructural génica realizada demostró que *E. canadensis* G7 contiene 11435 genes que codifican para proteínas. Dicho número está en el orden de sus pares del género *Echinococcus*; *E. multilocularis* (10552 genes) y *E. granulosus* G1 (10274 genes). La evaluación estadística de la anotación estructural de sus modelos génicos poseen valores que representan una alta confiabilidad para dichos genes y que a su vez están sustentados con evidencia transcripcional. Por otra parte, mediante la implementación de distintos métodos para la anotación funcional de los genes de *E. canadensis* G7 se asignaron términos GO al 60,1% del total de los genes cuyas categorías se distribuyeron acorde a lo esperado para el género *Echinococcus*. El mayor porcentaje de los genes está relacionado con la actividad hidrolasa y la actividad transferasa y al mismo tiempo poseen propiedades de unión a compuestos orgánicos cíclicos, heterocíclicos y a proteínas y también una combinación de estas características sugiriendo un rol conjunto en la función molecular. Asimismo, se identificaron elementos no codificantes para proteínas tales como los microARNs y ARNt. Para este último se determinó que los genes que codifican para ARNt disponen de una

organización en clústeres en el genoma de *Echinococcus*. Por otra parte la identificación de los microARNs de *E. canadensis* G7 permitió abordar estudios comparativos y de regulación de la expresión cuyos resultados se detallan en la sección 4.5.3).

4.2 Genómica comparativa de tres especies de *Echinococcus*

La genómica comparativa es un enfoque a gran escala que compara dos o más genomas para estudiar sus similitudes y diferencias con el fin de obtener información biológica significativa en los organismos de interés. Los estudios genómico comparativos llevados a cabo en esta tesis están enfocados en la comparación de tres especies de *Echinococcus*: *E. canadensis* G7, *E. multilocularis* y *E. granulosus* G1 y la descripción de la estructura de los genomas, organización del genoma a nivel de ADN y de genes, sintenia, regiones codificantes y genes ortólogos.

4.2.1 Ortología y Organización génica de *Echinococcus*

La disponibilidad de la versión v.4 del genoma de *E. multilocularis* dispuesto en cromosomas permitió el análisis por homología de las secuencias de los *contigs* de *E. canadensis* G7 con las secuencias de los cromosomas de *E. multilocularis*. Mediante este procedimiento los *contigs* de *E. canadensis* G7 y *E. granulosus* G1 se ordenaron en cromosomas artificiales basados en los cromosomas de *E. multilocularis*. El mapeo y la organización de los *contigs* de *E. canadensis* G7 sobre los cromosomas de *E. multilocularis* mostraron un porcentaje de identidad entre el 89,1% y 93,1% y un porcentaje de cobertura que va desde el 77,1% al 96,0% (Figura IV.6).

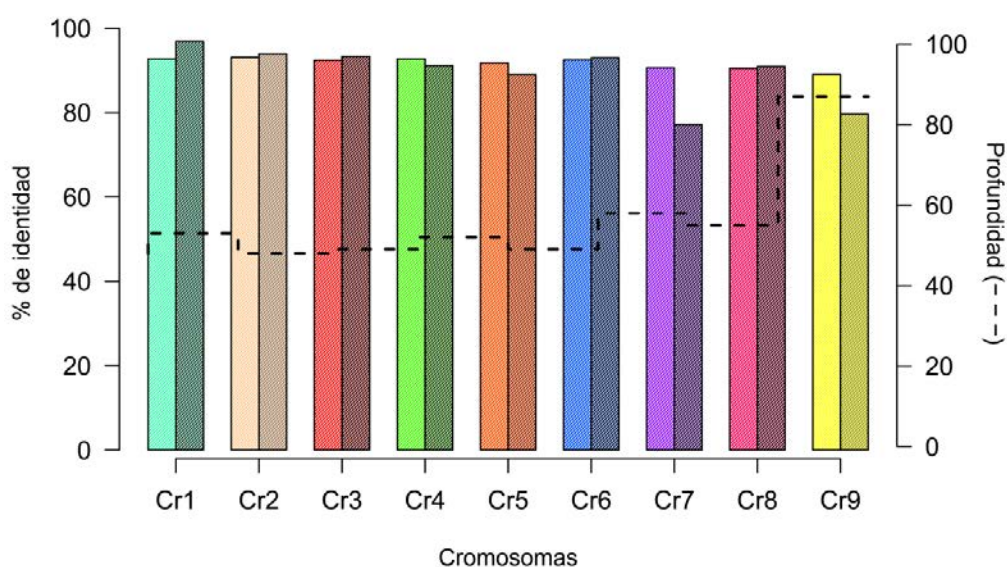


Figura IV.6: Comparación genómica total entre los cromosomas de *E. canadensis* G7 y *E. multilocularis* basados en los análisis del porcentaje de identidad (barras de la izquierda), porcentaje de cobertura (barras de la derecha) y cobertura en profundidad (línea punteada).

Una vez definidos los cromosomas de *E. canadensis* G7 y *E. granulosus* G1 se re-transfirió la anotación génica desde sus versiones dispuestas en *contigs* para analizar el grado de sintenia entre los genomas de las tres especies de *Echinococcus*. Para determinar el orden génico en las especies de *Echinococcus* se identificaron grupos de genes ortólogos uno a uno entre pares de especies de *E. multilocularis*, *E. canadensis* G7 y *E. granulosus* G1 utilizando el software OrthoMCL *gene clustering* (véase capítulo de Materiales y Métodos). Asimismo, se observó un alto grado de conservación en la estructura y organización de los genes entre las tres especies de *Echinococcus*. La relación entre los genes ortólogos puede observarse esquematizada en la [Figura IV.7](#).

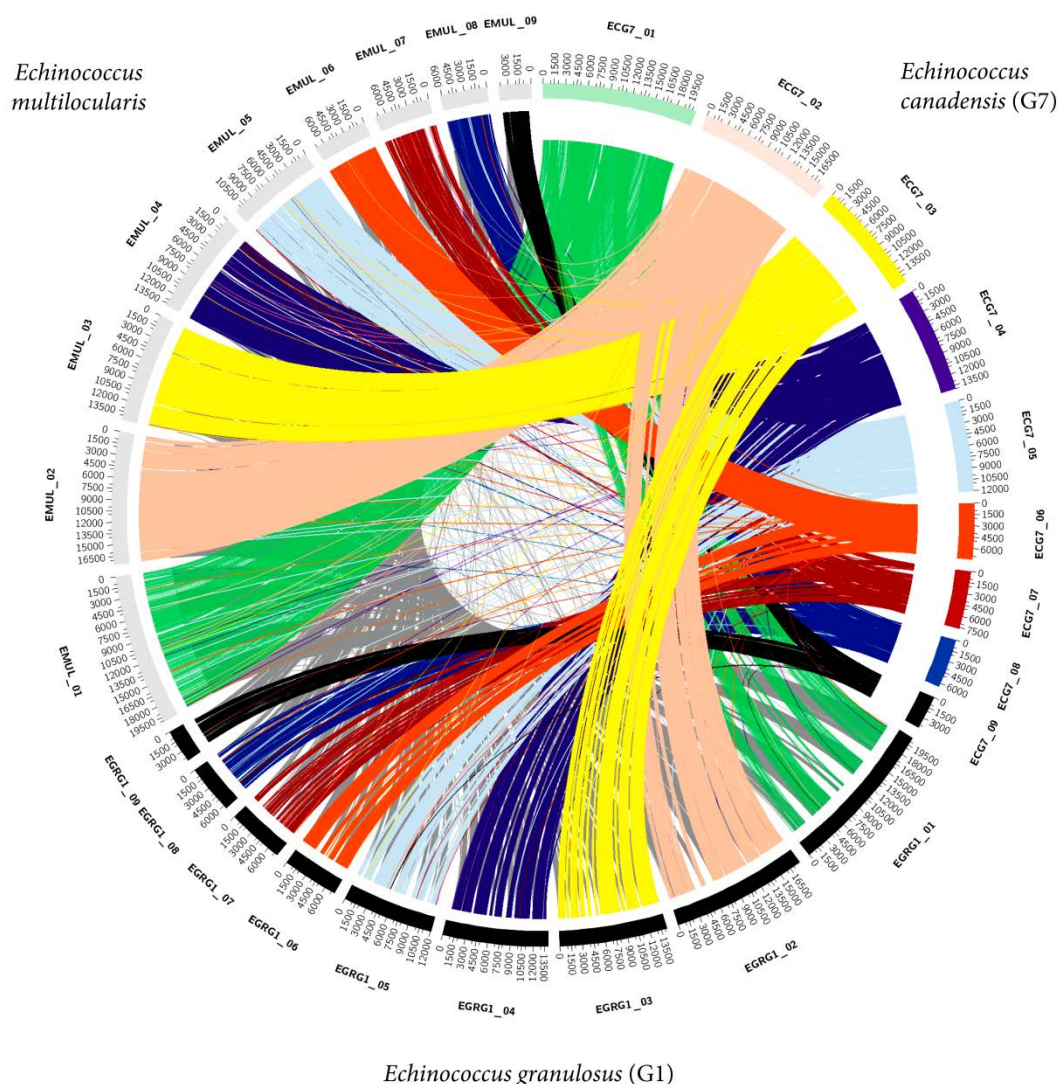


Figura IV.7: Representación esquemática de la sintenia entre los cromosomas de las tres especies de *Echinococcus* obtenida mediante "Circos Plot". Las líneas de colores conectan los ortólogos entre las tres especies de *Echinococcus* basados en los análisis de Ortología.

El rango de porcentaje de genes sinténicos es del 89,1% al 97,1% con un promedio del 94,6% entre *E. canadensis* G7 y *E. multilocularis* y de 95% a 99,7% con un promedio del 98,3% entre *E. canadensis* G7 y *E. granulosus* G1 ([Tabla IV.5](#)).

Tabla IV.5: Porcentaje de genes sinténicos entre las especies de *Echinococcus*.

Cromosomas de <i>E. canadensis</i> G7	1	2	3	4	5	6	7	8	9	Genes totales
Nº de Genes ortólogos en <i>E. multilocularis</i>	1725	1440	1175	1069	829	547	402	396	280	7863
Nº de genes sinténicos	1657	1370	1138	1010	739	507	371	371	272	7435
% de genes sinténicos	96%	95%	97%	94%	89%	93%	92%	94%	97%	95%
Nº de Genes ortólogos en <i>E. granulosus</i> G1	1040	1055	833	734	499	375	282	314	243	5375
Nº de genes sintenicos	1037	1050	830	720	474	364	269	301	239	5284
% de genes sintenicos	100%	100%	100%	98%	95%	97%	95%	96%	98%	98%

4.2.1.1 Análisis de ortología con otros organismos

Para llevar a cabo el análisis de ortología se construyeron grupos de ortólogos a partir de 14 proteomas de diferentes organismos los cuales incluyeron representantes de diversas clases: Mammalia, Actinopterygii, Leptocardii, Insecta, Chromadorea, Cestoda, Monogenea, Trematoda, Turbellaria, con especial énfasis en la inclusión de genomas de alta calidad del filo Platelminos: *E. canadensis* G7, *E. granulosus* G1, *E. multilocularis*, *H. microstoma*, *T. solium*, *G. salaris*, *S. mansoni* y *S. mediterránea*. Se obtuvieron un total de 39482 grupos de ortología, entre los cuales 16024 grupos contienen genes de platelmintos. Entre ellos 3570 grupos contienen genes que pertenecen a especies no parasitarias y 12454 grupos contienen genes que pertenecen a especies parasitarias. De los grupos de parásitos 5203 grupos contienen genes que están presentes en todas las especies de cestodos. Esta última comparte a su vez 3068 grupos de genes con especies de parásitos no cestodos (trematodos y monogeneos). Además este grupo de parásitos no cestodos está conformado por 1181 grupos de genes no compartidos con especies de cestodos. Finalmente se identificaron 881 grupos que contienen genes que son específicos de cestodos y 560 grupos que son específicos del género *Echinococcus*, el cual comprende 1768 genes de los cuales 581 genes pertenecen a *E. canadensis* G7. La clasificación de los grupos de ortología puede observarse en la [tabla IV.6](#) y [figura IV.8](#).

Tabla IV.6: Grupos de genes ortólogos*

	Criterio	Nº de grupos	Nº de genes totales
<i>Echinococcus</i>	exclusivo	560	1768
	No exclusivo	7029	23375
Cestodos	exclusivo	881	4709
	No exclusivo	5203	29715
Platelminos parásitos No cestodos	exclusivo	698	2538
	No exclusivo	1181	13387
Grupo de cestodos vs Humanos		2418	13387

*Se definió como exclusivo a los grupos cuyos genes están solo en un grupo (por ej. en el género *Echinococcus*) y No exclusivo a los grupos cuyos genes están en el grupo específico pero tienen además ortólogos en al menos una o más especies fuera de éste (por ej presentes en *Echinococcus* y en otro género).

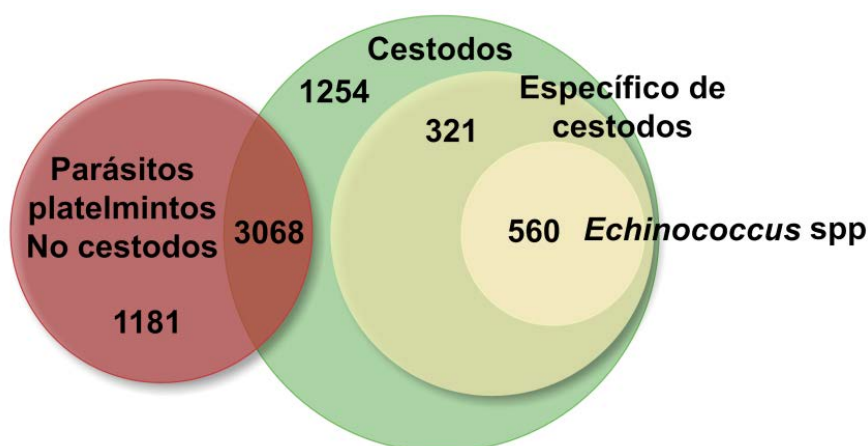


Figura IV.8: Diagrama de Venn del Número de grupos de ortología presentes en las categorías *Echinococcus*, cestodos y platelmintos parásitos.

Es interesante destacar que del total de los genes específicos del género *Echinococcus* solo el 20% de ellos tiene asignada una anotación funcional, mientras que el 80% restante están clasificados como proteínas hipotéticas. Entre los genes anotados, un total de 115 proteínas poseen dominios conservados y a 78 de ellas se les asignó términos GO. Un total de 76 genes específicos del género *Echinococcus* agruparon en 8 términos de la categoría Fm conformada en un 57% por la categoría “Unión” (54 genes) y en un 19% por la categoría “Actividad catalítica” (22 genes). Los siguientes dos términos más abundantes fueron "actividad de factor de transcripción de unión a ácido nucleico" (GO:0001071) (7,1%, 8 genes) y "actividad transportadora" (GO:0005215) (6,2%, 7 genes). En la categoría Pb 57 genes específicos del género *Echinococcus* agruparon en 21 categorías; de las cuales las cuatro más abundantes fueron "proceso celular" (GO:0009987) (14%, 51 genes), proceso metabólico" (GO:0008152) (11%, 41 genes); "regulación biológica" (GO:0065007) (8,5%, 35 genes) y "regulación del proceso biológico" (GO:0050789) (8%, 33 genes). Por otra parte; los genes agrupados en la categoría Cc se distribuyeron mayoritariamente entre las categorías asociadas a su localización celular y en organelas ([Figura IV.9](#)).

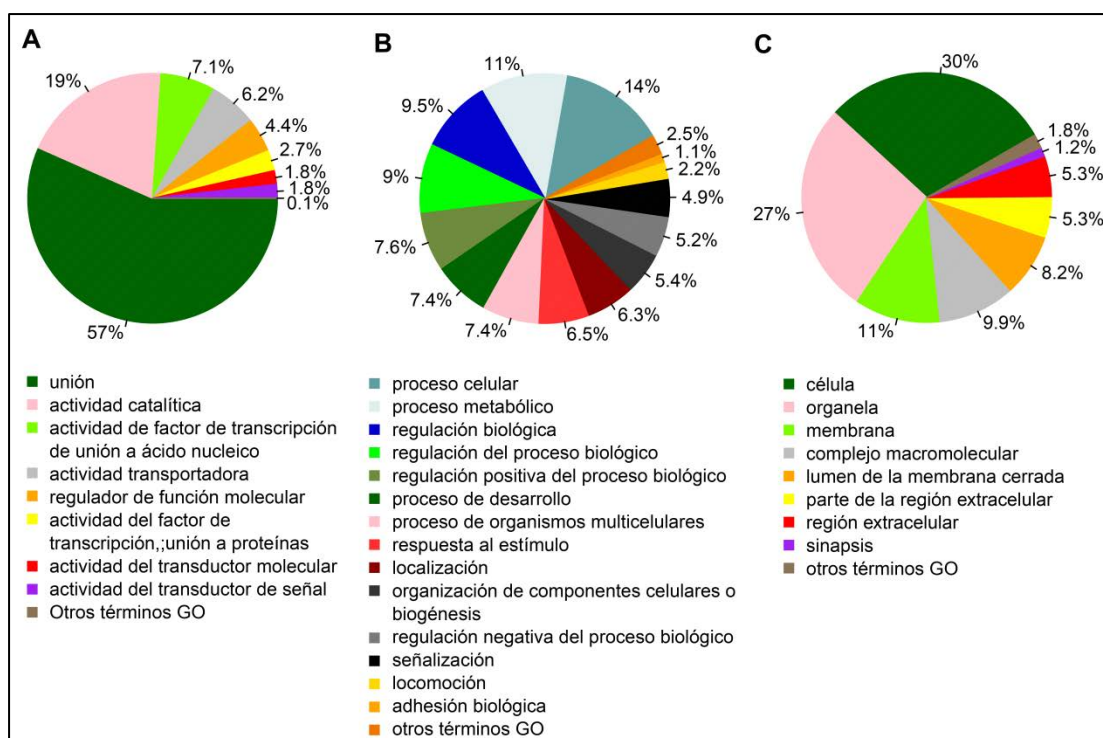


Figura IV.9: Anotación funcional de los genes específicos del género de *Echinococcus*. Porcentaje de términos GO asociados a la categoría A). "Función Molecular", B). "Proceso biológico"; C). Componente celular. Las categorías menores a 1% fueron sumadas y categorizadas como "otros términos GO".

4.2.2 Variantes genéticas entre las especies de *Echinococcus*: estudio de SNPs

Esta sección de la tesis está centrada en el estudio de SNPs entre las tres especies de *Echinococcus*. La obtención del genoma de *E. canadensis* G7 y la disponibilidad de los genomas y datos de secuenciación de especies filogenéticamente relacionadas permitió por primera vez llevar a cabo análisis de genómica comparativa utilizando la totalidad de las regiones genómicas y de este modo buscar variantes genéticas entre las especies de *Echinococcus* que revelen características o diferencias entre ellas haciendo foco en la variabilidad intraespecie e interespecie. Para la identificación de SNPs las lecturas de secuenciación de ADN genómico se mapearon contra los genomas de referencia. El estudio de variantes genómicas se realizó comparando todas las especies entre sí. Para ello en cada caso se tomó como referencia al genoma de una de las tres especies y se mapearon contra él las lecturas de las otras dos especies. Luego se repitió el procedimiento usando como referencia a otra de las tres especies y finalmente se integraron los datos obtenidos para las tres especies de *Echinococcus*. Para cada par analizado primero se mapearon las lecturas de una especie dada contra su propio genoma de referencia y luego se mapearon las lecturas de los demás genomas contra ese mismo genoma de referencia. Los porcentajes de mapeo y cobertura de los genomas fueron similares para las tres especies de *Echinococcus*. El esquema de mapeo y los porcentajes de lecturas mapeadas y porcentajes de cobertura se muestran en la [figura IV.10](#).

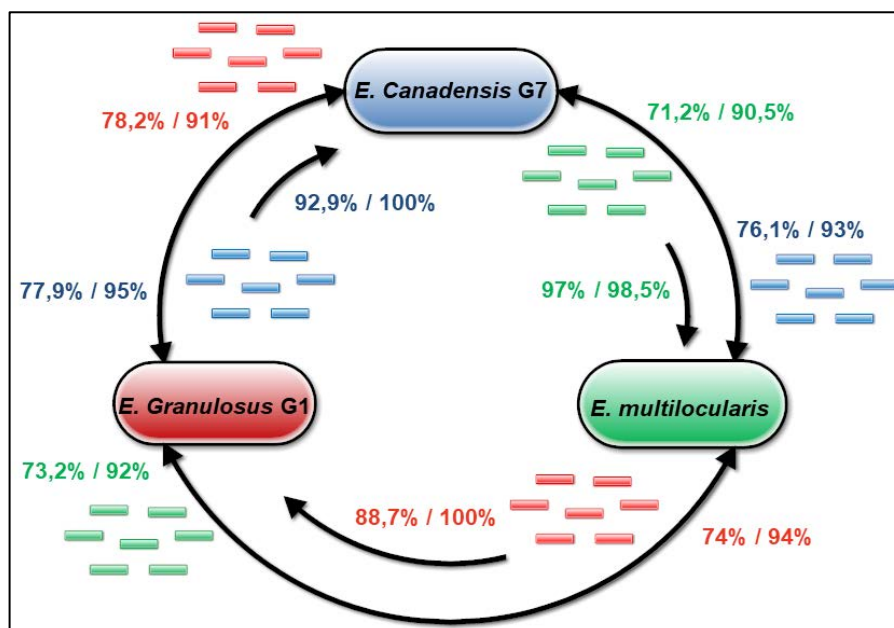


Figura IV.10: Diagrama del mapeo de las lecturas sobre los genomas de *Echinococcus*. Los pequeños rectángulos y recuadros del mismo color representan a los genomas y lecturas respectivamente. Los porcentajes separados por (/) indican Porcentaje de lecturas mapeadas / Porcentaje de cobertura de los genomas.

En cada par de especies analizadas se identificaron los sitios polimórficos homocigotas (SNPs homocigotas) y heterocigotas (SNPs heterocigotas) tanto para la especie que se usó como referencia (alelo de referencia) como para la especie alternativa (alelo alternativo). En primer lugar se evaluó la variación intra específica en cada genoma de *Echinococcus*, y se observó que el mayor número de sitios polimórficos intraespecíficos lo presentó el genoma de *E. granulosus* G1 con un total de 74585 SNPs, lo cual implica una tasa de 0,64 SNPs cada 1000 pb, seguido por el genoma de *E. canadensis* G7 con 9449 SNPs y una tasa de 0,09 SNPs cada 1000 pb y por último el genoma que presentó menor variabilidad intraespecífica fue el de *E. multilocularis* con un total 1658 SNPs y una tasa de 0,014 SNPs cada 1000 pb.

El análisis de polimorfismos entre las especies de *Echinococcus* reveló un mayor número de SNPs entre las especies de *E. canadensis* G7 y *E. granulosus* G1 que entre *E. canadensis* G7 y *E. multilocularis*. Entre *E. canadensis* G7 y *E. granulosus* G1 se identificaron un total de 788554 SNPs que comprendieron 1396 SNPs heterocigotas y 777710 SNPs homocigotas, mientras que al comparar *E. canadensis* G7 con *E. multilocularis* se encontraron un total de 327802 SNPs que incluyeron 215 SNPs heterocigotas y 327572 SNPs homocigotas. En ambos casos la relación de sustituciones por transición/transversión fue de 2,97. Por otro lado, al comparar el genoma de *E. granulosus* G1 con el de *E. multilocularis* se identificaron 332124 SNPs, de los cuales 6405 SNPs fueron heterocigotas y 325691 homocigotas con una relación transición/transversion de 2,9 ([Tabla IV.7](#)).

Tabla IV.7: Sitios polimórficos homocigotas y heterocigotas de *E. candensis* G7, *E. multilocularis* y *E. granulosus* G1.

Especies	Muestra	Ref/Alt:Hom/Hom	Ref/Alt:Hom/Het	Ref/Alt:Het/Het	Ref/Alt:Het/Hom	Alt:Het	Alt:Hom	Número total de SNPs
<i>E. multilocularis</i>	JAVA05	-	-	-	-	1264	394	1658
<i>E. canadensis</i> G7	PH14	-	-	-	-	8593	856	9449
<i>E. granulosus</i> G1	GH09/3	-	-	-	-	69169	5416	74585
<i>E. canadensis</i> G7 vs <i>E. multilocularis</i>	PH14 – JAVA05	7	0	1	7	215	327572	327802
<i>E. granulosus</i> G1 vs <i>E. multilocularis</i>	GH09/3 – JAVA05	11	0	0	17	6405	325691	332124
<i>E. canadensis</i> G7 vs <i>E. granulosus</i> G1	PH14 - H95/5	816	0	7	8625	1396	777710	788554

Ref/Alt:Hom/Hom: Sitios polimórficos homocigotas para el alelo de referencia y el alelo alternativo. Ref/Alt:Hom/Het: Sitios polimórficos homocigotas para el alelo de referencia y heterocigota para el alelo alternativo. Ref/Alt:Het/Het: Sitios polimórficos heterocigota para el alelo de referencia y el alelo alternativo. Ref/Alt:Het/Hom: Sitios polimórficos heterocigota para el alelo de referencia y homocigota para el alelo alternativo. Alt:Het: SNPs heterocigotas. Alt:Hom: SNPs homocigotas. El número de SNPs homocigotas y heterocigotas fue calculado para ambos alelos, el alelo de referencia y el alelo alternativo.

4.2.2.1 Localización y efecto de las variantes génicas en las regiones codificantes

Se evaluó la distribución de los SNPs entre las tres especies de *Echinococcus* según su localización en tres regiones genómicas determinadas: intrones, exones y regiones intergénicas. Mediante este análisis se observó que el número de SNPs es mayor en las regiones intergénicas y en los intrones que en las regiones codificantes. Dicho patrón fue observado en las comparaciones de las tres especies de *Echinococcus*. Además se observó que el número de SNPs en los intrones y en las regiones intergénicas también es mayor entre las especies de *E. canadensis* G7 y de *E. granulosus* G1, que entre *E. canadensis* G7 y *E. multilocularis* y que entre *E. granulosus* G1 y *E. multilocularis*. Sin embargo, en las regiones codificantes el número de SNPs es relativamente similar para las tres especies de *Echinococcus* ([Figura IV.11 A](#)).

Al evaluar la distribución de los SNPs en las regiones codificantes se observó que el porcentaje de sustituciones sinónimas fue mayor que el porcentaje de sustituciones no sinónimas en los tres pares de especies analizadas. El número de SNPs sinónimos entre las especies *E. canadensis* G7 y *E. granulosus* G1 fue de 58662, mientras que el número de SNPs no sinónimos fue de 45677. Por otro lado, entre las especies de *E. canadensis* G7 y *E. multilocularis* se observaron 62184 SNPs sinónimos y 38524 SNPs no sinónimos. Esto significa que hay 7153 SNPs no sinónimos más entre *E. canadensis* G7 y *E. granulosus* G1 que entre *E. canadensis* G7 y *E. multilocularis*. Finalmente entre *E. granulosus* G1 y *E. multilocularis* se identificaron 68963 SNPs sinónimos y 41817 SNPs no sinónimos ([Figura IV.11 B](#)). Para confirmar estos resultados se realizaron los mismos análisis utilizando el genoma de *E. granulosus* G1 obtenido por Zhang y col. [78], el cual consiste en un aislamiento de *E. granulosus* G1 diferente como así también una tecnología de secuenciación y estrategia de ensamblado independientes que el empleado por Tsai y col.[43]. Aun así los resultados obtenidos fueron prácticamente los mismos.

Posteriormente se comparó la distribución de SNPs que causan cambios en la secuencia de aminoácidos de las proteínas de *E. canadensis* G7 con sus correspondientes ortólogos de *E. multilocularis* y *E. granulosus* G1. Este análisis reveló una mayor variación aminoacídica entre proteínas de *E. canadensis* G7 y *E. granulosus* G1 (1,06 cambios cada 100 residuos de aminoácidos) que entre las proteínas de *E. granulosus* G1 y *E. multilocularis* (0,98 cambios cada 100 residuos de aminoácidos), y aun mayor que entre *E. canadensis* G7 y *E. multilocularis* (0,89 cambios cada 100 residuos de aminoácidos). Estas diferencias fueron evaluadas mediante el test de Anova, demostrándose que los resultados obtenidos fueron estadísticamente significativos ([Figura IV.11 C](#)).

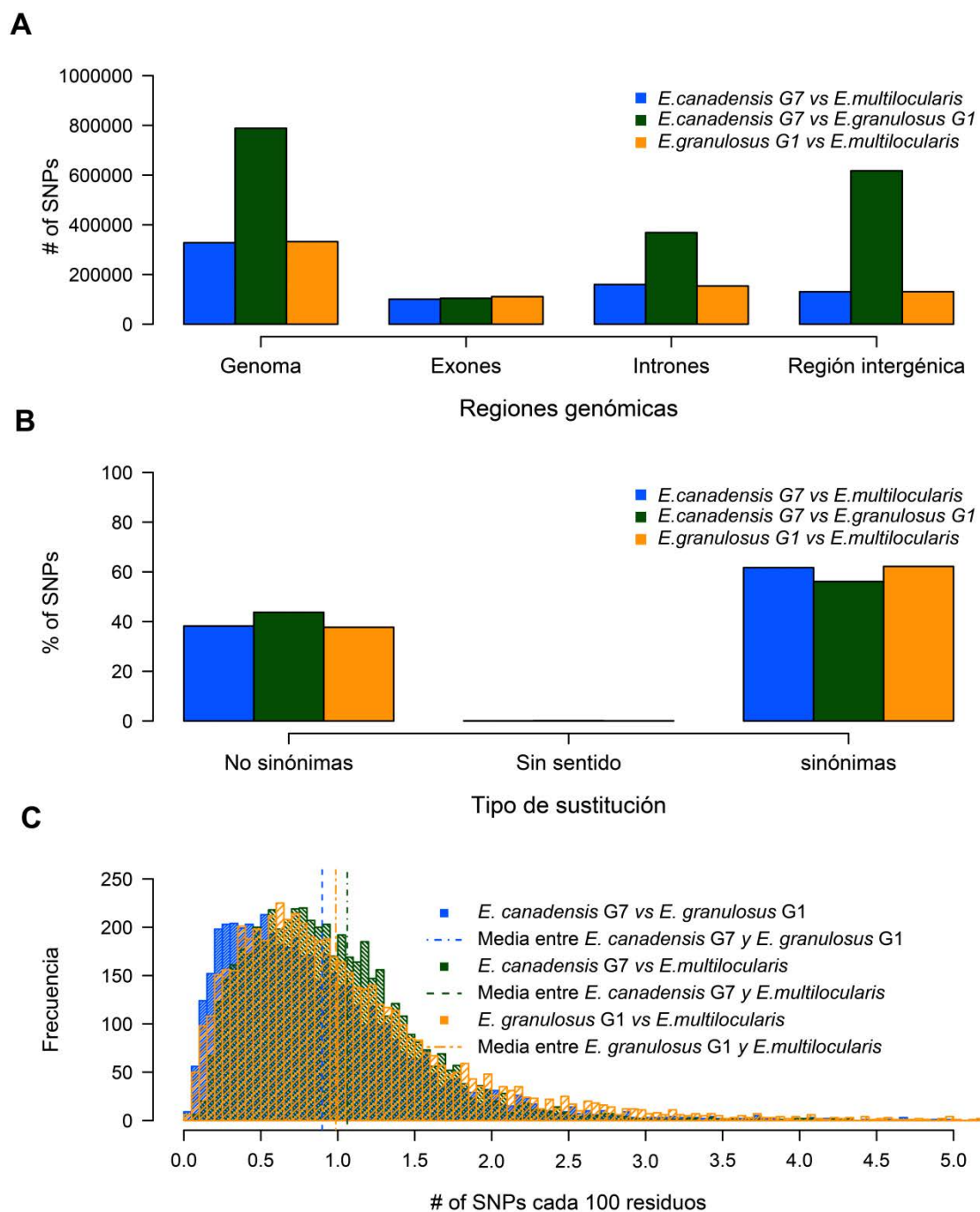
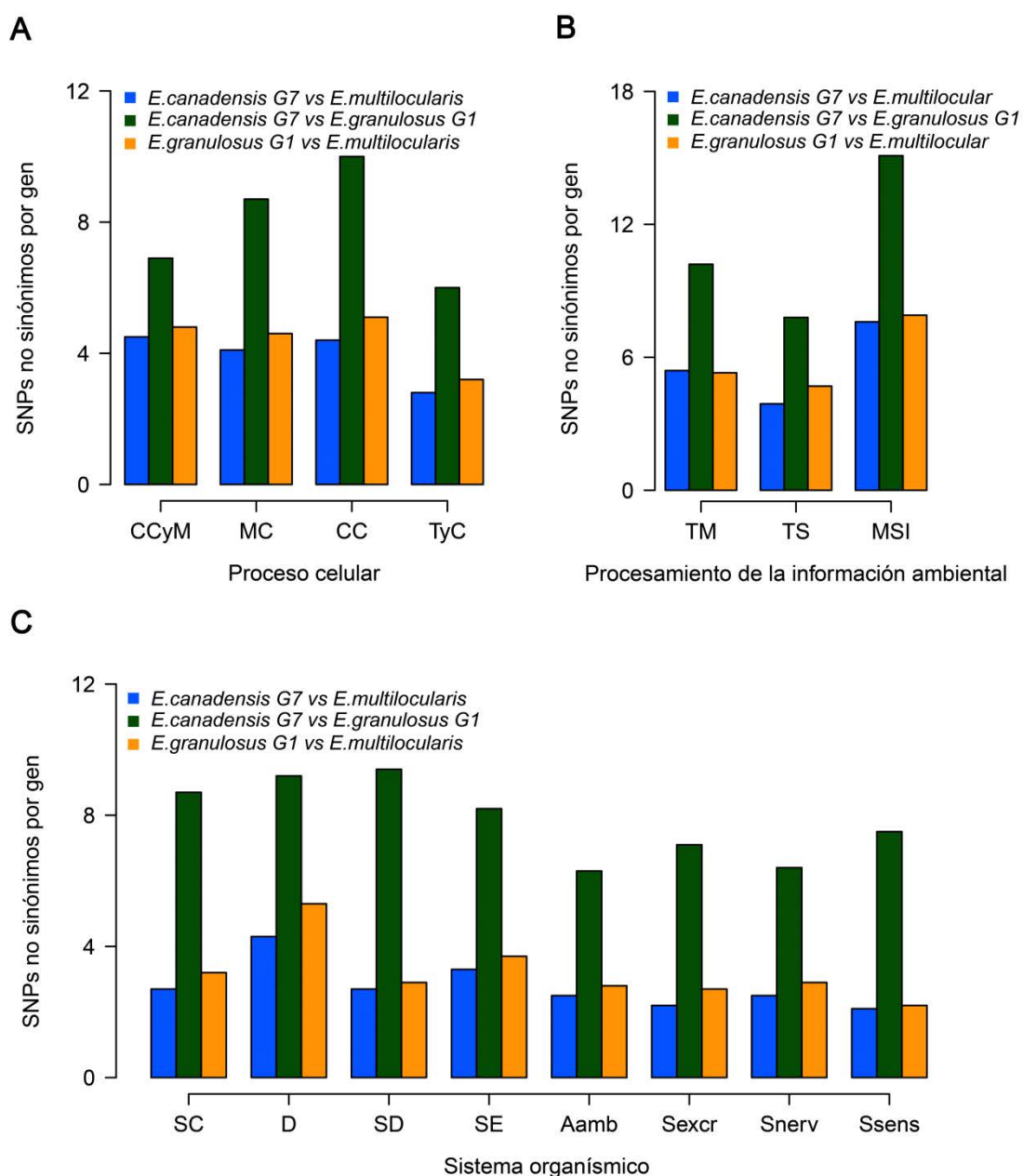


Figura IV.11: Distribución de SNPs entre las especies de *Echinococcus*. A). Número, localización, y distribución de SNPs en las estructuras genómicas. B). Tipo de sustitución causada por los SNPs en las regiones codificantes clasificadas en sinónimas, no sinónimas y sin sentido. C). Distribución de sustituciones no sinónimas cada 100 residuos de aminoácidos.

4.2.2.2 Distribución de SNPs según vías metabólicas KEGGs

La la presencia de SNPs en los genomas fue clasificada no solo por su localización genómica sino que también de acuerdo a su distribución en las vías metabólicas KEGGs con el fin de determinar su implicancia en genes que tengan un rol metabólico. Para ello se analizó la distribución de los SNPs en los genes de *Echinococcus* asociados a 31 categorías de las principales 5 vías metabólicas descriptas en la base de datos KEGGs. Si bien hay un mayor número de SNPs entre las especies de *E. canadensis* G7 y *E. granulosus* G1, los SNPs aparecen uniformemente distribuidos en todas las vías metabólicas. Tampoco se observa una distribución de SNPs diferencial entre los pares de especies analizadas. La distribución de SNPs de acuerdo a las vías metabólicas KEGGs se muestra en la [Figura IV.12](#) y [Tabla IV.8](#).



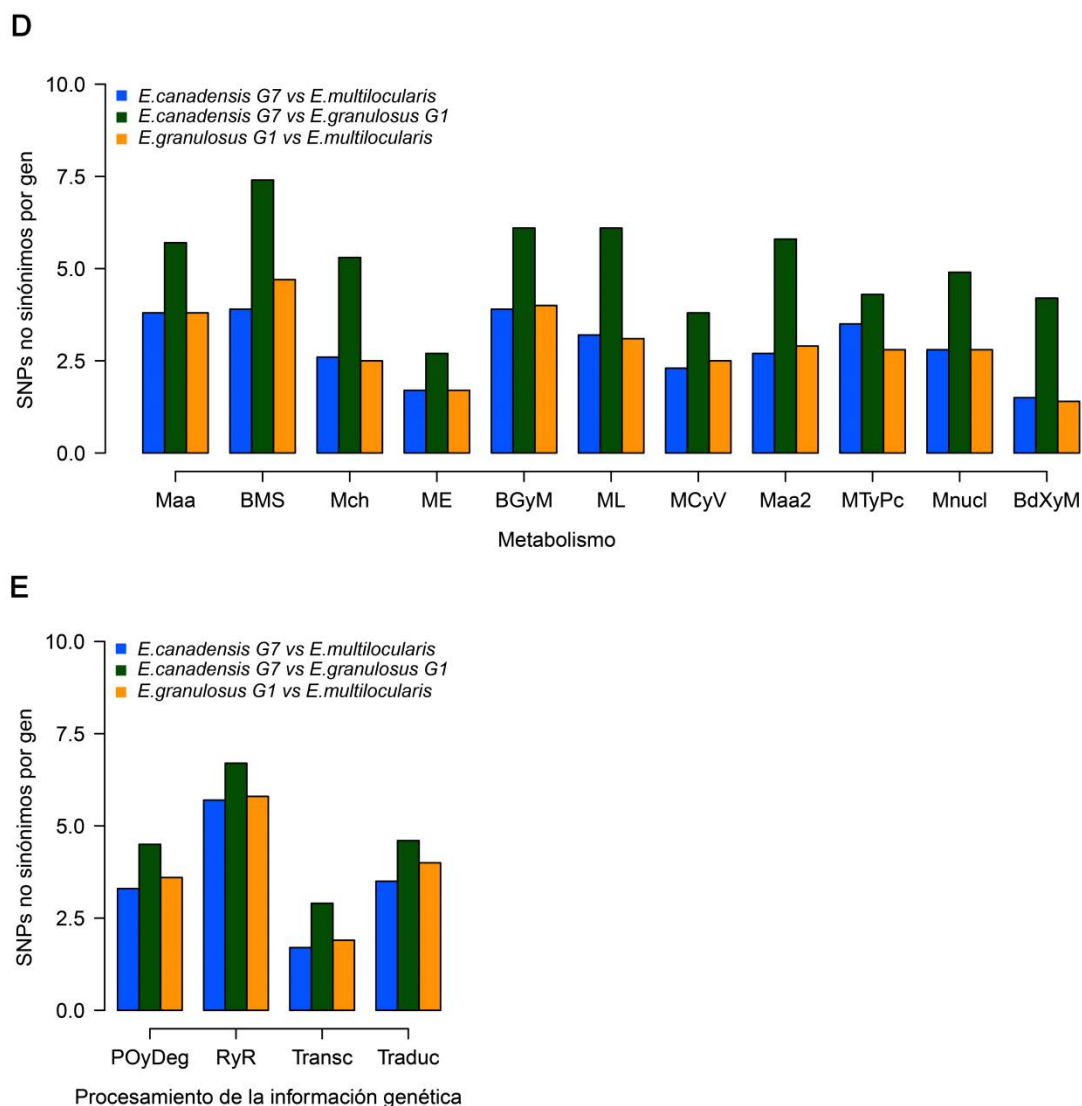


Figura IV.12: Distribución de SNPs en los genes de las tres especies de *Echinococcus* clasificados de acuerdo a las cinco vías metabólicas principales. A). Procesos celulares: CCyM: Desarrollo y muerte celular; MC: Motilidad celular; CC: Comunidad celular; TyC: Transporte y Catabolismo; B). Procesamiento de la información ambiental: TM: Transporte de membrana; TS: Transducción de señales; MSI: Moléculas de señalización e interacción; C). Sistema orgánico: SC: Sistema circulatorio; D: Desarrollo y muerte celular; SD: Sistema digestivo; SE: Sistema endocrino; Aamb: Adaptación ambiental; Sexcr: Sistema excretorio; Snerv: Sistema nervioso; Ssens: Sistema sensorial; D). Metabolismo: Maa: Metabolismo de amino ácidos; BMS: Biosíntesis de otros metabolitos secundarios; Mch: Metabolismo de carbohidratos; ME: Metabolismo de energía; BGyM: Biosíntesis de Glicano y metabolismo; ML: Metabolismo de lípidos; MCyV: Metabolismo de cofactores y vitaminas; Maa2: Metabolismo de otros amino ácidos; MTyPc: Metabolismo de terpenoides y policétidos; Mnucl: Metabolismo de ácidos nucleicos; BdXyM: Biodegradación de xenobioticos y metabolismo; E). Procesamiento de la información genética: POyDeg: Plegado, estructuración y degradación; RyR: Replicación y reparación; Transc: Transcripción; Traduc: Traducción.

Tabla IV.8: Vías metabólicas KEGGs en *Echinococcus*. Distribución y densidad de SNPs por gen.

Vías metabólicas	# Genes	EcG7vsEmul	SNPs/Genes	# Genes	EcG7vsEgrG1	SNPs/Genes	# Genes	EgrG1vsEmul	SNPs/Genes
Proceso celular	1252	4927	3,9	719	5451	7,6	1252	5512	4,4
Desarrollo y muerte celular	412	1869	4,5	274	1890	6,9	412	1987	4,8
Motilidad celular	90	370	4,1	52	452	8,7	90	413	4,6
Comunidad celular	366	1612	4,4	188	1882	10,0	366	1865	5,1
Transporte y Catabolismo	384	1076	2,8	205	1227	6,0	384	1247	3,2
Procesamiento de la información ambiental	1721	7051	4,1	994	8092	8,1	1721	8277	4,8
Transporte de membrana	20	108	5,4	10	102	10,2	20	105	5,3
Transducción de señales	1623	6347	3,9	936	7266	7,8	1623	7559	4,7
Moléculas de señalización e interacción	78	596	7,6	48	724	15,1	78	613	7,9
Enfermedades Humanas	3337	6858	2,1	1784	7548	4,2	3337	7930	2,4
Resistencia Antineoplástica	128	705	5,5	68	806	11,9	128	833	6,5
Cánceres: Visión general	628	2920	4,6	339	2973	8,8	628	3301	5,3
Cánceres: Tipo específico	425	2201	5,2	276	2383	8,6	425	2619	6,2
Enfermedades metabólicas y endócrino	187	594	3,2	126	718	5,7	187	645	3,4
Enfermedades inmunes	116	74	0,6	30	104	3,5	116	94	0,8
Enfermedades infecciosas: Parasitismo	167	364	2,2	71	564	7,9	167	438	2,6
Sistema Organísmico	2820	7659	2,7	1345	9583	7,1	2820	8846	3,1
Sistema circulatorio	183	492	2,7	79	687	8,7	183	587	3,2
Desarrollo y muerte celular	129	557	4,3	72	660	9,2	129	681	5,3
Sistema digestivo	229	614	2,7	84	788	9,4	229	666	2,9
Sistema endocrino	819	2703	3,3	380	3102	8,2	819	3048	3,7
Adaptación ambiental	82	209	2,5	40	251	6,3	82	229	2,8
Sistema excretorio	136	293	2,2	59	417	7,1	136	361	2,7
Sistema nervioso	452	1132	2,5	246	1567	6,4	452	1324	2,9
Sistema sensorial	136	279	2,1	50	375	7,5	136	299	2,2

Metabolismo	2704	4012	1,5	1768	4650	2,6	2704	4046	1,5
Metabolismo de amino ácidos	153	579	3,8	113	648	5,7	153	587	3,8
Biosíntesis de otros metabolitos secundarios	19	74	3,9	13	96	7,4	19	89	4,7
Metabolismo de carbohidratos	337	861	2,6	202	1073	5,3	337	842	2,5
Metabolismo de energía	147	251	1,7	108	295	2,7	147	254	1,7
Biosíntesis de Glicano y metabolismo	164	634	3,9	107	648	6,1	164	663	4,0
Metabolismo de lípidos	173	562	3,2	105	638	6,1	173	542	3,1
Metabolismo de cofactores y vitaminas	81	187	2,3	60	226	3,8	81	203	2,5
Metabolismo de otros amino ácidos	60	160	2,7	35	204	5,8	60	176	2,9
Metabolismo de terpenoides y policétidos	25	87	3,5	21	91	4,3	25	71	2,8
Metabolismo de ácidos nucleicos	187	519	2,8	127	618	4,9	187	525	2,8
Biodegradación de xenobioticos y metabolismo	66	98	1,5	27	113	4,2	66	94	1,4
Procesamiento de la información genética	1152	3904	3,4	906	4158	4,6	1152	4271	3,7
Plegado, estructuración y degradación	360	1188	3,3	272	1219	4,5	360	1309	3,6
Replicación y reparación	157	899	5,7	143	952	6,7	157	906	5,8
Transcripción	224	372	1,7	158	452	2,9	224	431	1,9
Traducción	411	1445	3,5	333	1535	4,6	411	1625	4,0

4.2.3 Conclusiones parciales

Los *contigs* de *E. canadensis* G7 fueron ordenados dando lugar a estructuras cromosómicas, que presentaron alto grado de identidad con los cromosomas de *E. multilocularis* permitiendo el análisis de sintenia. Mediante este análisis se demostró un alto grado de conservación en el número, tipo y orden de los genes codificantes entre los genomas de las tres especies de *Echinococcus*.

