



Universidad Nacional Autónoma de México

Instituto de Biotecnología

**Análisis del transcriptoma de la glándula de
veneno del alacrán *Centruroides noxius*
Hoffmann**

Tesis

que para obtener el grado de

Maestro en Ciencias Bioquímicas

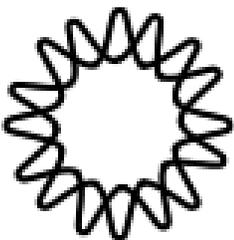
presenta:

Lic. en CG Martha Rosalía Rendón Anaya

Director de Tesis

Dr. Lourival D. Possani Postay

Cuernavaca Morelos, México. Agosto, 2011





Universidad Nacional
Autónoma de México

Dirección General de Bibliotecas de la UNAM

Biblioteca Central



UNAM – Dirección General de Bibliotecas
Tesis Digitales
Restricciones de uso

DERECHOS RESERVADOS ©
PROHIBIDA SU REPRODUCCIÓN TOTAL O PARCIAL

Todo el material contenido en esta tesis esta protegido por la Ley Federal del Derecho de Autor (LFDA) de los Estados Unidos Mexicanos (México).

El uso de imágenes, fragmentos de videos, y demás material que sea objeto de protección de los derechos de autor, será exclusivamente para fines educativos e informativos y deberá citar la fuente donde la obtuvo mencionando el autor o autores. Cualquier uso distinto como el lucro, reproducción, edición o modificación, será perseguido y sancionado por el respectivo titular de los Derechos de Autor.

Este proyecto se realizó en el departamento de Medicina Molecular y Bioprocesos del Instituto de Biotecnología de la Universidad Nacional Autónoma de México bajo la dirección del Dr. Lourival D. Possani Postay y el Dr. Alfredo Herrera Estrella (investigador titular en el Laboratorio Nacional de Genómica para la Biodiversidad). Durante el desarrollo del mismo, se contó con apoyo económico del Consejo Nacional de Ciencia y Tecnología (beca número 234807 a MRRA); el trabajo fue financiado con los proyectos DGAPA-UNAM IN204110 y el Instituto Bioclon S.A. de C.V a LDP.

Agradecimientos

Agradezco al Dr. Lourival Possani, tutor de este proyecto de maestría, el haberme permitido formar parte su grupo de investigación.

Agradezco el apoyo del Dr. Alfredo Herrera Estrella, quien fungió como co-tutor, y del Dr. Enrique Morett Sánchez, miembro del comité tutorial. Las aportaciones y críticas de ambos asesores durante el desarrollo de este proyecto fueron fundamentales para alcanzar los objetivos planteados.

A los Dres. Alejandro Alagón, Federico Sánchez, Miguel Ángel Cevallos, Enrique Merino y Lorenzo Segovia, miembros del jurado de examen, agradezco las excelentes observaciones y sugerencias para enriquecer y mejorar esta tesis.

Índice general

Resumen	1
1 Introducción	3
1.1 Venenos y Toxinas	3
1.2 Pirosecuenciación	7
1.3 Algoritmos de ensamblado	10
2 Antecedentes	12
2.1 Estado del arte en la genómica de los alacranes	13
2.2 Toxinas caracterizadas en el veneno de <i>Centruroides noxius</i>	14
2.3 Perfiles transcripcionales de glándulas de veneno de alacranes	15
3 Hipótesis	21
4 Objetivo general	21
5 Objetivos particulares	21
6 Metodología	22
6.1 Preparación de las librerías de cDNA	23
6.2 Pirosecuenciación por GS20-454, GS-FLX y FLX-Titanium	23
6.3 Ensamblado de novo de las lecturas de pirosecuenciación	24
6.4 Análisis cualitativo de la diversidad de transcritos	24
6.4.1 Búsquedas por homología	24
6.4.2 Búsqueda de dominios de proteínas	24
6.4.3 Búsqueda de péptidos señal	25
6.4.4 Análisis filogenético	25
6.4.5 Clasificación taxonómica y mapeo funcional	25
6.5 Análisis cuantitativo de la abundancia de transcritos	26
6.6 Procesamiento de archivos	27
7 Resultados y Discusión	28

7.1	Librerías de cDNA	28
7.2	Ensamblado	29
7.2.1	Newbler cDNA V2.3 y 2.5	29
7.2.2	MIRA EST	30
7.2.3	Newbler vs MIRA	31
7.2.4	Ensamblado global	34
7.3	Corridas Acumuladas	37
7.4	Análisis cualitativo de los transcritos de <i>Centruroides noxius</i>	39
7.4.1	Clasificación Taxonómica	39
7.4.2	Búsquedas por homología	41
7.4.3	Genes eucariotes esenciales	43
7.4.4	Mapeo de las secuencias en las rutas metabólicas de KEGG	45
7.4.5	Toxinas	46
7.4.5.1	Toxinas CAP	48
7.4.5.2	Péptidos con actividad antimicrobiana	50
7.4.5.3	Toxinas que afectan canales iónicos	51
7.4.5.4	Proteínas con dominio Kunitz	60
7.4.5.5	Fosfolipasas	61
7.4.5.6	Metaloproteasas	61
7.4.5.7	Otros componentes del veneno	65
7.5	Análisis cuantitativo de los niveles de expresión en la glándula de veneno en estados de actividad y mantenimiento	69
7.5.1	Prueba exacta de Fisher	70
7.5.2	Estadístico Q	70
7.5.3	Comparación entre estados de glándula	71
8	Conclusiones y perspectivas	77
9	Bibliografía	79
10	Material Suplementario	90

Índice de tablas

Tabla 1. Comparación de algunas plataformas de secuenciación masiva _____	9
Tabla 2. Cariotipos de algunas especies de alacranes _____	13
Tabla 3. Toxinas que actúan sobre canales iónicos caracterizadas en el veneno de <i>C. noxius</i> _____	14
Tabla 4. Perfiles transcripcionales de glándulas de veneno de algunas especies de alacrán _____	17
Tabla 5. Resultados del ensamblado del transcriptoma de <i>P. imperator</i> _____	19
Tabla 6. Número de lecturas por corrida de pirosecuenciación _____	29
Tabla 7. Datos cuantitativos de los cuatro ensamblados con Newbler y MIRA ____	32
Tabla 8. Resultados del ensamblado global con Newbler v2.5 cDNA _____	34
Tabla 9. Comparación pareada entre transcriptomas. Los transcriptomas de <i>I. scapularis</i> , <i>C. noxius</i> y <i>D. melanogaster</i> fueron comparados con dos estrategias de Blast (nucleótidos y proteínas) _____	43
Tabla 10. Genes eucariotes esenciales de copia única identificados en <i>C. noxius</i> (cobertura > 70%) _____	44
Tabla 11. Familias de toxinas identificadas en el transcriptoma de <i>C. noxius</i> ____	47
Tabla suplementaria 1. Genes eucariotes esenciales de copia única identificados en el transcriptoma de <i>C. noxius</i> _____	92
Tabla suplementaria 2. Isogrupos similares a toxinas _____	96

Índice de figuras

Figura 1. Sitios de convergencia neurotóxica _____	6
Figura 2. Principio básico de la pirosecuenciación _____	7
Figura 3. PCR en emulsión y pegado de las secuencias de cadena sencilla a las perlas de captura _____	8
Figura 4. Representación de las lecturas con grafos _____	11
Figura 5. Escala temporal de la evolución de los artrópodos _____	12
Figura 6. Frecuencia de los contigs, lecturas y singlets de <i>P. imperator</i> en rangos de longitud _____	19
Figura 7. Diagrama de flujo de la metodología _____	22
Figura 8. Detección de picos de profundidad _____	30
Figura 10. Características generales del ensamblado _____	35
Figura 11. Tipo de expresión de los transcritos _____	36
Figura 12. Número acumulado de secuencias ensambladas _____	38
Figura 13. Singlets y secuencias repetidas acumuladas _____	39
Figura 14. Clasificación taxonómica de los transcritos de <i>C. noxius</i> . _____	41
Figura 15. Secuencias ensambladas con homología en bases de datos. _____	42
Figura 16. Procesos metabólicos, celulares y de manejo de la información genética representados en el transcriptoma de <i>C. noxius</i> . _____	46
Figura 17. Alineamiento del alérgeno 5 de <i>T. serrulatus</i> con una secuencia de <i>C. noxius</i>	49
Figura 18. Alineamiento de la PBPO y 6 singlets de <i>C. noxius</i> _____	50
Figura 19. Filogenia de las secuencias similares a toxinas modificadoras de canales de sodio. _____	52
Figura 20. Alineamiento de las secuencias similares a LVP de <i>L. mucronatus</i> y <i>B.</i> <i>occitanus</i> _____	54
Figura 21. Alineamiento de los isotigs correspondientes a las β -toxinas de <i>C. noxius</i> ___	54
Figura 22. Alineamiento de las α y β toxinas identificadas a nivel transcripcional en <i>C.</i> <i>noxius</i> _____	55
Figura 23. Filogenia de las secuencias similares a toxinas bloqueadoras de canales de potasio. _____	56
Figura 24. Toxinas bloqueadoras de canales de potasio. _____	57
Figura 25. Alineamiento de los isotigs similares a toxinas de potasio tipo β . _____	58

Figura 26. Isotigs similares a toxinas bloqueadoras de canales de calcio _____	60
Figura 27. Alineamiento de isotigs con inhibidores de serin proteasas con dominios tipo Kunitz _____	60
Figura 28. Alineamiento del isotig similar a PLA2 secretorias de insectos _____	61
Figura 29. Filogenia de las secuencias similares a metaloproteasas de venenos de arácnidos _____	62
Figura 30. Alineamiento de una secuencia de <i>C. noxius</i> similar a la VMP1 del alacrán <i>M. eupeus</i> . _____	63
Figura 31. Alineamiento de los isotigs similares a la secuencia de antareasa de <i>T. serrulatus</i> . _____	64
Figura 32. Alineamiento de una secuencia de <i>C. noxius</i> similar a astacina de arácnidos _____	65
Figura 33. Alineamiento de una secuencia de <i>C. noxius</i> similar a la hialuronidasa del veneno de <i>M. martensii</i> _____	66
Figura 34. Alineamiento de un isotig similar a la proteína VP164 del veneno de <i>L. mucronatus</i> . _____	66
Figura 35. Alineamientos de isotigs similares a toxinas aisladas de venenos sin función descrita _____	68
Figura 36. Porcentaje de isogrupos con mayor expresión en glándula activa o en mantenimiento _____	71
Figura 37. Evaluación de los isogrupos con expresión diferencial consistente en los experimentos de pirosecuenciación _____	72
Figura 38. Isogrupos con diferencias de abundancia transcripcional validadas con el estadístico Q entre estados de glándula. _____	73
Figura 39. Categorías funcionales representadas de manera diferencial entre estados de glándula _____	75
Figura suplementaria 1. Géneros bacterianos representados en las secuencias ensambladas y los singlets de <i>C. noxius</i> _____	90
Figura suplementaria 2. Redes metabólicas representadas en el transcriptoma de <i>C. noxius</i> _____	91

Abreviaturas

Especies

ADO:	<i>Apis dorsala</i>
ASU:	<i>Anemonia sulcata</i>
BOC:	<i>Buthus occitanus tunetanus</i>
BOS:	<i>Buthus occitanus israelis</i>
CEL:	<i>Centruroides elegans</i>
CLL:	<i>Centruroides limpidus limpidus</i>
CNO:	<i>Centruroides noxius</i>
CSC:	<i>Centruroides sculpturatus</i>
CSS:	<i>Centruroides suffusus suffusus</i>
HJU:	<i>Hottentotta judaicus</i>
LIN:	<i>Loxosceles intermedia</i>
LHE:	<i>Latrodectus hesperus</i>
LMU:	<i>Lychas mucronatus</i>
MEU:	<i>Mesobuthus eupeus</i>
MMA:	<i>Mesobuthus martensii</i>
PGR:	<i>Parabuthus granulatus</i>
PSC:	<i>Parabuthus schlechteri</i>
RJU:	<i>Rhopalurus junceus</i>
TCO:	<i>Tityus costatus</i>
TDI:	<i>Tityus discrepans</i>
TSE:	<i>Tityus serrulatus</i>