Los análisis de ortología evidenciaron 560 grupos de proteínas que son específicas del género *Echinococcus* y 881 grupos específicos de la clase Cestoda y grupos de genes compartidos entre los miembros de las clases Cestoda, Trematoda y Monogenea. El grupo de ortología que se identificó como específico del género *Echinococcus* contiene un gran número de genes a los que no fue posible asignarles una anotación funcional. No obstante, los genes específicos de *Echinococcus* con términos GO asignados a la categoría Fm tuvieron porcentajes de términos GO diferentes respecto a los obtenidos para el repertorio total de genes. Si bien en ambos casos las dos categorías más abundantes fueron “Unión” y “Actividad catalítica”, el porcentaje de términos “Unión” fue mayor en el grupo de genes específicos de *Echinococcus*. A su vez, el término “Actividad de factor de transcripción de unión a ácidos nucleicos” gana dos niveles en abundancia sugiriendo que un porcentaje importante de genes específicos de género *Echinococcus* podrían tener actividad reguladora de la transcripción.

Los estudios de variantes genéticas entre los genomas de las especies de *Echinococcus* demostraron que el organismo que presenta mayor variabilidad intraespecífica es *E. granulosus* G1. Por otra parte, al realizar las comparaciones genómicas interespecie se demostró una mayor variabilidad entre los genomas de *E. canadensis* G7 y *E. granulosus* G1 que entre los genomas de *E. canadensis* G7 y *E. multilocularis*. Aún más, se determinó que el mayor número de SNPs se localiza en las regiones intergénicas y la mayoría de aquellos encontrados en las regiones codificantes generan mutaciones sinónimas. Por otra parte, la distribución de SNPs asociados a vías metabólicas sugiere que no hay una tendencia a la acumulación de SNPs en una vía metabólica particular.

4.3 Filogenia de las especies de *Echinococcus* disponibles actualmente

La relación filogenética entre las especies de *Echinococcus* es materia de discusión desde hace ya muchos años. Si bien los estudios filogenéticos han contribuido significativamente hacia una mejor comprensión de la taxonomía del género *Echinococcus*, no se han producido resultados que concuerden con las primeras clasificaciones fenotípicas llevadas a cabo por Rausch y col [235] en 1967 y ampliamente observadas en la actualidad donde las características morfológicas de *E. multilocularis* son claramente diferentes a las de *E. canadensis* G7 y *E. granulosus* G1.

Recientemente, se han reconstruido filogenias a partir de genomas mitocondriales completos [236], sin embargo aun así no se ha llegado a una concordancia respecto a dichas clasificaciones. En esta sección se estudió la relación filogenética de las especies de *Echinococcus* centrándonos particularmente en las tres especies con genomas completos y abordando el análisis desde tres perspectivas diferentes descriptas a continuación: genoma mitocondrial, genes de copia única y SNPs de genomas nucleares completos.

4.3.1 Reconstrucción de la filogenia de *Echinococcus* basada en genomas mitocondriales

El genoma mitocondrial ensamblado en esta tesis en conjunto con todos los genomas mitocondriales disponibles de las especies de *Echinococcus*, fueron utilizados para la reconstrucción filogenética. A partir del alineamiento de los genomas mitocondriales se construyó un árbol filogenético mediante el cual se pudo clasificar y discriminar entre las especies de *Echinococcus*. En dicho árbol se distinguen cuatro grupos monofiléticos. Un primer grupo que incluye a *E. canadensis* G6, G7, G8 y G10, *E. orteppi* G5 y *E. equinus* G4. Un segundo grupo que incluye a *E. granulosus* G1 y *E. felidis*. El tercer grupo que comprende a *E. multilocularis* y *E. shiquicus* y un último grupo que contiene a *E. oligarthrus* y *E. vogeli* (Figura IV.13). Los clústeres formados en dicho árbol separan claramente las especies de *E. canadensis* G7, *E. granulosus* G1 y *E. multilocularis* como fue descrito anteriormente [11,13].

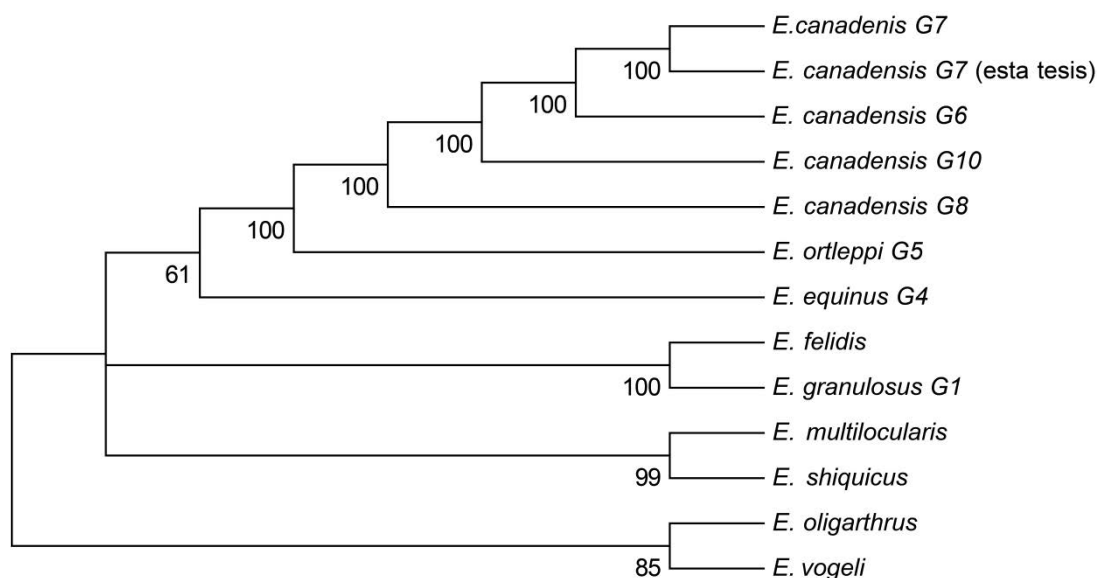


Figura IV.13: Reconstrucción de la filogenia de *Echinococcus* basado en genomas mitocondriales completos. Los genomas mitocondriales utilizados y sus códigos de acceso en GenBank fueron los siguientes: *E. oligarthrus* (NC_009461), *E. vogeli* (NC_009462), *E. shiquicus* (NC_009460), *E. orteppi* (NC_011122), *E. canadensis* G6 (NC_011121), *E. canadensis* G7 (AB235847); *E. canadensis* G8 (AB235848) [236]; *E. equinus*, (AB786665), *E. granulosus* (AB786664); *E. felidis* (AB732958), *E. canadensis* G10 (AB745463) [11]; *E. multilocularis* (NC_000928) [237]; *E. canadensis* G7 (PRJEB8992), (esta tesis).

4.3.2 Reconstrucción filogenética basada en genes de copia única

A partir de los grupos de ortología definidos anteriormente se identificaron 29 grupos de ortólogos que contienen genes de copia única y se los utilizó para la construcción de dos árboles filogenéticos. Un árbol usando las secuencias de aminoácidos correspondientes a las proteínas y otro utilizando las secuencias de nucleótidos correspondientes a la región codificante (CDS) de los genes de copia única. La topología de los árboles filogenéticos obtenidos fueron las esperadas para los filos Cordados, Artrópodos, Nematodos y Platelminos, ampliamente conocidos y descriptos previamente [43,233]. Dicha filogenia posiciona a las tres especies de *Echinococcus* en un *clúster* único pero diferente en cuanto a la especie basal de *Echinococcus* según se utilicen secuencias de aminoácidos o de nucleótidos ([Figura IV.14](#)).

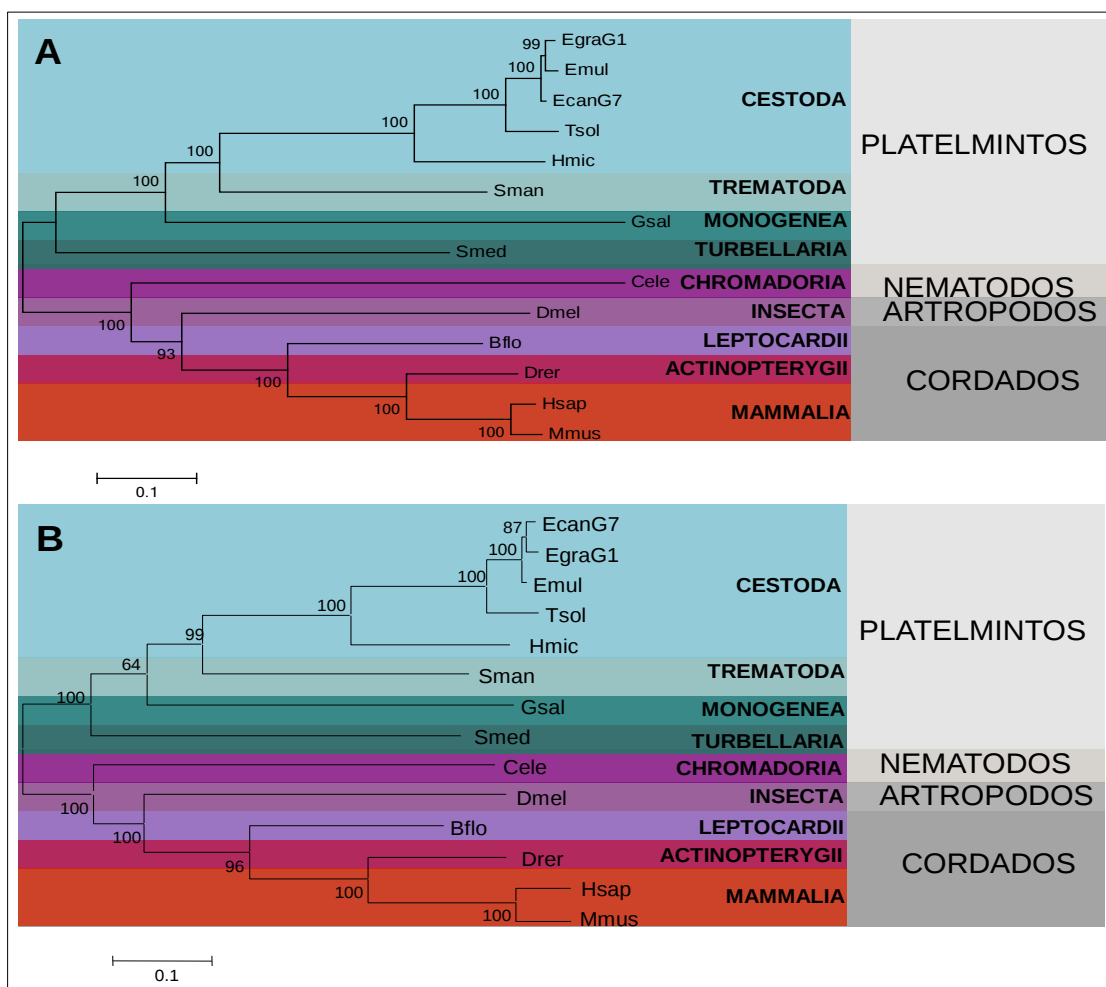


Figura IV.14: Reconstrucción de la filogenia de metazoos incluyendo a *Echinococcus* basado en genes de copia única de proteínas y CDSs: A) Árbol filogenético basado en el análisis de 14 grupos de ortología que incluyeron 29 secuencias de aminoácidos de cada organismo. B) Árbol filogenético basado en el análisis de 14 grupos de ortología que incluyeron 29 secuencias de CDSs de cada organismo. Hsap: *H. sapiens*; Mmus: *M. musculus*; Drer: *D. rerio*; Bflo: *B. floridae*; Dmel: *D. melanogaster*; Cele: *C. elegans*; EcanG7: *E. canadensis* G7; Egra: *E. granulosus* G1; Emul: *E. multilocularis*; Gsal: *G. salaris*; Hmic: *H. microstoma*; Sman: *S. mansoni*; Smed: *S. mediterranea*; TsM1: *T. solium*.

4.3.3 Reconstrucción filogenética basada en SNPs de genomas completos

Para comprender mejor la relación filogenética entre los taxones y evaluar la contribución de los SNPs a la diversidad genética entre las especies de *Echinococcus* se seleccionaron solo los sitios polimórficos homocigotas presentes entre las tres especies de *Echinococcus* y se los usó para llevar a cabo los análisis filogenéticos. Para ello se removieron los SNPs heterocigotas y los sitios polimórficos homocigotas se extrajeron y se concatenaron y el alineamiento de secuencias resultante se empleó para construir los árboles filogenéticos implementando el método de máxima verosimilitud. Este análisis se llevó a cabo para la totalidad del genoma y además clasificándolos por regiones genómicas de intrones, exones y regiones intergénicas. La topología del árbol mostró una mayor distancia filogenética entre *E. canadensis* G7 y *E. granulosus* G1 que entre *E. canadensis* G7 y *E. multilocularis* comprendiendo un total de 557254 sitios polimórficos homocigotas. Además *E. canadensis* G7 y *E. granulosus* G1 divergen de un nodo común cercano a *E. multilocularis* (Figura IV.15).

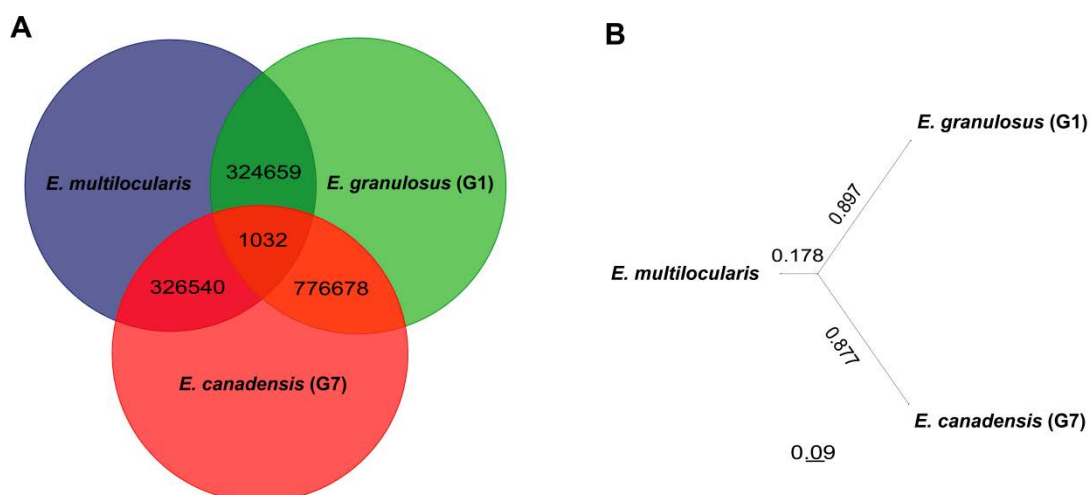


Figura 14.15: Análisis de SNPs totales en *Echinococcus*. A) Número de sitios con SNPs homocigotas entre los genomas de *E. canadensis* G7, *E. multilocularis* y *E. granulosus* G1. El número en la región solapante indica el número de SNPs entre las correspondientes especies. El número en la región de solapamiento de las tres especies indica el número de sitios trialélicos. B) Árbol filogenético basado en el análisis de SNPs totales utilizando el Método de Máxima Verosimilitud.

Complementariamente, se registró el número de sitios polimórficos compartidos entre las especies de *Echinococcus*. Para este propósito, se utilizó un genoma de referencia diferente en cada una de las tres rondas de mapeo, y se mapearon las lecturas de los otros dos genomas contra el genoma de referencia elegido. De este modo, al usar al genoma de *E. canadensis* G7 como referencia se encontraron 32142 sitios polimórficos compartidos, de los cuales el 92,57% (29756 SNPs) mostraron tener el mismo cambio. Por otra parte, al usar al genoma de *E. multilocularis* como referencia se identificaron 96135 sitios polimórficos compartidos, entre los cuales el

98,80% (94990 SNPs) mostraron el mismo cambio. Finalmente, al utilizar el genoma de *E. granulosus* G1 como referencia se detectaron 42866 sitios polimórficos compartidos de los cuales 91,63% (39281 SNPs) comparten además el mismo cambio. En efecto, se observó que la mayoría de los sitios polimórficos son únicos para cada par de especie analizada y que el mayor número de sitios compartidos se encuentra cuando se elige como referencia al genoma de *E. multilocularis* ([Figura IV.16](#)).

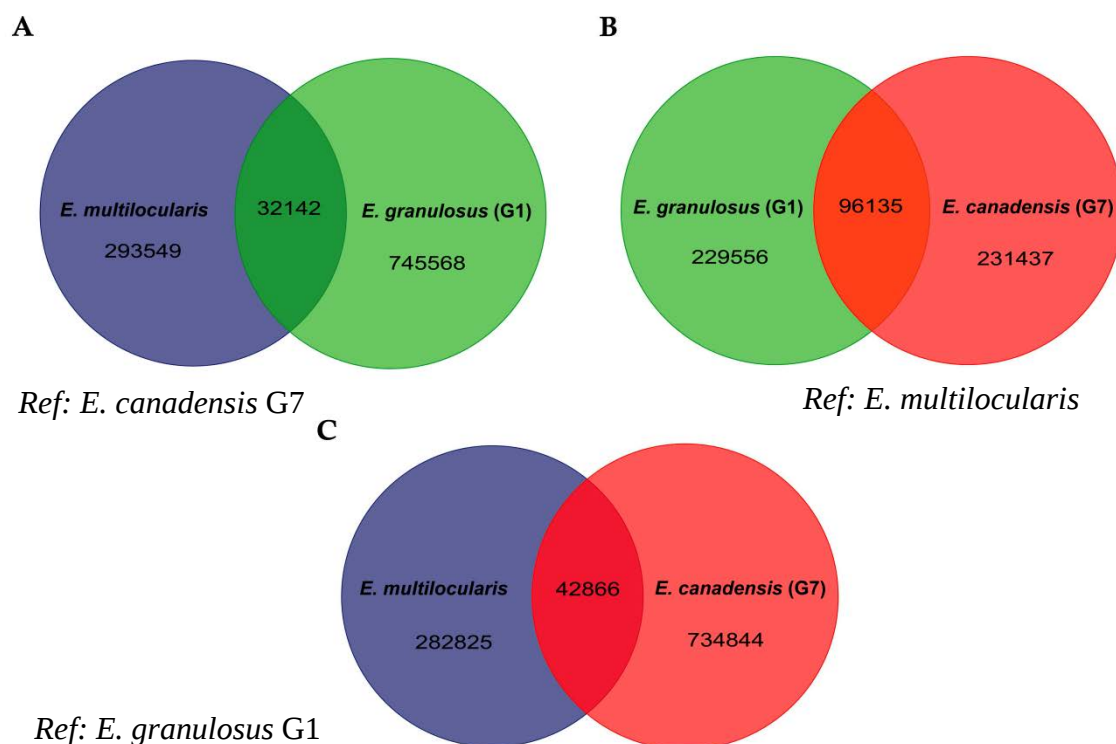


Figura IV.16: Diagrama de Venn de los sitios polimórficos compartidos entre las diferentes especies de *Echinococcus*. A). Número de SNPs entre cada especie usando al genoma de *E. canadensis* G7 como referencia. B). Número de SNPs entre cada especie usando al genoma de *E. multilocularis* como referencia. C). Número de SNPs entre cada especie usando al genoma de *E. granulosus* G1 como referencia. El número debajo del nombre de cada especie representa el número de SNPs con la especie tomada como referencia. El número en la región solapante indica el número de sitios polimórficos con el mismo polimorfismo en la misma posición entre las dos especies con respecto al genoma de referencia.

4.3.4 Validación de SNPs y nuevos marcadores moleculares

4.3.4.1 Validación de SNPs

Para validar y verificar los polimorfismos que fueron detectados mediante el mapeo de las lecturas y métodos de análisis de NGS, se seleccionaron exones de genes de *Echinococcus* conteniendo SNPs no sinónimos. Las regiones elegidas dentro de cada exón se seleccionaron con el fin de amplificar regiones con alto número de SNPs y a su vez localizadas entre regiones sin polimorfismos de manera que el diseño de los cebadores sea útil para amplificar las secuencias de todas las especies de *Echinococcus*. Las regiones seleccionadas fueron amplificadas por PCR utilizando como molde el

ADN genómico extraído de 10 aislamientos de parásitos de muestras independientes, tres de *E. canadensis* G7, tres de *E. granulosus* G1 y cuatro de *E. multilocularis*. Todos los SNPs seleccionados y detectados por métodos de análisis de secuencia de NGS fueron confirmados en todas las especies de *Echinococcus* analizadas ([Tabla IV.9](#)).

Tabla IV.9: Validación de SNPs, coordenadas, tipo nucleótido y tipo de cambio en genes de *Echinococcus*

Gen	Pos	Ref	Alt	Especie	Tipo	Cambio	Nombre de la proteína
EcanG7_09582 *	267	G	A	Emul	S	Ser85Ser	Proteína Argonauta-Piwi
EcanG7_09582 **	269	A	T	EgrG1	N	Lys86Met	
EcanG7_09582 *	293	T	A	EgrG1/Emul	N	Phe94Tyr	
EcanG7_09582 **	299	A	G	EgrG1	N	Glylu96Gly	
EcanG7_09582 **	329	A	G	EgrG1	N	Lys106Arg	
EcanG7_09582 *	340	G	A	EgrG1	N	Val110Ile	
EcanG7_05607 **	323	C	T	EgrG1	N	His100Tyr	Proteína Kinasa
EcanG7_05607 *	328	T	C	EgrG1	S	Ile101Ile	
EcanG7_05607 **	439	C	T	Emul	S	Tyr138Tyr	
EcanG7_02107 **	1275	C	A	Emul	S	Ile425Ile	Factor de elongación (EFG); GTP-binding domain;
EcanG7_02107 **	1281	T	C	Emul	S	Phe427Phe	
EcanG7_02107 **	1296	T	C	EgrG1/Emul	S	Gly432Gly	
EcanG7_02107 **	1328	G	A	EgrG1/Emul	N	Arg443Lys	
EcanG7_02107 **	1380	C	T	Emul	S	Tyr460Tyr	
EcanG7_02107 **	1431	G	A	EgrG1	S	Ala477Ala	
EcanG7_02097 **	68	A	G	EgrG1/Emul	N	His23Arg	Subunidad alfa/beta de proteosoma
EcanG7_02097 **	106	C	T	EgrG1/Emul	N	Leu36Phe	
EcanG7_02097 *	108	T	C	EgrG1	S	Leu36Phe	
EcanG7_02097 *	114	A	G	Emul	S	Leu38Leu	
EcanG7_02097 **	201	T	C	Emul	S	Tyr67Tyr	
EcanG7_02097 **	216	T	C	Emul	S	Gly72Gly	
EcanG7_02097 **	222	A	G	Emul	S	Glylu74Glu	
EcanG7_02097 **	249	C	T	Emul	S	Phe83Phe	
EcanG7_02097 **	408	C	T	EgrG1	S	Gly136Gly	
EcanG7_02097 **	412	G	A	Emul	N	Val138Met	
EcanG7_02097 **	556	A	G	EgrG1/Emul	N	Lys186Glu	Núcleotidodisulfuro de piridina oxidoductasa; Glutaredoxina
EcanG7_06983 *	881	A	G	Emul	N	Thr250Ala	
EcanG7_06983 **	893	G	A	EgrG1	N	Val254Ile	
EcanG7_06983 **	1726	C	T	EgrG1	S	Phe531Phe	
EcanG7_06983 **	1773	G	A	EgrG1	N	Arg547His	
EcanG7_06983 **	1783	G	T	Emul	S	Ala550Ala	

EcanG7_00768 **	4529	T	C	EgrG1/Emul	N	Ile1510Thr	
EcanG7_00768 *	4540	G	A	Emul	N	Ala1514Thr	
EcanG7_00768 **	4551	C	T	Emul	S	Ser1517Ser	
EcanG7_00768 **	4554	C	T	EgrG1	S	Ala1518Ala	
EcanG7_00768 *	4576	T	G	EgrG1/Emul	N	Ser1526Ala	
EcanG7_00768 *	4587	C	G	Emul	S	Val1529Val	Proteína MBD
EcanG7_00768 **	4588	T	G	Emul	N	Ser1530Ala	
EcanG7_00768 *	4607	C	T	Emul	N	Thr1536Ile	
EcanG7_00768 **	4765	G	A	EgrG1	N	Ala1589Thr	
EcanG7_00768 **	4782	A	G	Emul	S	Glylu1594Glu	
EcanG7_00768 **	4803	C	T	Emul	S	Tyr1601Tyr	
EcanG7_00768 **	441	G	T	EgrG1	N	Gln147His	
EcanG7_00768 *	459	G	A	Emul	S	Ala153Ala	
EcanG7_00768 **	466	A	G	EgrG1/Emul	N	Asn156Asp	
EcanG7_00768 *	483	T	C	EgrG1	S	His161His	Proteína MBD
EcanG7_00768 #	498	ccag	cCACcag	EgrG1/Emul	N	-/His167Gln	
EcanG7_00768 *	501	-	C,G	EcG7/EgrG1/Emul	N	-/His167Gln	
EcanG7_00768 *	540	G	A	Emul	S	Ala179Ala	

(**) SNPs identificados por métodos de análisis de secuencias de NGS y por el método de Sanger; (*) SNPs identificados por método de Sanger; (#) InDel identificados por el método de Sanger; (S/N) Sinónimos/no-Sinónimos respectivamente

4.3.4.2 Nuevos marcadores moleculares para la discriminación entre especies de *Echinococcus*

Basado en los análisis de validación de SNPs se seleccionaron los cebadores que amplificaron las regiones con mayor variabilidad genética entre las especies de *Echinococcus* y mayor tamaño de producto de amplificación de manera de analizar el mayor número de SNPs que permita diferenciar entre especies. De este modo las regiones seleccionadas correspondieron al gen ECANG7_02097 (568 pb) que codifica para la subunidad alfa/beta del proteosoma y el gen ECANG7_00768 (332 pb) que codifican para la proteína MBD específica de cestodos. Se amplificaron dichas regiones en *E. canadensis* G7, *E. granulosus* G1, *E. multilocularis*, *E. canadensis* G6, *E. orteppi* G5 y *E. oligarthrus*. Las secuencias analizadas fueron usadas para construir un árbol filogenético que permitió distinguir entre las distintas especies de *Echinococcus* ([Figura IV.17](#)).

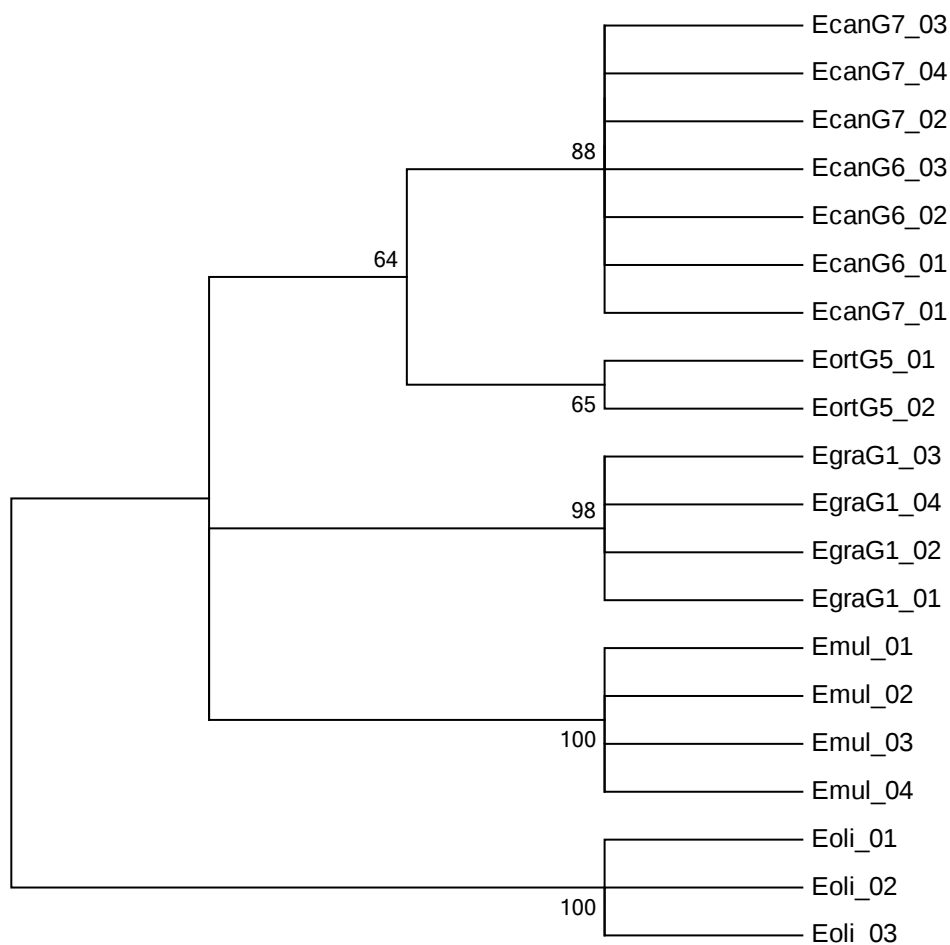


Figura IV.17: Nuevos marcadores moleculares nucleares. El análisis incluyó cinco especies de *Echinococcus* con réplicas biológicas. El árbol filogenético fue inferido utilizando el método de Máxima Verosimilitud implementado en el software MEGA5. Los números debajo de cada nodo indican el valor de *bootstrapping* mayores al 50% (para 500 réplicas). EcanG7: *E. canadensis* G7, EgraG1: *E. granulosus* G1, Emul: *E. multilocularis*, EcanG6: *E. canadensis* G6, EortG5: *E. orteppi* G5 y Eoli: *E. oligarthrus*

4.3.5 Conclusiones parciales

El estudio de la filogenia que relaciona a *E. canadensis* G7, *E. multilocularis* y *E. granulosus* G1 se abordó mediante tres análisis diferentes. Se reconstruyó la filogenia utilizando genomas mitocondriales completos que permitieron discriminar entre las especies conocidas de *Echinococcus*. Por otra parte se observó que la filogenia obtenida basada en genes de copia única no fue suficiente para determinar una relación filogenética entre estas tres especies de *Echinococcus* pero si para clasificarlos como especies diferentes. Por otra parte, el uso de SNPs para la reconstrucción de la filogenia y el análisis de los sitios polimórficos compartidos permitió determinar la relación filogenética entre estas tres especies de *Echinococcus* independientemente de las regiones genómicas utilizadas en el análisis. Los resultados del estudio del número polimorfismos y número de sitios polimórficos compartidos propone a *E. multilocularis* como la especie basal a *E. granulosus* G1 y *E. canadensis* G7. Esto sugiere que

E. multilocularis habría incorporado sustituciones hasta que ocurrió un fenómeno de especiación y luego *E. canadensis* G7 y *E. granulosus* G1 habrían divergido incorporando sustituciones independientes incrementando así su variabilidad genética. Asimismo, los marcadores moleculares identificados y analizados en esta tesis permiten diferenciar a todas las especies de *Echinococcus* analizadas hasta el momento, lo cual involucra a cuatro de las cinco especies que afectan a animales y al hombre en Argentina. Con estos resultados se aportan herramientas moleculares nuevas para la genotipificación de especies, en particular, marcadores moleculares nucleares que son escasos en la bibliografía y complementan a los marcadores mitocondriales y ribosómicos ampliamente usados [13,14,27,236].

4.4 Identificación y características particulares del repertorio génico de *Echinococcus canadensis* G7

Esta sección de la tesis está centrada en el estudio de genes de *Echinococcus* cuyos ortólogos están ausentes en otros organismos o bien poseen un alto grado de divergencia con ellos; con especial énfasis en posibles blancos terapéuticos para el control de la echinococcosis.

4.4.1 Genes como potenciales blancos de drogas o de importancia en el desarrollo del parásito

Además de los métodos quirúrgicos, los tratamientos actualmente disponibles para combatir la echinococcosis residen en quimioterapias utilizando compuestos derivados de benzimidazol dirigidos contra la beta-tubulina del parásito. Sin embargo como las beta-tubulinas de cestodos y humanos son similares solo pueden aplicarse en dosis parásitostáticas y frecuentemente producen efectos secundarios adversos. Previo a esta tesis nuestro grupo colaboró en la obtención de los genomas de *E. multilocularis* y *E. granulosus* G1 en los que se identificaron posibles blancos de drogas [43]. Los ortólogos a dichas proteínas fueron identificados entre el repertorio de genes de *E. canadensis* G7. Los análisis de homología de secuencia mostraron entre un 89,4% y un 100% de identidad con las proteínas de *E. granulosus* G1 y *E. multilocularis*, respectivamente. En cambio el porcentaje de identidad con las proteínas de *S. mansoni* fue entre 34,9% y el 97%. Los análisis de ortología revelaron que 20 de estos 21 blancos de droga propuestos poseen ortólogos en humanos por lo que una droga diseñada en base a estos blancos podría tener efectos adversos. Por este motivo, con el fin de obtener blancos de droga más eficientes y específicos contra los cestodos se seleccionaron candidatos basados en los siguientes criterios: estar presentes en todas las especies de cestodos, tener una alta conservación de secuencia entre los cestodos (umbral de puntuación MCL < 0,8) y estar ausentes o tener un alto grado de divergencia con respecto a los de los humanos.

Bajo estos criterios se seleccionaron un total de 42 proteínas de cestodos y los potenciales blancos se agruparon en 7 categorías: 1- antígenos: antígeno Taeniidae (AgB) y proteínas inmunogénicas Ts11; 2- péptidos de defensa antimicrobianos: péptido antimicrobiano; 3- neuropéptidos de señalización y hormonas peptídicas; 4- transporte: bomba de protones vacuolar (H⁺) ATPase; 5- proceso metabólico: dolicol fosfato manosil transferasa (DPM2) y mano EF de unión de calcio; 6- proceso de regulación de la transcripción: proteínas con dedos de zinc C2H2 y 7- proteínas hipotéticas conservadas. Del total de proteínas seleccionadas, solo 11 de ellas poseen anotación funcional y son las que se describen a continuación ([Tabla IV.10](#)).

Tabla IV.10: Nuevos blancos de drogas de *E.canadensis* G7 presentes en cestodos y ausentes o muy divergentes en humanos*

Categoría	Producto	<i>Ecanadensis</i> G7 IDs
Antígenos	Antígeno taeniidae (Antígeno B)	ECANG7_07838
	Proteína inmunogénica Ts11	ECANG7_01678
Defensa	Péptido antimicrobiano Taquistatina A	ECANG7_00862
Señalización	Neuropéptido	ECANG7_03703
		ECANG7_10139
	Hormona pancreática	ECANG7_09023
		ECANG7_05886
Transporte	ATPasa vacuolar (H ⁺) - subunidad G	ECANG7_02132
Proceso metabólico	Dolicol- fosfato manosiltransferasa (subunidad 3)	ECANG7_01023
	Proteína con dominio mano-EF	ECANG7_02884
Proceso de transcripción	Proteína de unión a CREB	ECANG7_05946
	Dedos de zinc tipo C2H2	ECANG7_07928

*Definidos por grupos de ortología con puntaje de umbral menor a 0,8 y presentes en todos los cestodos

4.4.1.1 Antígenos

El antígeno B (AgB) es uno de los principales antígenos que están presentes en el líquido hidatídico del quiste. Previamente hemos demostrado que es altamente polimórfico y que su perfil de expresión en *E. canadensis* G7 es diferente al de *E. granulosus* G1 [33,34]. Se han propuesto funciones importantes para este antígeno tales como la unión y transporte de lípidos [238] o la modulación de respuestas inflamatorias en las células [239]. En este estudio se identificó al AgB (ECANG7_07838) y a la proteína inmunogénica Ts11 (ECANG7_01678) en grupos de ortología específicos de cestodos. Las proteínas codificadas por los genes ortólogos de *T. solium* han sido encontrados además como productos de excreción/secreción en cultivos *in vitro* del metacestodo [240] y debido a la gran similitud en los sistemas de excreción/secreción de los helmintos es un candidato a ser estudiado más profundamente con un objetivo farmacológico.

4.4.1.2 Péptidos antimicrobianos

Entre el repertorio de genes de *E. canadensis* G7 se encontró el péptido antimicrobiano Taquistatina A (ECANG7_00862) el cual también está presente en todos los cestodos y se agrupa en el clúster específico de cestodos y por lo tanto está ausente en humanos. Este péptido está conformado por una secuencia de 92 aminoácidos con un péptido señal en el extremo N-terminal de 22 amino ácidos de longitud lo que sugiere que podría ser excretado o secretado. No existen muchas secuencias relacionadas depositadas en las bases de datos, las 11 secuencias de Taquistatina de organismos eucarióticos descritas pertenecen a artrópodos y la que posee mayor similitud (36% de identidad) a la de *E. canadensis* G7 es la Taquistatina A2 de *Tachypleus tridentatus* (Arthropoda; Q9U8X3) [241]. Esta proteína pertenece a la familia de las defensinas y posee una estructura secundaria que consiste de tres láminas beta estabilizadas por puentes disulfuro de cisteínas [242] y de un péptido señal. Las defensinas son abundantes y están altamente distribuidas en el reino animal y vegetal y se caracterizan por tener una estructura de tres láminas beta catiónicas anfipáticas y antiparalelas estabilizadas por puentes disulfuro [243]. El análisis estructural de la Taquistatina de *E. canadensis* G7 al igual que la del resto de los cestodos reveló la presencia de estructuras en láminas beta conteniendo residuos de cisteína que podrían participar en la estabilización de dicha estructura. Asimismo, y exhibe propiedades físico químicas descritas para péptidos antimicrobianos: carga neta +2, 10 kDa, ~50% de contenido de aminoácidos hidrofóbicos, la presencia de un péptido señal y un tamaño pequeño. Por otra parte, también se ha descrito recientemente que este péptido presenta un alto nivel de expresión, lo que podría indicar un rol biológico importante en el parásito [43,78].

4.4.1.3 Hormonas peptídicas

Las hormonas peptídicas y neuropéptidos forman parte de las moléculas de señalización de los metazoos más diverso funcional y estructuralmente. La familia de péptidos PP-fold es una de las familias de hormonas peptídicas capaz de unirse a varios receptores acoplados a proteína G, (GPCRs) [244]. Los GPCR están involucrados en muchas enfermedades y son blanco de aproximadamente el 40% de los fármacos usados en la actualidad para el tratamiento de enfermedades [245]. Este tipo de péptidos han sido identificados en *S. mediterranea* y *Schistosoma* sp. y la organización genómica de los genes que codifican para ellos parece estar conservada en platelmintos. Los miembros de la familia PP-fold se agruparon en un grupo de ortología específico de platelmintos junto con proteínas de la categoría “Péptidos de hormonas pancreáticas”. En ese grupo se encontraron 4 genes de *E. canadensis* G7. Dos de ellos, los genes ECANG7_09023 y ECANG7_05886 codifican para un péptido similar a hormonas pancreáticas cuyo dominio de unión a receptor y dimerización se encuentra poco conservado. Por otro lado para el gen ECANG7_09023, el dominio de unión a receptor está 100% conservado en *Echinococcus* pero posee solo el 55% de identidad de

secuencia con su correspondiente ortólogo en *H. microstoma* [43] y 60% de identidad con

S. mediterránea. En cambio para el gen ECANG7_05886 de *E. canadensis* G7 se observó que el dominio presenta un 100% de identidad con sus correspondientes ortólogos en *E. multilocularis* y *E. granulosus* G1, un 81% de identidad con el de *H. microstoma* [43] y un 75% de identidad con *Aplysia californica* (Lophochotrozoa/Mollusca/Gastropoda), el organismo invertebrado más relacionado sobre el cual se cuenta con información al respecto [246] y para todos ellos hay evidencia de transcripción que sustenta su expresión.

4.4.1.4 Transporte

La V-ATPasa es una bomba de protones de células eucarióticas que lleva a cabo actividades esenciales. Es la encargada de generar un potencial electroquímico mediante la formación de un gradiente de concentración de protones impulsado por la energía química del ATP [247]. Las V-ATPasas intervienen en funciones celulares muy importantes tales como la endocitosis, tráfico intracelular de membranas degradación de proteínas, y transporte acoplado a pequeñas moléculas y iones [248,249].

En nematodos se ha descrito que las proteínas V-ATPasas está involucrada en funciones tales como nutrición, osmoregulación, síntesis de la cutícula, y asociadas a funciones neurobiológicas y de reproducción [250]; sin embargo en platelmintos su rol aún no ha sido estudiado. En *E. canadensis* G7 se encontraron 3 genes que codifican para la V-ATPasa y sus correspondientes genes ortólogos en otros cestodos. Entre ellos el gen ECANG7_02132 codifica para una proteína que ha sido agrupada en un clúster específico de platelmintos y mostró un alto grado de conservación de secuencia y de su estructura secundaria (estructura alfa hélice en el extremo N-terminal) con la V-ATPasa del gusano del tabaco *Manduca sexta*, que es el organismo más cercano en el cual ha sido caracterizada esta proteína [251] ([Figura IV.18](#)).

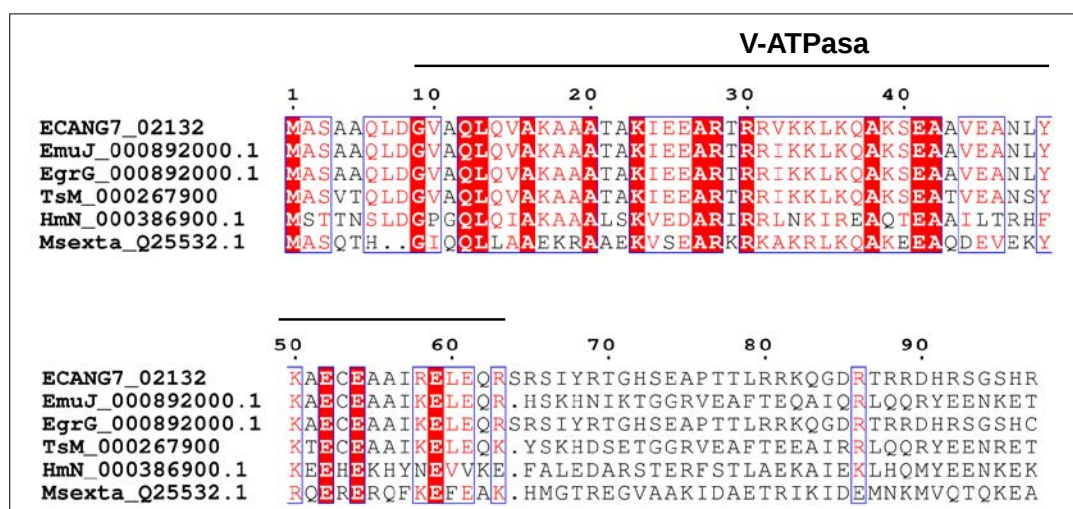
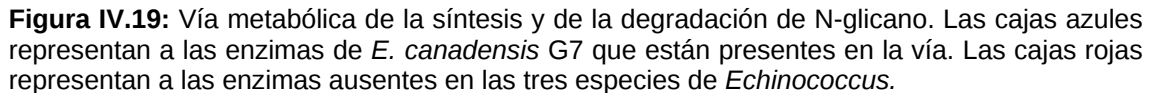


Figura IV.18: Alineamiento múltiple de la secuencias de proteínas V-ATPasa de cestodos y *M. sexta*.