Otras

AMP:	Péptidos anti-microbianos
CAP:	Cystein-rich secretory proteins (CRISP), Antigen 5 (Ag5), Pathogenesis-related (PR-1)
cDNA:	DNA complementario
ssDNA:	DNA de cadena sencilla
EST:	Expressed Sequence Tag
KEGG:	Kyoto Encyclopedia of Genes and Genomes
LCA:	Lowest Common Ancestor
LVPs:	Factores de activación de lipólisis
NJ:	Neighbor Joining
OLC:	Algoritmos de empalme (Overlap/Layout/Consensus)
PCR:	Reacción en cadena de la polimerasa
PLA2:	Fosfolipasa A2
aRNA:	RNA anti-sentido
SMasaD:	Esfingomielinasa D
VMPA:	Metaloproteasa presente en el veneno

Glosario

- Contig:** Secuencia ensamblada a partir de lecturas alineadas de manera contigua. Bajo los criterios de ensamblado de transcriptomas, se pueden considerar análogos de exones de un gen particular
- Debris:** Lecturas de baja calidad o longitud corta eliminados del ensamblado por MIRA.
- Isotig:** Secuencias ensamblada a partir de varias lecturas que alinean entre ellas; se puede considerar un análogo de una variante de *splicing* de un gen particular.
- Isogrupo:** Conjunto de isotigs o contigs conectados entre ellos por lecturas que divergen consistentemente hacia dos o más contigs diferentes; se puede considerar un análogo a un gen único. Ejemplo:

>isogrupo0001		numisotigs = 3		numcontigs = 5		(gen)
contig	0001	0002	0003	0004	0005	(exones)
isotig0003	-----	-----	-----	-----	-----	(variante de <i>splicing</i> 1)
isotig0004	-----	-----		-----	-----	(variante de <i>splicing</i> 2)
isotig0005	-----		-----	-----	-----	(variante de <i>splicing</i> 3)

- K-mero:** Motivos en una secuencia de DNA de “K” pares de bases de longitud
- Outlier:** Secuencia problemática que es excluida del ensamblado; esto puede deberse a que se trata de una secuencia quimérica o con regiones de baja calidad
- Repeat:** Secuencias cuya semilla alinea en más de 70% con al menos 70 lecturas más; estas secuencias se excluyen del ensamblado
- Singlet:** Secuencia que no alinea con ninguna lectura durante el ensamblado

Resumen

Los alacranes son quizás los animales terrestres más antiguos que se conocen. Existen más de 1500 especies alrededor del mundo, 200 de las cuales se han descrito en México (Hoffmann 2003). Al igual que otros animales venenosos, los alacranes poseen glándulas especializadas en la producción y secreción de veneno que se localizan en el último segmento postabdominal denominado telson. A nivel molecular y morfofisiológico, estos órganos han evolucionado por más de 400 millones de años para producir un arsenal de toxinas que actúan de manera selectiva sobre blancos exógenos con fines de predación y defensa.

En esta tesis se describen los resultados de un estudio transcriptómico realizado por pirosecuenciación a partir de RNA obtenido del alacrán de Nayarit, *Centruroides noxius*. Se hicieron tres experimentos independientes de pirosecuenciación con los sistemas GS20, GS-FLX y FLX-Titanium de 454, usando en cada uno de ellos tres librerías de cDNA construidas a partir de RNA extraído del cuerpo sin telson y de la glándula de veneno en estados activo y de mantenimiento. En total, el análisis incluyó más de tres millones de lecturas de longitudes variables (entre 100 y 350 pb), ensambladas en 26 672 isotigs de 950 nucleótidos de longitud promedio (aproximadamente 19 Mb de secuencia total) que se agrupan en 18 979 isogrupos y, en total, representan más del 80% de las lecturas aprovechadas. Más del 70% de los isogrupos o transcritos únicos se expresan tanto en el cuerpo como en la glándula de veneno; 24% se expresan de manera restringida en la glándula de veneno y el 3,5% únicamente en el cuerpo del alacrán.

Dentro de las secuencias glándula-específicas, se identificaron 72 isogrupos (0,4% del total de transcritos únicos ensamblados) similares a toxinas reportadas en otras especies de arácnidos y de anémonas marinas. Estos incluyen no sólo las toxinas caracterizadas a nivel bioquímico en el veneno de

Centruroides noxius, sino a transcritos similares a algunas toxinas bloqueadoras de canales de potasio tipo β , toxinas que afectan canales de calcio, zinc metaloproteasas, inhibidores de serinproteasas con dominios tipo Kunitz, alergenos, lipasas, péptidos antimicrobianos, entre otros componentes neurotóxicos cuya función y blanco específico no se han estudiado hasta el momento. Esta observación confirma que a lo largo de la historia evolutiva de estos arácnidos, han ocurrido diferentes eventos de reclutamiento y de duplicación de genes que han permitido generar una combinatoria de proteínas que los alacranes utilizan eficientemente en las glándulas de veneno.

La evaluación de los niveles de abundancia transcripcional reveló que el 3% y 2% del total de isogrupos ensamblados, mostraron mayor expresión en la glándula activa y en la glándula en estado de reposo, respectivamente. Entre estos transcritos, destacan 23 isogrupos similares a toxinas, 15 de los cuales se expresan preferentemente en la glándula activa y 8 en estado de reposo.

Los resultados obtenidos en este trabajo representan el primer estudio a gran escala que describe de manera integral el universo de ESTs presentes en el alacrán *Centruroides noxius*, un organismo de enorme relevancia médica y evolutiva.

1 Introducción

La construcción y tamizado de bancos de cDNA proveniente de las glándulas de veneno, así como caracterizaciones bioquímicas y recientemente, el uso de perfiles proteómicos de los venenos completos de diferentes especies de alacranes, han permitido identificar aproximadamente 600 secuencias similares a toxinas en diferentes especies de alacranes. Esta información ha ayudado a entender de manera parcial la complejidad de los venenos, dejando abierta la posibilidad de que aun estemos lejos de conocer la gama completa de péptidos activos que los conforman. Adicionalmente, es importante resaltar que se conoce poco sobre los procesos celulares que ocurren en el interior de la glándula y que están detrás de la conformación de una mezcla proteica tan diversa como lo es el veneno de estos artrópodos.

La implementación de estrategias de análisis a gran escala en el estudio de animales venenosos, que en conjunto se conoce como “venomics”, es una excelente alternativa para estudiar a fondo la complejidad biológica de estos organismos. En particular, un enfoque transcriptómico basado en metodologías eficientes como la pirosecuenciación, tiene un enorme potencial tanto en el campo de la toxinología como desde una perspectiva evolutiva, pues ofrece un panorama integral de las redes de interacciones comunes con otros eucariotes, así como de los procesos y de las familias de genes específicas de alacrán.

1.1 Venenos y Toxinas

La capacidad de producir y secretar veneno desarrollada por un gran número de animales, entre los que se encuentran algunos artrópodos, reptiles, mamíferos, cefalópodos y animales marinos, puede considerarse una importante innovación evolutiva, fundamentalmente utilizada para predación y defensa. Los venenos son mezclas complejas de sales, lípidos, poliaminas y diferentes

combinaciones de proteínas. Estas últimas son el resultado de eventos de reclutamiento, que típicamente involucran la duplicación de genes clave en procesos regulatorios y que, posterior a la duplicación, se expresan de manera selectiva en la glándula de veneno. Se han identificado estructuras proteicas comunes, altamente conservadas en los venenos (Fig.1), entre las que destacan las proteínas del grupo CAP [(*Cystein-rich secretory proteins* (CRISP), *Antigen 5* (Ag5), *Pathogenesis-related* (PR-1)], quitinasas, defensinas, hialuronidasas, Kunitz, lectinas, péptidos natriuréticos, proteasas, fosfolipasas A2 (PLA2), esfingomielinasa D (SMasaD), entre otras. Estos componentes actúan de manera específica sobre una gran variedad de blancos exógenos, como receptores de membrana, canales iónicos y otras proteínas citosólicas, y su efecto biológico reside en alguno de los siguientes mecanismos generales: daños estructurales causados por la catálisis de la hidrólisis de sustratos universalmente presentes (ej. SMasaD, PLA2, hialuronidasa); desbalance fisiológico o respuestas de corta duración causadas por el mimetismo de proteínas endógenas como si estas fueran sobre expresadas (ej. peptidasa de tipo S1); mimetismo de proteínas endógenas que actúan como inhibidores competitivos y causan alteraciones en respuestas fisiológicas (revisado por Fry *et al.* 2009). Todas las toxinas se generan a partir de precursores que muestran un péptido señal de secreción en el extremo N-terminal, el cual se escinde para dar lugar a la proteína madura cuya estructura terciaria se estabiliza generalmente debido a la formación de numerosos puentes disulfuro.

En el caso particular de los alacranes, los componentes del veneno mejor caracterizados son los péptidos que reconocen canales iónicos y receptores en membranas excitables. Estas toxinas se han clasificado en función de la especificidad de las especies a las que pueden afectar (mamíferos, insectos, crustáceos), de sus receptores blanco, la longitud de las secuencias (largas o cortas), el mecanismo de acción y sitios de unión (α y β). Las toxinas de cadena corta interfieren con el funcionamiento de los canales de potasio (Tytgat *et al.* 1999; García *et al.* 2001), mientras que las toxinas de cadenas más largas (59-76 residuos de longitud) modifican el funcionamiento de algunos canales de sodio

(Possani *et al.* 1999; Cestèle & Catterall, 2000; Ramírez-Domínguez *et al.* 2002; Gordon & Gurevitz, 2003). Dentro de esta última categoría existen dos subfamilias con diferentes sitios de unión al canal. Las de tipo alfa (α -NaTx) se unen al sitio 3 y disminuyen la velocidad del proceso de desactivación del canal al impedir el movimiento del segmento S4 durante la despolarización de la membrana. Las de tipo beta (β -NaTx) se unen al sitio 4 e inducen un cambio del potencial de activación del canal hacia potenciales de membrana hiperpolarizados. Otras toxinas actúan de manera específica sobre canales de calcio (Valdivia & Possani, 1998; Chuang *et al.* 1998; Olamendi-Portugal *et al.* 2002) y canales de cloro (DeBin *et al.* 1993).

De las estructuras antes mencionadas, también se han identificado (a nivel transcripcional y/o preteómico) algunas secuencias similares a defensinas (Conde *et al.* 2000; Torres-Larios *et al.* 2000; Corzo *et al.* 2001), hialuronidasas (Feng *et al.* 2010), fosfolipasas (Zamudio *et al.* 1997; Conde *et al.* 1999) y metaloproteasas (Fletcher *et al.* 2009). Por otra parte, se ha observado que algunos componentes proteicos del veneno tienen actividad analgésica (Guan *et al.* 2001) o pueden ser utilizados como agentes anti-epilépticos (Corona *et al.* 2003). En muchos de estos ejemplos se desconoce el papel que estas familias de toxinas puedan jugar en eventos de envenenamiento en mamíferos, pero es claro que la presencia de esta gama de toxinas en la glándula de veneno es el resultado de diferentes eventos de reclutamiento y duplicación de genes que han permitido generar una combinatoria de proteínas que los alacranes utilizan eficientemente con fines de depredación y defensa.

Se estima que en conjunto, estos componentes proteicos conforman entre 35 y 70% del veneno total de los alacranes extraído por estimulación eléctrica. Estas observaciones ponen en evidencia el enorme potencial de las toxinas de alacrán en el campo de la farmacología y de la biotecnología.