4.4.1.5 Metabolismo

En la categoría “metabolismo” se encontró una enzima específica de cestodos que estaría involucrada en procesos de glicosilación. En organismos eucarióticos la glicosilación es crítica para un correcto plegamiento de las proteínas y para la actividad de las enzimas. Además participa en el mecanismo de control de calidad de las proteínas, proceso en el que interviene el retículo endoplásmico (ER) asociado a vías de degradación (ERAD). Parte del proceso de glicosilación tiene lugar en el lumen del ER, donde las manosas y glucosas son transferidas a moléculas aceptoras. La manosilación en el lumen del ER es un paso común a cuatro vías de glicosilación: N-glicosilación, anclajes de proteína fosfatidilinositol (GPI), O-glicosilación de proteínas y C-glicosilación de proteínas [252]. La manosa dolicol fosfato es un donador de manosil central que interviene en las cuatro vías mencionadas y es sintetizado a partir de GDP-manosa y dolicol fosfato por la enzima Dolicol fosfato beta D manosil transferasa (DPM) (EC 2.4.1.83). En mamíferos esta enzima es un complejo formado por tres proteínas: DPM1 que posee el sitio catalítico, DPM2 y DPM3. La proteína DPM3 es una proteína de 92 aminoácidos de secuencia que se localiza en la región membranosa del ER [253], y es importante para la estabilización y localización de DPM1 que a su vez es estabilizada por DPM2. En *E. canadensis* G7 se encontraron dos subunidades del complejo DPM. La subunidad DPM1 codificada por el gen ECANG7_04674 que posee ortólogos en todos los metazoos analizados. Y la subunidad DPM3 codificada por el gen ECANG7_01023 que está presente solo en cestodos y que está conformado por dos regiones transmembrana y el dominio “DMP3 *superfamily*”. El alto porcentaje de conservación de secuencia en todas las especies de cestodos (90% entre las especies de *Echinococcus* y 65% con las otras especies de cestodos) y la baja conservación de secuencia con su contraparte humana y de otros organismos sugiere que se trata de una nueva clase de DPM3 ([Figura IV.19](#)). Debido a que la enzima DPM3 es el único miembro divergente de ésta vía metabólica en *Echinococcus* y cumple un rol importante a nivel regulatorio en la biosíntesis de precursores de N-glicano podría representar un blanco terapéutico.



Las proteínas dedos de Zinc son una clase de proteínas regulatorias que participan en una variedad de procesos celulares, tales como en el desarrollo, diferenciación celular y supresión de tumores entre otros. Los genes que codifican para proteínas dedos de zinc C2H2 son miembros de una de las más grandes y complejas súper familias de genes en los genomas de metazoos [254]. Algunos subgrupos linaje específico de proteínas portadoras de dedos de Zinc C2H2 pueden encontrarse tanto en levaduras como en nematodos, insectos y plantas [255]. Estas proteínas cumplen múltiples funciones; como por ejemplo la unión a ARN y como mediadores en la interacción entre proteínas. Sin embargo son mejores conocidas por su rol como proteínas de unión a secuencias específicas de ADN, particularmente en humanos donde se ha descrito que se unen específicamente a las secuencias de ADN metilado. Tales proteínas exhiben dominios de dedos de zinc que están organizados en repeticiones en tándem de dos, tres o más dedos de zinc conformando el dominio de unión a ADN de la proteína. En *E. canadensis* G7 se encontraron 124 genes que codifican para proteínas que contienen dominios C2H2. Entre todas ellas el gen ECANG7_07928 codifica para una proteína de 125 aminoácidos que contiene dominios dedos de zinc tipo C2H2 y están presentes exclusivamente en cestodos. Las proteínas de unión a secuencias específicas del ADN en conjunto con patrones característicos de metilación del ADN podrían tener un rol importante en la regulación de la transcripción génica.

4.4.2 Familias de genes expandidas

Las expansiones génicas son uno de los determinantes de la complejidad y del tamaño de un genoma. Los cambios en el número de copias pueden cambiar los niveles de expresión de los genes incluidos en las regiones de “número de copias variable”, permitiendo niveles de transcripción superiores o inferiores a los que mostrarían mediante el control de la transcripción de copias individuales. Las copias adicionales de genes también proporcionan redundancia en secuencia, de modo que algunas copias son pueden desarrollar funciones nuevas, mientras que otras copias mantienen la función original. Las duplicaciones génicas son la fuente de nuevos genes que confieren nuevos fenotipos y pueden contribuir a la evolución de la especiación [256]. En esta sección de la tesis se identificaron y analizaron familias de genes expandidos en *Echinococcus*.

Mediante la construcción y clasificación de los grupos de ortología se encontraron grupos conteniendo familias de genes para una misma especie. En *Echinococcus* se identificaron un total de 26 familias expandidas de entre 10 y 66 miembros, entre las cuales se encontró a la familia de proteínas de choque térmico “heat-shock protein 70 (Hsp70)”, presente en todos los platelmintos y previamente descrita como familia expandida por nuestro grupo [43]. También cabe destacar en esta sección a otras 3 familias de proteínas expandidas: las proteínas portadoras de motivos GPS, las enzimas de conjugación de ubiquitina y las proteínas Glicosil transferasas presentes solo en organismos cestodos. La proteína de motivos GPS está relacionada con la policistina-1, una proteína involucrada en la vía central de transducción de señales donde el motivo GPS es el responsable de las interacciones proteína-proteína. Las policistinas forman familias de proteínas compuestas por múltiples miembros en peces, invertebrados, mamíferos y humanos [257]. La enzima de conjugación de ubiquitina es conocida por su rol en la vía de ubiquitinación, en la modulación de la degradación de proteínas e interacciones proteína-proteína. El complejo de ubiquitinación UBC esta codificado por hasta 19 genes de *Echinococcus* cuyos alineamientos de las proteínas mostraron un alto grado de conservación del dominio “UBC superfamily” (PF00179) entre los miembros del grupo de cestodos. La tercera familia de proteínas expandidas es la de las glicosil transferasas, las cuales están involucradas en la biosíntesis y modificación de glicanos. Esta vía podría tener un rol muy importante en la biogénesis de la capa laminar acelular de *Echinococcus* que se encuentra enriquecida en carbohidratos y es una característica específica del género *Echinococcus* y uno de los principales rasgos morfológicos diferenciales entre las especies de *Echinococcus*.

4.4.3 Conclusiones parciales

Los análisis de los grupos de ortología e identidad de secuencia de las proteínas que han sido propuestas como blancos de drogas permitieron identificar un alto grado de similitud con las proteínas humanas, por lo cual las drogas cuyos blancos sean estas proteínas podrían causar efectos adversos en la salud humana. Por este motivo se

evaluaron proteínas candidatas a nuevos blancos de drogas que se agruparon en grupos de ortología específicos de cestodos y por lo tanto divergentes de cualquier homólogo en humanos. Estas proteínas en conjunto con ciertas características funcionales entre las que se destacan los péptidos antimicrobianos, V-ATPasas específicas del género, DPM3 y factores de transcripción son propuestas como nuevos blancos de drogas contra la echinococcosis.

4.5 Mecanismos de regulación génica en *Echinococcus*

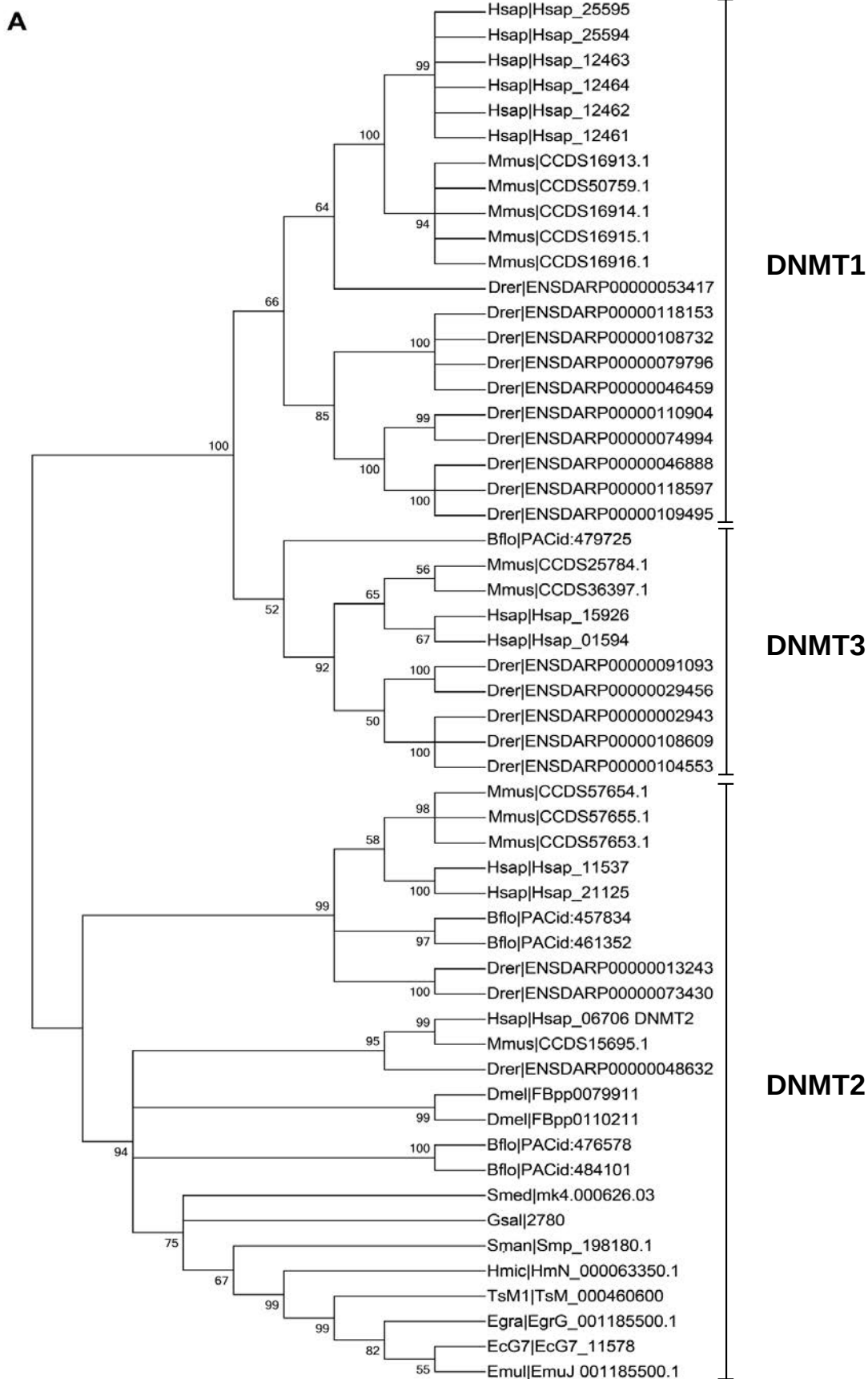
Los mecanismos de regulación génica en *Echinococcus*, así como la estructura genómica que da a lugar a la expresión y regulación de sus genes es un área aun no estudiada en este género ni en géneros relacionados. En esta sección de la tesis se aborda el estudio de tales características mediante la identificación de elementos genómicos estructurales, características epigenéticas y analizando estructuralmente uno de los principales efectores en la vía de pequeños ARN reguladores de la expresión génica.

4.5.1 Epigenética y sistema de metilación del ADN

Las ADN metil transferasas (DNMTs) y las proteínas con dominio de unión a methyl-CpG (MBDs) en conjunto con otras proteínas constituyen el sistema de metilación central de los metazoos, tanto en especies de vertebrados como en invertebrados. A su vez, la metilación de las citosinas del ADN en platelmintos es una característica epigenética muy conservada a través de este filo.

Las ADN metil transferasas de los metazoos son una familia de proteínas constituida por las DNMT1, DNMT2, DNMT3a/3b [258]. En el nuevo genoma generado en esta tesis se identificó el gen que codifica para la DNMT2 (ECANG7_11578) pero no se encontraron genes codificantes para DNMT1 ni para DNMT3 en ninguna de las especies de *Echinococcus*. Para confirmar la clasificación de la DNMT de *Echinococcus* se construyó un árbol filogenético con todas las proteínas pertenecientes al grupo de ortología de las DNMTs y se observó un clado que agrupó a la DNMT de *Echinococcus* junto con todas las proteínas DNMTs clase 2 de metazoos ([Figura IV.20 A](#)). Por otra parte entre el repertorio de genes de *E. canadensis* G7 se identificaron 2 genes que codifican para las proteínas MBD. Para corroborar la anotación se realizaron análisis de secuencias y filogenéticos del correspondiente grupo de ortología. En dichos análisis se corroboró la anotación y se observó que la proteína MDB codificada por el gen ECANG7_05304 posee ortólogos en platelmintos y otras especies de invertebrados mientras que la proteína MDB codificada por el gen ECANG7_00768 fue identificada como un nuevo tipo de MDB. Esta nueva proteína MDB posee un dominio de unión a CpG formado por 73 aminoácidos, un dominio PHD (PF00628) de 48 aminoácidos y un bromodominio (PF00439) en el extremo carboxilo terminal constituido por 86 residuos de aminoácidos que sería el responsable de unión al ADN. Los dominios identificados se encuentran conservados en cestodos y las nuevas

MBDs se agrupan filogenéticamente en un nodo específico de la clase Cestoda ([Figura IV.20 B y C](#)).



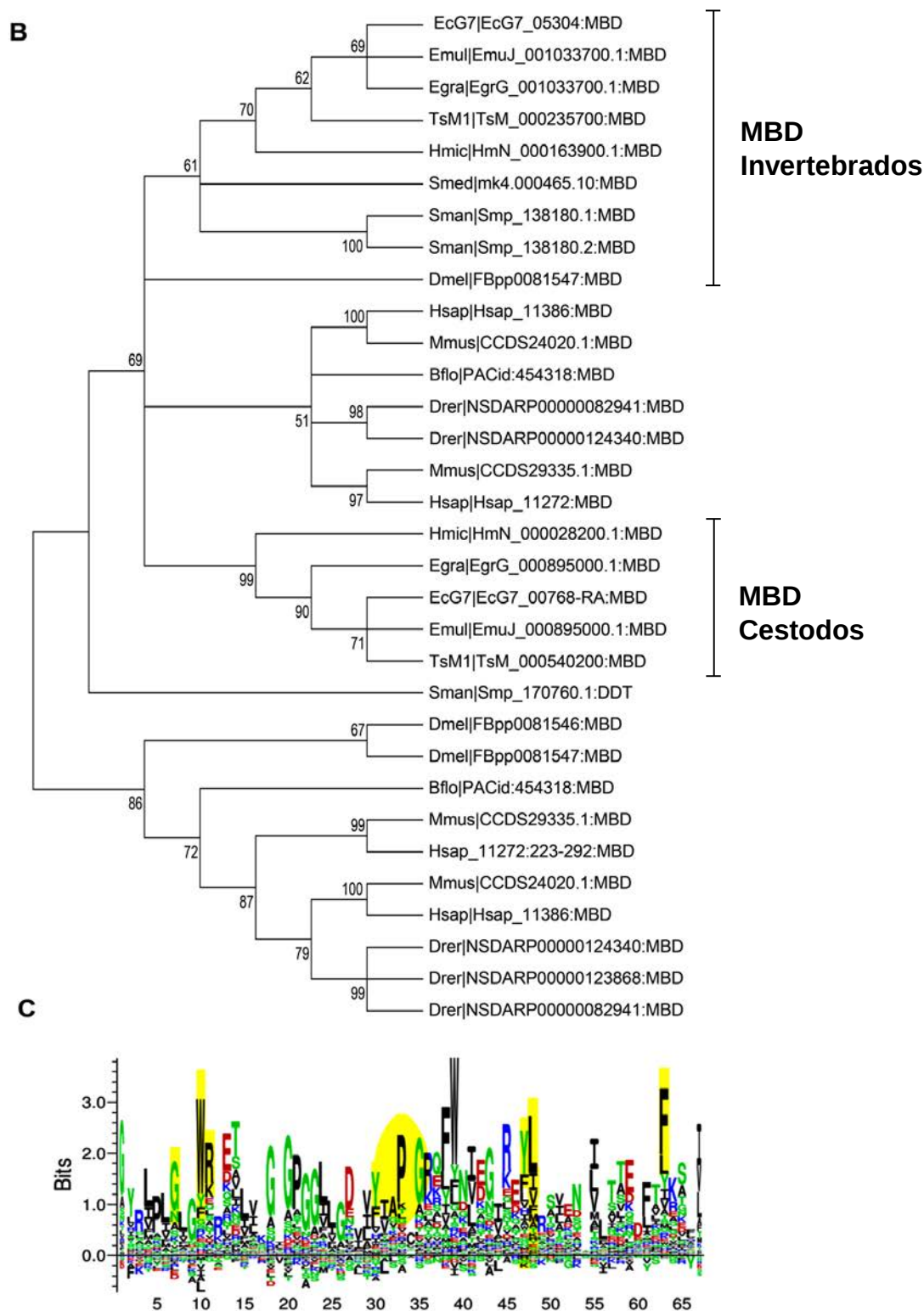


Figura IV.20: Proteínas DNMT y MBD de *Echinococcus* spp y su clasificación filogenética respecto a sus ortólogos en especies de metazoos. A). Árbol filogenético de las proteínas DNMTs de *Echinococcus* agrupadas en familias. B). Árbol filogenético de las proteínas MBD de *Echinococcus* y su clasificación en familias. Los recuadros señalan al grupo de proteínas conservado en organismos invertebrados y proteínas MBDs que fueron identificadas específicamente en cestodos. C). Logo de secuencia de aminoácidos involucrados en la formación Dominio MBD.

4.5.2 Identificación y localización genómica de Islas CpG

Las regiones promotoras en los genomas de *Echinococcus* han sido poco estudiadas hasta el momento; por ello en esta sección de la tesis se aborda el estudio de CGIs en los genomas de *Echinococcus* como una aproximación a la caracterización de regiones promotoras. Con este propósito se identificaron CGIs en las tres especies de *Echinococcus*, y se determinó su localización, el contenido de GC y la densidad de CGIs en los genomas completos, en las regiones codificantes y en regiones localizadas río arriba de los CDS.

En *E. canadensis* G7, *E. multilocularis*, y *E. granulosus* G1 se identificaron 4200, 4297 y 4249 CGIs respectivamente con una densidad de 36,5; 37,4 y 36,9 CGIs por Mb. En comparación con estudios llevados a cabo sobre genomas de mamíferos [259] los genomas de *Echinococcus* presentaron mayores niveles de densidad de CGIs y también una mayor relación de CpG observado/CpG esperado (CpG_{obs}/CpG_{esp}). Sin embargo la distribución de los tamaños de las CGIs fue similar a la de los mamíferos. Al evaluar la distribución de las CGIs en las regiones codificantes se observó que el número y densidad de CGIs es mucho mayor que en las regiones no codificantes. Este resultado también es similar al observado en los genomas de mamíferos.

Por otra parte, para identificar posibles regiones promotoras se analizaron las CGIs a una distancia de hasta 5 kb río arriba de las regiones codificantes y se identificaron 1363, 1637 y 1457 CGIs en los genomas de in *E. canadensis* G7, *E. multilocularis* y *E. granulosus* G1, respectivamente. La densidad de CGIs por Mb se calculó utilizando solamente las regiones en las cuales se encontraron CGIs y sus valores fueron 254,9; 253,1 y 241,7, respectivamente. Estos valores fueron ligeramente menores a los observados para las regiones codificantes pero mayores a las del resto del genoma ([Tabla IV.11](#)). Posteriormente se evaluó la correlación entre la densidad de CGIs y algunas características genómicas tales como el porcentaje de GC del genoma, tamaño de contig y la relación CpG_{obs}/CpG_{esp} en los tres genomas de *Echinococcus* ([Tabla IV.12](#)). Asimismo, se observó una correlación significativa entre la densidad de CGIs y el porcentaje de GC del genoma y también con el tamaño de contig. Sin embargo, se observó una correlación muy baja entre las CGIs y el cociente CpG_{obs}/CpG_{esp} . Adicionalmente se estimó la distribución de la densidad de CGIs por Mb de contig usando solo los contigs que contienen CGIs, mediante lo cual se observó que la mediana de densidad de CGIs en *E. canadensis* G7 y *E. granulosus* G1 es menor que la de *E. multilocularis* ([Figura IV.21](#)).

Tabla IV.11: Islas CG (CGIs) y su distribución en los genomas de *Echinococcus*

	Especies	<i>E. canadensis</i> G7	<i>E. multilocularis</i>	<i>E. granulosus</i> G1
Genoma	Tamaño (Mb)	115,0	115,0	115,0
	Número de contigs	9.330,0	1.815,0	1.288,0
	Contenido de GC (%)	41,9	42,3	41,9
CGIs en los genomas totales	Número de CGIs	4.200,0	4.297,0	4.249,0
	Densidad de CGIs / Mb	36,5	37,4	36,9
	Tamaño promedio de CGIs (pb)	739,4	952,1	754,9
	Contenido de GC	55,1	55,1	55,0
	CpG obs / CpG esp	0,9	0,8	0,9
	Número de CGIs	4.607,0	4.597,0	4.550,0
	Tamaño de región codificante (Mb)	16,3	16,0	15,3
CGIs en regiones codificantes	Densidad de CGIs / Mb	282,5	287,7	298,2
	Tamaño promedio de CGIs (pb)	2.965,0	2.502,9	2.625,6
	Contenido de GC	52,8	53,3	53,1
	CpG obs / CpG esp	0,9	0,9	0,9
	Número de CGIs	1.363,0	1.637,0	1.457,0
	Tamaño de región río arriba (Mb)	5,3	6,5	6,0
CGIs 5 kb río arriba de las regiones codificantes	Densidad de CGIs / Mb	254,9	253,1	241,7
	Tamaño promedio de CGIs (pb)	3.866,6	3.909,7	4.090,8
	Contenido de GC	43,8	44,1	43,3
	CpG obs / CpG esp	0,9	0,9	0,9
	Número de CGIs	1.363,0	1.637,0	1.457,0

Tabla IV.12: Correlación entre la densidad de CGI y las características genómicas en diferentes regiones del genoma de las tres especies de *Echinococcus*.

		CGIs en genomas completos		CGIs en regiones codificantes		CGIs 5 kb río arriba	
		r	P	r	P	r	P
<i>E. canadensis</i> G7	log tamaño de región	-0,826	2,2E-16	-0,723	2,2E-16	-0,562	2,2E-16
	Contenido de GC	0,782	2,2E-16	0,413	2,2E-16	0,305	2,2E-16
	Obs/Esp	-0,073	2,1E-06	-0,083	1,6E-08	0,305	2,2E-16
<i>E. multilocularis</i>	log tamaño de región	-0,698	2,2E-16	-0,672	2,2E-16	-0,496	2,2E-16
	Contenido de GC	0,541	2,2E-16	0,365	2,2E-16	0,285	2,2E-16
	Obs/Esp	-0,114	6,5E-14	-0,107	3,0E-13	0,285	2,2E-16
<i>E. granulosus</i> G1	log tamaño de región	-0,645	2,5E-15	-0,661	2,2E-16	-0,657	2,2E-16
	Contenido de GC	0,605	3,2E-10	0,337	2,2E-16	0,271	2,2E-16
	Obs/Esp	-0,044	5,4E-05	-0,058	9,5E-05	0,271	2,2E-16

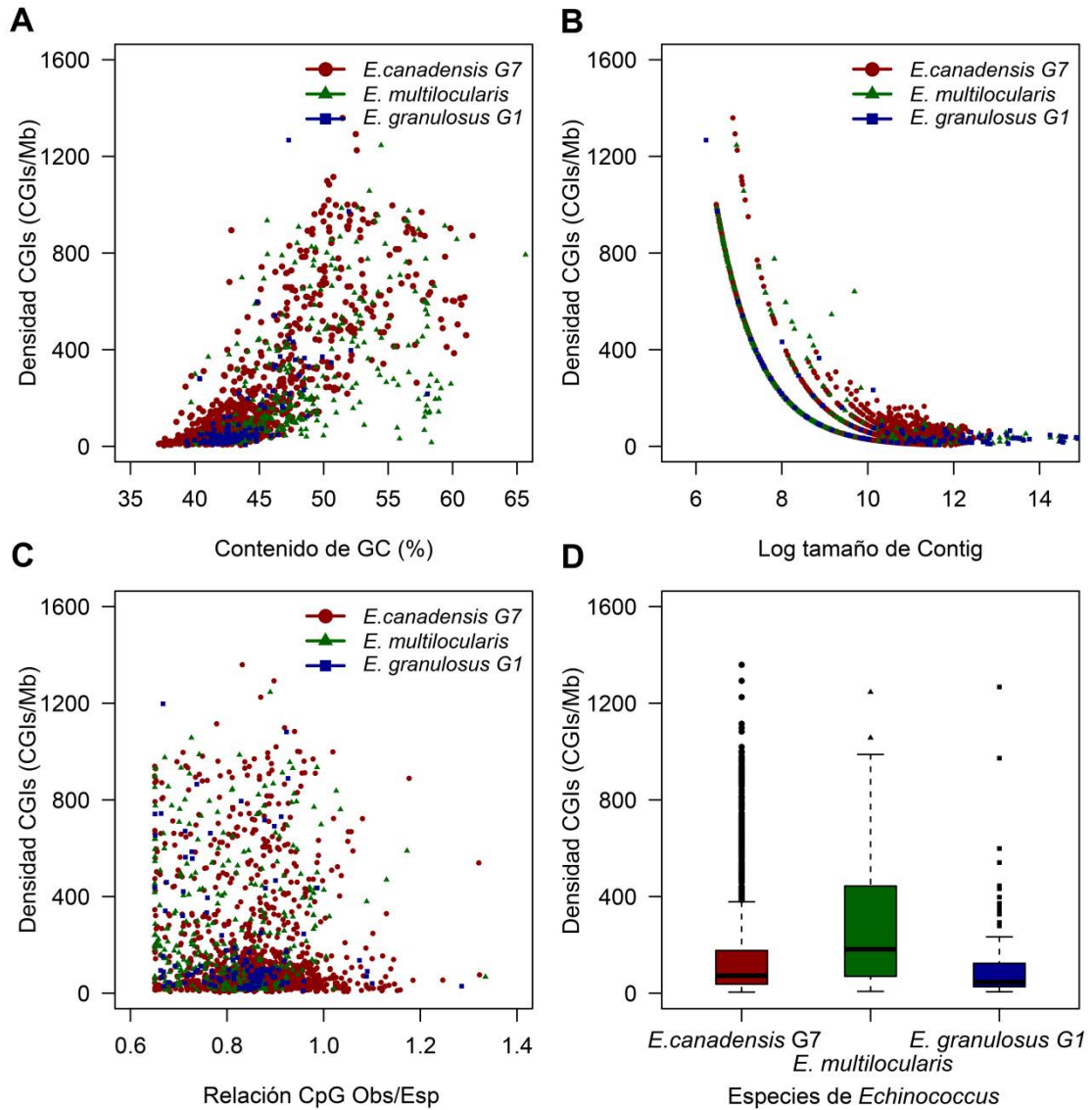


Figura IV.21: Islas CG en los genomas completos: Correlación entre la densidad de CGIs y los genomas entre las tres especies de *Echinococcus*. A). Densidad de CGIs (Mb) versus contenidos de GC (%) de los contigs. B). Densidad de CGIs (Mb) versus logaritmo del tamaño de los contigs. C). Densidad de CGIs (Mb) versus relación CpG_{obs}/CpG_{esp} . D). Densidad de CGIs (Mb) en cada especie de *Echinococcus*.

Por otra parte, la correlación entre la densidad de CGIs y el porcentaje de GC fue diferente entre las tres regiones analizadas (genoma completo, regiones codificantes y 5 kb río arriba de las regiones codificantes) observándose una mayor correlación entre la densidad de CGIs y el tamaño de los contigs de los genomas, seguida por la correlación entre CGIs y las regiones codificantes y en último lugar con respecto a su localización 5 kb río arriba de las regiones codificantes. Dichas diferencias se manifestaron en las tres especies de *Echinococcus* ([Figuras IV.22](#) y [Tablas IV.12](#)). También se observó que la media del tamaño de las CGIs es distinta entre las tres regiones genómicas, con un valor mayor para las regiones localizadas 5 kb río arriba de los CDS.

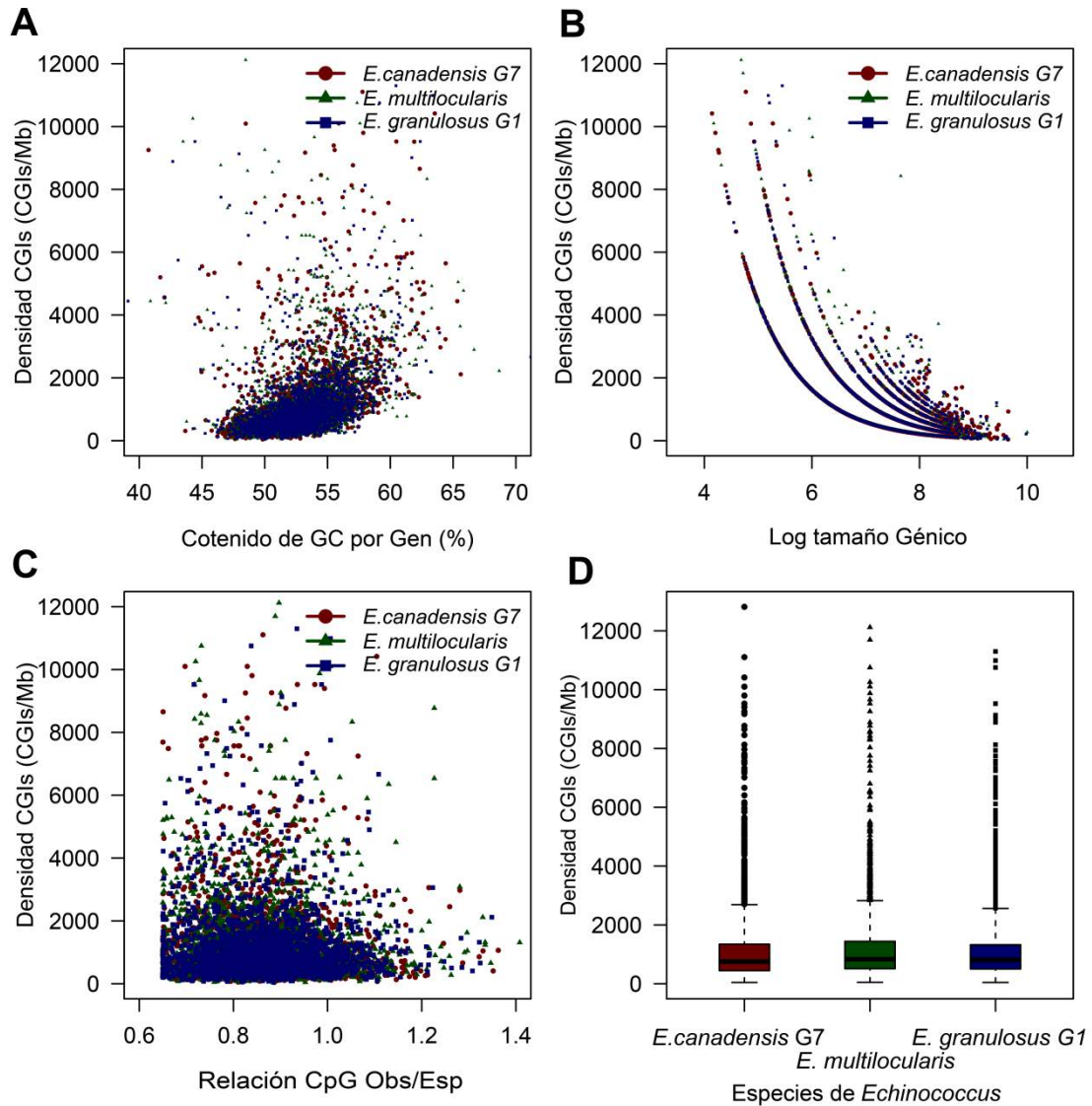


Figura IV.22: Islas CG en las regiones codificantes: Correlación entre la densidad de CGIs y las regiones codificantes en las tres especies de *Echinococcus*: A). Densidad de CGIs (Mb) versus contenidos de GC (%) de los contigs. B). Densidad de CGIs (Mb) versus logaritmo del tamaño de los contigs. C). Densidad de CGIs (Mb) versus relación CpG_{obs}/CpG_{esp} . D). Densidad de CGIs (Mb) en cada especie de *Echinococcus*.

En las regiones 5 kb río arriba de los CDS se observó que la distribución de la densidad de las CGIs en función del contenido de GC presenta cúmulos de puntos en valores de densidad específicos y en un rango variable del porcentaje de GC (entre 35% y 55%). En la [figura IV.23](#) se pueden distinguir grupos de regiones con tres valores diferentes de densidad de CGIs dados por la acumulación de puntos distribuidos horizontalmente (panel A) y en las curvas que se forman en el gráfico de la relación del tamaño de la regiones 5 kb río arriba con la densidad de las CGIs (panel B). Para estas regiones además se observó una correlación positiva entre la densidad de las CGIs y la relación CpG_{obs}/CpG_{esp} y un valor promedio de densidad de las CGIs de 254,9; 253,1 y 241,7 para *E. canadensis* G7, *E. multilocularis* y *E. granulosus* G1 respectivamente.

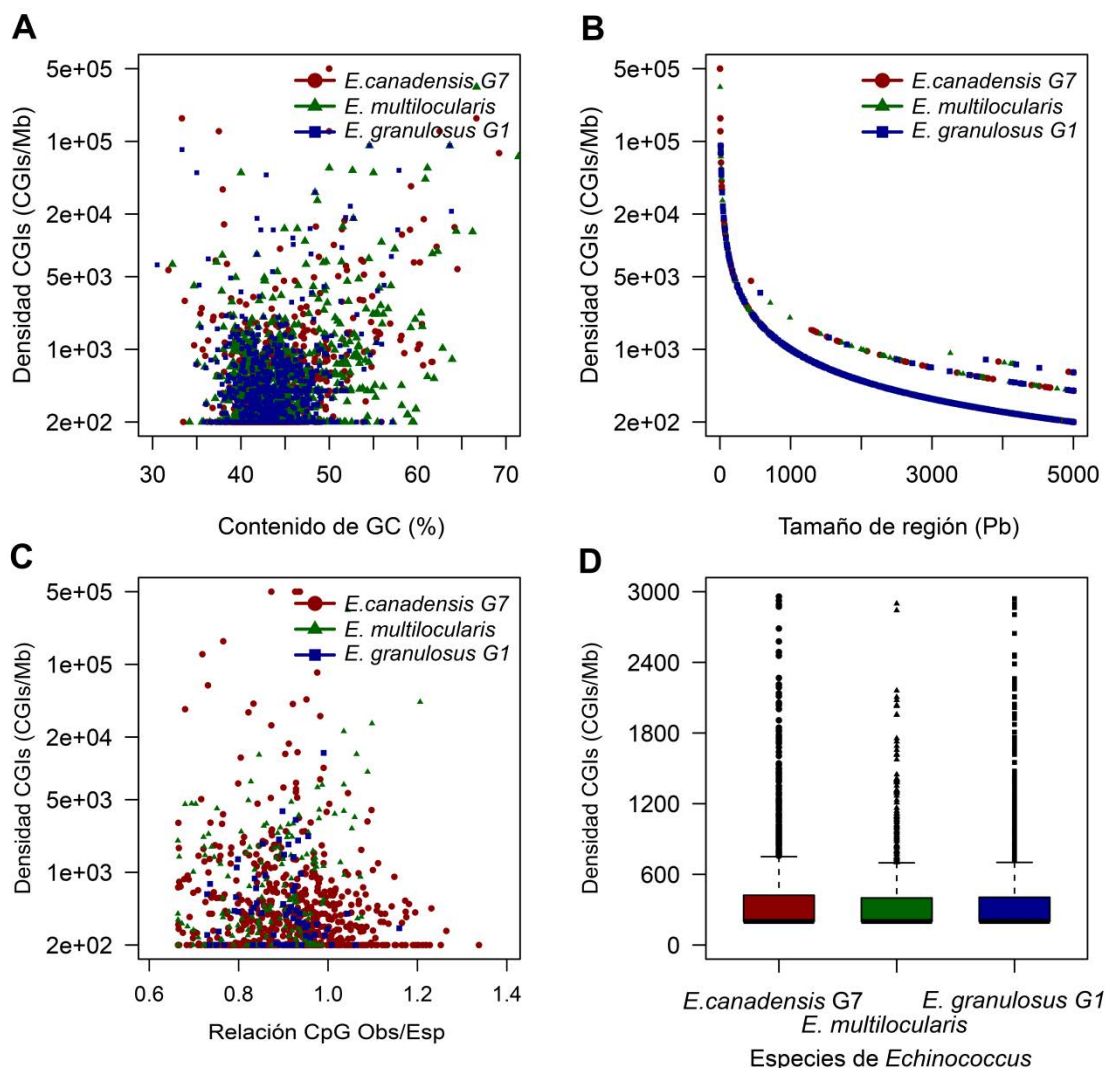


Figura IV.23: Islas CG en las regiones promotoras: Correlación entre la densidad de CGIs y las regiones de hasta 5 kb río arriba del codón de inicio ATG en las tres especies de *Echinococcus*: Correlación entre la densidad de CGIs y las regiones. A). Densidad de CGIs (Mb) versus contenidos de GC (%) de los contigs. B). Densidad de CGIs (Mb) versus logaritmo del tamaño de los contigs. C). Densidad de CGIs (Mb) versus relación CpG_{obs}/CpG_{esp} . D). Densidad de CGIs (Mb) en cada especie de *Echinococcus*.

Para identificar posibles regiones promotoras se analizaron los patrones de distancia de la localización de las CGIs desde el inicio de la traducción (ATG) de las regiones codificantes para las tres especies de *Echinococcus*. Mediante este análisis se identificaron un total de 1405, 1752 y 1457 CGIs precediendo a los genes de *E. canadensis* G7, *E. multilocularis* y *E. granulosus* G1, respectivamente y se observó que la mediana de la distancia de las CGIs desde el codón de inicio de la traducción fue de aproximadamente 2 kb y el rango para la mayoría de los genes alternó entre 500 y 3500 pb (Figura IV.24).

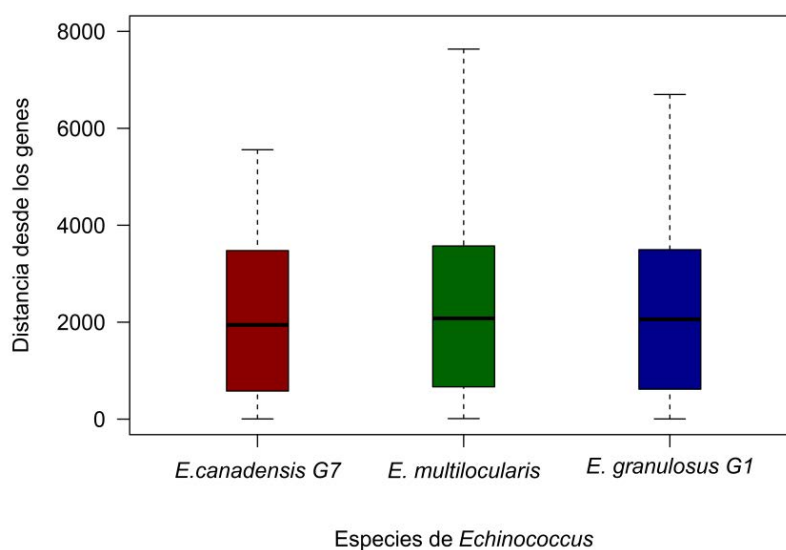


Figura IV.24: Localización de islas CG en *Echinococcus*. Distribución y mediana de la distancia de localización de las CGIs desde el codón de inicio de la traducción de cada gen.

Por otra parte también se identificaron CGIs que están diferencialmente distribuidas entre los genomas de las tres especies de *Echinococcus*. Se observó una distribución diferente que está dada por la presencia/ausencia de CGIs río arriba de determinados genes que son ortólogos entre las especies de *Echinococcus*. Estas observaciones sugieren que la expresión de determinados genes podría estar sujeta a una regulación diferencial dada por la localización genómica diferencial de CGIs río arriba de dichas regiones codificantes. El número de genes que están diferencialmente precedidos por las CGIs fueron 501 genes entre *E. canadensis* G7 y *E. granulosus* G1, 548 genes entre *E. canadensis* G7 y *E. multilocularis* y 357 genes entre *E. multilocularis* y *E. granulosus* G1. Cabe destacar que entre estos grupos de genes se encontraron algunos de los genes que han sido propuestos como blancos de drogas. Algunas estructuras genómicas representativas de estas características puede observarse en la [figura IV.25](#).

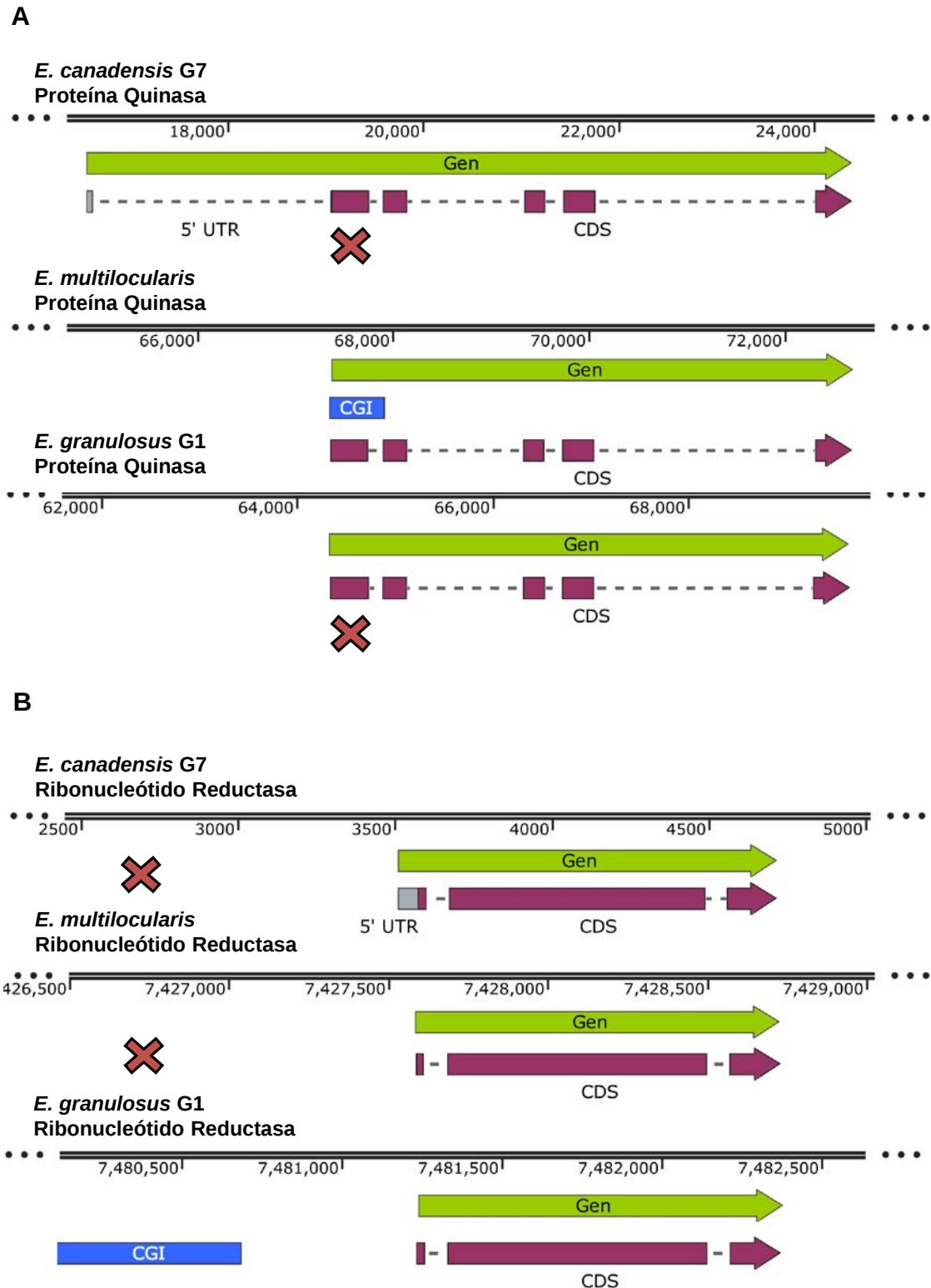


Figura IV.25: Contexto genómico diferencial dado por presencia/ausencia de CGIs río arriba de los genes en los genomas de las especies de *Echinococcus*: A). Proteína quinasa (ECANG7_03687). B). Ribonucleótido reductasa (ECANG7_06911).

4.5.3 Maquinaria de microARNs; disección de Proteínas

Argonautas - Modelado por Homología

Previamente a esta tesis, nuestro grupo de investigación determinó que las especies de *Echinococcus* poseen 4 proteínas Ago [43]. Los análisis filogenéticos y de ortología realizados mostraron la existencia de dos clados: uno de ellos correspondiente a las proteínas Ago conservadas en todos los metazoos analizados y el otro clado correspondiente a las "Ago/Piwi-like" similares entre sí pero distantes a las descriptas en mamíferos [43]. En *E. canadensis* G7 se identificaron los 4 miembros de la familia Ago, las cuales presentaron un alto porcentaje de identidad con sus respectivos ortólogos presentes en las demás especies de *Echinococcus*. Con el fin de identificar dominios y motivos conservados en estas proteínas y su posible interacción con los microARNs se obtuvieron las estructuras 3D de tres de las cuatro proteínas Ago mediante modelado por homología. En este análisis las proteínas Ago se clasificaron en 4 grupos (1 a 4), el primer grupo contiene a la proteína Ago conservada en metazoos que posee un alto porcentaje de identidad con las Ago-2 humanas. Los otros 3 grupos contienen a las proteínas Ago de *Echinococcus* del nuevo clado y si bien mostraron porcentajes de identidad de entre el 37% y 65% con la proteína Ago-2 humana, solo poseen ortólogos en otras especies de platelmintos, haciéndolas específicas del Filo. Los dominios identificados para cada proteína pueden observarse en la [Tabla IV.13](#).