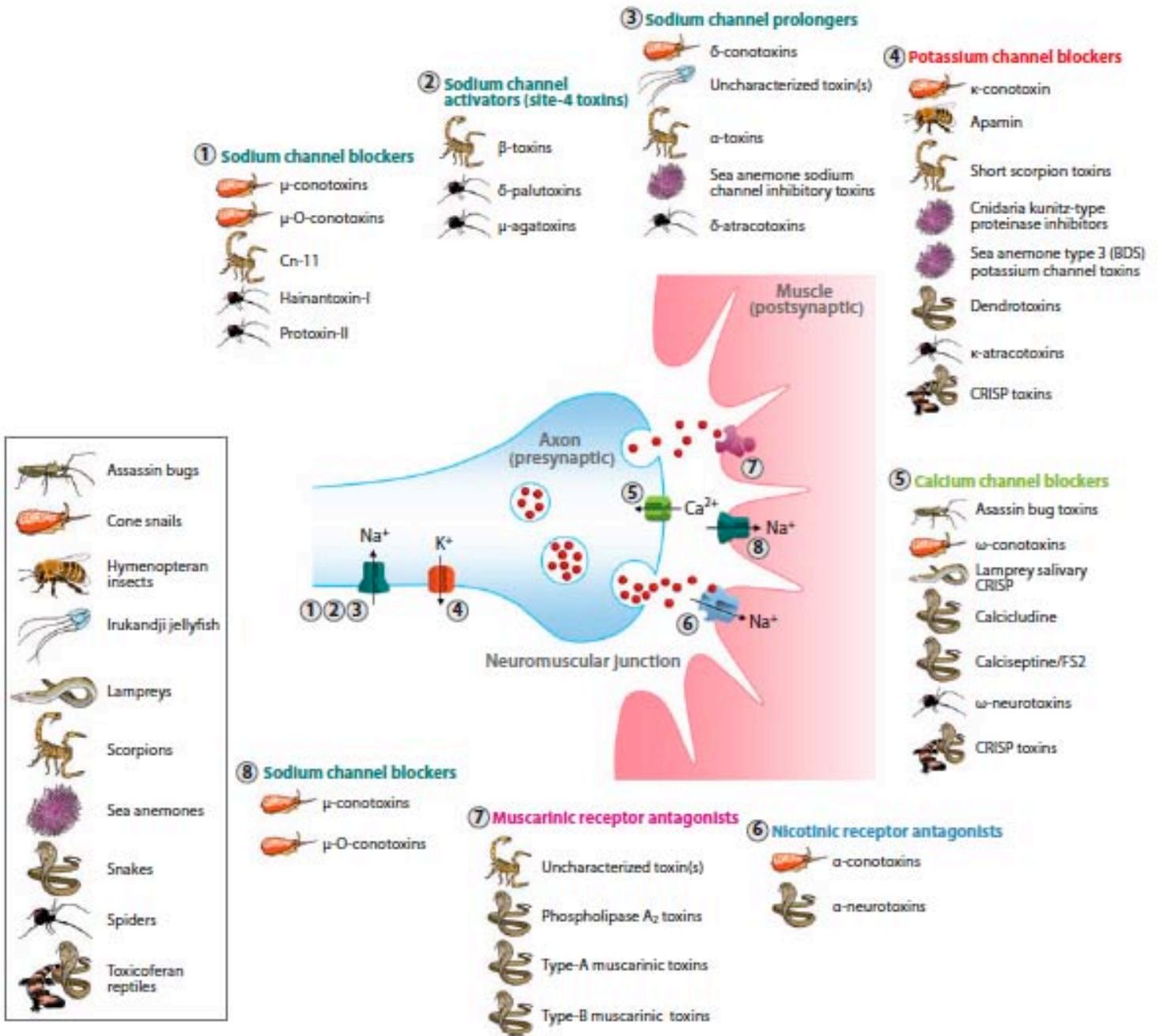


Figura 1. Sitios de convergencia neurotóxica (Fry *et al.* 2009)

1.2 Pirosecuenciación

La pirosecuenciación es una estrategia de secuenciación por síntesis, masiva y en paralelo de DNA, que se basa en la detección de luz producida cada vez que se incorpora un nucleótido complementario a la cadena molde o templado (Fig. 2). El protocolo de pirosecuenciación comprende cuatro pasos fundamentales: la generación de una librería de DNA de cadena sencilla (ssDNA); amplificación clonal en emulsión por reacción en cadena de la polimerasa (PCR) de la librería; secuenciación por síntesis y análisis de datos (Margulies *et al.* 2005).

El principio básico de esta metodología (Fig. 2), reside en aprovechar el pirofosfato (PPi) que se libera cada vez que se forma un nuevo enlace fosfodiéster cuando un nucleótido complementario a la cadena de DNA que funge como templado, se incorpora a la cadena en crecimiento. En presencia de adenosin 5' fosfosulfato (APS), la sulfurilasa transfiere el PPi al APS, liberando ATP y sulfato. El ATP generado permite que la luciferasa convierta la luciferina en oxiluciferina, que genera luz visible proporcional a la cantidad de ATP presente y por tanto, proporcional al número de nucleótidos incorporados.

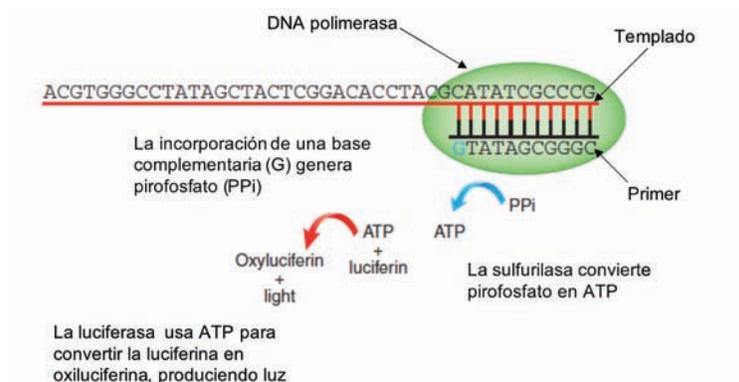


Figura 2. Principio básico de la pirosecuenciación. La incorporación de un nucleótido complementario libera un pirofosfato que la sulfurilasa convierte en ATP para ser usado por la luciferasa y producir una señal luminosa.