Tabla IV.13: Dominios identificados en las proteínas Ago de *Echinococcus*

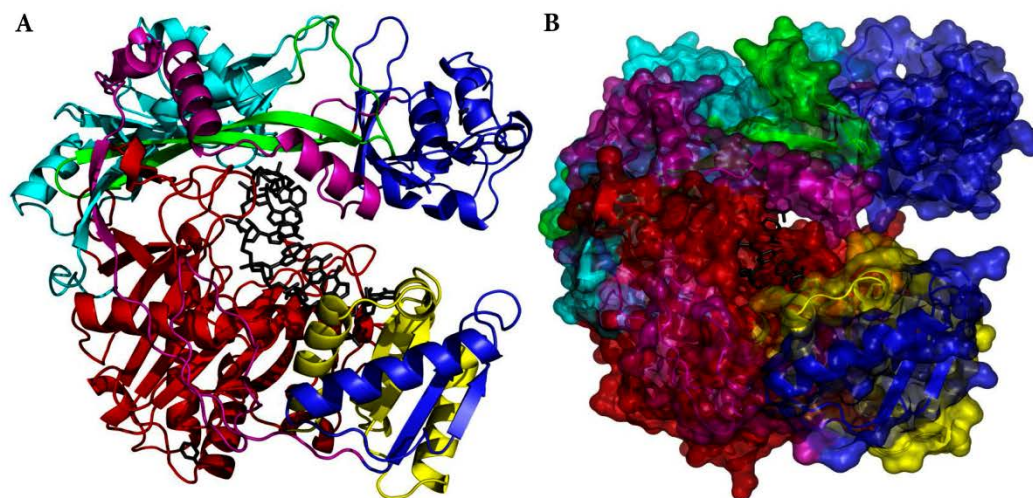
Grupo	Dominio	Pfam Id	Gen Id
Grupo1	Piwi	PF02171.14	ECANG7_05655 EmuJ_000346200 EgrG_000346200
	ArgoMid	PF16487.2	
	PAZ	PF02170.19	
	ArgoN	PF16486.2	
	ArgoL1	PF08699.7	
	ArgoL2	PF16488.2	
Grupo2	Piwi	PF02171.14	ECANG7_09582 EmuJ_000911600 EgrG_000911600
	PAZ	PF02170.19	
	ArgoN	PF16486.2	
	ArgoL1/ACT_6 (*)	PF08699.7 / PF13740.3 (*)	
	ArgoL2	PF16488.2	
Grupo 3	Piwi	PF02171.14	ECANG7_09607 EmuJ_000739100 EgrG_000739100
	PAZ	PF02170.19	
Grupo 4	Piwi	PF02171.14	ECANG7_09652 EmuJ_000911700 EgrG_000911700
	PAZ	PF02170.19	
	ArgoL1	PF08699.7	
	ArgoN	PF16486.2	

(*) EmuJ_000911600 de *E. multilocularis* contiene el dominio ArgoL1. ECANG7_09582 y EgrG_000911600 de *E. canadensis* G7 y *E. granulosus* G1 respectivamente contienen el dominio ACT_6

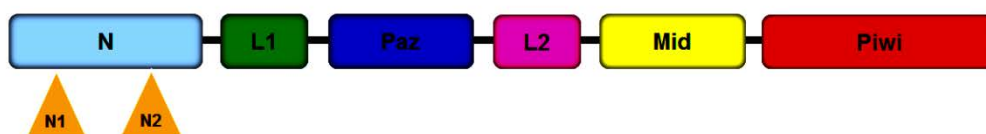
Cada una de las estructuras obtenidas exhibió una conformación RNAsaH-like tipo abrazadera que concuerda con las estructuras de las proteínas de la familia Ago/Piwi estudiadas hasta el momento [207,260] ([Figura IV.26 A y B](#)). Excepto las proteínas del grupo 3 que carecen del extremo N-terminal y de un segmento del dominio PAZ. Las proteínas Ago forman un canal o bolsillo en el cual reside el microARN e interactúa con ciertos residuos del dominio Piwi, Mid y PAZ y a través del cual el ARNm tiene acceso para aparearse con el microARN [207]. El dominio ArgoN localizado en el extremo N-terminal contribuye a la conformación abrazadera en el plegamiento de la misma, y se ha descrito que está involucrado en la actividad de clivaje del ARNm. A pesar de que a modo general los dominios se encuentran conservados en las proteínas de los grupos 1, 2 y 4 algunos residuos de aminoácidos son diferentes entre la Ago conservada del grupo 1 y las Ago del nuevo clado. Los dominios y residuos importantes para la unión y funcionalidad de las proteínas Ago se observan en la [Figura IV.26 C y D](#). La presencia de los motivos L1 y L2 adyacentes al dominio PAZ que consisten de dos láminas beta y dos alfa hélices respectivamente, confirman que las proteínas del grupo 1 de *Echinococcus* pertenecen a la familia Ago2. ([Figura IV.26 E](#)).

Las proteínas Ago 2 poseen el dominio DEDH y dos motivos definidos como N1 y N2 en su extremo N-terminal, los cuales ejercen la función de clivaje del ARNm y para el silenciamiento génico. Contrariamente a lo que ocurre en mamíferos, donde todas las proteínas Ago, salvo las Ago2, carecen del motivo DEDH encargado de la actividad de clivaje de los ARNm, todas las proteínas Ago de *Echinococcus* poseen dicho motivo DEDH en el dominio Piwi [260]. En la proteína Ago humana la metionina del motivo N1 está directamente involucrado en la actividad de clivaje del ARNm mientras que el motivo N2 que comprende a dos alfa hélices expuestas hacia la superficie de la proteína, está involucrada en activación del complejo de silenciamiento inducido por ARN (RISC) y en la remoción de la hebra pasajera [260]. En *Echinococcus*, la estructura secundaria del motivo N1 y N2 esta conservada, pero la metionina del motivo N1 en el extremo N-terminal está ausente en todas las proteínas ([Figura IV.26 F y G](#)). Además, las proteínas Ago del grupo 2 de *E. canadensis* G7 y *E. granulosus* G1 no presentan el dominio ArgoN encargado de otorgar especificidad y colaborar en el clivaje del ARNm junto con DEDH del dominio Piwi, sino que en su lugar poseen el dominio de unión a aminoácidos ACT_6 que regulan los dominios enzimáticos asociados por lo cual podría alterar la interacción con el complejo RISC.

Finalmente se identificaron residuos que conforman el motivo que une la región de siembra “Seed” del microARN. En las proteínas Ago del grupo 1 de *Echinococcus*, la secuencia de aminoácidos que constituye dicho motivo es QSKN y se encuentra conservado en relación a la Ago2 humana. En cambio, en las proteínas Ago del nuevo clado de *Echinococcus* (grupos 2-4) este motivo posee una secuencia de aminoácidos diferente, pero conservada entre los miembros del nuevo clado. La secuencia de aminoácidos de este motivo es KDGT, excepto para la proteína de *E. multilocularis* (EmuJ_000911600), la cual posee un ácido aspártico D en lugar de ácido glutámico E ([Figura IV.26 H](#)).



C

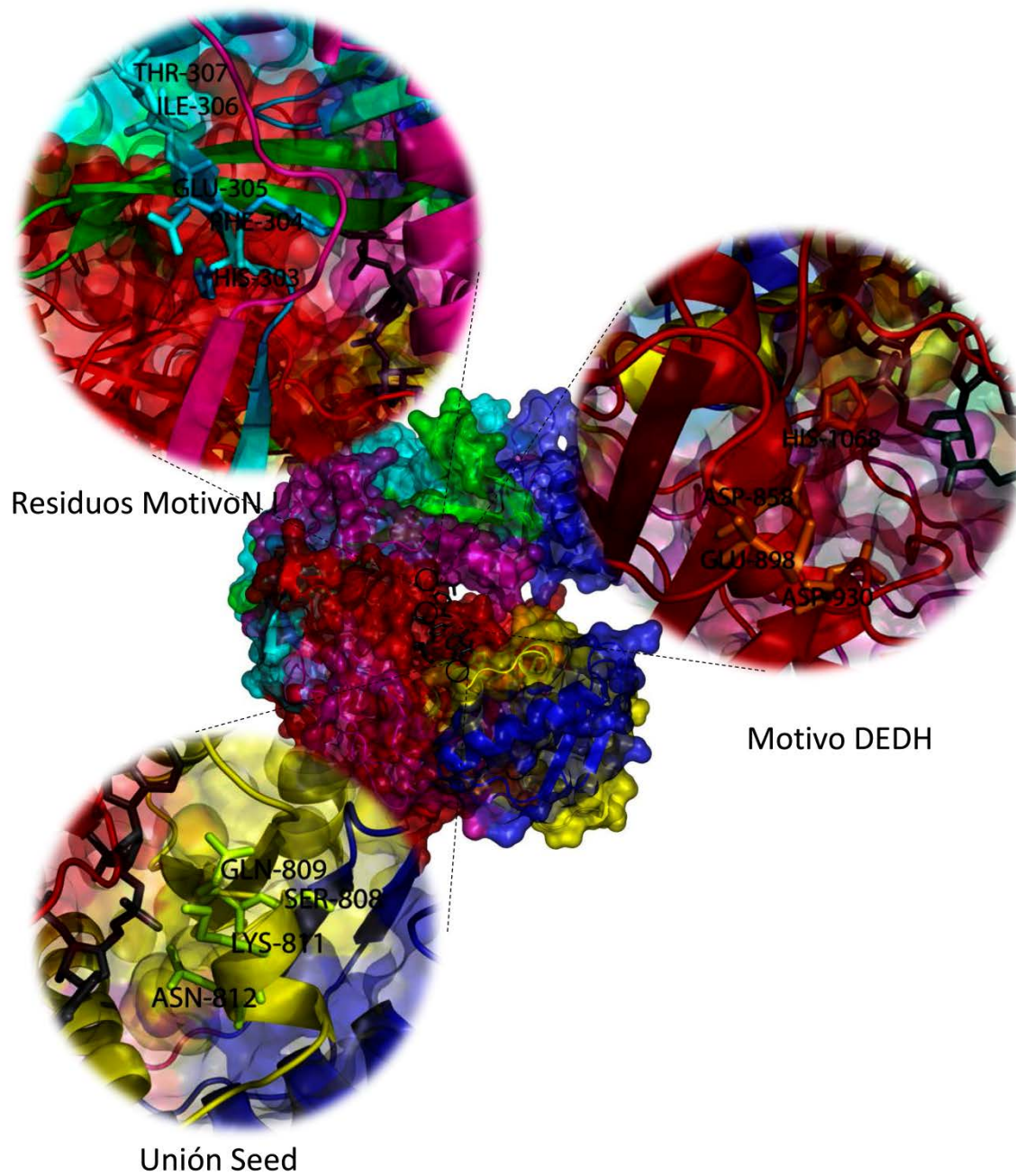


D

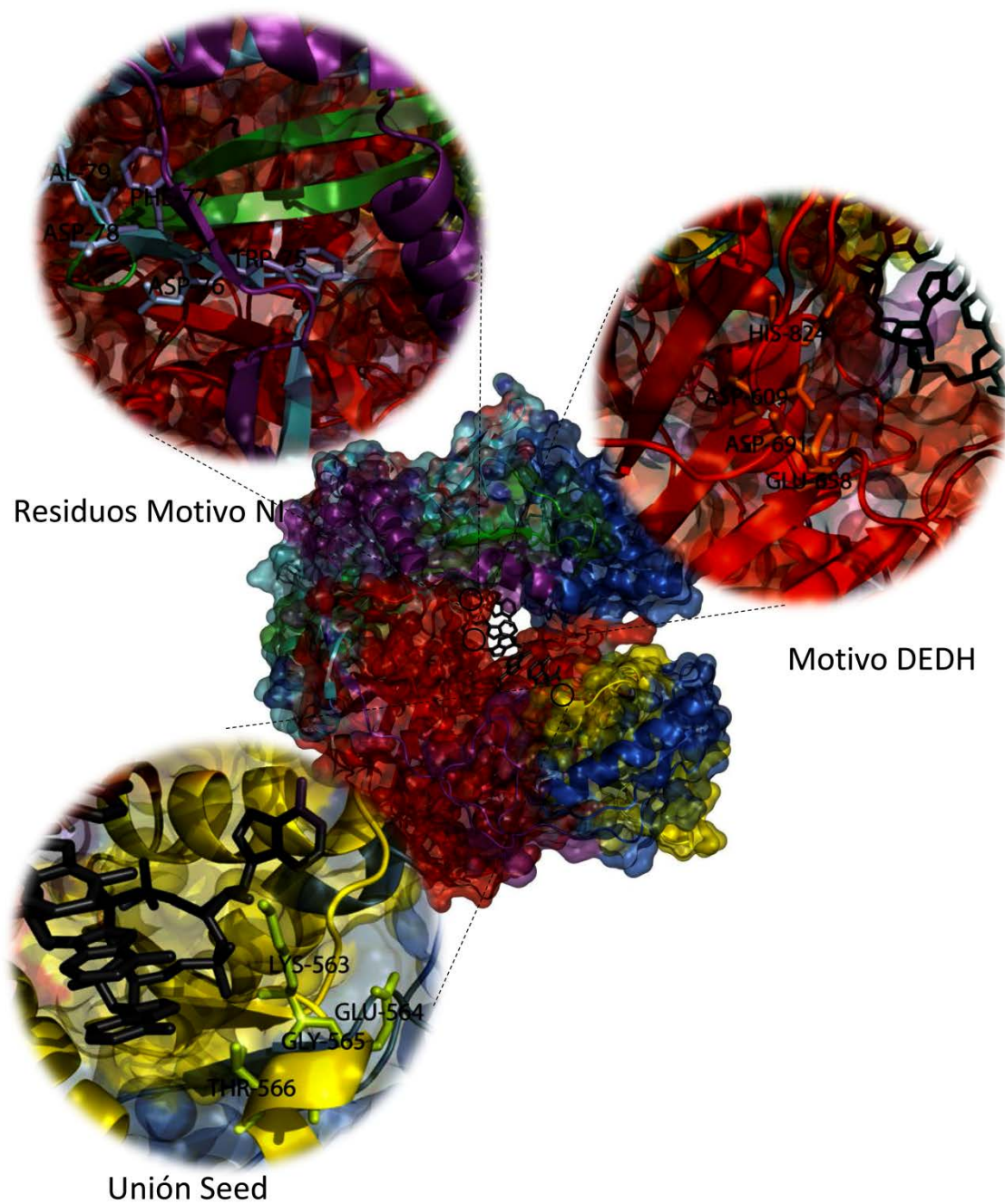
Grupo de proteínas Ago de *Echinococcus*. Residuos relevantes conservados y no conservados

Humano-Ago2	Humano-Ago1	Humano-Ago3	Humano-Ago4	Hiwi	Miwi	Group 1	Grupo 2	Grupo 3	Grupo 4	Dominios
Y529	+	+	+	+	+	+	+	+	+	PAZ
K533	+	+	+	+	+	+	+	+	+	
Q548	+	+	+	V	V	+	K	K	K	
N551	+	+	+	+	+	+	T	T	T	
L563	+	+	+	+	+	+	+	+	+	MID
K566	+	+	+	+	+	+	+	+	+	
K570	+	+	+	+	+	+	+	+	+	
Q633	R	R	R	D	D	S	D	D	D	PIWI
R710	+	+	+	+	+	+	+	+	+	
Q757	+	+	+	+	+	+	+	+	+	
R761	R	+	+	S	S	+	K	K	K	
R792	+	+	+	+	+	+	+	+	+	
C793	+	+	+	+	+	+	+	+	+	
S798	+	+	+	+	+	+	+	+	+	
Y804	+	+	+	+	+	+	+	+	+	
R812	+	+	+	+	+	+	+	+	+	

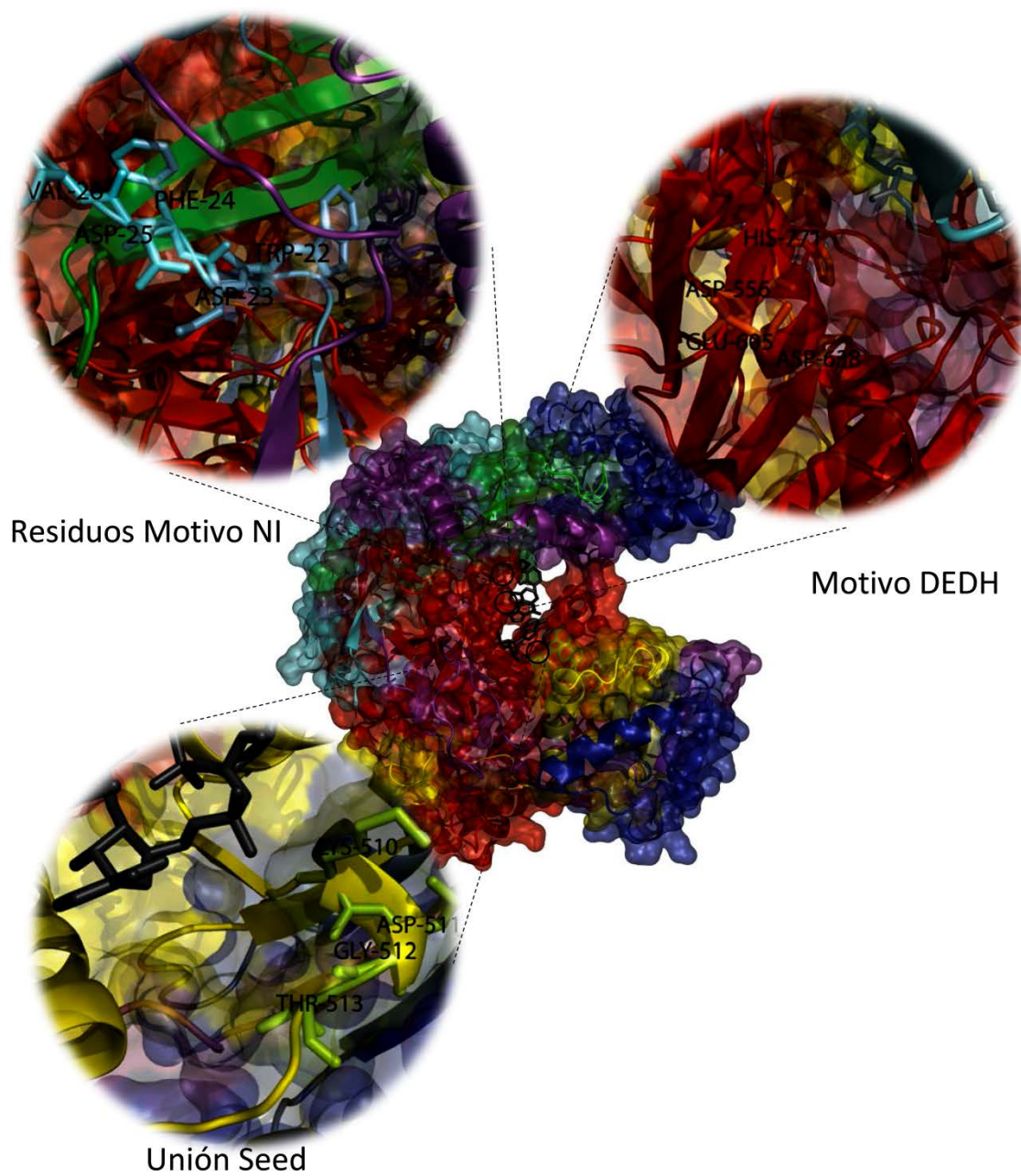
E



F



G



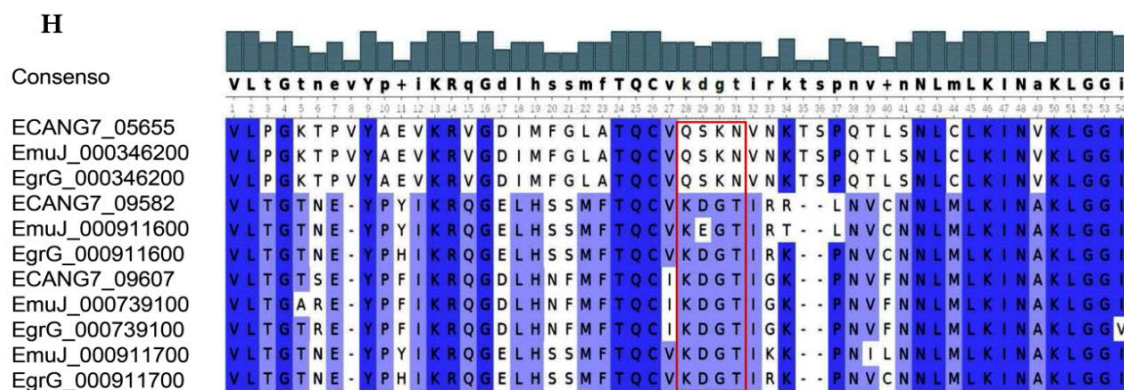


Figura IV.26: Modelado por homología de las proteínas argonautas de *Echinococcus*. A). Representación del plegado RNase-like de la proteína Ago representante del grupo 1 de *Echinococcus*. B). Representación de la superficie molecular de la proteína Ago representante del grupo 1 de *Echinococcus* y el bolsillo de unión con el microARN. C). Esquema de los dominios típicos de las proteínas Ago del grupo 1. D). Residuos de aminoácidos relevantes conservados y no conservados para la funcionalidad de las proteínas Ago de *Echinococcus*. La conservación de los dominios se encuentra indicada con el símbolo + y la sustitución del aminoácido por el código de letra. E). Representación *cartoon* de las proteínas Ago de *Echinococcus* del grupo 1. F). Representación *cartoon* de las proteínas Ago de *Echinococcus* del grupo 2. G). Representación *cartoon* de las proteínas Ago de *Echinococcus* del grupo 4. Los residuos involucrados en la unión del microARN. Las regiones ampliadas contienen los residuos involucrados en la unión del *Seed*, ARNm y la actividad de clivaje. H). Alineamiento de la secuencia de aminoácidos involucrados en unir el *Seed* del microARN en *Echinococcus*.

4.5.4 Conclusiones parciales

El estudio de las proteínas del sistema de metilación del ADN demostró que la familia de genes que están involucrados en este mecanismo cuenta con menos miembros que los que poseen los organismos mamíferos. A diferencia de ellos, *Echinococcus* solo cuenta con un solo tipo de DNMT con el rol de metilar las citosinas del ADN genómico. Asimismo, también posee menos miembros de las familias de proteínas MBDs que unen sitios metilados del ADN. Sin embargo en el contexto de esta tesis se identificó una nueva MBD que es específica de cestodos y posee dominios dedos de zinc de unión a ADN y dichos dominios se encuentran conservados entre las especies de cestodos. Por otra parte, el análisis de CGIs demostró que *Echinococcus* posee una mayor densidad de CGIs que las observadas en mamíferos y que ésta es mayor en las regiones codificantes. Además se observó una distribución diferencial de las CGIs entre los genomas de *Echinococcus* exhibiendo un patrón de distancia de entre 500 y 3500 pb río arriba de la posición de inicio de la traducción, sugiriendo su asociación a regiones promotoras. En cuanto a los análisis basados en el modelado por homología de las proteínas Ago se observó que al menos tres de ellas serían capaces de unir microARNs. La conservación del motivo DEDH en el dominio PIWI sugiere que varias de las proteínas Ago presentes en *Echinococcus* poseen actividad de clivaje. Asimismo el cambio de determinados residuos localizados en los dominios MID y N-terminal y en la región de unión al *Seed* por residuos con propiedades fisicoquímicas similares sugiere que existe conservación de la actividad y funcionalidad de las proteínas Ago del nuevo clado. Las proteínas Ago funcionales sugieren la existencia de un mecanismo de regulación por pequeños ARNs aún desconocido.

4.6 Estudio del uso de codones en *Echinococcus*

El estudio de los patrones del uso de codones, que son la unidad mínima que determina la expresión génica, contribuye notablemente a la comprensión de las características básicas de la organización molecular de los genomas de un organismo. Cualquier similitud en los patrones del uso de codones revela cierto grado de relación biológica, adaptación ambiental o evolución entre los organismos. Generalmente el contenido de GC total de un genoma muestra una gran similitud con el contenido de GC de su región codificante [261]. En muchos organismos procariotas y eucariotas el uso de codones, así como el uso de aminoácidos están altamente influenciados por el sesgo composicional de GC [262,263]. Debido al carácter degenerado del código genético, la tercera base de los codones es menos restrictiva para la codificación de un aminoácido que las primeras dos bases. La tercera posición en el codón se conoce como la posición de balanceo y la mayoría de los aminoácidos permiten la sustitución de codones sinónimos dado por un cambio de base en la tercera posición del codón [264]. En esta sección de la tesis se aborda el estudio del patrón del uso de codones de los genes de *Echinococcus*.

4.6.1 Contenido de nucleótidos en los codones en genes de *Echinococcus*

En primer lugar se realizó el análisis del contenido nucleotídico de los CDSs de *Echinococcus*, los cuales mostraron un mayor porcentaje de GC con respecto al porcentaje total observado en sus genomas y similar entre las tres especies de *Echinococcus*. Los porcentajes de GC variaron entre el 33,3% y el 60,4% con un promedio de 49,8% para *E. canadensis* G7, entre el 35,7% y el 59,8% con un promedio de 49,6% para *E. multilocularis* y entre el 34,2% y el 63,4% con un promedio de 49,8% para *E. granulosus* G1 ([Figura IV.27](#)).

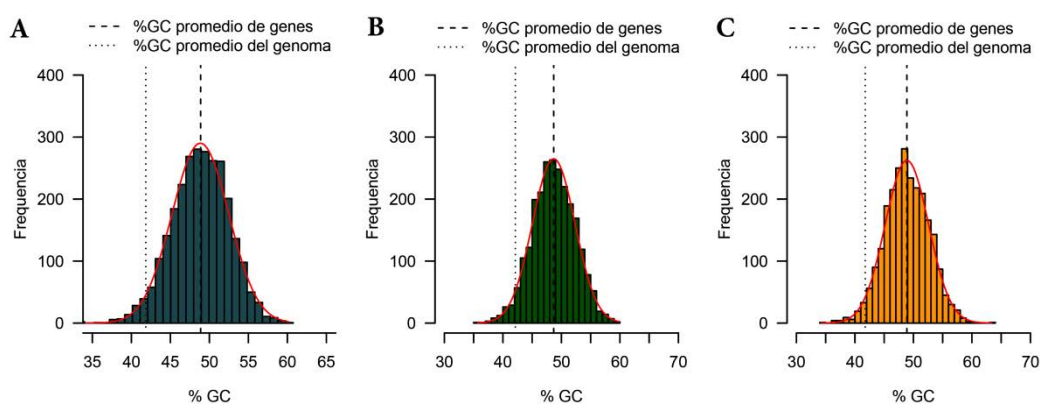


Figura IV.27: Distribución de los genes de *Echinococcus* según el contenido de GC. El histograma representa la frecuencia de genes. La línea punteada fina representa la media del porcentaje de GC del genoma. La línea punteada gruesa representa la media del porcentaje de GC de los genes. A). *E. canadensis* G7. B). *E. multilocularis*. C). *E. granulosus* G1

Para el estudio del uso de codones se analizaron un total de 1849107, 1425315 y 1436461 codones de *E. canadensis* G7, *E. multilocularis* y *E. granulosus* G1 respectivamente y se determinó la composición de nucleótidos de cada posición del triplete que conforma a los codones. Se observó que tanto para la posición 1, 2 y 3 no existen diferencias en la composición de nucleótidos con las respectivas posiciones de las tres especies de *Echinococcus* (Figura IV.28), y que además los porcentajes de GC en cada posición son diferentes a los observados en los genomas completos donde se ve un menor contenido de GC; lo cual es coherente con el resultado de menor contenido de GC en el genoma completo y la media del porcentaje de GC de los genes. Asimismo; la composición de nucleótidos en cada una de las tres posiciones del triplete es diferente entre sí. Para la tercera posición del triplete, la posición menos restringente, hay preferencia por T/C. En la primera posición el nucleótido de preferencia parece ser una G y en la segunda posición, la posición más restringente, el uso de nucleótidos preferente es A/T. Asimismo se observó que la composición de G y C en la primera y segunda posición es claramente superior a la que presenta la secuencia genómica no codificante.

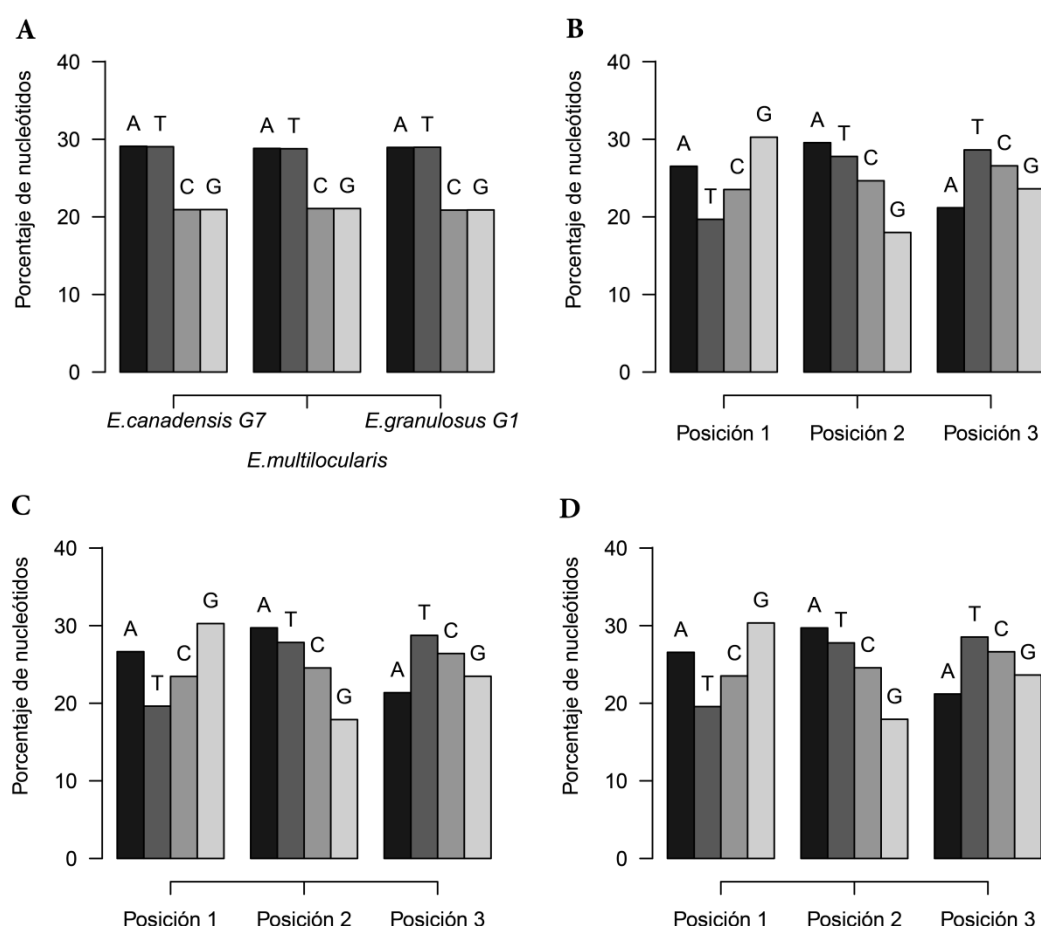


Figura IV.28: Media del porcentaje de nucleótidos: A). Porcentaje de nucleótidos totales en los genomas de *Echinococcus*. B). Porcentaje de nucleótidos por posición en el codón en *E. canadiensis* G7. C). Porcentaje de nucleótidos por posición en el codón en *E. multilocularis*. D). Porcentaje de nucleótidos por posición en el codón en *E. granulosus* G1.

En las tres especies el mayor porcentaje de GC se observa en la primera posición del codón (GC1) con un 53,7%, seguido de la tercera posición (GC3) con un 48,6% y por último la segunda posición del codón (GC2) con un 42,5%, aproximadamente ([tabla IV.14](#)).

Tabla IV.14: Porcentajes de GC por posición en los codones

Porcentaje (%)	<i>E. canadensis</i> G7	<i>E. multilocularis</i>	<i>E. granulosus</i> G1
GC1	53,79	53,72	53,86
GC2	42,65	42,44	42,51
GC3	48,64	48,31	48,71
A	29,10	28,82	28,95
T	29,04	28,78	28,97
C	20,93	21,07	20,87
G	20,94	21,07	20,89
GC genes	49,80	49,60	49,80
GC genómico	41,87	42,14	41,76

%GC en el primer codón (GC1); %GC en el segundo codón (GC2); %GC en el tercer codón (GC3)

Por otra parte; al comparar el contenido de GC3 en genes altamente expresados con genes que poseen baja expresión (ver sección 4.6.4.6) se observó que no presentan un sesgo composicional de nucleótidos; sin embargo los genes con baja expresión exhiben un mayor rango en su porcentaje de GC y una mediana ligeramente inferior que los genes con alto nivel de expresión ([Figura IV.29](#)).

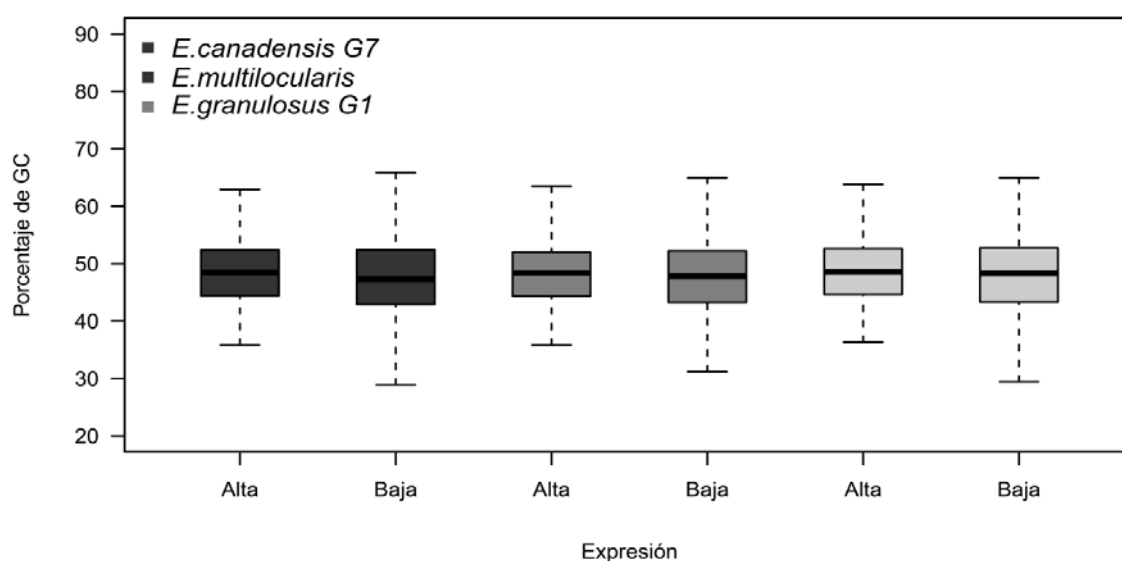


Figura IV.29: Gráficos de cajas del contenido de GC en genes de alta y baja expresión en las tres especies de *Echinococcus*.

4.6.2 Uso de codones sinónimos relativos en las especies de *Echinococcus*

Para el estudio del uso de codones se evaluó el patrón de codones sinónimos relativos (RSCU) en las tres especies de *Echinococcus*. En ausencia de un sesgo en el uso de codones, los valores RSCU deben ser igual a 1. Un codón que se utiliza con menor frecuencia que la esperada tendrá un valor de RSCU inferior a 1 y por el contrario un codón que se utiliza con mayor frecuencia que la esperada posee valores mayores que 1. En la [tabla IV.15](#) se muestra el uso de los codones sinónimos relativos para los genes de *Echinococcus*. Se puede observar que el uso de codones con un alto sesgo en las tres especies de *Echinococcus* fueron: GCU (Ala), CGU, CGC, CGA (Arg), AUU (Ile), GGU (Gly), CUU, CUC y CUG (Leu), UCC (Ser), ACC (Thr), UAC (Tyr), GUG (Val). Sin embargo, la mayoría de los codones posee valores de RSCU cercanos a 1 indicando que no poseen un fuerte sesgo de uso de codones.

Tabla IV.15: Frecuencia relativa del uso de codones sinónimos (RSCU) en las tres especies de *Echinococcus*.

Especies		<i>E. canadensis</i> G7		<i>E. multilocularis</i>		<i>E. granulosus</i>	
Aas	Codón	RSCU	Frecuencia	RSCU	Frecuencia	RSCU	Frecuencia
Ala	GCU*	1,2	41506	1,21	32115	1,2	32066
Ala	GCC	1,15	39819	1,15	30351	1,16	30958
Ala	GCA	0,92	31704	0,94	24929	0,92	24670
Ala	GCG	0,73	25100	0,7	18508	0,72	19352
Arg	CGU*	1,33	25818	1,32	19690	1,32	19905
Arg	CGC*	1,21	23452	1,2	17964	1,21	18272
Arg	CGA*	1,19	23222	1,2	17888	1,21	18254
Arg	CGG	0,75	14576	0,74	11080	0,75	11302
Arg	AGA	0,8	15596	0,82	12197	0,8	12127
Arg	AGG	0,72	13932	0,72	10708	0,71	10741
Asn	AAU	1,09	40335	1,1	31593	1,09	31411
Asn	AAC	0,91	33543	0,9	26044	0,91	26308
Asp	GAU	1,12	52181	1,12	40450	1,12	40687
Asp	GAC	0,88	41289	0,88	31646	0,88	32264
Cys	UGU	0,99	20195	0,99	15388	0,99	15463
Cys	UGC	1,01	20435	1,01	15679	1,01	15721
Gln	CAA	0,96	35184	0,97	27472	0,96	27501
Gln	CAG	1,04	38333	1,03	29414	1,04	29729
Glu	GAA	0,94	54686	0,95	42447	0,94	42597
Glu	GAG	1,06	61673	1,05	47028	1,06	47868
Gly	GGU*	1,36	35442	1,36	27193	1,36	27543
Gly	GGC	1,11	28973	1,11	22136	1,11	22564
Gly	GGA	0,97	25290	0,98	19677	0,98	19752
Gly	GGG	0,55	14359	0,55	11083	0,55	11112
His	CAU	0,96	21941	0,97	16949	0,96	16980

His	CAC	1,04	23644	1,03	18063	1,04	18293
Ile	AUU*	1,32	39932	1,34	31326	1,32	31008
Ile	AUC	1,11	33738	1,09	25541	1,12	26192
Ile	AUA	0,57	17139	0,57	13264	0,56	13096
Leu	UUA	0,5	15307	0,49	11716	0,49	11663
Leu	UUG	1,08	33240	1,09	25935	1,07	25733
Leu	CUU*	1,27	39219	1,3	30944	1,28	30769
Leu	CUC*	1,37	42228	1,34	31944	1,37	32738
Leu	CUA	0,62	19019	0,62	14866	0,62	14774
Leu	CUG*	1,18	36322	1,16	27666	1,17	28038
Lys	AAA	0,96	42364	0,97	33076	0,96	32999
Lys	AAG	1,04	45723	1,03	35382	1,04	35783
Met	AUG	1	37610	1	29240	1	29311
Phe	UUU	1	37052	1,02	28940	1	28495
Phe	UUC	1	36872	0,98	27987	1	28671
Pro	CCU	1,1	28823	1,11	22317	1,1	22330
Pro	CCC	1,06	27821	1,05	21139	1,07	21687
Pro	CCA	1,11	28968	1,13	22757	1,11	22664
Pro	CCG	0,73	19103	0,71	14223	0,72	14754
Ser	UCU	1,14	32114	1,15	24827	1,13	24798
Ser	UCC*	1,23	34705	1,22	26503	1,23	27021
Ser	UCA	1,02	28882	1,05	22710	1,03	22444
Ser	UCG	0,87	24385	0,85	18349	0,87	18992
Ser	AGU	0,94	26522	0,93	20235	0,93	20429
Ser	AGC	0,8	22493	0,81	17465	0,81	17635
TER	UAA*	1,39	1277	1,41	1135	1,4	1141
TER	UAG	0	0	0	0	0	0
TER	UGA*	1,61	1477	1,59	1283	1,6	1301
Thr	ACU	1,13	30710	1,14	23855	1,13	23842
Thr	ACC*	1,17	31623	1,17	24540	1,18	24818
Thr	ACA	0,96	26004	0,97	20279	0,96	20305
Thr	ACG	0,74	19917	0,72	14996	0,72	15237
Trp	UGG	1	19897	1	15440	1	15543
Tyr	UAU	0,82	20009	0,82	15468	0,82	15523
Tyr	UAC*	1,18	28605	1,18	22043	1,18	22223
Val	GUU	1,16	34046	1,17	26372	1,15	26217
Val	GUC	1,02	30050	1,01	22875	1,03	23350
Val	GUA	0,56	16512	0,56	12702	0,56	12733
Val	GUG*	1,26	37171	1,25	28283	1,26	28764

Los codones con valores de RSCU > 1 indican que son usados más frecuentemente en *Echinococcus*. Los codones indicados con (*) representan a los codones sinónimos con cuyos valores de RSCU son mayores a 1,15.

Por otra parte también se observó un uso preferencial de los codones de terminación de la traducción dado por los codones UAA y UGA mientras que el codón UAG está totalmente ausente ([Figura IV.30](#)).

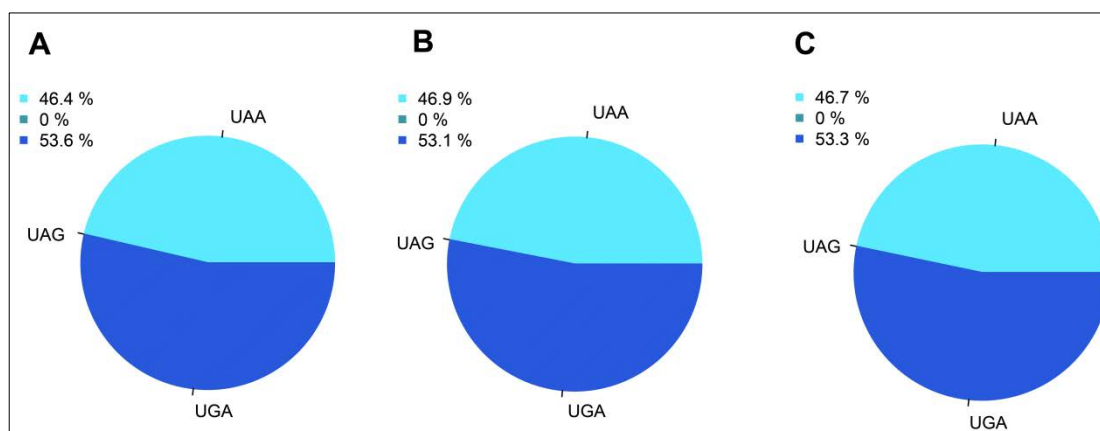


Figura IV.30: Porcentajes del uso de codones de terminación de la traducción en las tres especies de *Echinococcus*. A). *E. canadensis* G7. B). *E. multilocularis*. C). *E. granulosus* G1

Al comparar el RSCU entre las tres especies de *Echinococcus* se observó que no existen diferencias evidentes en el uso de codones y exhiben en forma general el mismo patrón RSCU. Adicionalmente se determinó la correlación de RSCU entre las tres especies de *Echinococcus* y se observó un alto índice de correlación positiva de 0,998 (p-valor < 0,01) en el uso de codones sinónimos para los tres pares de las especies analizadas ([Figura IV.31](#)).

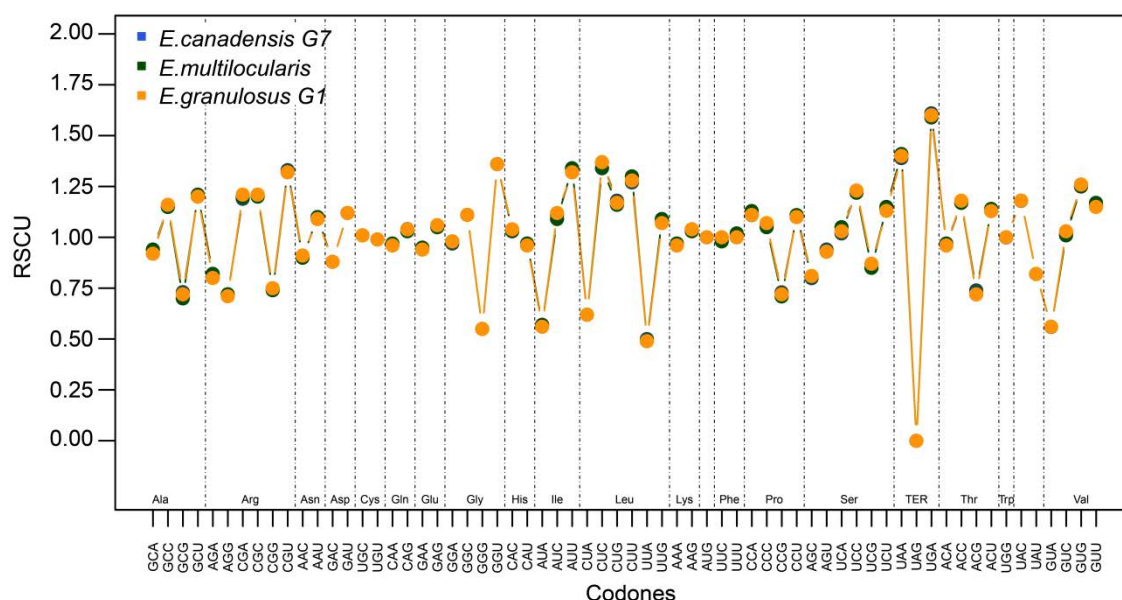


Figura IV.31: Valores de RSCU por codón y agrupados según el aminoácido para el cual codifican en las tres especies de *Echinococcus*.

4.6.3 Sesgo y variabilidad en el uso de codones en *Echinococcus*

Para determinar si existe o no un sesgo en el uso de codones en *Echinococcus* se analizó el número efectivos de codones (ENc) de los genes de las tres especies de *Echinococcus*. En un rango de valores de 20 a 61, los valores más bajos indican un mayor sesgo en el uso de codones. Los valores de ENc inferiores a 36 indican un fuerte sesgo en el uso de codones [224].

Se observó que los valores de ENc de los genes de *E. canadensis* G7 varió entre 30,5 y 61 con un promedio de ENc de 57. Los valores de ENc de los genes de *E. multilocularis* varió entre 31,6 y 61 con un promedio de ENc de 57,1. Y por último los valores de ENc de los genes de *E. granulosus* G1 varió entre 34,2 y 61 con un promedio de ENc de 57. Además, se observó que un grupo menor de genes presentan un alto sesgo en el uso de codones, los cuales están representados por las líneas que se alejan de la distribución normal, y cuyos valores de ENc son menores a 50 ([Figura IV.32](#)).

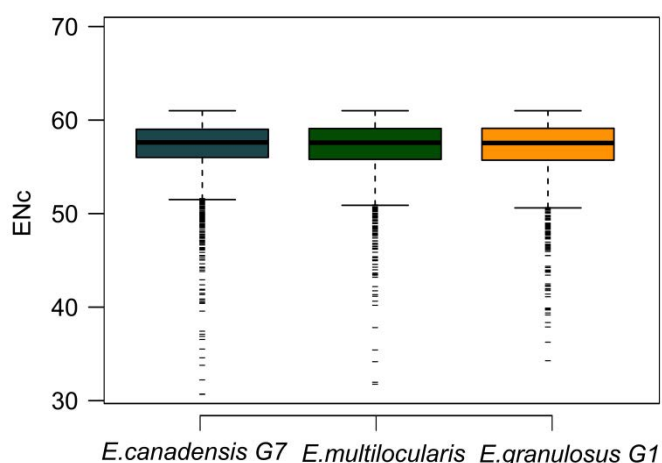


Figura IV.32: Gráficos de cajas de la distribución de genes según su valor ENc. Las líneas por debajo de la pestaña inferior son los genes que caen por debajo de los valores normales observados para los genes de *Echinococcus*.

Para estimar la frecuencia de los genes que poseen un sesgo en su uso de codones se relacionó los valores de ENc observados con los esperados si el uso de codones fuese aleatorio. La relación entre la frecuencia de ENc esperada y observada se muestra en la [figura IV.33](#). Se observó un pico de frecuencia máxima en 0,04 y una distribución que va desde valores negativos hacia valores positivos (-0,1 a 0,25) lo cual indica que la mayoría de los genes poseen un ENc ligeramente diferente al esperado. Sin embargo existen grupos de genes que poseen un mayor sesgo en el uso de codones (valores más positivos) y otro grupo de genes cuyo ENc es igual o incluso mayor que el esperado (valores más negativos).

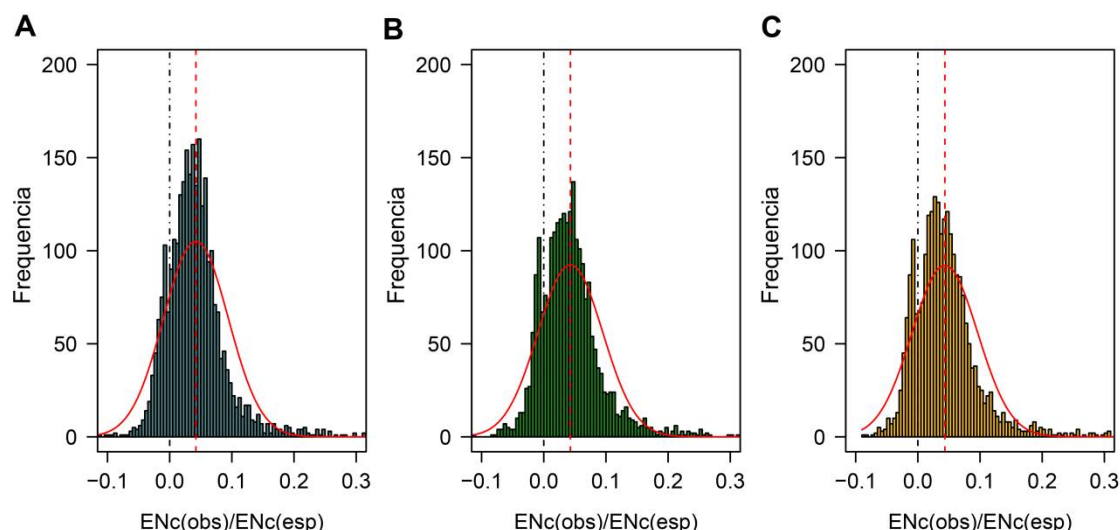


Figura IV.33: Distribución de los genes de acuerdo a la relación entre el ENc observado sobre el ENc esperado de A). *E. canadensis* G7; B). *E. multilocularis*; C). *E. granulosus* G1. Las abscisas están representadas por la relación $(ENC_{esp} - ENC_{obs}) / ENC_{esp}$.

4.6.4 Estudio de los fenómenos responsables en el sesgo del uso de codones

Habiendo determinado que existe una proporción de genes que poseen un sesgo en su uso de codones, se estudió cuales son los fenómenos responsables que dan lugar a la variación del uso de codones en *Echinococcus*. Para ello se realizaron varios análisis que se describen a continuación.

4.6.4.1 Análisis de Neutralidad

Se caracterizó la correlación entre las tres posiciones de los tripletes de los codones sinónimos de acuerdo a su composición nucleotídica. Se graficó la relación entre el contenido de GC en las primeras dos posiciones de los codones (P12) y el contenido de GC de la tercera posición de los codones (P3); ([Figura IV.34](#))

Dado que P3 es la posición menos restrictiva y por lo tanto la más dependiente del sesgo mutacional, si un valor P12 fuese tan neutro como P3 frente a la selección, todos los puntos deberían estar distribuidos a lo largo de la línea diagonal de pendiente igual a 1, indicando que el sesgo en el uso de codones solo está determinado por un sesgo mutacional. Así pues, si la pendiente es inferior a 1 esto indica que el grado de neutralidad es inferior a la de P3. Por lo tanto, la pendiente proporciona una medida de la neutralidad relativa de P12 con respecto a la de P3.

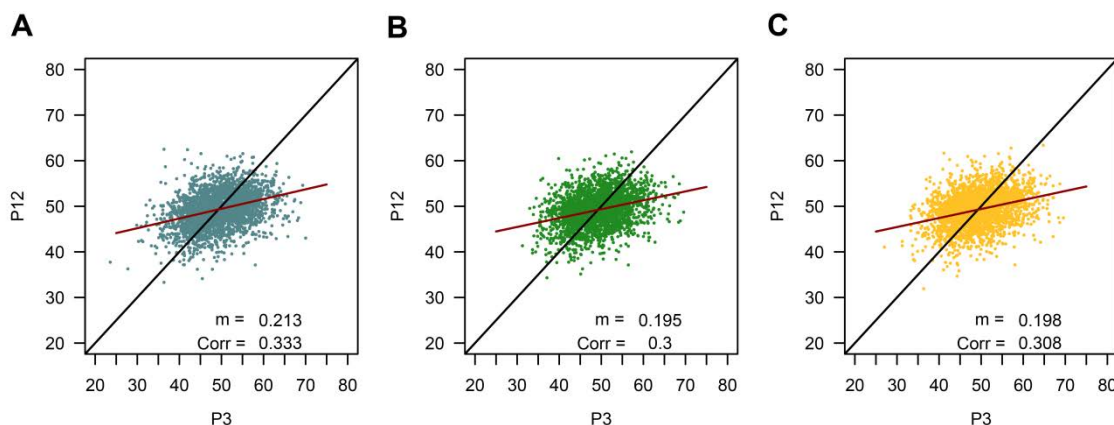


Figura IV.34: Gráficos de neutralidad según la relación entre la composición de GC en la primera y segunda posición de los codones sinónimos (P12) con la composición de GC en la tercera posición de los codones sinónimos (P3). A). *E. canadensis* G7, B). *E. multilocularis*. C). *E. granulosus* G1. P12 representa la media de GC en la primera y segunda posición de los codones sinónimos (GC1s y GC2s), mientras que P3 representa el contenido de GC en la tercera posición de los codones sinónimos (GC3s). La línea roja muestra la regresión lineal de P12 frente a P3. La intersección de la curva de regresión y la diagonal, y representa el punto en el que P12 es igual a P3.

El coeficiente de regresión (pendiente de la recta) proporciona una estimación del grado de neutralidad de P12 relativo al de P3 [228]. En otras palabras si la distribución de los genes da lugar a una recta de pendiente igual a 1, existe una presión de mutación neutral. Por el contrario, una regresión lineal con pendiente nula indicaría que el genoma de un organismo habría sido moldeado bajo una presión mutacional no neutral [228]. Estos resultados mostraron una alta correlación positiva entre P12 y P3 para las tres especies de *Echinococcus* con p-valores significativos cercanos a 0. Asimismo, también se observó que la pendiente de regresión de las curvas dadas por la distribución de los genes de *Echinococcus* es menor a uno, indicando que el grado de neutralidad relativa que interviene en la formación del uso de codones es del 21,3%, 19,5% y 21,8% para *E. canadensis* G7, *E. multilocularis* y *E. granulosus* G1 respectivamente mientras que el 78,7%, 81,5% y 78,2% restantes se debe a la contribución de una presión de selección.

4.6.4.2 Evaluación de la regla de paridad 2 de Chargaff (PR2)

La ley de Chargaff se basa en la relación cuantitativa de los nucleótidos que forman la doble hélice del ADN y establece que la cantidad de adenina (A) es igual a la cantidad de timina (T) y la cantidad de guanina (G) es igual a la cantidad de citosina (C), es decir, el número total de bases purinas es igual al número total de bases pirimidinas ($A+G = C+T$). En el contexto del uso de codones, si el uso de codones de los genes de un organismo está determinado solo por un sesgo mutacional, las proporciones $G + C$ de los genes deberían ser usadas proporcionalmente a las de $A + T$ en las familias de codones sinónimos. Por el contrario si existiera una fuerza de selección las proporciones de nucleótidos de la tercera posición del triplete no necesariamente serán proporcionales entre sí [229].

En este análisis se observó que la mayoría de los genes se distribuyeron en el cuadrante inferior izquierdo ([Figura IV.35](#)) lo que implica que C y T (pirimidinas) se utilizaron con mayor frecuencia que G y A (purinas) en los codones de las tres especies de *Echinococcus*. Este resultado generó un vector que apunta hacia el cuadrante inferior izquierdo en la posición (0,468; 0,438), proporcionando evidencia adicional de que otros factores, además del sesgo mutacional, contribuyen a moldear el sesgo en el uso de codones.

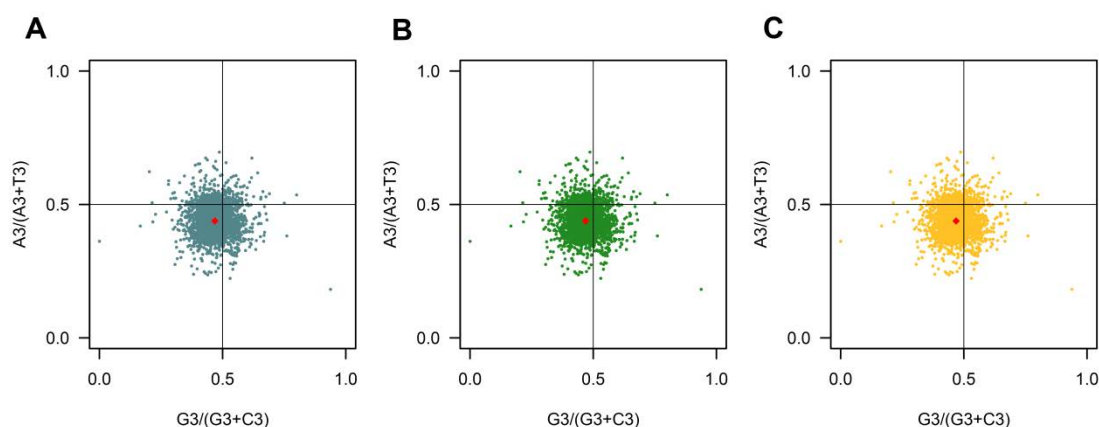


Figura IV.35: Distribución gráfica PR2 de A). *E. canadensis* G7, B). *E. multilocularis*. C). *E. granulosus* G1. La distribución está dada por $[A3/ (A3 + T3)$ contra $G3/ (G3 + C3)]$. El círculo abierto rojo indica la media de la distribución.

4.6.4.3 Relación entre ENc y GC3s

Para mostrar la relación funcional entre los valores ENc y GC3s se generó la curva estándar que se obtendría si el sesgo en el uso de codones estuviese influenciado solo por un sesgo mutacional en lugar de una presión de selección. En los genes cuyo uso de codones está limitado sólo por un sesgo de mutación G + C, los valores de ENc se encuentran sobre o muy cercanos a los de la curva estándar. Sin embargo, la presencia de otros fenómenos, tales como los efectos de la selección provocan que los valores se desvíen considerablemente por debajo de la curva esperada de GC3s.

En la representación gráfica entre ENc y GC3s ([Figura IV.36](#)) la mayoría de los genes de las especies de *Echinococcus* se ubican bajo la curva de valores esperados del ENc, mientras que sólo un pequeño número se distribuyó a lo largo o por encima de ella, demostrando que la mayoría de los genes de *Echinococcus* utilizan codones regulados por otros factores que no dependen estrictamente del contenido de GC. Estos resultados apoyan las observaciones anteriores que sugieren que no sólo la mutación sino también otros factores, como la selección traduccional u otra fuerza de selección, pueden estar involucrados en determinar la variación del uso de codones.

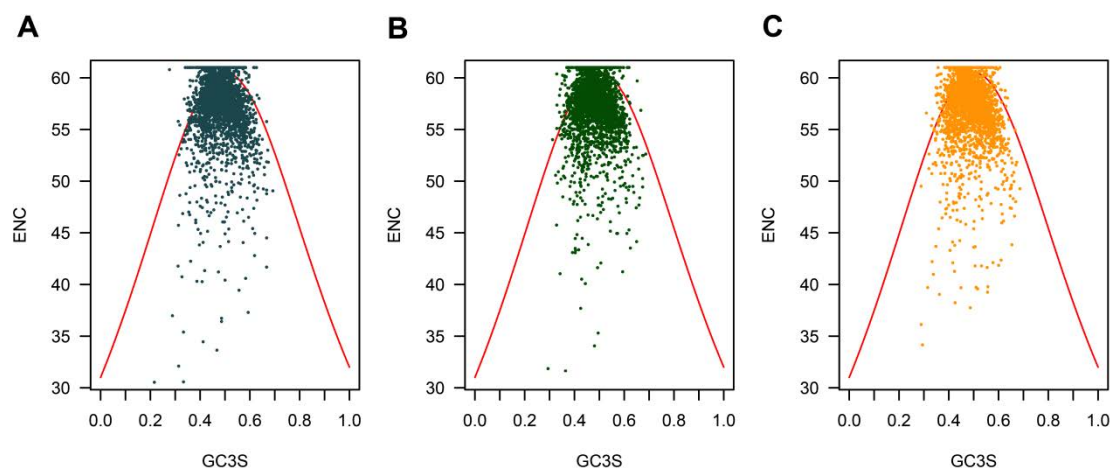


Figura IV.36: Distribución del número efectivo de codones (ENC) y GC3s de genes de A). *E. canadensis* G7, B). *E. multilocularis*. C). *E. granulosus* G1. La línea sólida (roja) indica el valor ENC esperado si el sesgo en el uso de codones se debe únicamente a GC3s.

4.6.4.4 Análisis de los principales factores responsables del sesgo en el uso de codones

Para explorar la variación de RSCU en los genes de *Echinococcus* y determinar cuáles son los principales factores responsables de dicha variación se aplicaron análisis de correspondencia y multivariados que permitan explicar como influyen dichos factores sobre los patrones de uso de codones en los genes de *Echinococcus*.

Se seleccionaron como ejes significativos los componentes principales (CPs) con variaciones mayores que la varianza máxima de las variables originales [223]. Los análisis de componentes principales (ACP) basados en la adaptabilidad relativa mostraron que el primer componente principal (CP1) explicó el 9,64% de la variación total de RSCU mientras que los otros tres CPs representaron el 5,22%, 3,68% y el 3,39% de los datos en *E. canadensis* G7. Para *E. multilocularis* CP1 explicó el 8,9% de la variación total de RSCU mientras que CP2, CP3 y CP4 representaron el 5,34%, 3,64% y el 3,43% de la variabilidad total respectivamente. Por otra parte para *E. granulosus* G1, CP1, CP2, CP3 y CP4 explicaron el 9,27%; 5,23%; 3,66% y 3,47% de la variación total de RSCU, respectivamente ([Figura IV.37](#)). El 64% del total de los datos que conforman la variabilidad de RSCU se distribuyó en los primeros 20 CPs.

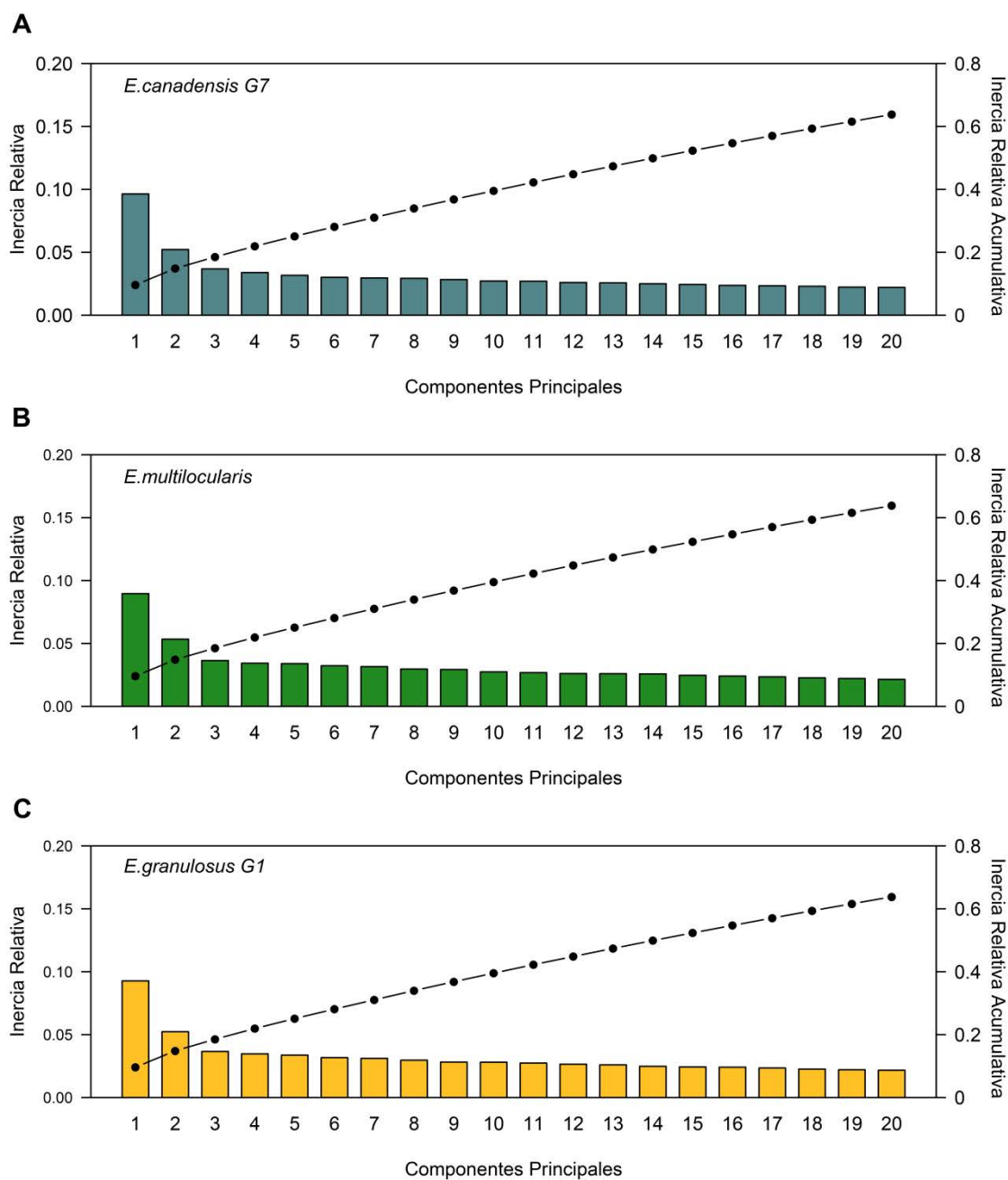


Figura IV.37: La inercia relativa (barras) y acumulativa (líneas con puntos) del total de 20 factores del análisis de componentes principales (PCA) utilizando la distribución de adaptabilidad relativa. A). *E. canadensis* G7, B). *E. multilocularis*. C). *E. granulosus* G1.

Por otra parte, mediante los análisis de correspondencia del uso de codones sinónimos se distinguió la variabilidad de los mismos dispuestos a lo largo de los ejes de los CPs. La posición de los codones en los ejes principales posicionan más cerca a los que presentan características más similares y se observó que su distribución en el espacio de los CPs 1 y 2 está agrupada en principalmente por su contenido nucleotídico en la última base de los codones ([Figura IV.38](#)).

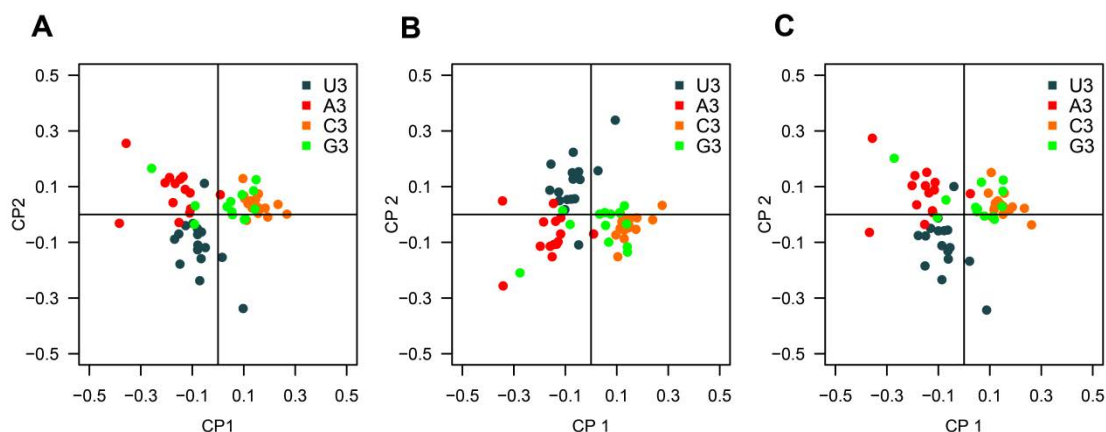


Figura IV.38: Distribución de los codones a lo largo de los ejes de los componentes principales en A). *E. canadensis* G7, B). *E. multilocularis*. C). *E. granulosus* G1. Los codones están representados por puntos en diferentes colores según la base en la última posición del codón.

Para estimar cuales son los factores que más intervienen en moldear el uso de codones de los genes de las especies de *Echinococcus* y cuáles de ellos están representados en distintas proporciones en los CPs se calcularon las correlaciones estadísticas de Spearman y sus p-valores entre los primeros cuatro CPs y los parámetros moleculares que pueden ser responsables del sesgo en el uso de codones tales como GC, GC3s, A3s, T3s, C3s, G3s, tamaño del gen, aromaticidad, hidrofobicidad y el nivel de expresión estimado por el índice CAI. ([Tabla IV.16](#))

El análisis de correlación multivariable permitió obtener una mejor comprensión de la relación entre el sesgo relativo del uso de codones y la composición de nucleótidos de los genes. Basado en la significancia del coeficiente de correlación de Spearman entre los CPs y el ENc y entre el ENc y los valores de los distintos parámetros medidos en cada grupo de genes de *Echinococcus* (teniendo en cuenta un $p < 0,05$) se seleccionaron los CPs y los factores que introducen mayor variabilidad al RSCU y se graficaron las posiciones de los genes en los CPs de manera que permita visualizar dicha variabilidad en el uso de codones.

Tabla IV.16a: Correlaciones entre los factores medidos de los genes de *E. canadensis* G7 y los 4 primeros componentes principales

	CP1	CP2	CP3	CP4	T3s	C3s	A3s	G3s	CAI	CBI	Fop	ENc	GC3s	GC	L_sym	L_aa	Hf	Aromo
CP1																		
CP2	0,06																	
CP3	0,06	-0,04																
CP4	0,00	0,00	0,02															
T3s	-0,57	-0,70	-0,01	0,03														
C3s	0,83	0,26	0,00	0,12	-0,65													
A3s	-0,78	0,27	-0,28	0,05	0,19	-0,56												
G3s	0,33	0,18	0,40	-0,27	-0,34	0,09	-0,33											
CAI	0,77	-0,23	-0,05	0,04	-0,16	0,63	-0,72	0,19										
CBI	0,94	0,16	0,09	0,00	-0,66	0,85	-0,76	0,41	0,71									
Fop	0,94	0,15	0,09	0,00	-0,65	0,85	-0,75	0,41	0,72	0,99								
ENc	-0,24	0,14	0,07	-0,06	0,01	-0,19	0,27	0,05	-0,38	-0,20	-0,21							
GC3s	0,89	0,29	0,21	-0,06	-0,76	0,82	-0,71	0,54	0,60	0,95	0,94	-0,14						
GC	0,77	0,15	0,11	-0,08	-0,61	0,58	-0,64	0,32	0,51	0,78	0,77	-0,14	0,77					
L_sym	0,25	-0,03	-0,08	-0,04	-0,12	0,11	-0,20	0,08	0,18	0,22	0,22	0,04	0,19	0,35				
L_aa	0,25	-0,03	-0,08	-0,04	-0,13	0,11	-0,20	0,08	0,18	0,22	0,22	0,04	0,20	0,35				
Hf	0,11	-0,08	-0,03	0,05	0,01	0,18	-0,28	-0,22	0,04	0,11	0,10	-0,04	0,07	-0,07	-0,02	-0,02		
Arom	-0,04	0,02	-0,05	0,03	0,08	0,16	0,00	-0,18	-0,01	-0,02	-0,03	0,03	-0,03	-0,31	-0,22	-0,21	0,44	1

Los índices de correlación de Spearman en color rojo: no significativos ($p\text{-val} > 0,05$); en color azul significativos ($p < 0,05$); en color negro significativos ($p < 0,01$)

Tabla IV.16b: Correlaciones entre los factores medidos de los genes de *E. multilocularis* y los 4 primeros componentes principales

	CP1	CP2	CP3	CP4	T3s	C3s	A3s	G3s	CAI	CBI	Fop	ENc	GC3s	GC	L_sym	L_aa	Hf	Aromo
CP1																		
CP2	-0,04																	
CP3	0,05	0,03																
CP4	-0,02	0	0,02															
T3s	-0,52	0,71	0	-0,04														
C3s	0,8	-0,27	-0,09	-0,13	-0,62													
A3s	-0,78	-0,26	-0,27	0,13	0,14	-0,54												
G3s	0,32	-0,21	0,5	0,05	-0,33	0,06	-0,32											
CAI	0,73	0,26	-0,1	-0,1	-0,06	0,57	-0,7	0,13										
CBI	0,93	-0,16	0,03	-0,06	-0,62	0,85	-0,75	0,34	0,68									
Fop	0,92	-0,16	0,04	-0,06	-0,59	0,86	-0,73	0,36	0,69	0,99								
ENc	-0,21	-0,17	0,11	0,07	-0,05	-0,16	0,26	0,07	-0,38	-0,19	-0,19							
GC3s	0,87	-0,32	0,2	-0,06	-0,73	0,81	-0,7	0,53	0,53	0,93	0,92	-0,09						
GC	0,76	-0,14	0,12	0,01	-0,58	0,55	-0,63	0,3	0,49	0,76	0,72	-0,11	0,75					
L_sym	0,25	0,03	-0,04	0,08	-0,12	0,1	-0,19	0,05	0,18	0,21	0,19	0,04	0,18	0,35				
L_aa	0,25	0,03	-0,04	0,08	-0,12	0,1	-0,19	0,05	0,18	0,22	0,2	0,04	0,18	0,35				
Hf	0,15	0,1	-0,06	-0,08	-0,01	0,2	-0,31	-0,21	0,04	0,15	0,13	-0,05	0,1	-0,03	-0,02	-0,02		
Arom	-0,03	-0,03	-0,08	-0,06	0,08	0,18	-0,02	-0,16	-0,02	0,00	0,02	0,01	-0,02	-0,30	-0,21	-0,20	0,42	1

Los índices de correlación de Spearman en color rojo no significativos (p-val > 0,05); en color azul significativos (p-val < 0,05); en color negro significativos (p-val < 0,01)

Tabla IV.16c: Correlaciones entre los factores medidos de los genes de *E. granulosus* G1 y los 4 primeros componentes principales

	CP1	CP2	CP3	CP4	T3s	C3s	A3s	G3s	CAI	CBI	Fop	ENc	GC3s	GC	L_sym	L_aa	Gravy	Aromo
CP1																		
CP2	0,05																	
CP3	0,04	-0,04																
CP4	0,01	0	-0,02															
T3s	-0,55	-0,7	0,02	0,09														
C3s	0,81	0,24	-0,03	0,25	-0,64													
A3s	-0,77	0,26	-0,26	-0,07	0,17	-0,56												
G3s	0,33	0,22	0,36	-0,37	-0,35	0,08	-0,32											
CAI	0,78	-0,23	-0,05	0,13	-0,15	0,63	-0,72	0,18										
CBI	0,93	0,14	0,05	0,07	-0,64	0,85	-0,76	0,37	0,73									
Fop	0,93	0,14	0,05	0,08	-0,62	0,86	-0,74	0,38	0,74	0,99								
ENc	-0,25	0,15	0,07	-0,14	-0,01	-0,2	0,29	0,06	-0,38	-0,23	-0,22							
GC3s	0,88	0,31	0,17	-0,02	-0,75	0,82	-0,7	0,54	0,6	0,93	0,93	-0,13						
GC	0,76	0,14	0,09	-0,09	-0,6	0,56	-0,64	0,31	0,53	0,76	0,75	-0,15	0,76					
L_sym	0,26	-0,03	-0,05	-0,11	-0,13	0,11	-0,19	0,06	0,2	0,22	0,22	0,03	0,19	0,36				
L_aa	0,26	-0,03	-0,05	-0,11	-0,13	0,12	-0,19	0,06	0,2	0,23	0,22	0,03	0,19	0,36				
Hf	0,13	-0,08	-0,02	0,13	-0,02	0,2	-0,3	-0,21	0,02	0,15	0,1	-0,04	0,1	-0,04	-0,03	-0,02		
Arom	-0,03	0,03	-0,03	0,14	0,07	0,19	-0,02	-0,17	-0,03	-0,01	0,01	0,01	-0,02	-0,3	-0,22	-0,22	0,42	1

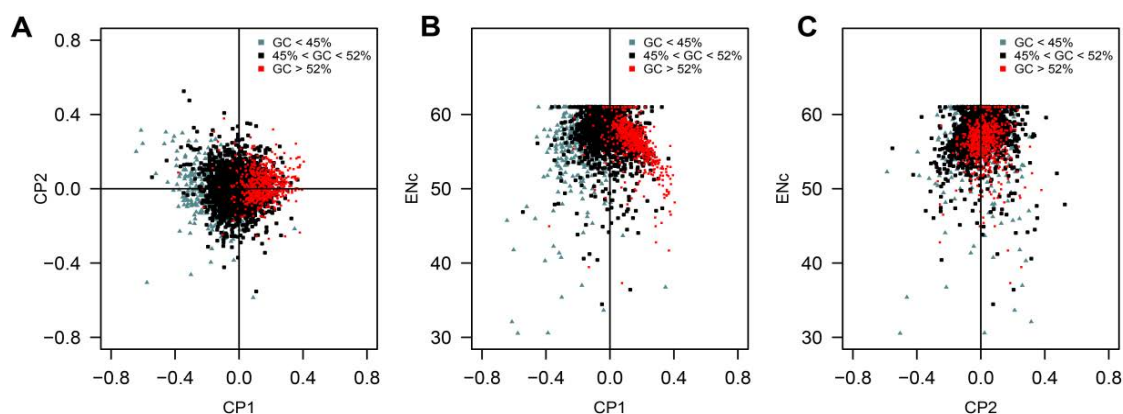
Los índices de correlación de Spearman en color rojo no significativos ($p\text{-val} > 0,05$); en color azul significativos ($p\text{-val} < 0,05$); en color negro significativos ($p\text{-val} < 0,01$)

Al evaluar las correlaciones entre los CPs y ENc y los diferentes parámetros calculados para los genes de *Echinococcus*, se observó que los dos primeros componentes (CP1 y CP2) fueron los que mostraron una mayor correlación con ENc con p-valores menores a 0,01. Asimismo, para el CP3 y ENc casi no se observa correlación excepto para *E. multilocularis* que muestra una ligera correlación positiva significativa. Por otra parte, para el CP4 y ENc el único que exhibe una correlación significativa es *E. granulosus* G1 cuyo valor es negativo (Tabla IV.16).

Basado en las correlaciones significativas observadas en las tablas anteriores, particularmente las correlaciones entre ENc con el porcentaje de GC, el porcentaje de GC3s, los valores de CAI, la hidrofobicidad y la aromaticidad es que se eligieron dichos parámetros para evaluar la influencia que poseen sobre el sesgo del uso de codones en los genes de *Echinococcus*. Para ello estos parámetros fueron integrados a los ACP y se analizó la distribución de los genes en las coordenadas en el espacio de los dos primeros CPs.

4.6.4.5 Efecto composicional de los genes sobre el sesgo del uso de codones

Para estudiar el efecto del contenido de GC sobre el sesgo del uso de codones se aplicaron los análisis de correspondencia de manera similar a los realizados para el análisis del uso de codones sinónimos. Se evaluó la distribución de los genes en el espacio multidimensional de acuerdo a la composición de nucleótidos total y a la composición de nucleótidos en la tercera posición de los codones. Para cada caso se obtuvo la distribución de los genes según su contenido GC a lo largo de los dos primeros CPs (CP1 vs CP2), ENc y CP1 y ENc y CP2 (Figura IV.39).



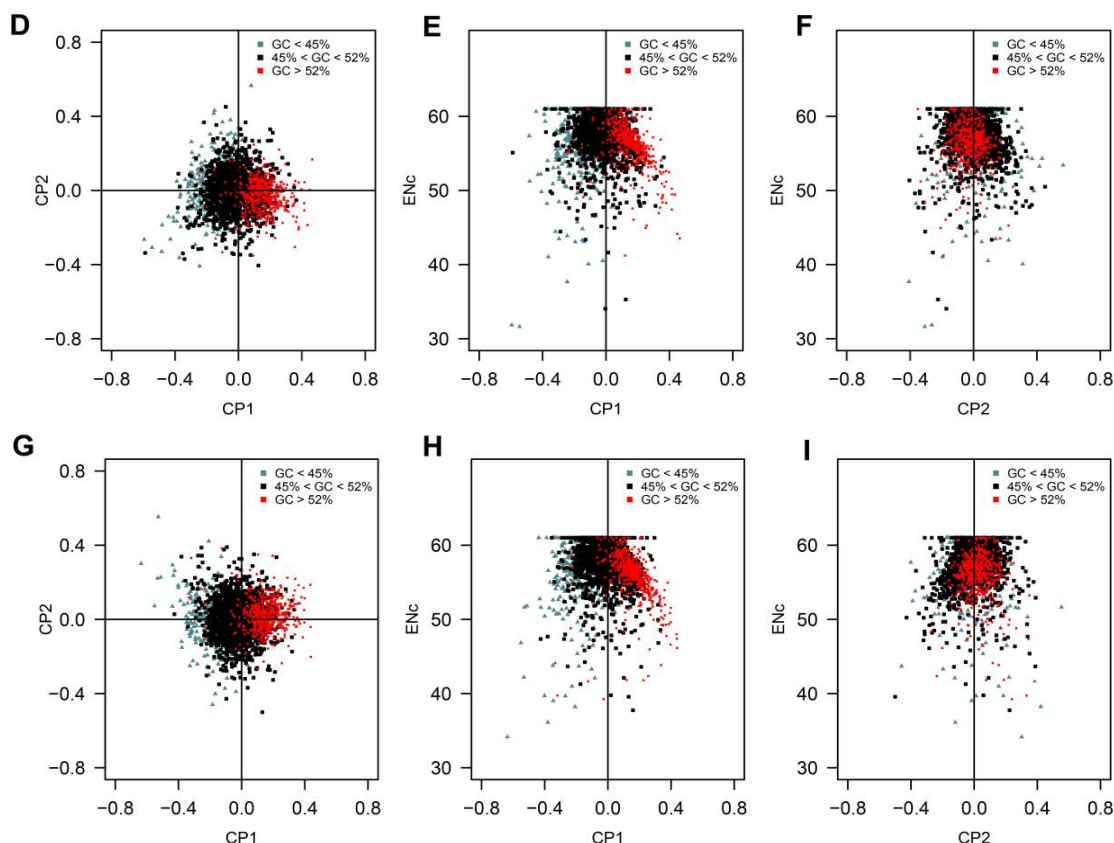


Figura IV.39: Análisis de correspondencia. Distribución de los genes en las coordenadas de los componentes principales en relación con ENc. A). Distribución de genes de *E. canadensis* G7 en los CP1 y CP2. B). Distribución de genes de *E. canadensis* G7 en CP1 y valores de ENc. C). Distribución de genes de *E. canadensis* G7 en CP2 y valores de ENc. D). Distribución de genes de *E. multilocularis* en los CP1 y CP2. E). Distribución de genes de *E. multilocularis* en CP1 y valores de ENc. F). Distribución de genes de *E. multilocularis* en CP2 y valores de ENc. G). Distribución de genes de *E. granulosus* G1 en los CP1 y CP2. H). Distribución de genes de *E. granulosus* G1 en CP1 y valores de ENc. I). Distribución de genes de *E. granulosus* G1 en CP2 y valores de ENc. Los puntos representan la distribución de los genes de cada especie coloreados de acuerdo al contenido GC total de los genes.

Se observó que el CP1 es el componente que mejor refleja la variabilidad de los genes de acuerdo a su contenido total de GC. Además se observó que los genes se distribuyeron a lo largo del CP1 posicionando en el extremo izquierdo a aquellos genes con menor contenido de GC y en el extremo derecho a los de mayor contenido de GC (panel A, D y G) para las tres especies de *Echinococcus*. Al graficar ENc vs CP1 y agrupando a los genes según su contenido de GC se observó una ligera tendencia de que los genes con mayor porcentaje de GC poseen menor ENc, sin embargo hay grupos de genes con valores de ENc muy bajos que poseen contenido de GC menores al 45%. Al comparar entre las tres especies de *Echinococcus* no se observaron diferencias en la distribución de los genes relacionados a los valores de GC. Sin embargo se observa una distribución más compacta en los genes de *E. canadensis* G7 y *E. multilocularis*. Al realizar este análisis según su contenido de GC3s se observó que los genes se distribuyeron principalmente sobre el CP1. Los genes con menor proporción de GC3s se distribuyeron hacia el extremo izquierdo del CP1 y los genes con mayor proporción GC3s hacia el extremo derecho. Al graficar ENc vs CP1 se observó una dispersión similar a la anterior, sin embargo hay una mayor distribución de genes que poseen

valores de ENC inferiores y proporciones de GC3s mayores. Asimismo, en *E. canadensis* G7 y *E. multilocularis* la distribución es un poco más compacta y hacia valores de ENC inferiores pudiendose observar grupos de genes de bajo contenido GC3s ([Figura IV.40](#))

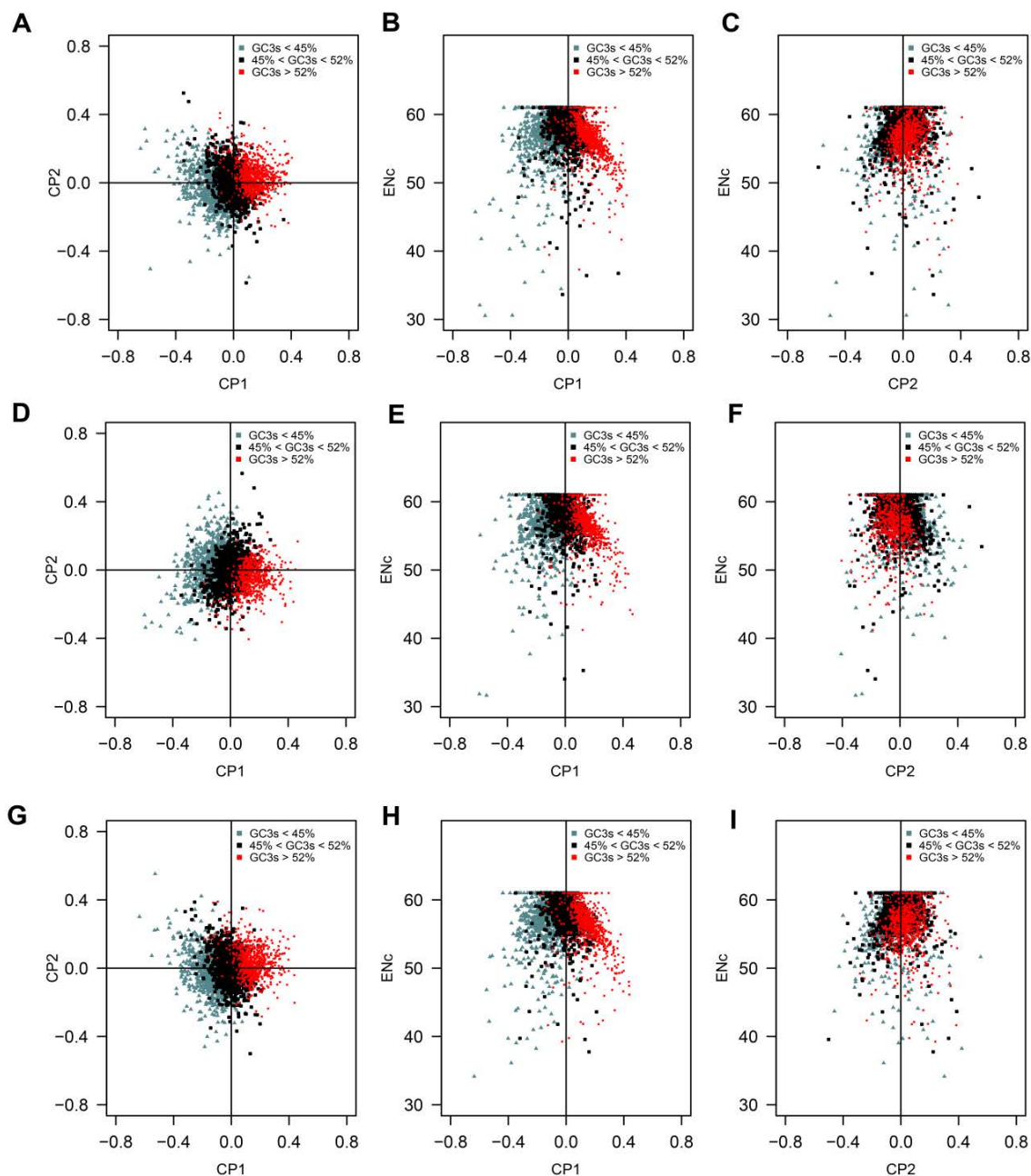


Figura IV.40: Analisis de correspondencia. Distribución de los genes en las coordenadas de los componentes principales en relación con ENC. A). Distribución de genes de *E. canadensis* G7 en los CP1 y CP2. B). Distribución de genes de *E. canadensis* G7 en CP1 y valores de ENC. C). Distribución de genes de *E. canadensis* G7 en CP2 y valores de ENC. D). Distribución de genes de *E. multilocularis* en los CP1 y CP2. E). Distribución de genes de *E. multilocularis* en CP1 y valores de ENC. F). Distribución de genes de *E. multilocularis* en CP2 y valores de ENC. G). Distribución de genes de *E. granulosus* G1 en los CP1 y CP2. H). Distribución de genes de *E. granulosus* G1 en CP1 y valores de ENC. I). Distribución de genes de *E. granulosus* G1 en CP2 y valores de ENC. Los puntos representan la distribución de los genes de cada especie coloreados de acuerdo al contenido GC en la tercera posición de los codones sinónimos que conforman a los genes.

Por otra parte, al analizar ENc vs CP2 (paneles C, F e I) se observó que los genes se distribuyeron de manera más compacta con centro en el 0 del eje del CP2. Además la dispersión de los genes en los gráficos de GC vs CP2 fue similar a la de los gráficos de GC3s vs CP2. No obstante, se distinguió una mayor distribución de genes con porcentajes de GC3s mayores al 52% que la que pudo observarse al evaluar el porcentaje de GC génico total. Asimismo, los genes de las especies de *E. canadensis* G7 y *E. multilocularis* se distribuyeron a lo largo del CP2 según el porcentaje GC3s con una mayor dispersión que la que se observó para los genes de *E. granulosus* G1. Estas observaciones fueron sustentadas mediante una correlación significativa entre los valores de GC y PC2 y entre los valores de GC3s y PC2 con p-valores menores a 0,01 ([Tabla IV.16](#)).