Para la construcción de la librería, la muestra de DNA o cDNA debe ser fragmentada, ligada a adaptadores (necesarios para pasos posteriores de purificación, amplificación y secuenciación) en los extremos 5' y 3' y separada en segmentos de cadena sencilla. Estos segmentos se inmovilizan en perlas de

captura bajo condiciones que favorecen la unión de un solo segmento de la librería por perla. Posteriormente, las perlas se encierran en gotas que funcionan como “microreactores” en una emulsión en aceite que contiene los reactivos necesarios para la amplificación clonal por PCR (Fig. 3). Con este paso de amplificación, cada perla se enriquece con millones de copias de un solo tipo de secuencia molde. La emulsión se rompe y las hebras de DNA se desnaturalizan; las perlas que llevan unidas las secuencias templado de cadena sencilla se depositan en pozos en una placa de fibra óptica. Adicionalmente, en cada pozo se introducen perlas a las cuales se han fijado las enzimas necesarias para la reacción de secuenciación. El sistema de flujo del secuenciador deja pasar sucesivamente los nucleótidos de manera individual en un orden fijo, así como otros sustratos y amortiguadores necesarios para las reacciones. La señal luminosa producida, proporcional al número de bases incorporadas en cada flujo de nucleótidos, se capta por una cámara CCD. De esta forma, la combinación de la intensidad luminosa y la información de las posiciones de cada pozo a lo largo de la placa, permite determinar la secuencia de cada fragmento de la librería (Margulies *et al.* 2005).

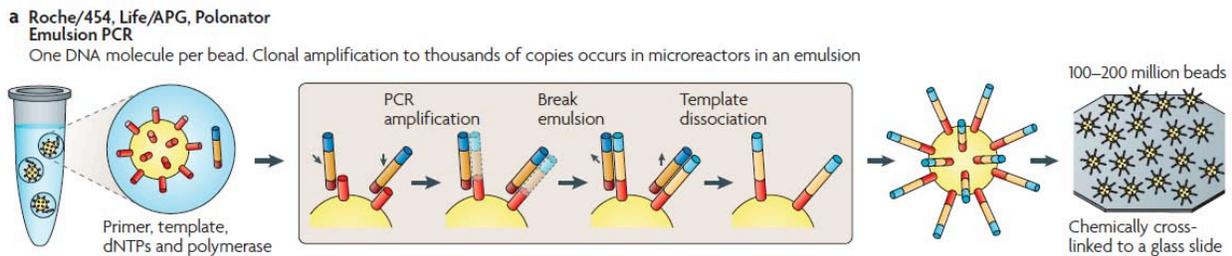


Figura 3. PCR en emulsión y pegado de las secuencias de cadena sencilla a las perlas de captura (Metzker, 2009)

Esta forma de secuenciación masiva presenta muchas ventajas respecto a otros sistemas recientemente desarrollados. En primer lugar, es importante resaltar que no se requiere una amplificación de la muestra de DNA previa a la construcción de la librería. Aplicando este principio a estudios de transcriptómica, se vuelve evidente que el número de lecturas obtenidas refleja de manera directa la abundancia del transcrito en las condiciones de estudio. Por ello, es posible

tener un panorama cualitativo y cuantitativo de las secuencias presentes al momento de la extracción del material genético. Otra ventaja importante sobre otros métodos de secuenciación masiva, es que permite hacer ensamblados *de novo* de manera confiable (Huse *et al.* 2007). En efecto, las lecturas cortas generadas por equipos como Illumina o Solid, se ensamblan principalmente sobre anclas de secuencia conocidas. Esto se vuelve una limitante cuando no se tiene información previa de la muestra de DNA o del organismo de interés. Por ello, las lecturas de hasta 400 nucleótidos que se obtienen por pirosecuenciación permiten generar ensamblados más robustos sin necesidad de tener conocimiento previo de la muestra que se desea secuenciar (Tabla 1).

Tabla1. Comparación de algunas plataformas de secuenciación masiva

Plataforma	Tipo de secuenciación	Longitud de las lecturas (pb)	Gb generadas por corrida	Algunas aplicaciones / ventajas	Acceso web
Roche 454 GS FLX Titanium	PS	~400	~0.5	Secuenciación de genomas complejos <i>de novo</i> , transcriptómica	454.com
Illumina	TR	100	10-95*	Re-secuenciar genomas conocidos, análisis de	www.illumina.com
AB SOLiD	SL	75	30 – 50*	polimorfismos	www.appliedbio systems.com
Helicos	TR	35	21 – 35*	Permite secuenciar moléculas individuales de DNA	www.helicosbio.com
Ion Torrent	DH (pH)	~200	0.1 – 1*	Secuenciación de genomas y transcriptomas pequeños de baja complejidad	www.iontorrent.com

*Gb generadas en función de los tiempos de corrida. PS: pirosecuenciación; TR: terminadores reversibles; SL: secuenciación por ligación; DH: detección de iones H⁺

1.3 Algoritmos de ensamblado

El uso cada vez más frecuente de plataformas de secuenciación a gran escala ha hecho indispensable el desarrollo de algoritmos de ensamblado, capaces de soportar números elevados de lecturas de longitudes cortas, tanto para ensamblados *de novo*, como para mapear secuencias sobre anclas genómicas conocidas. En general, estos algoritmos se clasifican en tres tipos: “glotones” (*greedy*) basados en gráficos; de empalme y los que utilizan la aproximación de Brujin (Miller *et al.* 2010; Kumar *et al.* 2010). Los dos primeros se han usado ampliamente para ensamblar secuencias obtenidas con Sanger y pirosecuenciación, mientras que el tercer tipo ha probado ser mas útil para lecturas cortas obtenidas con sistemas como Illumina o Solid.

El primer tipo de algoritmo aplica como principio básico agregar una a una, nuevas lecturas a una secuencia de base, repitiendo esta operación hasta que no es posible seguir añadiendo lecturas a la secuencia de partida. Para ello, se utilizan gráficos, donde cada nodo representa una lectura y los vértices corresponden a los empalmes que hay entre las secuencias, los cuales se califican haciendo alineamientos pareados (Fig. 4). Este tipo de algoritmos, simplifican la construcción de los gráficos, tomando en cuenta únicamente los empalmes mejor calificados, según diferentes funciones de evaluación. Este es el caso de MIRA (Chevreux, *et al.* 2004), cuya aplicación en el ensamblado de las lecturas de *C. noxius* será discutida más adelante.

Los algoritmos de empalme OLC (por su nombre en inglés: *Overlap/Layout/Consensus*) como Newbler o CABOG de Celera, tienen dos etapas fundamentales: se hacen alineamientos pareados entre todas las lecturas de entrada para generar “unitigs” con empalmes de alta calidad. Este paso se vuelve eficiente al aplicar un algoritmo heurístico de extensión de semilla, que busca k-meros (motivos de longitud k) en las lecturas, para definir candidatos para alineamientos posteriores entre aquellas secuencias que comparten dicha semilla. La longitud de la semilla, el tamaño mínimo del empalme, así como el

porcentaje mínimo de identidad del alineamiento, son los tres criterios más importantes para lograr mayor robustez en la generación de unitigs. Posteriormente, la operación de búsqueda se repite entre los unitigs obtenidos, para formar contigs más largos y finalmente, scaffolds o metacontigs.

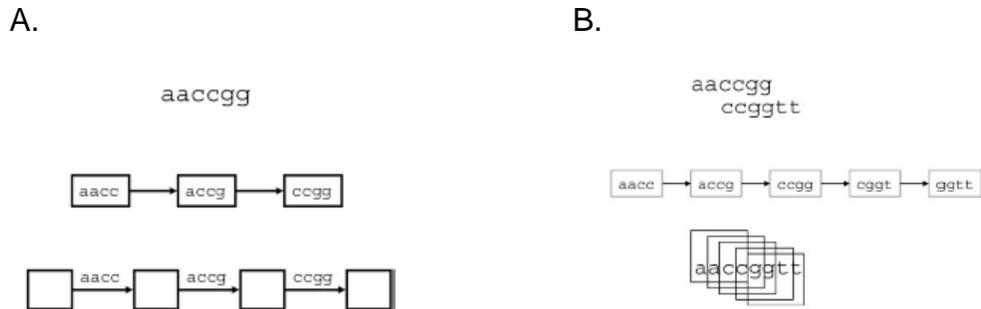


Figura 4. Representación de las lecturas con grafos. A. Una lectura está representada por K-meros de longitud $K=4$. El grafo tiene un nodo para cada K-mero de la lectura, y un vértice que une cada pareja de K-meros que se empalman en $K-1$ bases. Alternativamente, cada vértice representa un K-mero y los nodos el empalme de $K-1$ bases. B. Un alineamiento pareado se representa con K-meros que incluyen ambas lecturas (Miller *et al.* 2010).

2 Antecedentes

La importancia en términos de salud pública de los alacranes, los cuales causan alrededor de un cuarto de millón de casos de envenenamiento en Norteamérica cada año (Boyer *et al.* 2009), ha sesgado el estudio de los venenos hacia la caracterización de péptidos con efectos neurotóxicos. Si bien hemos logrado describir de manera parcial la complejidad composicional de los venenos a nivel proteómico y bioquímico, la información referente a los procesos celulares y de regulación que dan origen a esta mezcla proteica, la organización genómica de estos organismos y las posibles variaciones inter-especie, es muy limitada. Tomando en cuenta que los primeros fósiles taxonómicamente identificables de animales terrestres en conjunto con análisis filogenómicos (Jeyaprakash *et al.* 2009; Pisan *et al.* 2004), han fechado el origen de los alacranes hace más de 400 millones de años (Fig. 5) y que el tiempo de divergencia entre los arácnidos y los insectos ocurrió hace más de 700 millones de años, es claro que el estudio de los alacranes debe abordarse desde perspectivas médicas, farmacológicas y evolutivas.

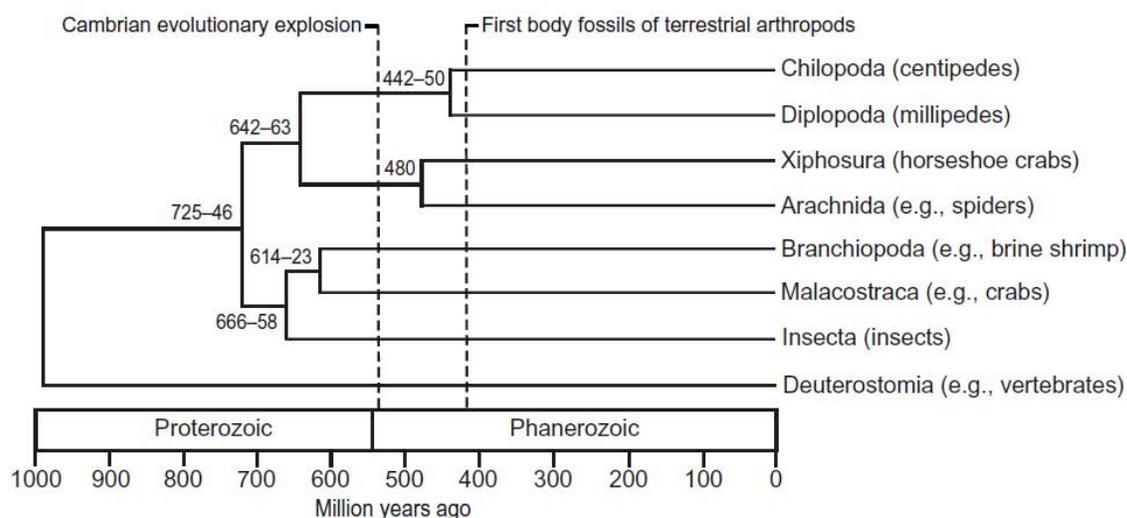


Figura 5. Escala temporal de la evolución de los artrópodos (Pisan *et al.* 2004).

2.1 Estado del arte en la genómica de los alacranes

La información referente al tamaño del genoma del alacrán, el número de cromosomas y la ploidía de estos organismos se ha discutido en algunos reportes sin llegar a un consenso. En particular, los cariotipos hasta el momento estudiados de las diferentes familias de alacranes varían enormemente. Dentro de la familia Buthidae, los géneros *Tityus* y *Mesobuthus* son los mejor estudiados en este sentido. Se estima que el genoma haploide del alacrán *Mesobuthus martensii* Karsch comprende cerca de 600 Mpb, el cual se calculó por citometría de flujo usando células del aparato reproductor femenino y masculino (Li *et al.* 2009). Los cromosomas varían entre familias de alacranes tanto en número (de $2n < 10$ hasta $2n > 100$) (Schneider *et al.* 2009 y 2010) como en morfología (Tabla 2).