4.6.4.6 Efecto del nivel de expresión de los genes sobre el sesgo del uso de codones

Para explorar la relación entre el nivel de expresión génico y el sesgo en el uso de codones de *Echinococcus* se calculó el índice de adaptación de codones (CAI), y otras características tales como ENc y el porcentaje de GC, los cuales fueron integrados en los ACPs. En primer lugar para verificar la relación entre el índice CAI y los niveles de expresión de los genes se correlacionaron los valores obtenidos de CAI con los valores de RFPKM de *E. multilocularis* obtenidos por Tsai y col [43] observándose un índice de correlación positivo significativo de 0,32 (p-valor < 0.005). Los análisis de correlación multivariados mostraron una alta correlación positiva del índice CAI con los dos primeros CPs para las tres especies de *Echinococcus*, ([Tabla IV.16](#)). A su vez el índice CAI también presentó una correlación significativa con ENc, %GC, %GC3s y el tamaño de los genes. La distribución de los genes basado en sus valores de CAI y ENc mostraron una correlación negativa y al graficarlos agrupados según su contenido de GC se observó una distribución de los genes con mayor contenido de GC hacia valores mayores de CAI; lo que concuerda con los valores positivos del coeficiente de correlación de Spearman entre el %GC y %GC3s con el índice CAI ([Figura IV.41](#)). Al explorar los datos de los ACP se observó que CP1 contribuye mejor a la dispersión de los genes a lo largo de su eje. Asimismo se observó un grupo de genes que se agrupan hacia valores mayores del índice CAI con mayor porcentaje de GC. Al comparar entre las tres especies de *Echinococcus* se observó que *E. multilocularis* es la especie cuyos genes presentan menor dispersión y hacia valores de CAI ligeramente menores. Asimismo, se observó una correlación positiva significativa entre CAI y Fop y se determinó que al menos un 5% de los genes de *Echinococcus* poseen codones optimizados ([Tabla IV.16](#)).

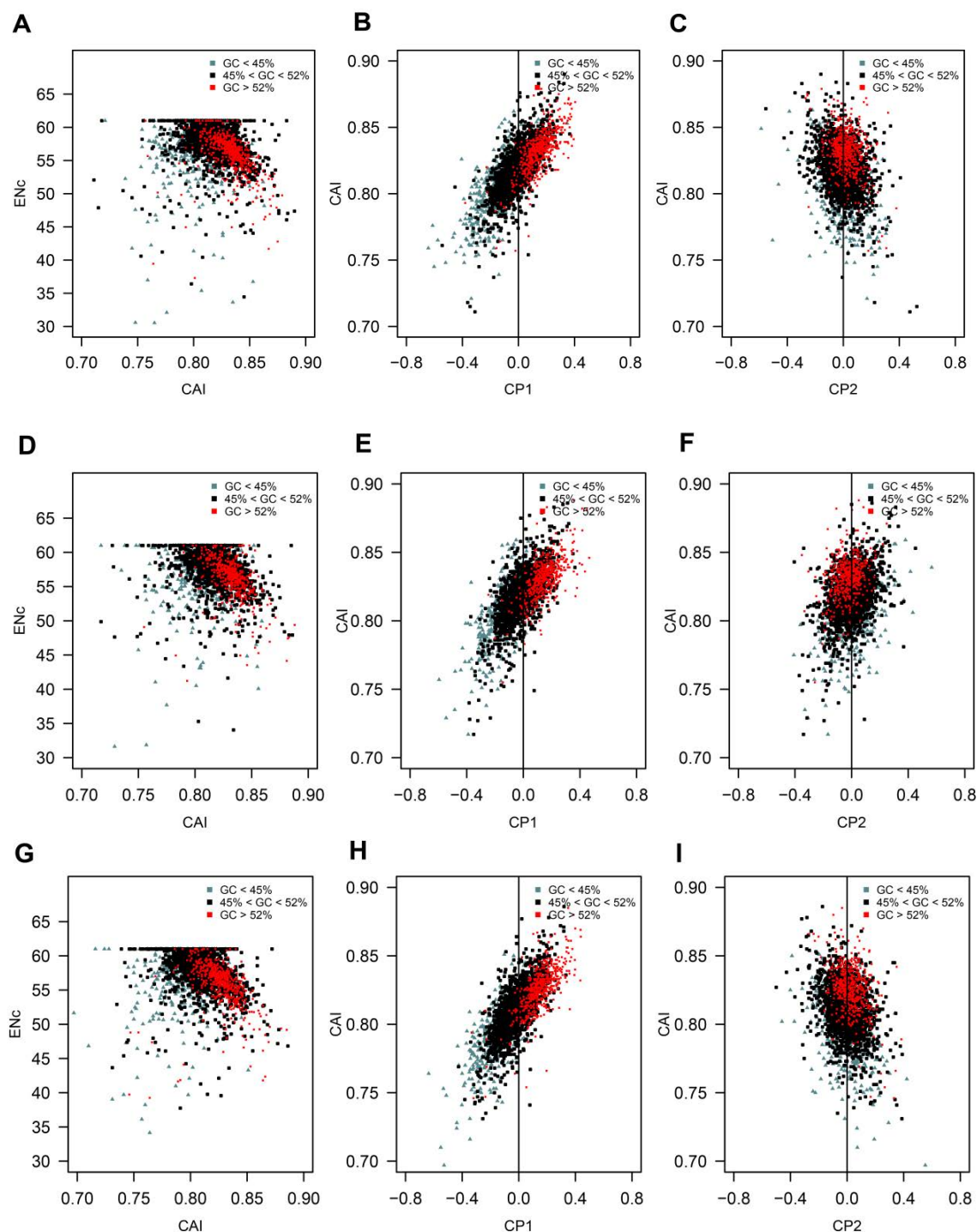
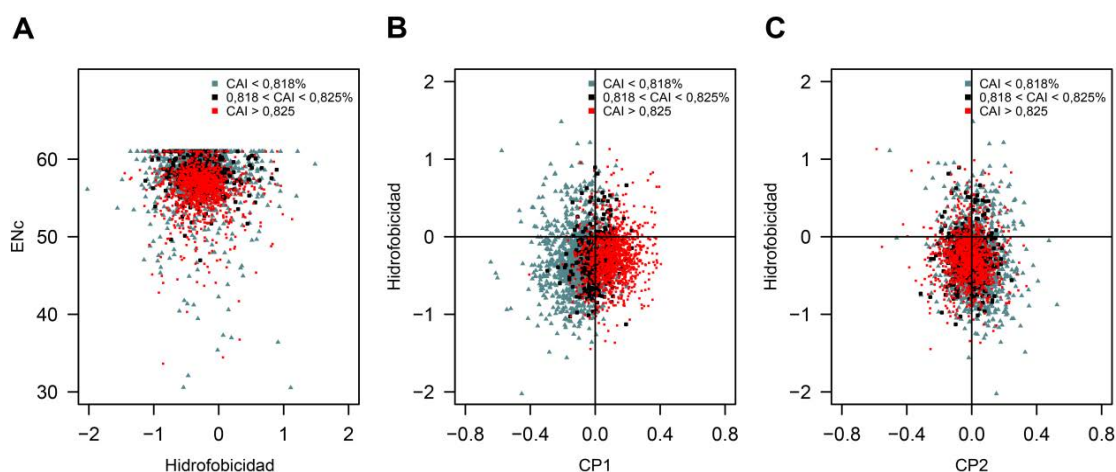


Figura IV.41: Relación entre ENC y CAI y análisis de correspondencia. Distribución de los genes de acuerdo a sus valores de ENC, CAI y contenido de GC. A). Distribución de genes de *E. canadensis* G7 entre CAI y ENC B). Distribución de genes de *E. canadensis* G7 en CP1 y valores de CAI. C). Distribución de genes de *E. canadensis* G7 en CP2 y valores de CAI. D). Distribución de genes de *E. multilocularis* entre CAI y ENC. E). Distribución de genes de *E. multilocularis* en CP1 y valores de CAI. F). Distribución de genes de *E. multilocularis* en CP2 y valores de CAI. G). Distribución de genes de *E. granulosus* G1 entre CAI y ENC. H). Distribución de genes de *E. granulosus* G1 en CP1 y valores de CAI. I). Distribución de genes de *E. granulosus* G1 en CP2 y valores de CAI. Los puntos representan la distribución de los genes de cada especie coloreados de acuerdo al contenido GC total de los genes

4.6.4.7 Efecto de la hidrofobicidad de las proteínas codificadas sobre el sesgo del uso de codones

Numerosos estudios han demostrado que la hidrofobicidad (Hf) y aromaticidad (Arom) de las proteínas codificadas por los genes juegan un papel importante en la formación del uso de codones de algunas especies. Para investigar si el sesgo en el uso de codones está influenciado por estos factores, se realizó un análisis de correlación para evaluar si los valores de los índices Hf estaban relacionados con los valores ENc. Los análisis de correlación entre la Hf de cada proteína y el valor ENc mostraron bajos coeficientes de correlación pero fueron significativamente correlacionados. La aromaticidad de las proteínas no se correlacionó significativamente con ENc ($r = 0.03$, 0.01 y 0.01 con P-valores > 0.05) para *E. canadensis* G7, *E. multilocularis* y *E. granulosus* G1 respectivamente. Los resultados de estos análisis indicaron que la variación en el uso de codones está relacionada con el grado de hidrofobicidad de las proteínas, pero no con la aromaticidad de los aminoácidos aromáticos que conforman a las proteínas. Al graficar ENc vs Hf y agrupados por sus valores de CAI se observó una distribución de los genes más compacta que en los casos anteriores, sin que pueda observarse una dependencia clara del ENc respecto a los valores de Hf. Por otra parte, se observó que la distribución de los genes con mayores valores de CAI se localizaron mayoritariamente hacia valores menores de Hf ($HF \sim -0.5$) ([Figura IV.42](#)). Al explorar los datos de los ACP se observó que el CP1 contribuye mejor a la dispersión de los genes a lo largo de su eje.



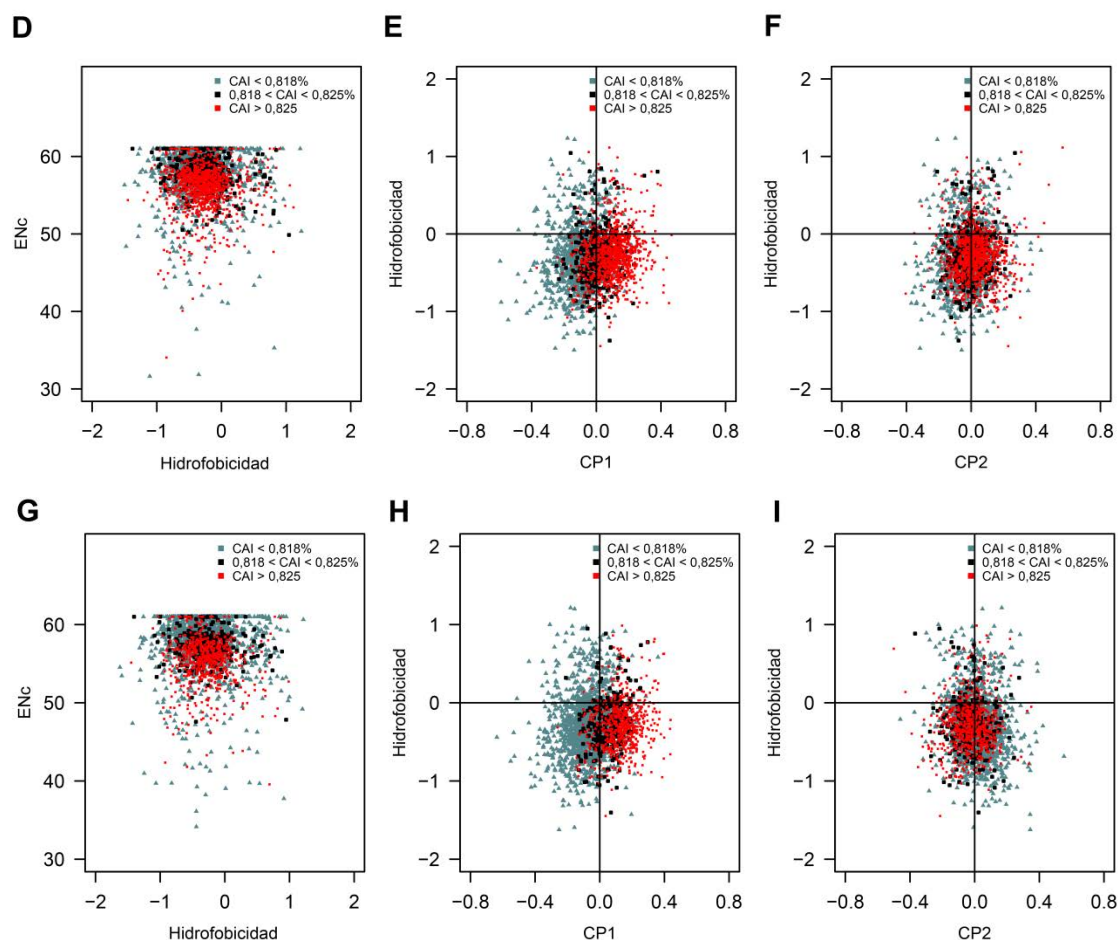


Figura IV.42: Relación entre ENc y el índice de hidrofobicidad (Hf) y análisis de correspondencia. Distribución de los genes de acuerdo a sus valores de ENc, Hf. y el índice CAI. A). Distribución de genes de *E. canadensis* G7 entre Hf y ENc B). Distribución de genes de *E. canadensis* G7 en CP1 y valores de Hf. C). Distribución de genes de *E. canadensis* G7 en CP2 y valores de Hf. D). Distribución de genes de *E. multilocularis* entre Hf y ENc. E). Distribución de genes de *E. multilocularis* en CP1 y valores de Hf. F). Distribución de genes de *E. multilocularis* en CP2 y valores de Hf. G). Distribución de genes de *E. granulosus* G1 entre Hf y ENc. H). Distribución de genes de *E. granulosus* G1 en CP1 y valores de Hf. I). Distribución de genes de *E. granulosus* G1 en CP2 y valores de Hf. Los puntos representan la distribución de los genes de cada especie coloreados de acuerdo al índice CAI de los genes.

4.6.4.8 Efecto del tamaño de los genes sobre el sesgo del uso de codones

Para estudiar el efecto que pueda tener el tamaño de los genes o proteínas (L_{aa}) sobre el sesgo del uso de codones se realizó un análisis de correlación entre L_{aa} y ENc, obteniéndose un índice de correlación significativo de 0,04 con p-valores < 0,05 ([Tabla IV.16](#)). Al graficar el ENc respecto de L_{aa} y agrupando según los valores de CAI, se observó que los genes de menor tamaño mostraron valores variables de ENc y que los genes de mayor tamaño poseen valores de ENc mayores. No obstante, las proteínas de menor tamaño mostraron valores mayores de CAI; ([Figura IV.43](#)). Se observó además que el CP1 contribuyó más eficientemente a la dispersión de los genes según su tamaño y agrupó a los que poseen valores de CAI menores hacia su extremo izquierdo y los que poseen valores de CAI mayores hacia el extremo derecho. Por el contrario CP2 no logró dispersar los genes según su tamaño o los valores de CAI.

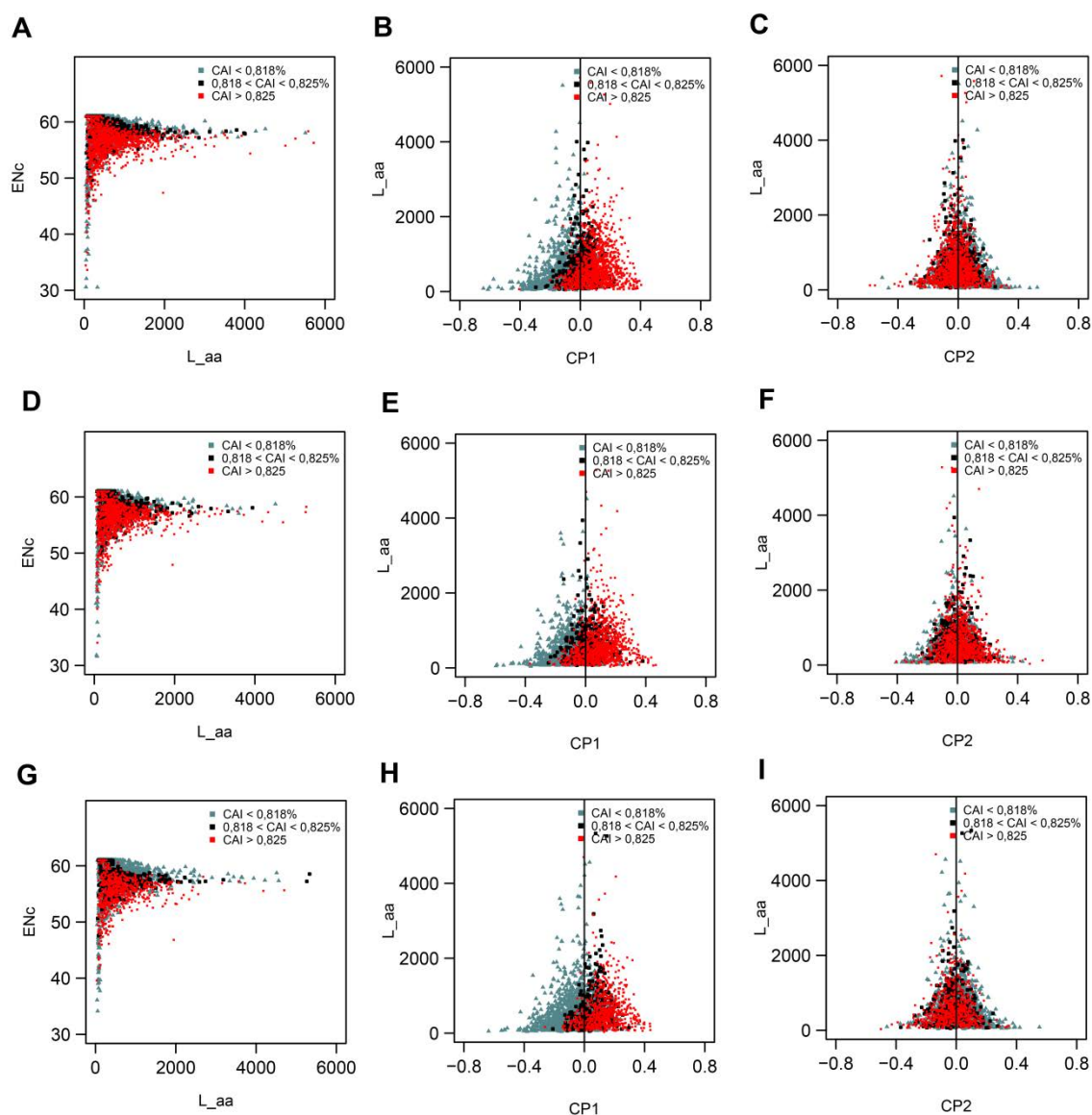


Figura IV.43: Relación entre ENc y el tamaño de las proteínas (L_{aa}) y análisis de correspondencia. Distribución de los genes de acuerdo a sus valores de ENc, Hf y el índice CAI. A). Distribución de genes de *E. canadensis* G7 entre L_{aa} y ENc B). Distribución de genes de *E. canadensis* G7 en CP1 y valores de L_{aa}. C). Distribución de genes de *E. canadensis* G7 en CP2 y valores de L_{aa}. D). Distribución de genes de *E. multilocularis* entre L_{aa} y ENc. E). Distribución de genes de *E. multilocularis* en CP1 y valores de L_{aa}. F). Distribución de genes de *E. multilocularis* en CP2 y valores de L_{aa}. G). Distribución de genes de *E. granulosus* G1 entre L_{aa} y ENc. H). Distribución de genes de *E. granulosus* G1 en CP1 y valores de L_{aa}. I). Distribución de genes de *E. granulosus* G1 en CP2 y valores de L_{aa}. Los puntos representan la distribución de los genes de cada especie coloreados de acuerdo al índice CAI de los genes.

4.6.5 Análisis global e integrado del efecto de los distintos factores sobre el sesgo en el uso de codones en las especies de *Echinococcus*

Para obtener una visión global e integrada del efecto de los distintos factores sobre el sesgo en el uso de codones en las especies de *Echinococcus* se analizó la distribución de diferentes tipos de genes en los dos primeros CPs. Para ello se seleccionaron los genes de cada especie que exhibieron características que se alejaron de la distribución normal: contenido de GC mayor al 55% y menor al 42%; contenido de GC3s mayor al 60% y menor al 38%; aromaticidad mayor a 0.15 y menor a 0,035; hidrofobicidad mayor a 0,46 y menor a -0,9, ENc menor a 50 e igual a 61 y CAI mayor a 0,85 y menor a 0,75. Bajo estos criterios se seleccionaron un total de 850, 755 y 801 genes de *E. canadensis* G7, *E. multilocularis* y *E. granulosus* G1, respectivamente. El número de genes de cada especie agrupados según las características mencionadas puede observarse en la [tabla IV.17](#) y la distribución total de los genes y la media de la distribución de cada grupo anteriormente mencionado se representó en un gráfico de distribución en componentes principales ([Figura IV.44](#)).

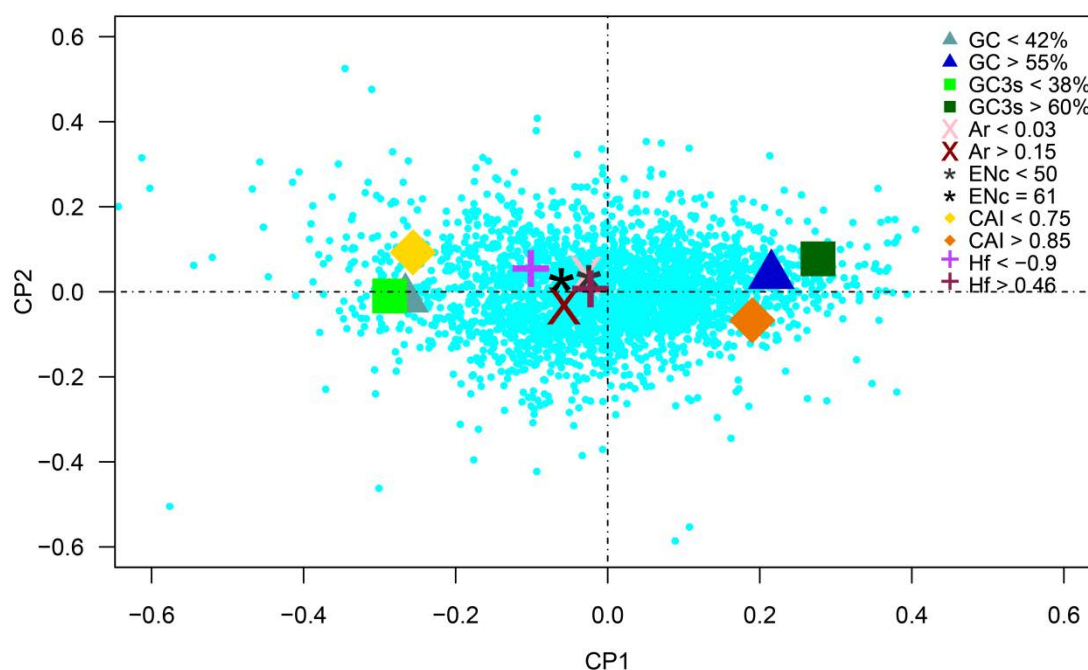
Tabla IV.17: Número de genes con características alejadas de la distribución normal de las especies de *Echinococcus*

	<i>E. canadensis</i> G7	<i>E. multilocularis</i>	<i>E. granulosus</i> G1
GC < 42%	85	73	70
GC > 55%	115	97	113
Arom < 0,035	50	55	54
Arom > 0,15	56	49	49
GC3s < 0,38	96	83	72
GC3s > 0.6	87	66	90
ENc < 50	99	86	94
ENc = 61	245	225	233
CAI < 0,78	113	90	174
CAI > 0,85	109	93	52
Hf < -0,9	91	80	94
Hf > 0,46	99	104	109
Total de Genes	850	755	801

Se observó que la distribución de los genes fue mayor a lo largo de CP1, el cual separó a los grupos de genes mayormente según su contenido de GC y GC3s (menor porcentaje de GC hacia valores negativos y mayor porcentaje de GC hacia valores positivos) y según su valor de CAI (menores valores de CAI hacia valores negativos y mayores valores de CAI hacia valores positivos). Por su parte el CP2 dispersa en menor grado dichas características. La mayoría de los genes con mayor contenido GC y GC3s se posicionaron en el cuadrante superior derecho para *E. canadensis* y *E. granulosus* G1 mientras que para *E. multilocularis* se localizaron en el cuadrante inferior derecho. Por

el contrario los genes con menor contenido GC y GC3s se posicionaron en el cuadrante inferior izquierdo para *E. canadensis* y *E. granulosus* G1 mientras que para *E. multilocularis* se localizaron en el cuadrante superior izquierdo. De modo similar la mayoría de los genes con mayor índice CAI se posicionaron en el cuadrante inferior derecho y los genes con menor índice CAI se localizaron en el cuadrante superior izquierdo para *E. canadensis* y *E. granulosus* G1 mientras que para *E. multilocularis* los genes con menor y mayor CAI se posicionan en el cuadrante inferior izquierdo y superior derecho respectivamente. Estas observaciones desiguales en *E. multilocularis* concuerdan con la correlación negativa que exhibió el CP2 con tales características. Los genes con valores altos y bajos de aromaticidad e hidrofobicidad se posicionaron cerca de la intersección de los CP1 y CP2. Los genes con mayor hidrofobicidad se localizaron hacia valores mayores en el CP1 en las tres especies de *Echinococcus*, mientras que los genes con mayor aromaticidad se distribuyeron hacia valores menores del CP1 pero a una distancia que fue menor para *E. multilocularis* y *E. canadensis* G7 que para *E. granulosus* G1. En efecto, la media de los genes de mayor hidrofobicidad y menor aromaticidad se superpusieron para *E. granulosus* G1 coincidiendo también con la media de los genes con menor ENc. Por su parte; la media de la distribución de los genes con menor ENc se localizó hacia valores mayores en el CP1. Al igual que lo observado para las demás características el CP2 en *E. multilocularis* mostró una distribución contraria a la observada en *E. granulosus* G1 y *E. canadensis* G7 acorde a la correlación negativa estimada previamente.

A



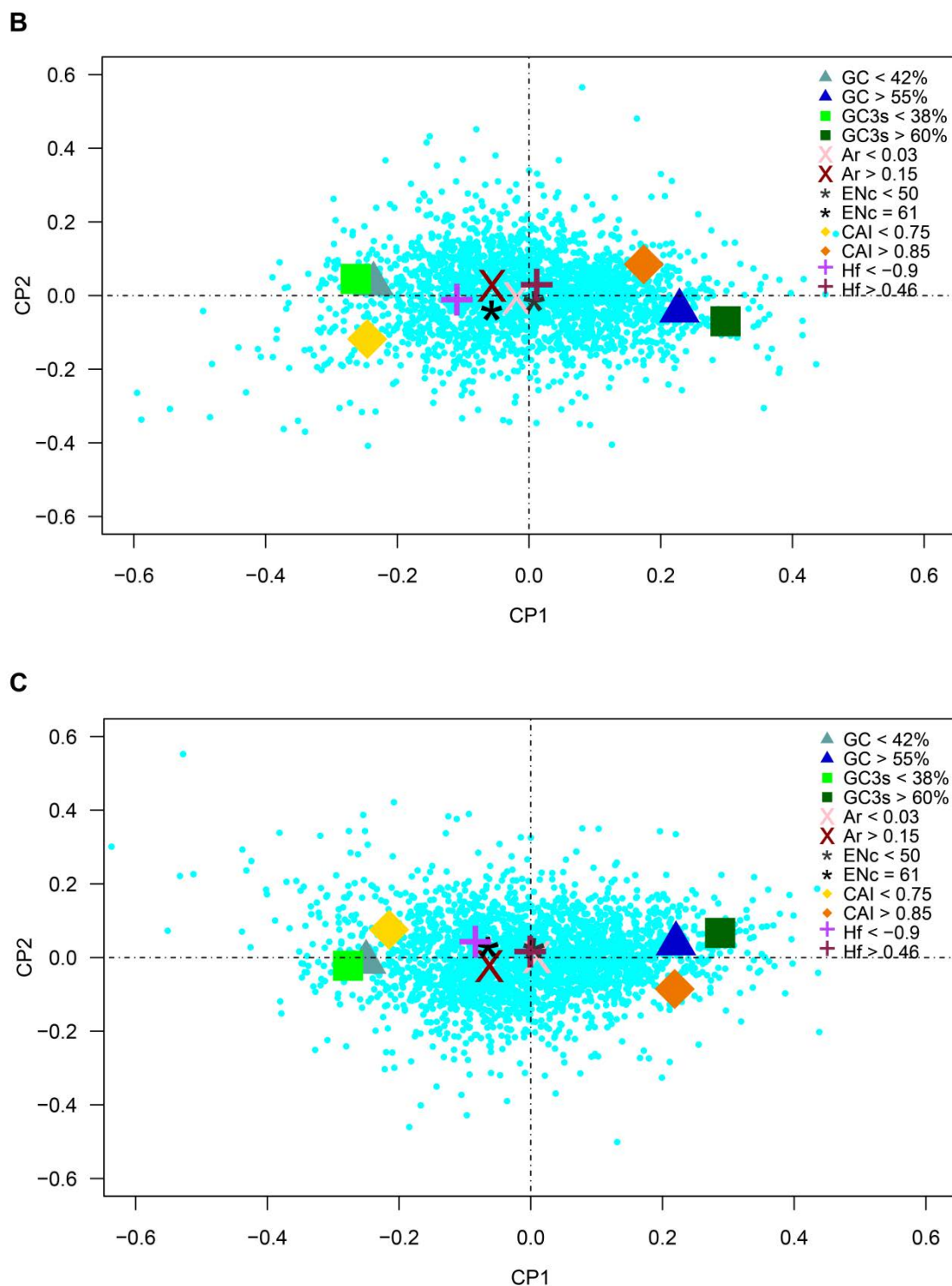
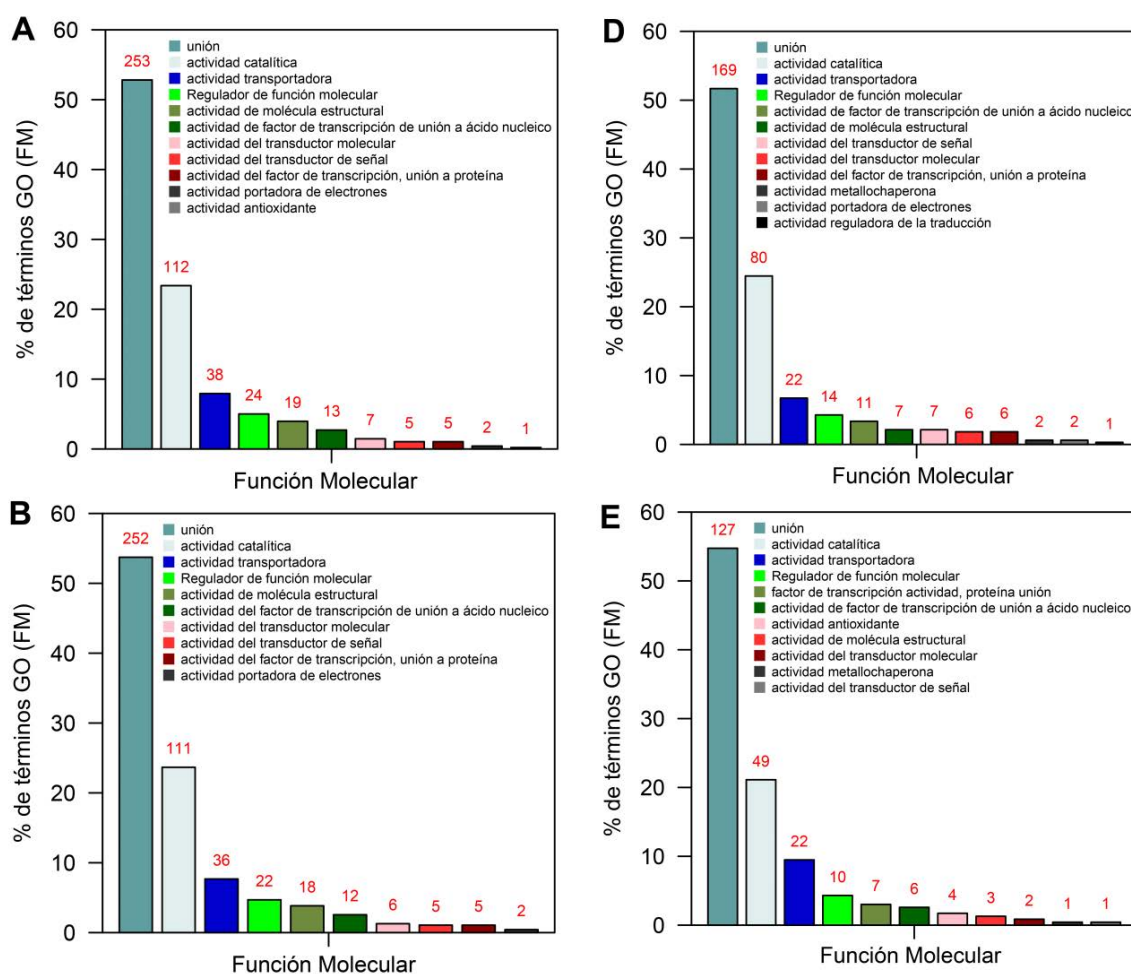


Figura IV.44: Representación de la distribución de los genes de *Echinococcus* en un gráfico de componentes principales para A). *E. canadensis* G7; B). *E. multilocularis*; C). *E. granulosus* G1. Las figuras referenciadas en los gráficos representan la media de la distribución de los genes agrupados de acuerdo a la tabla IV.17. Los puntos celestes representan la distribución total de los genes de cada especie.

4.6.5.1 Tipos de genes involucrados en la variabilidad del uso de codones identificados por ACP

Para conocer la función biológica de los genes cuya distribución fue analizada en los estudios de ACP anteriores se los relacionó con los términos GO de Función molecular (Fm). De los 850 genes de *E. canadensis* G7 519 de ellos poseen términos asociados a la categoría Fm. Del total de 755 genes de *E. multilocularis*, 458 poseen términos asociados a la categoría Fm y para *E. granulosus* G1 472 genes fueron asociados a términos GO de la categoría Fm de un total de 801 genes.

La distribución de los genes en las categorías GO de Fm fue similar a la distribución total donde los términos más abundantes fueron “Unión” y “Actividad catalítica” ([Figura IV.45. A, B y C](#)). También se evaluó la distribución de los genes que no quedaron agrupados según la [tabla IV.17](#) pero sí lo hicieron sus ortólogos y se observó una distribución similar a la anteriormente descrita. La distribución y número de genes en cada categoría de FM puede observarse en la [figura IV.45. D, E y F](#).



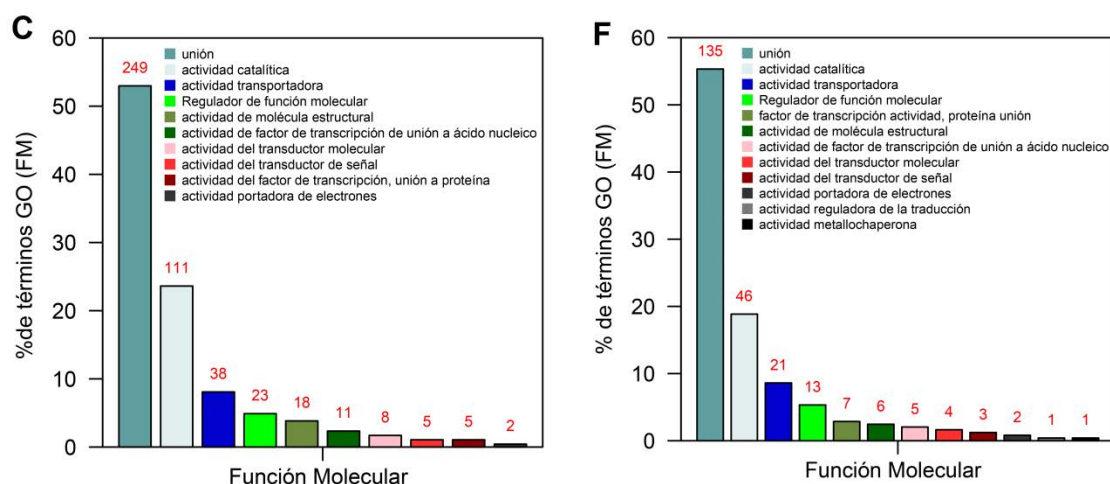


Figura IV.45: Distribución de los genes de *Echinococcus* agrupados por ACP en las categorías principales de los términos GO de FM. Los números sobre las barras indica el número de genes.

Entre el grupo de genes con un alto sesgo en el uso de codones se identificaron genes que mostraron valores de ENc por debajo de 36. Estos genes involucraron a 5 genes de *E. canadensis* G7, 4 genes de *E. multilocularis* y solo 1 gen de *E. granulosus* G1. De las 10 proteínas, 6 de ellas están anotadas como proteínas hipotéticas y las 4 restantes poseen la siguiente anotación funcional: proteína de unión a SH₃ rica en ácido glutámico (ECANG7_07463), proteína de ensamblaje de la citocromo C oxidasa PET191 (ECANG7_03640 y EmuJ_000818300) y proteína transmembrana 93 (EmuJ_000847600).

4.6.6 Conclusiones parciales

En esta sección de la tesis se estudió el uso de codones en tres especies de *Echinococcus*, su variabilidad entre los distintos genes dentro de una misma especie, entre los genes de las tres especies de *Echinococcus*, y los fenómenos evolutivos y factores biológicos que influyen en el sesgo del uso de codones. En las tres especies de *Echinococcus* se observó que la media del contenido de GC en la región codificante es mayor que la del genoma y similar entre los genes de las tres especies. No obstante; la tercera posición (GC3s) que es la menos restrictiva para la codificación de aminoácidos tiene un porcentaje promedio de GC de 48,7% para *E. canadensis* G7 y *E. granulosus* G1 y 48,3% para *E. multilocularis*. Además el rango en el contenido de GC para esta posición es mayor para los genes con bajo nivel de expresión pero la mediana es mayor para los genes con alto nivel de expresión. Por otra parte se observó que la fuerzas evolutivas que moldean el uso de codones se deben en un ~80% a los fenómenos de selección mientras que el 20% restante se debe a un sesgo mutacional. Las representaciones graficas de PR2 y GC3s vs ENc confirmaron estas observaciones al posicionar a la mayoría de los genes por debajo de la curva que se esperaba obtener si el sesgo en el uso de codones solo estuviese influenciado por el contenido de GC, es decir si el sesgo en el uso de codones dependiera solamente del sesgo mutacional. Los análisis de correspondencia de los 4 componentes principales permitieron explicar

aproximadamente el 22% de la variabilidad total del RSCU. Adicionalmente se observó una correlación negativa significativa entre ENc y CAI y entre ENc y Fop y una correlación positiva entre CAI y Fop y que a su vez correlaciona con un mayor contenido de GC en la tercera posición de los codones indicando que los genes con mayor nivel de expresión tendrían un mayor sesgo de uso de codones al mismo tiempo que poseen un mayor porcentaje de GC. Por otra parte, la hidrofobicidad y aromaticidad de las proteínas parecen cumplir un rol menos importante. No obstante, se observó que dentro de cada especie existen grupos de genes con alto y bajo sesgo en el uso de codones. Los genes que pudieron ser discriminados por los ACP no mostraron un patrón definido o enriquecimiento en términos GO para determinar su funcionalidad.



DISCUSIÓN Y CONCLUSIONES

Capítulo V

Discusión y Conclusiones

5.1 Genoma de *Echinococcus canadensis* G7

E. canadensis G7 es uno de los parásitos más ampliamente distribuidos en el mundo responsable del 30% de los casos de echinococcosis humana. A pesar de que en los últimos años se han secuenciado y obtenido varios genomas de platelmintos, este parásito ha sido poco estudiado.

Como sucede con muchos proyectos genoma, aún los genomas más completos de platelmintos (*E. multilocularis* y *S. mansoni*) poseen numerosos “gaps” que no han podido completarse. Más aún, el estudio de genomas individuales todavía no ha permitido describir completamente características que están involucradas en la capacidad infectiva, desarrollo o supervivencia de estas especies. Y tampoco ha podido explicarse o encontrar la funcionalidad a muchos de sus genes, los cuales en su mayoría son altamente divergentes a los presentes en otros organismos eucariotas y no presentan dominios conocidos que permitan inferir la función de las proteínas para las que codifican. Por otra parte, basados en los genomas de platelmintos disponibles es que se han descrito o inferido características generales de los platelmintos que están lejos de representar a la gran variedad de especies que los contienen. Y no solo características generales, sino también anotaciones, predicciones génicas y anotaciones funcionales de genes los cuales en muchos casos están sujetos a errores ya sea por predicción equivocada, por extrapolación de características o por transferencia de anotación. Aun luego de la publicación inicial de un genoma, los genomas *draft* deben ser sometidos a mejoras constantes basadas no solo en métodos computacionales sino también funcionales, sin embargo, la mayoría de los genomas de helmintos publicados sólo están siendo sometidos a mejoras impulsadas por análisis bioinformáticos [265].

Generalmente, los proyectos genoma involucran a consorcios o equipos interdisciplinarios especializados que intervienen en las distintas etapas del proyecto, ya sea durante la obtención del material de partida, en la obtención del genoma, la anotación, y las etapas de curación entre otros y pueden involucrar interacciones con grupos externos o comités de nomenclatura génica [266]. En esta tesis se secuenció, ensambló y anotó el genoma de *E. canadensis* G7 realizando todas las etapas que implica un proyecto genoma. El tamaño del genoma de *E. canadensis* G7 fue de 115 Mb, similar al observado en *E. multilocularis* [43] y *E. granulosus* G1 [43,78]. La calidad de las lecturas, la cobertura y profundidad que se logró en la secuenciación del ADN genómico permitió realizar un ensamblado *de novo*. Esta metodología posee

notables ventajas respecto al uso de un genoma de referencia ya que pueden obtenerse secuencias nuevas de regiones que no están disponibles o son diferentes respecto al genoma de referencia. Su calidad basada en la medida del N50 fue alta y similar a la obtenida para la primera versión del genoma de *S. mansoni* por un consorcio internacional [40].

Durante la anotación de un genoma entran en juego numerosos factores que determinan que estrategia elegir. Obviamente, las consideraciones económicas suponen importantes limitaciones para la disponibilidad de recursos, de potencia computacional, y para la generación de datos experimentales que proporcionen evidencia para la construcción de modelos y por supuesto la estrategia también depende del propósito de la anotación [266]. Un objetivo común de los proyectos de secuenciación dentro de la biología evolutiva es encontrar genes que hayan sido sometidos a una selección positiva. En este escenario, se da mucha importancia a la identificación de las secuencias codificantes de proteínas dejando de lado otros elementos tales como los ARN no codificantes (micro ARNs, ARN de transferencia, ARN ribosomal, entre otros). Por otra parte cuanto más importante sea un genoma para la comunidad científica en general, más recursos se destinan a su anotación; por lo tanto, los genomas humanos, de ratones, *A. thaliana*, *C. elegans* y *D. melanogaster* han sido objeto de proyectos de anotación a gran escala durante muchos años, en los que han participado numerosos institutos científicos y centros de secuenciación [266] y han dado lugar a los llamados “organismos modelo” los cuales en muchos casos no son útiles al momento de abordar un determinado proyecto genoma. Incluso los proyectos de anotación más grandes aún no pueden describir los genomas en su totalidad, y los investigadores a menudo deben producir sus propios modelos de anotación [266]. Es por ello que para realizar la predicción génica de *E. canadensis* G7 fue necesario entrenar los algoritmos de anotación con datos de CDS curados y específicos del género *Echinococcus* extraídos de bases de datos de ESTs y de los proyectos genoma de especies relacionadas. El entrenamiento y la construcción de los modelos de anotación fue crucial para la correcta asignación de la estructura génica del genoma de *E. canadensis* G7. El genoma de *E. canadensis* G7 contiene 11435 genes; todos ellos con evidencia transcripcional y con una densidad génica de 99 genes / Mb; casi tres veces superior al trematodo *S. mansoni* y la turbelaria *S. mediterránea* que poseen ~30 genes / Mb y por debajo la monogenea *G. salaris* con 129 genes / Mb. Las diferencias en la densidad de genes se explica principalmente por la diferencias en la cantidad de ADN no codificante, más que por las variaciones en el número de genes. Por ejemplo, el genoma humano cuyo tamaño es de 30 veces mayor que el de *Echinococcus* posee una densidad génica de 6,5 genes / Mb con apenas 21000 genes [267].

Los métodos más populares para la anotación funcional de genes suelen basarse en la evaluación de la similitud de secuencia con proteínas de función conocida utilizando BLAST [268] o mediante el uso de Blast2GO [196] que proporciona una anotación de ontología génica [197] y funciones basadas en dominios de la base de datos InterPro [269] que incluye información de varias bases de datos como PROSITE [191], PRINTS [192], InterProScan [270] o alternatively, se pueden transferir las anotaciones desde una especie relacionada [271]. En la presente tesis, la combinación

de las metodologías implementadas para la anotación funcional de los genes de *E. canadensis* G7 otorgó una mayor eficiencia en el proceso de anotación génica en comparación con las técnicas usadas para la anotación de los genes de *G. salaris* [233] y *S. mansoni* v1 [40,272] lográndose entre un 5% y 15% más de genes anotados. La integración de los análisis de los grupos de ortología de especies de platelmintos parásitos (cestodos, monogenea y trematodos) a los protocolos de anotación génica fue fundamental para lograr un mayor porcentaje de anotación.

5.2 Análisis de ortología y genómica comparativa de las especies de *Echinococcus*

Las diferencias en las estructuras de los dominios de las proteínas codificadas por genes considerados ortólogos son muy importantes desde el punto de vista funcional, particularmente en los organismos eucariotas multicelulares, en los que la creación de un dominio a menudo conduce a estructuras de dominios específicas de linaje y altamente complejas [273]. La importancia de los análisis de ortología radica en que los genes ortólogos son responsables de funciones equivalentes en diferentes organismos y constituyen la base conceptual de la anotación funcional de los genomas que es esencial para la biología experimental en la era post genómica. Es muy común equiparar la ortología a la equivalencia funcional y por lo tanto buscar ortólogos y extrapolar la función que posee a un gen ortólogo dado de cualquier especie; este enfoque disminuye notablemente con el aumento de la distancia evolutiva entre los genomas comparados, pero es válida para organismos filogenéticamente relacionados [273]. Los grupos de ortología contruidos en esta tesis contribuyeron no solo a la asignación de las anotaciones funcionales de varios genes de *E. canadensis* G7; sino que también pusieron en evidencia la falta de bases de datos y elementos con información biológica que permitan describir las características estructurales y funcionales de los genes de las especies de *Echinococcus*, y también de cestodos y platelmintos en general. Solo el 20% de los genes específicos de *Echinococcus* poseen dominios conocidos mientras que el 80% restante son proteínas hipotéticas con evidencia transcripcional. Por lo tanto; sería sumamente importante llevar a cabo experimentos que permitan describir la estructura o función de las proteínas hipotéticas ya que representan un área aún no explorada en absoluto pero que son, llamativamente conservadas en las tres especies de *Echinococcus* e incluso muchas de ellas también se encuentran conservadas en cestodos y platelmintos. Por otra parte, la mayoría de los genes que si poseen dominios conocidos mostraron un enriquecimiento en los términos “Unión” y “Actividad de factor de transcripción de unión a ácidos nucleicos” sugiriendo que podrían tener una actividad reguladora de la transcripción. A diferencia de la baja conservación en la organización genómica que se ha observado entre *S. mansoni* y *E. multilocularis*; como era de esperar, las tres especies de *Echinococcus* mostraron un alto porcentaje de sintenia evidenciando una conservación en estructura y orden de los genes entre estas especies, permitiendo la realización de un análisis profundo de

genómica comparativa con el fin de evaluar la variabilidad genética entre las tres especies de *Echinococcus*. Los resultados de los análisis de genómica comparativa basados en la identificación de SNPs demostraron que la frecuencia de SNPs intraespecíficos de *E. granulosus* G1 es más de 10 veces mayor que la observada en los genomas de *E. canadensis* G7 y de *E. multilocularis*. Dado que los genomas de *E. granulosus* G1 y *E. canadensis* G7 se obtuvieron a partir de ADN de protoescolices provenientes de un único quiste que tiene origen a partir de un único huevo de *Echinococcus*, y que en cuyo interior los protoescolices se desarrollan de manera asexual (clonal), se esperaba ausencia o escasa variabilidad genómica intra especie [78]. Sin embargo, los resultados obtenidos en esta tesis demostraron que el genoma de *E. granulosus* G1 presenta gran variabilidad intra específica a pesar de las características de su desarrollo y en comparación a lo observado en *E. canadensis* G7 y en *E. multilocularis*. El estudio de la diversidad genética entre especies de *Echinococcus* también mostró resultados inesperados ya que se observó un mayor grado de variabilidad entre *E. canadensis* G7 y *E. granulosus* G1, a pesar de que ambas especies pertenecen al complejo *E. granulosus* s.l. y presentan un fenotipo similar del estadio de metacestodo y a su vez muy diferente al ciclo de vida y fenotipo del metacestodo de *E. multilocularis*. El grado de variabilidad observado entre las tres especies de *Echinococcus* es de 3 a 7 veces superior al que se ha estimado entre los trematodos *S. mansoni* y *S. japonicum* cuya variabilidad fue evaluada también en el estadio larval [274]. Por otra parte a pesar de existir más del doble de SNPs entre *E. canadensis* G7 y *E. granulosus* G1 que entre *E. canadensis* G7 y *E. multilocularis*, la frecuencia de SNPs en las regiones codificantes fue similar para todas las especies, indicando que las regiones intergénicas no sufren una presión de selección tan alta y son propensas a acumular mutaciones mientras que las regiones codificantes conservan la secuencia entre las tres especies de *Echinococcus*. Asimismo, la distribución de SNPs no sinónimos en vías metabólicas KEGGs demuestra que no existe acumulación de SNPs en alguna vía metabólica en particular.

5.3 Relación filogenética entre tres especies de *Echinococcus*

Hasta el momento no hay una clasificación taxonómica certera de las especies de *Echinococcus* [13–15]. En esta tesis los estudios filogenéticos basados en tres métodos diferentes permitieron reconstruir la relación filogenética entre las tres especies de *Echinococcus* cuyos genomas han sido secuenciados completamente. En primer lugar, el análisis filogenético a partir de genomas mitocondriales completos permitió confirmar la monofilia de la especie *E. canadensis*. En segundo lugar, la topología del árbol filogenético construido a partir de genes de copia única de diversos filos de metazoos basado tanto en secuencias de aminoácidos como en secuencia de nucleótidos demostró el estatus de especies diferentes entre *E. canadensis* G7 y *E. granulosus* G1, tópico aún en discusión por parte de algunos investigadores dado que son especies

críticas desde un punto de vista morfológico. Si bien las filogenias basadas en genes de copia única fueron informativas, se observaron algunas diferencias en cuanto a la topología del árbol según se utilizaran las secuencias de nucleótidos o de proteínas, sugiriendo que el análisis basado en estas secuencias no es lo suficientemente robusto. Una posible explicación es que los genes de copia única representan un subgrupo de datos que puede sesgar el resultado debido al escaso número de proteínas elegidas y de esta forma la distancia genética real podría resultar sobre o subestimada. En tercer lugar, los análisis filogenéticos usando los datos de secuenciación genómica y la distribución de SNPs en genomas completos aportaron el nivel de detalle genético requerido para resolver la relación filogenética entre estas tres especies de *Echinococcus*. Estos resultados demostraron que *E. canadensis* G7 y *E. granulosus* G1 son filogenéticamente más distantes uno de otro debido a un alto grado de variabilidad genética que presentan y sugiere que *E. multilocularis* sería la especie ancestral de ambas especies. Esta hipótesis también es apoyada por los resultados obtenidos al analizar el número de sitios polimórficos homocigotas compartidos con el mismo tipo de sustitución nucleotídica. Dado que el mayor número de sitios polimórficos compartidos con el mismo polimorfismo se observó entre los genomas de *E. canadensis* G7 y *E. granulosus* G1 al usar como referencia al genoma de *E. multilocularis*, se reforzó la hipótesis de que *E. multilocularis* sería la especie basal entre estas tres especies y que *E. canadensis* G7 y *E. granulosus* G1 luego del fenómeno de especiación habrían acumulado mutaciones independientes incrementando su diversidad genética. Este resultado es acorde a resultados previos obtenidos por Saarma y col [14] quienes analizaron un subgrupo de secuencias exónicas, intrónicas e intergénicas tomadas al azar y observaron que *E. multilocularis* era la especie basal a *Echinococcus granulosus sensu lato*. Por otra parte, la validación de los SNPs corroboró el grado de confianza asignado a los SNPs identificados y permitió desarrollar nuevos marcadores moleculares para diferenciar entre las especies de *Echinococcus* que infectan a animales y al hombre en Argentina; en particular, los marcadores moleculares nucleares que son escasos en la bibliografía y complementan a los marcadores mitocondriales y ribosomales ampliamente utilizados. Es evidente que la secuenciación y obtención de los genomas es crucial no solo para exponer las diferencias genéticas y características de las distintas especies, sino también para definir con mayor precisión la filogenia y la historia evolutiva de estos parásitos y de muchas otras especies. Por este motivo, es necesaria la secuenciación de genomas de otros cestodos y de un mayor número de especies de *Echinococcus* para comprender con mayor certeza la historia evolutiva de estos organismos.

5.4 Blancos terapéuticos

Los análisis de las secuencias y de los dominios de las proteínas identificadas como potenciales blancos terapéuticos mostraron diferencias en su secuencia entre las distintas especies de *Echinococcus*. Cabe destacar que tanto *E. canadensis* G7 como *E. granulosus* G1 son las especies más frecuentes en los casos de echinococcosis en humanos y son simpátricas en muchos países del mundo incluyendo la Argentina [7].

Debido a ello, es que esta información es relevante y debe ser considerada para el desarrollo de drogas anti parasitarias de manera que las mismas sean efectivas en regiones donde coexisten las diferentes especies. En la actualidad solo existen dos métodos para el tratamiento de la echinococcosis humana: la extirpación quirúrgica del quiste y el uso de benzimidazoles. Sin embargo ambas técnicas presentan resultados poco satisfactorios [275] y es por ello que la WHO recomienda el desarrollo de nuevos fármacos (<http://www.who.int/mediacentre/factsheets/fs377/en/>). En este estudio se identificaron nuevos blancos terapéuticos contra la echinococcosis, que podrán ser evaluados en mediano o corto plazo. Dichos candidatos se encuentran altamente conservados en cestodos pero ausentes o con un alto grado de divergencia en humanos, lo cual favorecería la especificidad y por lo tanto se reducirían los efectos adversos que generan las drogas utilizadas en la actualidad. En particular, los péptidos antimicrobianos codificados en el genoma de *E. canadensis* G7 son ortólogos a péptidos que son secretados por parásitos helmintos y poseen un rol determinante en la modulación del sistema inmune del hospedador. Algunos estudios llevados a cabo con *F. hepática* han descripto la secreción de un péptido, que exhibe características bioquímicas y funcionales que son similares a la de las catelicidina LL-37 de humanos y modula la respuesta celular innata simulando la función de los péptidos antimicrobianos del hospedador [276–279]. La conservación de la arquitectura de los dominios de los péptidos de *Echinococcus* sugiere que se trata de péptidos con actividad antimicrobiana y podrían cumplir un rol inmunoregulatorio como el propuesto por Zheng y col. [78] para *F. hepática*. Por otra parte, estos péptidos también podrían estar implicados en la interacción parásito-hospedero cumpliendo un rol importante como mecanismo de evasión. Otros blancos propuestos en esta tesis están relacionados con procesos de señalización, transcripción, transporte y el metabolismo de glicanos entre los cuales este último podría ser particularmente importante, ya que estaría asociado al desarrollo del metacestodo mediante la formación de la capa laminar cuya estructura se encuentra enriquecida en glicanos. La capa laminar es una característica específica del género *Echinococcus*, y su rol se asocia a la supervivencia del parásito en el hospedero por largos períodos de tiempo (años) sin despertar la respuesta inmune, característica peculiar de la echinococcosis humana. La modulación de la interfaz parásito-hospedero depende de la integridad del quiste hidatídico que estaría garantizada por la capa laminar, protegiéndolo del sistema inmune del hospedador al ofrecer una barrera inmunológicamente inerte [1]. Por este motivo, el diseño de blancos terapéuticos o métodos que inhiban la actividad de las enzimas que intervienen en la síntesis de glicanos y particularmente la que es específica de *Echinococcus* podría ser un método eficaz para impedir el desarrollo del quiste hidatídico. Como producto de esta tesis se cuenta con una lista priorizada de genes específicos de cestodos que será necesario evaluar tanto *in silico* como en experimentos húmedos para corroborar su utilidad como blancos de drogas permitiendo así su validación.

5.5 Regulación de la expresión génica

En esta tesis se abordó por primera vez el estudio de la distribución de las CGIs en los genomas de *Echinococcus*. Las CGIs son consideradas marcadores genéticos debido a su alta conservación y especificidad de especie, por lo que se espera que exista una correlación positiva con la densidad génica la cual fue observada en los genomas de *Echinococcus* exhibiendo un mayor número y mayor densidad de CGIs en las regiones codificantes respecto a las regiones intergénicas. Un patrón similar es observado en los genomas de diversas especies de mamíferos sugiriendo que es un fenómeno extendido para especies eucarióticas [259,280]. No obstante, a pesar de que el porcentaje de GC total de los genomas es similar (~41%), los genomas de *Echinococcus* mostraron una mayor densidad de CGIs que los genomas de mamíferos, lo cual concuerda con los bajos niveles de metilación que se han observado en algunos genomas de platelmintos [83] y organismos modelo no mamíferos, tales como gusanos redondos y moscas [281] ya que, como se ha informado en muchos estudios; las CGIs tienden a desaparecer a lo largo de la evolución por un mecanismo de metilación *de novo* de los dinucleótidos CpG, que subsecuentemente mutan a TpGs o CpAs debido a una alta tasa de transición dependiente de la metilación del ADN [280,282–285]. Asimismo, la mayor densidad de CGIs y menor grado de metilación observados en *Echinococcus* también concuerda con el menor número de miembros de las familias de DNMTs y MBDs presentes en las especies de cestodos. En otras palabras, dado que los proteomas de mamíferos contienen muchos miembros de las familias de proteínas involucradas en la metilación del ADN, es esperable encontrar una menor densidad de CGIs en comparación con los genomas de las especies de *Echinococcus*. Por el contrario, el alto grado de la densidad de CGIs ~37 CGIs/Mb en *Echinococcus* puede ser explicado por la ausencia de otros miembros de las familias de MBDs y DNMTs. Al no poseer varios miembros de enzimas que metilan el ADN, éste se encuentra poco metilado por lo cual la tasa de transición C por T será menor dando lugar a la conservación de las CGIs. La baja tasa de metilación del ADN también ha sido observada en otros miembros del filo platelmintos tales como *Protopolystoma xenopodi*, *Polycelis nigra* y *Macrostomum lignano* [83,286]. Dado que *Echinococcus* posee un número reducido de las proteínas de la familia de las DNMTs y una nueva proteína MBD, que es altamente divergente respecto a las MBDs de mamíferos, lo convierte en un modelo muy interesante para el estudio de la evolución del sistema de metilación y regulación epigenética. Asimismo, debido a que se ha descrito que la mayoría de los promotores de los genes residen en las CGIs [280], se utilizó el patrón de distribución de las CGIs para identificar estas regiones regulatorias en *Echinococcus*. Mediante este estudio se evidenció que las CGIs se localizan en un rango de distancias característico en los genomas de *Echinococcus* y que la mayoría de ellas se sitúan en aproximadamente 2 kb río arriba respecto del codón de inicio ATG de los genes sugiriendo que podrían albergar regiones con motivos regulatorios de la expresión génica. Además la distinta distribución de las CGIs río arriba de los genes ortólogos de las tres especies de *Echinococcus* sugiere que los genes que poseen la misma función en

las diferentes especies estarían regulados diferencialmente por la presencia/ausencia de este tipo de estructuras genómicas. La gran diversidad de promotores eucarióticos ha sido uno de los principales obstáculos para la caracterización de regiones promotoras en platelmintos las cuales solo han sido caracterizadas en algunos estudios particulares tales como la de los genes que intervienen en el sistema glutatión-tioredoxina en cestodos [287]. Asimismo, el patrón de distribución de CGIs puede utilizarse para identificar regiones promotoras ya sea experimentalmente o *in silico*. Por lo tanto, los análisis realizados en esta tesis resultan en un gran avance para entender los mecanismos moleculares de la transcripción y provee información muy valiosa para el estudio y la manipulación génica de estos organismos.

Otro mecanismo regulatorio de la expresión génica se encuentra asociado a la expresión de pequeños ARN capaces de llevar a cabo el silenciamiento postranscripcional de los genes. Previamente, en *E. canadensis* G7 hemos descripto por primera vez la expresión diferencial de microARNs durante su desarrollo [122]. El mecanismo de acción de estos microARNs se encuentra íntimamente relacionado a las proteínas Ago [115,288]. Es por ello que en esta tesis se analizaron exhaustivamente todas las proteínas de la familia Ago de *Echinococcus* confirmando la existencia de un nuevo clado de proteínas Ago y su posible interacción con los microARNs. La conservación de la estructura, de los dominios y de varios residuos específicos de las proteínas Ago del nuevo clado sugiere que serían capaces de unir ARNs pequeños y que dichos residuos participarían en la estabilización del ARNm blanco. Además la presencia de los residuos que conforman el motivo DEDH podría otorgarle la actividad de clivaje necesaria para la degradación del ARNm. Por otra parte, los cambios específicos de los residuos que conforman el motivo de la proteína que une a la región de siembra *Seed* del microARN, y su conservación en las proteínas del nuevo clado sugieren que podría existir una especificidad distintiva de unión a ARNs pequeños. En conjunto estas características que podría significar un mecanismo novedoso de la regulación de la expresión génica. Más aun, dado que las especies de *Echinococcus* carecen de las clásicas proteínas Piwi [43] y aparentemente tampoco expresan moléculas piARNs [122], estas proteínas serían particularmente importantes para estudiar pequeños ARN reguladores aún desconocidos e identificar las moléculas que interactúan con las nuevas proteínas Ago identificadas en esta tesis.

5.6 Uso de codones en *Echinococcus*

Los estudios sobre los patrones en el uso del codones han contribuido a comprender las características básicas de la organización molecular de los genomas, la expresión génica heteróloga [134], la predicción de los niveles de expresión génica [289], la función génica [290] y la posición de los genes en los cromosomas [291], y también han revelado información sobre la evolución molecular de genes individuales en muchos organismos [292]. Sin embargo, la mayoría de los estudios sobre patrones del uso de codones se centran en algunos organismos modelo o patógenos tales como *Caenorhabditis* sp, *Drosophila* sp, *Arabidopsis* sp [143],

Saccharomyces sp [293], *Giardia lamblia* [294] y *Borrelia burgdorferi* [295] y recientemente se ha estudiado este fenómeno en los cestodos *T. solium* [296], *T. multiceps* [297] y *T. saginata* [298] y el trematodo *S. japonicum* [299]. Aunque los mecanismos de traducción de las proteínas permanecen altamente conservados a través de los taxones de los organismos, existe una gran variación a nivel del reconocimiento entre el codón y el anti codón, tanto dentro como entre las especies. Esta variación se refleja en el uso de codones, el número de copias de los ARNt y la redundancia y la degeneración de los anti codones [141]. El sesgo en uso de codones es una característica generalizada en los genomas de muchos organismos que está profundamente influenciado por fenómenos evolutivos, y resulta básicamente del equilibrio entre el sesgo mutacional y la selección natural. Dentro de la selección natural la selección traduccional es particularmente importante y está asociada a la eficiencia y a la precisión de la traducción, la cual se explica principalmente en base a los genes de ARNt codificados en los genomas donde la coevolución entre el uso de codones y la abundancia de los ARNt isoaceptores desempeña un rol decisivo [130,131]. La extensión del sesgo en uso de codones tiende a ser específica en grupos individuales de genes dentro de un genoma y está determinada por múltiples factores. En esta tesis se determinó el grado de contribución que posee el sesgo mutacional y la presión de selección natural en los genes de tres especies de *Echinococcus*, partiendo de una suposición general de que la mutación, la selección y la deriva aleatoria representan las tres fuerzas principales involucradas que dan lugar al sesgo en el uso del codones [149–154]. Para que estos estudios sean inequívocos es necesario contar con un grupo de genes curados que sean representativos de lo que ocurre en la naturaleza de la composición molecular de los organismos. Por lo tanto, para este estudio se seleccionaron estrictamente genes con ortólogos en las tres especies de *Echinococcus* y cuya estructura génica se encuentre bien definida. La composición de nucleótidos se considera uno de los factores más importantes que determina el uso de codones de los genes, con un contenido de GC que refleja la tendencia general de la mutación de los codones en relación a su contexto genómico [297]. Los cambios de base en la primera y en la segunda posición de los codones provocan cambios en la secuencia de aminoácidos, mientras que la tercera posición del codón rara vez induce dicha variación de secuencia. Se reconoce generalmente que la tercera posición del codón está sujeta a una presión de selección inferior en comparación con la de la primera y segunda posición de los codones. Por lo tanto, el análisis de correlación de ENc-GC3s y el estudio de la composición de nucleótidos son de vital importancia para la elucidación de los patrones del sesgo del uso de codones en muchos organismos [297,300]. Estos análisis llevados a cabo sobre los genes de *Echinococcus* indicaron que existe un ligero sesgo del uso de codones con un valor promedio de ENc igual a 57. Además se observó que hay diferencias significativas en el sesgo del uso de codones entre los genes de *Echinococcus*. Principalmente los genes con mayor contenido de GC tanto total como en la tercera posición de los codones poseen un mayor sesgo del uso de codones, sin embargo en genes con valores altos de ENc, el contenido nucleotídico es variable indicando que la composición nucleotídica tiene un efecto importante sobre la elección del uso de codones solamente en algunos genes. El análisis de correlación entre ENc y

GC3s demostró que el sesgo mutacional desempeña un papel menor en la elección del uso de los codones. Los análisis de ENc proporcionan un método para cuantificar el sesgo en el uso de codones sinónimos; sin embargo, este análisis por sí solo no es suficiente para determinar las contribuciones exactas que ejercen la selección natural y la presión mutacional direccional [300,301]. Por este motivo se realizaron análisis de neutralidad y se aplicó la regla de paridad 2 de Chargaff para proporcionar información más precisa sobre los factores que influyen en el sesgo del uso de codones. Según estos análisis, la presión de selección es la fuerza evolutiva principal que moldea el uso de codones en las tres especies de *Echinococcus* con una contribución de casi el 80% mientras que el restante 20% es atribuible al sesgo mutacional. Por lo tanto, la presión de selección traduccional o la contribución de una presión de selección externa o una combinación de ambas podrían estar ejerciendo efectos significativos sobre el sesgo del uso de codones en estas especies de parásitos. A diferencia de lo encontrado en *Echinococcus*, los estudios realizados en mamíferos, levaduras y moscas [154,219] han mostrado un rango más amplio de los valores de ENc y GC3s con una distribución de genes cuyo sesgo en el uso de codones depende claramente del sesgo composicional de nucleótidos indicando un mayor nivel de sesgo en los genes con porcentajes de CG3 mayores [219,228,302]. En efecto, se cuenta con muy poca evidencia de que la variación del sesgo en el uso de codones haya sido formada por la selección en favor de la eficiencia de la traducción [303]. Aunque en muchas especies la selección natural asociada a favorecer la precisión y eficiencia de la traducción influye considerablemente a moldear el uso de codones, no siempre es tan evidente la forma en la que se está llevando a cabo la selección del uso de codones y por lo tanto no es posible explicar la variación total del uso de codones. Por eso mismo es común observar que algunos genes poseen un uso de codones que está determinado principalmente por mutación y deriva, mientras que otros exhiben un uso de codones que surge del equilibrio entre el sesgo mutacional y las presión de selección [161,304,305].

Por otra parte, el patrón del sesgo del uso de codones parece estar conservado en platelmintos. Los análisis del uso de codones realizado en los cestodos *T. saginata* [298], *T. multiceps* [297] y *T. solium* [296] y el trematodo *S. japonicum* [299] han mostrado resultados similares a los de *Echinococcus* en cuanto a los valores medios de ENc (~57) y a la distribución de los genes en las representaciones ENc-GC3. Todos ellos exhiben bajos niveles en el sesgo del uso de codones utilizando casi aleatoriamente todos los codones sinónimos. Sin embargo el grado con el que la presión de selección contribuye a moldear el uso de codones es diferente en algunos de estos organismos, siendo mucho menor en *S. japonicum* (~20%) y en *T. solium* (~50%). Aunque este patrón pueda significar que los codones están poco optimizados para la mayoría de los genes de platelmintos, este fenómeno podría otorgarles ventajas para usar casi cualquier ARNt isoceptor en algún momento determinado dependiendo de la disponibilidad y abundancia de los ARNt y podría tratarse de una ventaja adaptativa en el ambiente del hospedador en el que se desarrollan.

No obstante; estudios realizados en *Drosophila* sp y *Caenorhabditis* sp han demostrado que los genes altamente expresados poseen mayor sesgo en el uso de codones explicando que esto se debe al uso preferencial de codones sinónimos óptimos

en genes fundamentales para asegurar una mayor efectividad y precisión de la traducción de dichos genes [143,306]. Este mismo fenómeno podría estar sucediendo en un número importante de genes de *Echinococcus* (al menos un 5% de ellos), dado que se observó una correlación negativa significativa entre ENc y CAI y entre ENc y Fop y una correlación positiva entre CAI y Fop y que a su vez correlaciona con un mayor contenido de GC en la tercera posición de los codones. Esto sugiere que los genes con mayor nivel de expresión tienen un mayor sesgo en el uso de codones y que estarían optimizados para su traducción. Asimismo, dado que estos genes presentan un sesgo composicional de nucleótidos; un fenómeno de coevolución o de balance entre el sesgo mutacional y la presión de selección podría estar determinando el patrón de uso de codones de estos genes.

Los análisis de componentes principales permitieron separar a lo largo sus ejes a los genes con las características más diferentes, principalmente basados en su contenido de GC total y por posición y según el nivel de expresión estimado, mientras que otros factores tales como el tamaño del gen, la aromaticidad y la hidrofobicidad mostraron no tener una incidencia notable sobre el uso de codones. Diversos estudios en distintos organismos han determinado los factores que pueden influir en el sesgo del uso de codones; estos incluyen: conversión génica sesgada [307,308]; expresión génica sesgada por sexo [309]; estructura del ARNm [226,310], las tasas de recombinación, la estabilidad del ARN [154,310–313], la estructura de la proteína [314]; tamaño de la población [315]; edad evolutiva de los genes [316]; estrés ambiental [317] y disponibilidad de nitrógeno [318]. Se cree además que el tamaño de la población pueden influir en el sesgo del uso de codones entre las distintas especies y entre miembros de una misma especie [130,306,319,320]. Es evidente que el sesgo en uso de codones es un fenómeno complejo que involucra muchos factores dando como resultado un patrón característico de cada grupo de organismos o de cada especie. Muchos de estos factores son difíciles de abordar en *Echinococcus* dada la complejidad de su ciclo de vida. Sin embargo, los análisis llevados a cabo en esta tesis revelaron por primera vez el patrón del uso de codones en tres especies de *Echinococcus* y su conservación en platelmintos, lo cual proporciona la base para investigaciones adicionales que contribuyan a la identificación de nuevos genes, como así también a la ingeniería genética molecular y estudios evolutivos en este filo y más aún a la relación parásito-hospedero y determinar qué factores del hospedador son relevantes para moldear el sesgo en el uso de codones.

5.7 Conclusión general

La disponibilidad de un nuevo genoma es fundamental para estudiar la biología de un organismo patógeno, y más aún cuando éste presenta limitaciones para su mantenimiento y manipulación en el laboratorio. Por estas razones, la obtención del genoma de *E. canadensis* G7 resulta de gran importancia para el estudio y comprensión de su biología. Con el desarrollo de las nuevas tecnologías de secuenciación, el volumen de datos de secuencias genómicas ha aumentado considerablemente y su análisis representa un gran desafío, requiriendo el análisis y manipulación de datos biológicos u ómicos a gran escala, lo que comprende la evaluación de la calidad de los datos, el análisis integral, la interpretación de los resultados y la presentación de los mismos, y constituye en muchos casos el cuello de botella para la eficiente extracción de información biológica relevante. Los datos generados en esta tesis y las metodologías implementadas no solo permitieron el estudio de uno de los máximos responsables de la echinococcosis, sino que también contribuyeron al estudio de otros genomas de platelmintos, tales como los análisis de expresión y de genes blanco de los microARNs de *Echinococcus* spp y *Taenias* spp y la anotación génica de *M. corti* los cuales fueron publicados en revistas internacionales. También a nivel internacional se colaboró en la construcción de la base de datos WormBaseParaSite, especializada en parásitos platelmintos y nematodos. A su vez, esta tesis fue el pilar fundamental para la construcción del nodo bioinformático IMPaM permitiendo la construcción local de bases de datos espejo de las principales bases de datos internacionales de secuencias de nucleótidos, proteínas, dominios proteicos; como así también la implementación de herramientas para el análisis específico de genomas de parásitos platelmintos. Estas plataformas han representado grandes avances para el estudio de este tipo de organismos y han acompañado el desarrollo de técnicas llevadas a cabo en nuestro laboratorio con el propósito de estudiar la biología de estos parásitos y desarrollar herramientas que permitan el control de estas enfermedades infecciosas.

Referencias

1. Thompson RCA. Biology and Systematics of *Echinococcus*. Adv. Parasitol 95: 65-109. Elsevier Ltd; 2017.
2. Thompson R, McManus DP. Aetiology: parasites and life-cycles. In: Eckert J, MA G, Meslin F, Pablowski Z, editors. WHO/OIE Man. Echinococcosis humans Anim. a public Heal. Probl. Glob. concern. Paris; 2001.
3. Zhang W, Wen H, Li J, Lin R, McManus DP. Immunology and immunodiagnosis of cystic echinococcosis: An update. Clin. Dev. Immunol. Hindawi; 2012; 2012:165-71.
4. Díaz A, Casaravilla C, Allen JE, Sim RB, Ferreira AM. Understanding the laminated layer of larval *Echinococcus* II: Immunology. Trends Parasitol. 2011; 27:264-73.
5. Galindo M, Schadebrodt G, Galanti N. *Echinococcus granulosus*: Cellular territories and morphological regions in mature protoscoleces. Exp. Parasitol. 2008; 119:524-33.
6. Morgan UM, Xiao L, Fayer R, Lal AA, Thompson RCA. Variation in *Cryptosporidium*: Towards a taxonomic revision of the genus. Int. J. Parasitol. 1999; 29:1733-51.
7. Cucher MA, Macchiaroli N, Baldi G, Camicia F, Prada L, Maldonado L, *et al.* Cystic echinococcosis in South America: systematic review of species and genotypes of *Echinococcus granulosus sensu lato* in humans and natural domestic hosts. Medline. 2016; 21:166-75.
8. Bowles J, Blair D, McManus DP. Genetic variants within the genus *Echinococcus* identified by mitochondrial DNA sequencing. Mol. Biochem. Parasitol. 1992; 54:165-73.
9. Bowles J, McManus DP. NADH dehydrogenase 1 gene sequences compared for species and strains of the genus *Echinococcus*. Int. J. Parasitol. 1993; 23:969-72.
10. Thompson RCA. The taxonomy, phylogeny and transmission of *Echinococcus*. Exp. Parasitol. 2008. p. 439-46.
11. Nakao M, Yanagida T, Konyaev S, Lavikainen A, Odnokurtsev VA, Zaikov VA, *et al.* Mitochondrial phylogeny of the genus *Echinococcus* (Cestoda: Taeniidae) with emphasis on relationships among *Echinococcus canadensis* genotypes. Parasitology. 2013; 140:1625-36.
12. Xiao N, Qiu J, Nakao M, Li T, Yang W, Chen X, *et al.* *Echinococcus shiquicus* n. sp., a taeniid cestode from Tibetan fox and plateau pika in China. Int. J. Parasitol. 2005; 35:693-701.
13. Nakao M, Lavikainen A, Yanagida T, Ito A. Phylogenetic systematics of the genus *Echinococcus* (Cestoda: Taeniidae). Int. J. Parasitol. Australian Society for Parasitology Inc.; 2013; 43:1017-29.
14. Saarma U, Jogisalu I, Moks E, Varcasia A, Lavikainen A, Oksanen A, *et al.* A novel phylogeny for the genus *Echinococcus*, based on nuclear data, challenges relationships based on mitochondrial evidence. Parasitology. 2009; 136:317.
15. Yanagida T, Lavikainen A, Hoberg EP, Konyaev S, Ito A, Sato MO, *et al.* Specific status of *Echinococcus canadensis* (Cestoda: Taeniidae) inferred from nuclear and mitochondrial gene sequences. Int. J. Parasitol. 2017.
16. Jenkins DJ, Lievaart JJ, Boufana B, Lett WS, Bradshaw H, Armua-Fernandez MT. *Echinococcus granulosus* and other intestinal helminths: Current status of

- prevalence and management in rural dogs of eastern Australia. *Aust. Vet. J.* 2014; 92:292-8.
17. McManus DP. Current status of the genetics and molecular taxonomy of *Echinococcus* species. *Parasitology.* 2013; 140:1617-23.
 18. McManus DP, Thompson RCA. Molecular epidemiology of cystic echinococcosis. *Parasitology.* 2003; 127:S37-51.
 19. Budke CCM, Deplazes P, Torgerson PPR. Global socioeconomic impact of cystic echinococcosis. *Emerg Infect Dis.* 2006; 12:296-303.
 20. Torgerson PR. The emergence of echinococcosis in central Asia. *Parasitology.* 2013; 140:1667-73.
 21. Lorenzo C, Ferreira HB, Monteiro KM, Rosenzvit M, Kamenetzky L, Garcı HH, *et al.* Comparative Analysis of the Diagnostic Performance of Six Major *Echinococcus granulosus* Antigens Assessed in a Double-Blind , Randomized Multicenter Study. *J. Clin. Microbiol.* 2005; 43:2764-70.
 22. Moro P, Schantz PM. Echinococcosis: a review. *Int. J. Infect. Dis.* 2009; 13:125-33.
 23. Avila HG, Santos GB, Cucher MA, Macchiaroli N, Pérez MG, Baldi G, *et al.* Implementation of new tools in molecular epidemiology studies of *Echinococcus granulosus sensu lato* in South America. *Parasitol. Int.* 2017; 66:250-7.
 24. Arrabal JP, Avila HG, Rivero MR, Camicia F, Salas MM, Costa SA, *et al.* *Echinococcus oligarthrus* in the subtropical region of Argentina: First integration of morphological and molecular analyses determines two distinct populations. *Vet. Parasitol.* 2017; 240:60-7.
 25. Alvarez Rojas CA, Romig T, Lightowlers MW. *Echinococcus granulosus sensu lato* genotypes infecting humans - review of current knowledge. *Int. J. Parasitol.* 2014; 44:9-18.
 26. Debiaggi MF, Soriano SV, Pierangeli NB, Lazzarini LE, Pianciola LA, Mazzeo ML, *et al.* Genetic characterization of human hydatid cysts shows coinfection by *Echinococcus canadensis* G7 and *Echinococcus granulosus sensu stricto* G1 in Argentina. *Parasitol. Res. Springer Berlin Heidelberg*; 2017; 116:2599-604.
 27. Kamenetzky L, Gutierrez AM, Canova SG, Haag KL, Guarnera E a, Parra A, *et al.* Several strains of *Echinococcus granulosus* infect livestock and humans in Argentina. *Infect. Genet. Evol.* 2002; 2:129-36.
 28. Schneider R, Gollackner B, Schindl M, Tucek G, Auer H. *Echinococcus canadensis* G7 (pig strain): An underestimated cause of Cystic Echinococcosis in Austria. *Am. J. Trop. Med. Hyg.* 2010; 82:871-4.
 29. Dybicz M, Gierczak A, Dabrowska J, Rdzanek Ł, Michałowicz B. Molecular diagnosis of cystic echinococcosis in humans from central Poland. *Parasitol. Int.* 2013; 62:364-7.
 30. Cucher M, Mourglia-Ettlin G, Prada L, Costa H, Kamenetzky L, Poncini C, *et al.* *Echinococcus granulosus* pig strain (G7 genotype) protoscoleces did not develop secondary hydatid cysts in mice. *Vet. Parasitol.* 2013; 193:185-92.
 31. Eckert J, Thompson RCA, Lymbery AJ, Pawlowski ZS, Gottstein B, Morgan UM. Further evidence for the occurrence of a distinct strain of *Echinococcus granulosus* in European pigs. *Parasitol. Res.* 1993; 79:42-8.
 32. Rosenzvit MC, Canova SG, Kamenetzky L, Guarnera EA. *Echinococcus granulosus*: intraspecific genetic variation assessed by a DNA repetitive element. *Parasitology.* 2001; 123:381-8.

33. Kamenetzky L, Muzulin PM, Gutierrez AM, Angel SO, Zaha A, Guarnera EA, *et al.* High polymorphism in genes encoding antigen B from human infecting strains of *Echinococcus granulosus*. *Parasitology*. 2005; 131:805.
34. Muzulin PM, Kamenetzky L, Gutierrez AM, Guarnera EA, Rosenzvit MC. *Echinococcus granulosus* antigen B gene family: Further studies of strain polymorphism at the genomic and transcriptional levels. *Exp. Parasitol.* 2008; 118:156-64.
35. Chow C, Gauci CG, Vural G, Jenkins DJ, Heath DD, Rosenzvit MC, *et al.* *Echinococcus granulosus*: variability of the host-protective EG95 vaccine antigen in G6 and G7 genotypic variants. *Exp. Parasitol.* 2008; 119:499-505.
36. Alvarez Rojas CA, Gauci CG, Lightowlers MW. Antigenic differences between the EG95-related proteins from *Echinococcus granulosus* G1 and G6 genotypes: Implications for vaccination. *Parasite Immunol.* 2013; 35:99-102.
37. Maule AG, Marks NJ. Parasitic flatworms molecular biology, biochemistry, immunology and physiology. 2006.
38. Rohde K. Platyhelminthes. *Encycl. Life Sci.* 2001; 1-6.
39. Zhang Z, Wilson DE, Reeder DM, Wen T, Amrine JW, Beron P, *et al.* Animal biodiversity : An outline of higher-level classification and survey of taxonomic richness. *Zootaxa*. Magnolia Press; 2011.
40. Berriman M, Haas BJ, LoVerde PT, Wilson RA, Dillon GP, Cerqueira GC, *et al.* The genome of the blood fluke *Schistosoma mansoni*. *Nature*. NIH Public Access; 2009; 460:352-8.
41. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* Nature Publishing Group; 2016; 17:333-51.
42. Zerlotini A, Aguiar ERGR, Yu F, Xu H, Li Y, Young ND, *et al.* SchistoDB: An updated genome resource for the three key *schistosomes* of humans. *Nucleic Acids Res.* 2013; 41:D728-31.
43. Tsai IJ, Zarowiecki M, Holroyd N, Garciarrubio A, Sanchez-Flores A, Brooks KL, *et al.* The genomes of four tapeworm species reveal adaptations to parasitism. *Nature*. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2013; 496:57-63.
44. Spakulova M, Orosova M, Mackiewicz JS, Špakulová M, Orosova M, Mackiewicz JS. Cytogenetics and chromosomes of tapeworms (Platyhelminthes, Cestoda). *Adv. Parasitol.* 2011.
45. Ellis JT, Morrison DA. *Schistosoma mansoni*: patterns of codon usage and bias. *Parasitology*. 1995; 110:Pt 1:53-60.
46. Musto H, Romero H, Rodríguez-Maseda H. Heterogeneity in codon usage in the flatworm *Schistosoma mansoni*. *J. Mol. Evol.* 1998; 46:159-67.
47. Fernández V, Zavala A, Musto H. Evidence for translational selection in codon usage in *Echinococcus* spp. *Parasitology*. 2001; 123:203-9.
48. Simpson AJG, Sher A, McCutchan TF. The genome of *Schistosoma mansoni*: Isolation of DNA, its size, bases and repetitive sequences. *Mol. Biochem. Parasitol.* 1982; 6:125-37.
49. Cox GS, Phares CK, Schmidt RA. Molecular characterization of the *Spirometra mansonioides* genome: Renaturation kinetics, methylation, and hybridization to human cDNA probes. *BBA - Gene Struct. Expr.* 1990; 1049:134-44.
50. Marx KA, Bizzaro JW, Blake RD, Hsien Tsai M, Feng Tao L. Experimental DNA melting behaviour of the three major *Schistosoma* species. *Mol. Biochem. Parasitol.* 2000; 107:303-7.

51. Brindley PJ, Laha T, McManus DP, Loukas A. Mobile genetic elements colonizing the genomes of metazoan parasites. *Trends Parasitol.* 2003; 19:79-87.
52. DeMarco R, Kowaltowski AT, Machado AA, Soares MB, Gargioni C, Kawano T. Saci-1, -2, and -3 and Perere, four novel retrotransposons with high transcriptional activities from the human parasite *Schistosoma mansoni*. *J. Virol.* 2004; 78:2967-78.
53. Fernández C, Gregory WF, Loke P, Maizels RM. Full-length-enriched cDNA libraries from *Echinococcus granulosus* contain separate populations of oligo-capped and trans-spliced transcripts and a high level of predicted signal peptide sequences. *Mol. Biochem. Parasitol.* 2002; 122:171-80.
54. Barsiené J V. The chromosome sets of trematodes. *Parazitologija.* 1993; 27:336-53.
55. Petkeviciute R. A chromosome study in the progenetic cestode *Cyathocephalus truncatus* (Cestoda: Spathebothriidea). *Int. J. Parasitol.* 1996; 26:1211-6.
56. Sharbel TF, Beukeboom LW, Pijnacker LP. Multiple supernumerary chromosomes in the pseudogamous parthenogenetic flatworm *Polycelis nigra*: lineage markers or remnants of genetic leakage? *Genome.* 1997; 40:850-6.
57. Beukeboom LW, Sharbel TF, Michiels NK. Reproductive modes, ploidy distribution, and supernumerary chromosome frequencies of the flatworm *Polycelis nigra* (Platyhelminthes: Tricladida). *Hydrobiologia.* 1998; 383:277-85.
58. Roca JR, Ribas M, Baguñà J. Distribution, ecology, mode of reproduction and karyology of freshwater planarians (Platyhelminthes; Turbellaria; Tricladida) in the springs of the central Pyrenees. *Ecography (Cop.). Blackwell Publishing Ltd;* 1992; 15:373-84.
59. Sluys R, Ribas M, Baguna J. Taxonomy, Ecology, and Karyology of a New Species of *Phagocata* from Spain, with a Discussion on the Phylogenetic Systematics of the Genus *Phagocata* Sl (Platyhelminthes, Tricladida, Paludicola). *Can. J. Zool. Can. Zool.* 1995; 73:557-68.
60. Zadesenets KS, Schärer L, Rubtsov NB. New insights into the karyotype evolution of the free-living flatworm *Macrostomum lignano* (Platyhelminthes, Turbellaria). *Sci. Rep. Nature Publishing Group;* 2017; 7:6066.
61. Primrose SB, Twyman RM. 1 - Principles of Genome Analysis and Genomics. Blackwell Pub; 2003; 3rd Ed.
62. Hardison RC. Comparative genomics. *PLoS Biol.* 2003; 1:e58.
63. Fang G, Bhardwaj N, Robilotto R, Gerstein MB. Getting started in gene orthology and functional analysis. *PLoS Comput. Biol. Public Library of Science;* 2010; 6:e1000703.
64. Tatusov RL. A Genomic Perspective on Protein Families. *Science (80-.).* 1997; 278:631-7.
65. Fleischmann R, Adams M, White O, Clayton R, Kirkness E, Kerlavage A, *et al.* Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science (80-.).* 1995; 269:496-512.
66. Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, *et al.* Life with 6000 Genes. *Science (80-.).* 1996; 274:546-67.
67. Edwards YJK. Bioinformatics and Functional Genomics. Briefings Funct. Genomics Proteomics. Wiley-Liss, Inc; 2004.
68. Bult CJ, White O, Olsen GJ, Zhou L, Fleischmann RD, Sutton GG, *et al.* Complete Genome Sequence of the Methanogenic Archaeon, *Methanococcus jannaschii*. *Science (80-.).* 1996; 273:1058-73.

69. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, *et al.* Initial sequencing and analysis of the human genome. *Nature*. 2001; 409:860-921.
70. Iacobuzio-Donahue CA, Ashfaq R, Maitra A, Adsay NV, Shen-Ong GL, Berg K, *et al.* Highly Expressed Genes in Pancreatic Ductal Adenocarcinomas: A Comprehensive Characterization and Comparison of the Transcription Profiles Obtained from Three Major Technologies. *Cancer Res*. 2003; 63:8614-22.
71. Eichinger L, Pachebat JA, Glöckner G, Rajandream M-A, Sugang R, Berriman M, *et al.* The genome of the social amoeba *Dictyostelium discoideum*. *Nature*. 2005; 435:43-57.
72. Gibbs RA, Weinstock GM, Metzker ML, Muzny DM, Sodergren EJ, Scherer S, *et al.* Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature*. 2004; 428:493-521.
73. Chinwalla AT, Cook LL, Delehaunty KD, Fewell GA, Fulton LA, Fulton RS, *et al.* Initial sequencing and comparative analysis of the mouse genome. *Nature*. 2002; 420:520-62.
74. Wood V, Gwilliam R, Rajandream M-A, Lyne M, Lyne R, Stewart A, *et al.* The genome sequence of *Schizosaccharomyces pombe*. *Nature*. 2002; 415:871-80.
75. Rubin GM. Comparative Genomics of the Eukaryotes. *Science* (80-.). 2000; 287:2204-15.
76. Feuk L, Carson AR, Scherer SW. Structural variation in the human genome. *Nat. Rev. Genet*. 2006; 7:85-97.
77. Worthington J, Eyre S. Principles of genetic epidemiology. *Rheumatol*. Sixth Ed. 2014; 1-2:80-5.
78. Zheng H, Zhang W, Zhang L, Zhang Z, Li J, Lu G, *et al.* The genome of the hydatid tapeworm *Echinococcus granulosus*. *Nat. Genet*. Nature Publishing Group; 2013; 45:1168-75.
79. Parkinson J, Wasmuth JD, Salinas G, Bizarro C V., Sanford C, Berriman M, *et al.* A Transcriptomic Analysis of *Echinococcus granulosus* Larval Stages: Implications for Parasite Biology and Host Adaptation. *PLoS Negl. Trop. Dis*. 2012; 6:e1897.
80. Monteiro KM, de Carvalho MO, Zaha A, Ferreira HB. Proteomic analysis of the *Echinococcus granulosus* metacestode during infection of its intermediate host. *Proteomics*. 2010; 10:1985-99.
81. Cui SJ, Xu LL, Zhang T, Xu M, Yao J, Fang CY, *et al.* Proteomic characterization of larval and adult developmental stages in *Echinococcus granulosus* reveals novel insight into host-parasite interactions. *J. Proteomics*. 2013; 84:158-75.
82. Brehm K, Jensen K, Frosch M. mRNA trans-splicing in the human parasitic cestode *Echinococcus multilocularis*. *J. Biol. Chem*. 2000; 275:38311-8.
83. Geyer KK, Chalmers IW, MacKintosh N, Hirst JE, Geoghegan R, Badets M, *et al.* Cytosine methylation is a conserved epigenetic feature found throughout the phylum Platyhelminthes. *BMC Genomics*. BMC Genomics; 2013; 14:462.
84. Anderson L, Pierce RJ, Verjovski-Almeida S. *Schistosoma mansoni* histones: From transcription to chromatin regulation; An in silico analysis. *Mol. Biochem. Parasitol*. 2012; 183:105-14.
85. Cucher M, Prada L, Mourglia-Ettlin G, Dematteis S, Camicia F, Asurmendi S, *et al.* Identification of *Echinococcus granulosus* microRNAs and their expression in different life cycle stages and parasite genotypes. *Int. J. Parasitol*. Australian Society for Parasitology Inc.; 2011; 41:439-48.

86. Moore LD, Le T, Fan G. DNA Methylation and Its Basic Function. Neuropsychopharmacology. Nature Publishing Group; 2013; 38:23-38.
87. Feng J, Chang H, Li E, Fan G. Dynamic expression of de novo DNA methyltransferases Dnmt3a and Dnmt3b in the central nervous system. J. Neurosci. Res. 2005; 79:734-46.
88. Goll MG. Methylation of tRNA^{Asp} by the DNA Methyltransferase Homolog Dnmt2. Science (80-.). 2006; 311:395-8.
89. Schaefer M, Lyko F. Solving the Dnmt2 enigma. Chromosoma. Springer-Verlag; 2010; 119:35-40.
90. Bird A, Taggart M, Frommer M, Miller OJ, Macleod D. A fraction of the mouse genome that is derived from islands of nonmethylated, CpG-rich DNA. Cell. 1985; 40:91-9.
91. Nan X, Meehan RR, Bird A. Dissection of the methyl-CpG binding domain from the chromosomal protein MeCP2. Nucleic Acids Res. 1993; 21:4886-92.
92. Hendrich B, Bird A. Identification and Characterization of a Family of Mammalian Methyl-CpG Binding Proteins. Mol. Cell. Biol. 1998; 18:6538-47.
93. Lewis JD, Meehan RR, Henzel WJ, Maurer-Fogy I, Jeppesen P, Klein F, *et al.* Purification, sequence, and cellular localization of a novel chromosomal protein that binds to Methylated DNA. Cell. 1992; 69:905-14.
94. Meehan RR, Lewis JD, McKay S, Kleiner EL, Bird AP. Identification of a mammalian protein that binds specifically to DNA containing methylated CpGs. Cell. 1989; 58:499-507.
95. Sarraf SA, Stancheva I. Methyl-CpG binding protein MBD1 couples histone H3 methylation at lysine 9 by SETDB1 to DNA replication and chromatin assembly. Mol. Cell. 2004; 15:595-605.
96. Ng HH, Zhang Y, Hendrich B, Johnson C a, Turner BM, Erdjument-Bromage H, *et al.* MBD2 is a transcriptional repressor belonging to the MeCP1 histone deacetylase complex. Nat. Genet. 1999; 23:58-61.
97. Geyer KK, Rodríguez López CM, Chalmers IW, Munshi SE, Truscott M, Heald J, *et al.* Cytosine methylation regulates oviposition in the pathogenic blood fluke *Schistosoma mansoni*. Nat. Commun. 2011; 2:424.
98. Hendrich B, Tweedie S. The methyl-CpG binding domain and the evolving role of DNA methylation in animals. Trends Genet. 2003; 19:269-77.
99. Hendrich B, Hardeland U, Ng HH, Jiricny J, Bird A. The thymine glycosylase MBD4 can bind to the product of deamination at methylated CpG sites. Nature. 1999; 401:301-4.
100. Jørgensen HF, Ben-Porath I, Bird AP. Mbd1 is recruited to both methylated and nonmethylated CpGs via distinct DNA binding domains. Mol. Cell. Biol. 2004; 24:3387-95.
101. Saxonov S, Berg P, Brutlag DL. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. Proc. Natl. Acad. Sci. 2006; 103:1412-7.
102. Medvedeva Y a, Fridman M V, Oparina NJ, Malko DB, Ermakova EO, Kulakovskiy I V, *et al.* Intergenic, gene terminal, and intragenic CpG islands in the human genome. BMC Genomics. 2010; 11:48.
103. Brenet F, Moh M, Funk P, Feierstein E, Viale AJ, Socci ND, *et al.* DNA methylation of the first exon is tightly linked to transcriptional silencing. Papavasiliou N, editor. PLoS One. Public Library of Science; 2011; 6:e14524.

104. Maunakea AK, Nagarajan RP, Bilenky M, Ballinger TJ, D'Souza C, Fouse SD, *et al.* Conserved role of intragenic DNA methylation in regulating alternative promoters. *Nature*. 2010; 466:253-7.
105. Illingworth RS, Gruenewald-Schneider U, Webb S, Kerr ARW, James KD, Turner DJ, *et al.* Orphan CpG Islands Identify numerous conserved promoters in the mammalian genome. Reik W, editor. *PLoS Genet*. 2010; 6:e1001134.
106. Meissner A, Mikkelsen TS, Gu H, Wernig M, Hanna J, Sivachenko A, *et al.* Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature*. 2008; 454:766-70.
107. Eckhardt F, Lewin J, Cortese R, Rakyan VK, Attwood J, Burger M, *et al.* DNA methylation profiling of human chromosomes 6, 20 and 22. *Nat. Genet*. 2006; 38:1378-85.
108. Rakyan VK, Hildmann T, Novik KL, Lewin J, Tost J, Cox A V., *et al.* DNA methylation profiling of the human major histocompatibility complex: A pilot study for the Human Epigenome Project. Peter B. Becker, editor. *PLoS Biol*. 2004; 2:e405.
109. Choi JK. Contrasting chromatin organization of CpG islands and exons in the human genome. *Genome Biol*. 2010; 11:R70.
110. Ramirez-Carrozzi VR, Braas D, Bhatt DM, Cheng CS, Hong C, Doty KR, *et al.* A Unifying Model for the Selective Regulation of Inducible Transcription by CpG Islands and Nucleosome Remodeling. *Cell*. 2009; 138:114-28.
111. Tazi J, Bird A. Alternative chromatin structure at CpG islands. *Cell*. 1990; 60:909-20.
112. Croce CM, Calin GA. miRNAs, cancer, and stem cell division. *Cell. Cell Press*; 2005; 122:6-7.
113. Meister G, Tuschl T. Mechanisms of gene silencing by double-stranded RNA. *Nature*. 2004; 431:343-9.
114. Lachman HM, Nolan KA, Mohr P, Saito T, Volavka J. Association between catechol O-methyltransferase genotype and violence in *schizophrenia* and schizoaffective disorder. *Am. J. Psychiatry*. 1998; 155:835-7.
115. Bartel DP. MicroRNAs: Genomics, Biogenesis, Mechanism, and Function. *Cell*. 2004; 116:281-97.
116. Kim VN. MicroRNA biogenesis: coordinated cropping and dicing. *Nat. Rev. Mol. Cell Biol*. 2005; 6:376-85.
117. Ameres SL, Zamore PD. Diversifying microRNA sequence and function. *Nat. Rev. Mol. Cell Biol. Nature Publishing Group*; 2013; 14:475-88.
118. Ha M, Kim VN. Regulation of microRNA biogenesis. *Nat. Rev. Mol. Cell Biol. Nature Publishing Group*; 2014; 15:509-24.
119. Lee RC, Feinbaum RL, Ambros V. The *C. elegans* heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-14*. *Cell*. 1993; 75:843-54.
120. Niwa R, Slack FJ. The evolution of animal microRNA function. *Curr. Opin. Genet. Dev*. 2007; 17:145-50.
121. Cucher M, Macchiaroli N, Kamenetzky L, Maldonado L, Brehm K, Rosenzvit M. High-throughput characterization of *Echinococcus* spp. metacestode miRNomes. *Int. J. Parasitol. Australian Society for Parasitology Inc.*; 2015; 45:253-67.
122. Macchiaroli N, Cucher M, Zarowiecki M, Maldonado L, Kamenetzky L, Rosenzvit MC. microRNA profiling in the zoonotic parasite *Echinococcus canadensis* using a high-throughput approach. *Parasit. Vectors*. 2015; 8:83.

123. Filipowicz W, Bhattacharyya SN, Sonenberg N. Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? *Nat. Rev. Genet.* 2008; 9:102-14.
124. Hutvagner G. A Cellular Function for the RNA-Interference Enzyme Dicer in the Maturation of the let-7 Small Temporal RNA. *Science* (80-.). 2001; 293:834-8.
125. Lau NC. An Abundant Class of Tiny RNAs with Probable Regulatory Roles in *Caenorhabditis elegans*. *Science* (80-.). 2001; 294:858-62.
126. Hershberg R, Petrov DA. Selection on Codon Bias. *Annu. Rev. Genet. Annual Reviews*; 2008; 42:287-99.
127. Sharp PM, Emery LR, Zeng K. Forces that influence the evolution of codon bias. *Philos. Trans. R. Soc. B Biol. Sci. The Royal Society*; 2010; 365:1203-12.
128. Plotkin JB, Kudla G. Synonymous but not the same: the causes and consequences of codon bias. *Nat. Rev. Genet. Nature Publishing Group*; 2011; 12:32-42.
129. Sharp PM, Matassi G. Codon usage and genome evolution. *Curr. Opin. Genet. Dev.* 1994; 4:851-60.
130. Behura SK, Severson DW. Codon usage bias: Causative factors, quantification methods and genome-wide patterns: With emphasis on insect genomes. *Biol. Rev.* 2013; 88:49-61.
131. Ikemura T. Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: A proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. *J. Mol. Biol.* 1981; 151:389-409.
132. Grantham R, Gautier C, Gouy M, Mercier R, Pavé A. Codon catalog usage and the genome hypothesis. *Nucleic Acids Res.* 1980; 8:197.
133. Bermudez-Santana C, Attolini C, Kirsten T, Engelhardt J, Prohaska SJ, Steiglele S, *et al.* Genomic organization of eukaryotic tRNAs. *BMC Genomics. BioMed Central*; 2010; 11:270.
134. Kane JF. Effects of rare codon clusters on high-level expression of heterologous proteins in *Escherichia coli*. *Curr. Opin. Biotechnol. Elsevier Current Trends*; 1995; 6:494-500.
135. Gustafsson C, Govindarajan S, Minshull J. Codon bias and heterologous protein expression. *Trends Biotechnol.* 2004; 22:346-53.
136. Fletcher SP, Muto M, Mayfield SP. Optimization of recombinant protein expression in the chloroplasts of green algae. *Adv. Exp. Med. Biol. New York, NY: Springer New York*; 2007. p. 90-8.
137. McArthur IV GH, Fong SS. Toward engineering synthetic microbial metabolism. *J. Biomed. Biotechnol. Hindawi*; 2010; 2010:459760.
138. Mueller S, Papamichail D, Coleman JR, Skiena S, Wimmer E. Reduction of the rate of poliovirus protein synthesis through large-scale codon deoptimization causes attenuation of viral virulence by lowering specific infectivity. *J. Virol. American Society for Microbiology (ASM)*; 2006; 80:9687-96.
139. Coleman JR, Papamichail D, Skiena S, Fletcher B, Wimmer E, Mueller S. Virus Attenuation by Genome-Scale Changes in Codon Pair Bias. *Science* (80-.). 2008; 320:1784-7.
140. Fan RLY, Valkenburg SA, Wong CKS, Li OTW, Nicholls JM, Rabadan R, *et al.* Generation of Live Attenuated Influenza Virus by Using Codon Usage Bias. Perlman S, editor. *J. Virol.* 2015; 89:10762-73.
141. Krakauer D.C., Jansen V.A. NM. Biological Evolution and Statistical Physics. Lässig M, Valleriani A, editors. *Lect. Notes Phys. Berlin, Heidelberg: Springer Berlin Heidelberg*; 2002.

142. Duret L. Evolution of synonymous codon usage in metazoans. *Curr. Opin. Genet. Dev.* 2002; 12:640-9.
143. Duret L, Mouchiroud D. Expression pattern and, surprisingly, gene length shape codon usage in *Caenorhabditis*, *Drosophila*, and *Arabidopsis*. *Proc. Natl. Acad. Sci. U. S. A. National Academy of Sciences*; 1999; 96:4482-7.
144. Comeron JM, Kreitman M, Aguad M. Natural selection on synonymous sites is correlated with gene length and recombination in *Drosophila*. *Genetics*. 1999; 151:239-49.
145. Marais G, Mouchiroud D, Duret L. Does recombination improve selection on codon usage? Lessons from nematode and fly complete genomes. *Proc. Natl. Acad. Sci.* 2001; 98:5688-92.
146. Qin H, Wu WB, Comeron JM, Kreitman M, Li WH. Intragenic spatial patterns of codon usage bias in prokaryotic and eukaryotic genomes. *Genetics. Genetics Society of America*; 2004; 168:2245-60.
147. Spencer PS, Barral JM. Genetic Code Redundancy and Its Influence on the Encoded Polypeptides. *Comput. Struct. Biotechnol. J. Research Network of Computational and Structural Biotechnology*; 2012; 1:e201204006.
148. Quax TEF, Claassens NJ, Söll D, van der Oost J. Codon Bias as a Means to Fine-Tune Gene Expression. *Mol. Cell. NIH Public Access*; 2015; 59:149-61.
149. Rocha EPC. Codon usage bias from tRNA's point of view: Redundancy, specialization, and efficient decoding for translation optimization. *Genome Res. Cold Spring Harbor Laboratory Press*; 2004; 14:2279-86.
150. Akashi H, Kliman RM, Eyre-Walker A. Mutation pressure, natural selection, and the evolution of base composition in *Drosophila*. *Genetics. Kluwer Academic Publishers*; 1998; 102-103:49-60.
151. Kliman RM, Hey J. The effects of mutation and natural-selection on codon bias in the genes of *Drosophila*. *Genetics*. 1994; 137:1049-56.
152. Sharp PM, Matassi G. Codon usage and genome evolution. *Curr. Opin. Genet. Dev.* 1994; 4:851-60.
153. Bulmer M. The selection-mutation-drift theory of synonymous codon usage. *Genetics*. 1991; 129:897-907.
154. Powell JR, Moriyama EN. Evolution of codon usage bias in *Drosophila*. *Proc. Natl. Acad. Sci. U. S. A. Colloquium Paper*; 1997; 94:7784-90.
155. Novoa EM, Ribas de Pouplana L. Speeding with control: Codon usage, tRNAs, and ribosomes. *Trends Genet.* 2012; 28:574-81.
156. Behura SK, Stanke M, Desjardins CA, Werren JH, Severson DW. Comparative analysis of nuclear tRNA genes of *Nasonia vitripennis* and other arthropods, and relationships to codon usage bias. *Insect Mol. Biol.* 2010; 19:49-58.
157. Akashi H. Translational selection and yeast proteome evolution. *Genetics*. 2003; 164:1291-303.
158. Coghlan A, Wolfe KH. Relationship of codon bias to mRNA and concentration protein length in *Saccharomyces cerevisiae*. *Yeast*. 2000; 16:1131-45.
159. Grosjean H, Fiers W. Preferential codon usage in prokaryotic genes: the optimal codon-anticodon interaction energy and the selective codon usage in efficiently expressed genes. *Gene*. 1982; 18:199-209.
160. Bennetzen JL, Hall BD. Codon selection in yeast. *J. Biol. Chem.* 1982; 257:3026-31.
161. Sharp PM, Li WH. An evolutionary perspective on synonymous codon usage in unicellular organisms. *J. Mol. Evol.* 1986; 24:28-38.

162. Carlini DB, Stephan W. In vivo introduction of unpreferred synonymous codons into the *drosophila Adh* gene results in reduced levels of ADH protein. *Genetics*. 2003; 163:239-43.
163. Ikemura T. Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes. Differences in synonymous codon choice patterns of yeast and *Escherichia coli* with reference to the abundance of isoaccepting transfer RNAs. *J. Mol. Biol.* 1982; 158:573-97.
164. Kanaya S, Yamada Y, Kudo Y, Ikemura T. Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of *Bacillus subtilis* tRNAs: Gene expression level and species-specific diversity of codon usage based on multivariate analysis. *Gene*. 1999; 238:143-55.
165. Bulmer M. Coevolution of codon usage and transfer RNA abundance. *Nature*. 1987. p. 728-30.
166. Kamenetzky L, Canova SG, Guarnera EA, Rosenzvit MC. *Echinococcus granulosus*: DNA Extraction from Germinal Layers Allows Strain Determination in Fertile and Nonfertile Hydatid Cysts. *Exp. Parasitol.* 2000; 95:122-7.
167. Bolger AM, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*. 2014; 30:2114-20.
168. Zerbino DR, Birney E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* 2008; 18:821-9.
169. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, *et al.* Erratum: SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience*. 2015; 4:30.
170. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, *et al.* SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J. Comput. Biol.* 2012; 19:455-77.
171. Safonova Y, Bankevich A, Pevzner PA. DIPSPADES: Assembler for highly polymorphic diploid genomes. *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*. 2014; 8394 LNBI:265-79.
172. Tsai IJ, Otto TD, Berriman M. Improving draft assemblies by iterative mapping and assembly of short reads to eliminate gaps. *Genome Biol.* 2010; 11:R41.
173. Otto TD, Sanders M, Berriman M, Newbold C. Iterative correction of reference Nucleotides (iCORN) using second generation sequencing technology. *Bioinformatics*. 2010; 26:1704-7.
174. Assefa S, Keane TM, Otto TD, Newbold C, Berriman M. ABACAS: Algorithm-based automatic contiguation of assembled sequences. *Bioinformatics*. 2009; 25:1968-9.
175. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUASt: Quality assessment tool for genome assemblies. *Bioinformatics*. 2013; 29:1072-5.
176. Parra G, Bradnam K, Korf I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics*. 2007; 23:1061-7.
177. Holt C, Yandell M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics*. BioMed Central Ltd; 2011; 12:491.
178. Stanke M, Waack S. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics*. 2003; 19:ii215-25.
179. Korf I. Gene finding in novel genomes. *BMC Bioinformatics*. 2004; 5:59.
180. Slater G, Birney E, Box G, Smith T, Waterman M, Altschul S, *et al.* Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*. 2005; 6:31.

181. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol. BioMed Central Ltd*; 2013; 14:R36.
182. Trapnell C, Pachter L, Salzberg SL. TopHat: Discovering splice junctions with RNA-Seq. *Bioinformatics*. 2009; 25:1105-11.
183. Lomsadze A, Burns PD, Borodovsky M. Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. *Nucleic Acids Res*. 2014; 42:e119.
184. Lomsadze A, Ter-Hovhannisyan V, Chernoff YO, Borodovsky M. Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids Res*. 2005; 33:6494-506.
185. Smit Arian HR. RepeatModeler - 1.0.8. Institute for Systems Biology; 2008.
186. Smit AFA, Hubley R GP. RepeatMasker 4.0.6. 2010.
187. Lowe TM, Eddy SR. TRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res*. 1996; 25:955-64.
188. Stemmer K, Burghardt I, Mayer C, Reinicke GB, Wgele H, Tollrian R, *et al.* Morphological and genetic analyses of xeniid soft coral diversity (*Octocorallia; Alcyonacea*). *Org. Divers. Evol*. 2013; 13:135-50.
189. Keibler E, Brent MR. Eval: a software package for analysis of genome annotations. *BMC Bioinformatics*. 2003; 4:50.
190. Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, *et al.* InterProScan: Protein domains identifier. *Nucleic Acids Res*. 2005; 33:W116-20.
191. Sigrist CJA, Cerutti L, De Castro E, Langendijk-Genevaux PS, Bulliard V, Bairoch A, *et al.* PROSITE, a protein domain database for functional characterization and annotation. *Nucleic Acids Res. Oxford University Press*; 2009; 38:D161-6.
192. Attwood TK, Bradley P, Flower DR, Gaulton A, Maudling N, Mitchell AL, *et al.* PRINTS and its automatic supplement, prePRINTS. *Nucleic Acids Res. Oxford University Press*; 2003; 31:400-2.
193. Letunic I, Bork P. 20 years of the SMART protein domain annotation resource. *Nucleic Acids Res*. 2017;
194. Finn RD, Coghill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, *et al.* The Pfam protein families database: Towards a more sustainable future. *Nucleic Acids Res. Oxford University Press*; 2016; 44:D279-85.
195. Gough J, Karplus K, Hughey R, Chothia C. Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *J. Mol. Biol. Academic Press*; 2001; 313:903-19.
196. Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M. Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*. 2005; 21:3674-6.
197. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, *et al.* Gene Ontology: tool for the unification of biology. *Nat. Genet*. 2000; 25:25-9.
198. Li L, Stoeckert CJJ, Roos DS. OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes -- Li *et al.* 13 (9): 2178 -- *Genome Research. Genome Res*. 2003; 13:2178-89.
199. Fischer S, Brunk BP, Chen F, Gao X, Harb OS, Iodice JB, *et al.* Using OrthoMCL to assign proteins to OrthoMCL-DB groups or to cluster proteomes into new ortholog groups. *Curr. Protoc. Bioinforma*. 2011; Chapter 6:Unit 6.12.1-19.

200. Kanehisa M, Sato Y, Morishima K. BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences. *J. Mol. Biol.* 2016; 428:726-31.
201. Finn RD. Pfam: clans, web tools and services. *Nucleic Acids Res.* 2006; 34:D247-51.
202. Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJE. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat. Protoc.* 2015; 10:845-58.
203. Biasini M, Bienert S, Waterhouse A, Arnold K, Studer G, Schmidt T, *et al.* SWISS-MODEL: Modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic Acids Res.* 2014; 42:W252-8.
204. Arnold K, Bordoli L, Kopp J, Schwede T. The SWISS-MODEL workspace: A web-based environment for protein structure homology modelling. *Bioinformatics.* 2006; 22:195-201.
205. Kiefer F, Arnold K, Künzli M, Bordoli L, Schwede T. The SWISS-MODEL Repository and associated resources. *Nucleic Acids Res.* 2009; 37:D387-92.
206. Guex N, Peitsch MC, Schwede T. Automated comparative protein structure modeling with SWISS-MODEL and Swiss-PdbViewer: A historical perspective. *Electrophoresis.* 2009; 30:S162-73.
207. Elkayam E, Kuhn CD, Tocilj A, Haase AD, Greene EM, Hannon GJ, *et al.* The structure of human argonaute-2 in complex with miR-20a. *Cell.* 2012; 150:100-10.
208. Takai D, Jones PA. Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proc. Natl. Acad. Sci.* 2002; 99:3740-5.
209. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, *et al.* Circos: An information aesthetic for comparative genomics. *Genome Res.* 2009; 19:1639-45.
210. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat. Methods.* 2012; 9:357-9.
211. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics.* 2009; 25:1754-60.
212. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, *et al.* The variant call format and VCFtools. *Bioinformatics.* 2011; 27:2156-8.
213. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin).* 2012; 6:80-92.
214. Otto TD, Dillon GP, Degraeve WS, Berriman M. RATT: Rapid Annotation Transfer Tool. *Nucleic Acids Res.* 2011; 39:1-7.
215. Nakao M, Sako Y, Yokoyama N, Fukunaga M, Ito A. Mitochondrial genetic code in cestodes. *Mol. Biochem. Parasitol.* 2000; 111:415-24.
216. Jeanmougin F, Thompson JD, Gouy M, Higgins DG, Gibson TJ. Multiple sequence alignment with Clustal X. *Trends Biochem. Sci.* 1998; 23:403-5.
217. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, *et al.* The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 2010; 20:1297-303.
218. Price MN, Dehal PS, Arkin AP. Fasttree: Computing large minimum evolution trees with profiles instead of a distance matrix. *Mol. Biol. Evol.* Oxford University Press; 2009; 26:1641-50.
219. Peden JF. Analysis of codon usage. *Biosystems.* 1999.

220. Wan XF, Zhou J, Xu D. CodonO: A new informatics method for measuring synonymous codon usage bias within and across genomes. *Int. J. Gen. Syst.* 2006; 35:109-25.
221. Angellotti MC, Bhuiyan SB, Chen G, Wan XF. CodonO: Codon usage bias analysis within and across genomes. *Nucleic Acids Res.* 2007; 35:132-6.
222. Suzuki H, Saito R, Tomita M. A problem in multivariate analysis of codon usage data and a possible solution. *FEBS Lett.* 2005; 579:6499-504.
223. Sharp PM, Li WH. The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.* 1987; 15:1281-95.
224. Wright F. The “effective number of codons” used in a gene. *Gene.* 1990; 87:23-9.
225. Lee S, Weon S, Lee S, Kang C. Relative codon adaptation index, a sensitive measure of codon usage bias. *Evol. Bioinformatics.* 2010; 2010:47-55.
226. Hartl DL, Moriyama EN, Sawyer SA. Selection intensity for codon bias. *Genetics.* 1994; 138:227-34.
227. Lobry JR, Gautier C. Hydrophobicity, expressivity and aromaticity are the major trends of amino-acid usage in 999 *Escherichia coli* chromosome-encoded genes. *Nucleic Acids Res.* 1994; 22:3174-80.
228. Sueoka N, Kawanishi Y. DNA G + C content of the third codon position and codon usage biases of human genes. *Gene.* 2000; 261:53-62.
229. Sueoka N. Intrastrand parity rules of DNA base composition and usage biases of synonymous codons. *J. Mol. Evol.* 1995; 40:318-25.
230. Sueoka N. Directional mutation pressure and neutral molecular evolution. *Proc. Natl. Acad. Sci. U. S. A.* 1988; 85:2653-7.
231. Sueoka N. Directional mutation pressure, selective constraints, and genetic equilibria. *J. Mol. Evol.* 1992; 34:95-114.
232. Yandell M, Ence D. A beginner’s guide to eukaryotic genome annotation. *Nat. Rev. Genet.* Nature Publishing Group; 2012; 13:329-42.
233. Hahn C, Fromm B, Bachmann L. Comparative genomics of flatworms (Platyhelminthes) reveals shared genomic features of ecto- and endoparasitic neodermata. *Genome Biol. Evol.* 2014; 6:1105-17.
234. The RNACentral Consortium. RNACentral: a comprehensive database of non-coding RNA sequences. *Nucleic Acids Res.* Oxford University Press; 2017; 45:D128-34.
235. Rausch RL, Journal T, Jun N. A Consideration of Intraspecific Categories in the Genus *Echinococcus Rudolphi* , 1801 (Cestoda : Taeniidae); *Genus.* 1967; 53:484-91.
236. Nakao M, Mcmanus DP, Schantz PM, Craigs PS, Ito A. A molecular phylogeny of the genus *Echinococcus* inferred from complete mitochondrial genomes. *Parasitology.* 2007; 134:713.
237. Nakao M. The complete mitochondrial DNA sequence of the cestode *Echinococcus multilocularis* (Cyclophyllidea: Taeniidae). *Mitochondrion.* 2002; 1:497-509.
238. Silva-Álvarez V, Franchini GR, Pórfido JL, Kennedy MW, Ferreira AM, Córscico B. Lipid-Free Antigen B Subunits from *Echinococcus granulosus*: Oligomerization, Ligand Binding, and Membrane Interaction Properties. *PLoS Negl. Trop. Dis.* 2015; 9:e0003552.
239. Silva-Álvarez V, Folle AM, Ramos AL, Kitano ES, Iwai LK, Corraliza I, *et al.* *Echinococcus granulosus* Antigen B binds to monocytes and macrophages modulating cell response to inflammation. *Parasit. Vectors.* 2016; 9:69.

240. Victor B, Dorny P, Kanobana K, Polman K, Lindh J, Deelder AM, *et al.* Use of expressed sequence tags as an alternative approach for the identification of *Taenia solium* metacestode excretion/secretion proteins. BMC Res. Notes. 2013; 6:224.
241. Osaki T, Omotezako M, Nagayama R, Hirata M, Iwanaga S, Kasahara J, *et al.* Horseshoe crab hemocyte-derived antimicrobial polypeptides, tachystatins, with sequence similarity to spider neurotoxins. J. Biol. Chem. 1999; 274:26172-8.
242. Fujitani N, Kawabata SI, Osaki T, Kumaki Y, Demura M, Nitta K, *et al.* Structure of the antimicrobial peptide tachystatin A. J. Biol. Chem. 2002; 277:23651-7.
243. Ganz T. Defensins: antimicrobial peptides of innate immunity. Nat. Rev. Immunol. 2003; 3:710-20.
244. Collins JJ, Hou X, Romanova E V., Lambrus BG, Miller CM, Saberi A, *et al.* Genome-wide analyses reveal a role for peptide hormones in planarian germline development. PLoS Biol. 2010; 8:e1000509.
245. Filmore D. It's a GPCR world. J. Mod. Drug Discov. 2004; 7:24-7.
246. Rajpara SM, Garcia PD, Roberts R, Eliassen JC, Owens DF, Maltby D, *et al.* Identification and molecular cloning of a neuropeptide y homolog that produces prolonged inhibition in *aplysia* neurons. Neuron. 1992; 9:505-13.
247. Hinton A, Bond S, Forgac M. V-ATPase functions in normal and disease processes. Pflugers Arch. Eur. J. Physiol. 2009; 457:589-98.
248. Jefferies KC, Cipriano DJ, Forgac M. Function, structure and regulation of the vacuolar (H⁺)-ATPases. Arch. Biochem. Biophys. 2008; 476:33-42.
249. Forgac M. The vacuolar H⁺-ATPase of clathrin-coated vesicles is reversibly inhibited by S-nitrosoglutathione. J. Biol. Chem. 1999; 274:1301-5.
250. Knight AJ, Behm CA. Minireview: The role of the vacuolar ATPase in nematodes. Exp. Parasitol. 2012; 132:47-55.
251. Lepier A, Gräf R, Azuma M, Merzendorfer H, Harvey WR, Wieczorek H. The peripheral complex of the tobacco hornworm V-ATPase contains a novel 13-kDa subunit G. J. Biol. Chem. 1996; 271:8502-8.
252. Kinoshita T, Fujita M, Maeda Y. Biosynthesis, remodelling and functions of mammalian GPI-anchored proteins: Recent progress. J. Biochem. 2008; 144:287-94.
253. Maeda Y, Tanaka S, Hino J, Kangawa K, Kinoshita T. Human dolichol-phosphate-mannose synthase consists of three subunits, DPM1, DPM2 and DPM3. EMBO J. 2000; 19:2475-82.
254. Knight RD, Shimeld SM. Identification of conserved C2H2 zinc-finger gene families in the Bilateria. Genome Biol. BioMed Central; 2001; research0016.1-0016.8.
255. Seetharam A, Bai Y, Stuart GW. A survey of well conserved families of C2H2 zinc-finger genes in Daphnia. BMC Genomics. BioMed Central; 2010; 11:276.
256. Soukup SW. Evolution by gene duplication. S. Ohno. Springer-Verlag, New York. 1970. 160 pp. Teratology. Wiley Subscription Services, Inc., A Wiley Company; 1974; 9:250-1.
257. Collins CA, Brown EJ. Cytosol as battleground: Ubiquitin as a weapon for both host and pathogen. Trends Cell Biol. 2010; 20:205-13.
258. Bogdanović O, Veenstra GJC. DNA methylation and methyl-CpG binding proteins: Developmental requirements and function. Chromosoma. 2009; 118:549-65.
259. Han L, Su B, Li W-H, Zhao Z. CpG island density and its correlations with genomic features in mammalian genomes. Genome Biol. 2008; 9:R79.

260. Schürmann N, Trabuco LG, Bender C, Russell RB, Grimm D. Molecular dissection of human Argonaute proteins by DNA shuffling. *Nat. Struct. Mol. Biol.* 2013; 20:818-26.
261. Palidwor GA, Perkins TJ, Xia X. A general model of Codon bias due to GC mutational bias. Stekel DJ, editor. *PLoS One. R Foundation for Statistical Computing*; 2010; 5:e13431.
262. Brbić M, Piškorec M, Vidulin V, Kriško A, Šmuc T, Supek F. The landscape of microbial phenotypic traits and associated genes. *Nucleic Acids Res.* 2016; 44:10074-90.
263. Esposito LA, Gupta S, Streiter F, Prasad A, Dennehy JJ. Evolutionary interpretations of mycobacteriophage biodiversity and host-range through the analysis of codon usage bias. *Microb. genomics.* 2016; 2:e000079.
264. Yu H, Li Q. Mutation and selection on the wobble nucleotide in tRNA anticodons in marine bivalve mitochondrial genomes. Lightowlers B, editor. *PLoS One. ASM Press*; 2011; 6:e16147.
265. Maldonado LL, Assis J, Araújo FMG, Salim ACM, Macchiaroli N, Cucher M, Kamenetzky L. The *Echinococcus canadensis* (G7) genome: a key knowledge of parasitic platyhelminth human diseases. *BMC Genomics.* 2017; 18:204.
266. Mudge JM, Harrow J. The state of play in higher eukaryote gene annotation. *Nat. Rev. Genet. Nature Publishing Group*; 2016; 17:758-72.
267. Crámer P, Iusem ND. Mitos y verdades del genoma humano. *Med. Cold Spring Harbor Laboratory Press*; 2013; 73:379-83.
268. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, *et al.* BLAST+: architecture and applications. *BMC Bioinformatics. BioMed Central*; 2009; 10:421.
269. Mitchell A, Chang HY, Daugherty L, Fraser M, Hunter S, Lopez R, *et al.* The InterPro protein families database: The classification resource after 15 years. *Nucleic Acids Res. Oxford University Press*; 2015; 43:D213-21.
270. Zdobnov EM, Apweiler R. InterProScan - an integration platform for the signature-recognition methods in InterPro. *Bioinformatics.* 2001; 17:847-8.
271. Huerta-Cepas J, Szklarczyk D, Forslund K, Cook H, Heller D, Walter MC, *et al.* EGGNOG 4.5: A hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res. Oxford University Press*; 2016; 44:D286-93.
272. Protasio A V., Tsai IJ, Babbage A, Nichol S, Hunt M, Aslett MA, *et al.* A systematically improved high quality genome and transcriptome of the human blood fluke *Schistosoma mansoni*. Hoffmann KF, editor. *PLoS Negl. Trop. Dis. Public Library of Science*; 2012; 6:e1455.
273. Gabaldón T, Koonin E V. Functional and evolutionary implications of gene orthology. *Nat. Rev. Genet. Nature Publishing Group*; 2013; 14:360-6.
274. Zhou Y, Zheng H, Chen Y, Zhang L, Wang K, Guo J, *et al.* The *Schistosoma japonicum* genome reveals features of host-parasite interplay. *Nature. Nature Publishing Group*; 2009; 460:345-51.
275. Pan W, Shen Y, Han X, Wang Y, Liu H, Jiang Y, *et al.* Transcriptome Profiles of the Protoscoleces of *Echinococcus granulosus* Reveal that Excretory-Secretory Products Are Essential to Metabolic Adaptation. *PLoS Negl. Trop. Dis.* 2014; 8:1-15.
276. Hewitson JP, Grainger JR, Maizels RM. Helminth immunoregulation: The role of parasite secreted proteins in modulating host immunity. *Mol. Biochem. Parasitol.* 2009; 167:1-11.

277. Nono JK, Pletinckx K, Lutz MB, Brehm K. Excretory/secretory-products of *Echinococcus multilocularis* larvae induce apoptosis and tolerogenic properties in dendritic cells in vitro. PLoS Negl. Trop. Dis. 2012; 6.
278. Robinson MW, Donnelly S, Hutchinson AT, To J, Taylor NL, Norton RS, *et al.* A family of helminth molecules that modulate innate cell responses via molecular mimicry of host antimicrobial peptides. PLoS Pathog. 2011; 7:e1002042.
279. Thivierge K, Cotton S, Schaefer DA, Riggs MW, To J, Lund ME, *et al.* Cathelicidin-like Helminth Defence Molecules (HDMs): Absence of Cytotoxic, Anti-microbial and Anti-protozoan Activities Imply a Specific Adaptation to Immune Modulation. PLoS Negl. Trop. Dis. 2013; 7:e2307.
280. Han L, Zhao Z. Contrast features of CpG islands in the promoter and other regions in the dog genome. Genomics. Elsevier Inc.; 2009; 94:117-24.
281. Yi S. Birds do it, bees do it, worms and ciliates do it too: DNA methylation from unexpected corners of the tree of life. Genome Biol. 2012; 13:174.
282. Jiang C, Han L, Su B, Li WH, Zhao Z. Features and trend of loss of promoter-associated CpG islands in the human and mouse genomes. Mol. Biol. Evol. 2007; 24:1991-2000.
283. Matsuo K, Clay O, Takahashi T, Silke J, Schaffner W. Evidence for erosion of mouse CpG islands during mammalian evolution. Somat. Cell Mol. Genet. 1993; 19:543-55.
284. Zhao Z, Jiang C. Methylation-dependent transition rates are dependent on local sequence lengths and genomic regions. Mol. Biol. Evol. 2007; 24:23-5.
285. Fryxell KJ, Moon WJ. CpG mutation rates in the human genome are highly dependent on local GC content. Mol. Biol. Evol. 2005; 22:650-8.
286. Wasik K, Gurtowski J, Zhou X, Ramos OM, Delás MJ, Battistoni G, *et al.* Genome and transcriptome of the regeneration-competent flatworm, *Macrostomum lignano*. Proc. Natl. Acad. Sci. 2015; 112:12462-7.
287. Otero L, Bonilla M, Protasio A V, Fernández C, Gladyshev VN, Salinas G. Thioredoxin and glutathione systems differ in parasitic and free-living platyhelminths. BMC Genomics. 2010; 11:237.
288. Bartel DP. MicroRNAs: Target Recognition and Regulatory Functions. Cell. 2009; 136:215-33.
289. Naya H, Romero H, Carels N, Zavala A, Musto H. Translational selection shapes codon usage in the GC-rich genome of *Chlamydomonas reinhardtii*. FEBS Lett. 2001; 501:127-30.
290. Lin K. Conserved codon composition of ribosomal protein coding genes in *Escherichia coli*, *Mycobacterium tuberculosis* and *Saccharomyces cerevisiae*: lessons from supervised machine learning in functional genomics. Nucleic Acids Res. Oxford University Press; 2002; 30:2599-607.
291. Salamov AA, Solovyev V V. Ab initio gene finding in *Drosophila* genomic DNA. Genome Res. Cold Spring Harbor Laboratory Press; 2000; 10:516-22.
292. Cai MS, Cheng AC, Wang MS, Zhao LC, Zhu DK, Luo QH, *et al.* Characterization of synonymous codon usage bias in the duck plague virus UL35 gene. Intervirology. 2009; 52:266-78.
293. Kliman RM, Irving N, Santiago M. Selection conflicts, gene expression, and codon usage trends in yeast. J. Mol. Evol. Springer-Verlag; 2003; 57:98-109.
294. Lafay B, Sharp PM. Synonymous codon usage variation among *Giardia lamblia* genes and isolates. Mol. Biol. Evol. 1999; 16:1484-95.

295. Ran W, Kristensen DM, Koonin E V. Coupling between protein level selection and codon usage optimization in the evolution of bacteria and archaea. *MBio*. National Academy of Sciences; 2014; 5:e1001107.
296. Yang X, Ma X, Luo X, Ling H, Zhang X, Cai X. Codon usage bias and determining forces in *Taenia solium* genome. *Korean J. Parasitol.* 2015; 53:689-97.
297. Huang X, Xu J, Chen L, Wang Y, Gu X, Peng X, *et al.* Analysis of transcriptome data reveals multifactor constraint on codon usage in *Taenia multiceps*. *BMC Genomics*. BMC Genomics; 2017; 18:308.
298. Yang X, Luo X, Cai X. Analysis of codon usage pattern in *Taenia saginata* based on a transcriptome dataset. *Parasit. Vectors.* 2014; 7:527.
299. Mazumder GA, Uddin A, Chakraborty S. Comparative analysis of codon usage pattern and its influencing factors in *Schistosoma japonicum* and *Ascaris suum*. *Acta Parasitol.* 2017; 62:748-61.
300. Jia X, Liu S, Zheng H, Li B, Qi Q, Wei L, *et al.* Non-uniqueness of factors constraint on the codon usage in *Bombyx mori*. *BMC Genomics*. BioMed Central; 2015; 16:356.
301. Kawabe A, Miyashita NT. Patterns of codon usage bias in three dicot and four monocot plant species. *Genes Genet. Syst. The Genetics Society of Japan*; 2003; 78:343-52.
302. D'Onofrio G, Mouchiroud D, Aïssani B, Gautier C, Bernardi G. Correlations between the compositional properties of human genes, codon usage, and amino acid composition of proteins. *J. Mol. Evol.* 1991; 32:504-10.
303. Sharp PM, Stenico M, Peden JF, Lloyd AT, Centre QM. Codon usage: mutational bias, translational selection, or both? *Biochem Soc Trans.* 1993; 21:835-41.
304. Berg OG, Martelius M. Synonymous substitution-rate constants in *Escherichia coli* and *Salmonella typhimurium* and their relationship to gene expression and selection pressure. *J. Mol. Evol.* Springer-Verlag; 1995; 41:449-56.
305. Bulmer M. Are codon usage patterns in unicellular organisms determined by selection-mutation balance? *J. Evol. Biol.* Blackwell Science Ltd; 1988; 1:15-26.
306. Akashi H. Inferring weak selection from patterns of polymorphism and divergence at "silent" sites in *Drosophila* DNA. *Genetics*. Genetics Society of America; 1995; 139:1067-76.
307. Marais G, Mouchiroud D, Duret L. Neutral effect of recombination on base composition in *Drosophila*. *Genet. Res.* 2003; 81:S0016672302006079.
308. Galtier N. Gene conversion drives GC content evolution in mammalian histones. *Trends Genet.* 2003; 19:65-8.
309. Hambuch TM, Parsch J. Patterns of synonymous codon usage in *Drosophila melanogaster* genes with sex-biased expression. *Genetics*. Genetics Society of America; 2005; 170:1691-700.
310. Carlini DB, Chen Y, Stephan W. The relationship between third-codon position nucleotide content, codon bias, mRNA secondary structure and gene expression in the drosophilid alcohol dehydrogenase genes *Adh* and *Adhr*. *Genetics*. Genetics Society of America; 2001; 159:623-33.
311. Akashi H. Codon bias evolution in *Drosophila*. *Population genetics of mutation-selection drift*. *Gene*. 1997; 205:269-78.
312. Chen Y, Carlini DB, Baines JF, Parsch J, Braverman JM, Tanda S, *et al.* RNA secondary structure and compensatory evolution. *Genes Genet. Syst.* 1999; 74:271-86.

313. Moriyama EN, Powell JR. Gene length and codon usage bias in *Drosophila melanogaster*, *Saccharomyces cerevisiae* and *Escherichia coli*. Nucleic Acids Res. Oxford University Press; 1998; 26:3188-93.
314. Oresic M, Dehn MHH, Korenblum D, Shalloway D. Tracing specific synonymous codon-secondary structure correlations through evolution. J. Mol. Evol. 2003; 56:473-84.
315. Berg OG. Selection intensity for codon bias and the effective population size of *Escherichia coli*. Genetics. 1996; 142:1379-82.
316. Prat Y, Fromer M, Linial N, Linial M. Codon usage is associated with the evolutionary age of genes in metazoan genomes. BMC Evol. Biol. 2009; 9:285.
317. Goodarzi H, Torabi N, Najafabadi HS, Archetti M. Amino acid and codon usage profiles: Adaptive changes in the frequency of amino acids and codons. Gene. 2008; 407:30-41.
318. Seward EA, Kelly S. Dietary nitrogen alters codon bias and genome composition in parasitic microorganisms. Genome Biol. 2016; 17:226.
319. Dedon PC, Begley TJ. A system of RNA modifications and biased codon use controls cellular stress response at the level of translation. Chem. Res. Toxicol. American Chemical Society; 2014; 27:330-7.
320. McVean GAT, Vieira J. Inferring parameters of mutation, selection and demography from patterns of synonymous site evolution in *Drosophila*. Genetics. Genetics Society of America; 2001; 157:245-57.