Tabla 2. Cariotipos de algunas especies de alacranes

Especie	Numero diploide	Tipo de cromosomas
Buthidae		
<i>Androctonus australis, bicolor</i>	24	Holocéntrico
<i>Centruroides exilicauda, vittatus</i>	26	Holocéntrico
<i>Tityus serrulatus</i>	12	Probablemente holocéntrico
Iuroidea		
<i>Hadrurus hirsutus</i>	~100	-----
Scorpionidae		
<i>Heterometrus gravimanus</i>	112	Monocéntrico
<i>Pandinus imperator</i>	120	M, A
Urodacidae		
<i>Urodacus armatus</i>	124, 144	M, A, T
Liochelidae		
<i>Liocheles australasiae</i>	54 - 64	M, T

M: metacéntrico; A: acrocéntrico; T: telocéntrico

Si bien el cariotipo de *C. noxius* no se ha estudiado, podemos inferir basados en los datos de especies de la familia Buthidae (Tabla 2), que el número de cromosomas se encuentra alrededor de $2n \sim 26$ y que muy probablemente estos son de tipo holocéntrico.

2.2 Toxinas caracterizadas en el veneno de *Centruroides noxius*

La relevancia médica de las neurotoxinas, así como su utilidad como herramientas farmacológicas en el estudio del funcionamiento de canales iónicos, ha sesgado el estudio del veneno de los alacranes hacia la caracterización de dichos péptidos (Rodríguez de la Vega *et al.* 2010). Un claro ejemplo es el caso del alacrán *Centruroides noxius*, que es la especie de mayor toxicidad en México y cuyos componentes neurotóxicos se han estudiado bioquímicamente. Como se muestra en la tabla 3, en esta especie se han aislado tanto toxinas moduladoras de la función de canales de sodio (NaTx) (Ramírez-Dominguez *et al.* 2002; del Río-Portilla *et al.* 2004) como bloqueadoras de canales de potasio (KTx) (Nieto *et al.* 1996; García-Valdés 2001; Corona *et al.* 2002), con blancos de acción en mamíferos, insectos y crustáceos.

Tabla 3. Toxinas que actúan sobre canales iónicos caracterizadas en el veneno de *C. noxius*

Tipo de Toxinas	Blanco / Familia	Nombre
Toxinas modificadoras de canales de sodio (NaTx)	Específicas para mamíferos / beta toxinas	Cn2, Cn3, Cn4, Cn6-9
	Específicas para insectos y crustáceos / alfa-beta toxinas	Cn1, Cn5, Cn10, Cn11, Cn12
Toxinas bloqueadoras de canales de potasio (KTx)	Canales de tipo Maxi-K en mamíferos / toxinas de cadena corta	α -1.11 Slotoxina
	Mamíferos y grillos / toxinas de cadena corta	α -2.1 Noxiustoxina α -2.4 Noxiustoxina2
	Canales de tipo Shaker B y canales de potasio tipo Kv1.1, Kv1.2 y Kv1.3 / toxinas de cadena corta	α -10.1 Cobatoxina α -10.2 Cobatoxina
	Canales de tipo K ⁺ -Erg en distintas especies / Ergtoxinas	γ -1.1 CnErg1 γ -3.1 CnErg2 γ -4.13 CnErg3 γ -4.11 CnErg4 γ -4.2 CnErg5

Si bien es cierto que el estudio de los animales venenosos se ha concentrado fundamentalmente en la identificación y caracterización bioquímica de moléculas con posibles implicaciones médicas y farmacológicas, las aproximaciones “ómicas” se han convertido en herramientas clave para entender la complejidad de estos organismos. La transcriptómica en particular, se ha usado de manera recurrente para entender diferentes funciones de las glándulas de veneno de algunas especies de alacranes como discutiremos en las siguientes secciones.

2.3 Perfiles transcripcionales de glándulas de veneno de alacranes

Dada la complejidad composicional de los venenos de los alacranes observada en experimentos de separación por cromatografía líquida de alta presión (HPLC) (revisado por Rodríguez de la Vega *et al.* 2010), así como la falta de información referente a los procesos celulares detrás del ensamblado del arsenal de toxinas, en los últimos años se ha visto un creciente interés por abordar el estudio de los alacranes con base en el análisis de las secuencias expresadas en la glándula de veneno, denominadas ESTs (por sus siglas en inglés “Expressed Sequence Tags”). Sin duda, un estudio pionero en el uso de esta aproximación fue la secuenciación de ESTs de la glándula del alacrán mexicano *Hadrurus gertschi* (Schwartz *et al.* 2007), en el que se obtuvieron 147 secuencias de calidad que permitieron estudiar el contexto celular de la glándula de veneno. Otros estudios similares se han hecho en especies tanto de la familia Buthidae (*Lychas mucronatus*, Ruiming *et al.* 2010; *Buthus occitanus israelis*, Kozminsky-Atias *et al.* 2008; *Tityus discrepans*, D’Suze *et al.* 2009) como de las familias Scorpionidae y Euscorpidae (*Scorpiops jendeki*, Ma *et al.* 2009; *Heterometrus petersii*, Ma *et al.* 2010), mostrando diferencias importantes en los perfiles transcripcionales en términos de abundancia y diversidad de secuencias similares a toxinas (Tabla 4). Estos reportes comparten una base metodológica: las librerías de cDNA se construyeron a partir de RNA proveniente de glándulas entre 2 y 5 días después de la extracción de veneno por estimulación eléctrica, lo cual implica que la

glándula está comprometida en el proceso de regeneración del veneno. Esta condición resulta ser muy útil para explorar la diversidad de transcritos de toxinas que no necesariamente se expresan de manera activa una vez que el veneno ha sido producido. En efecto, el análisis de las glándulas en estado activo o de regeneración mostró un importante enriquecimiento de los transcriptomas con secuencias similares a toxinas, teniendo entre 50 -78% en el caso de la familia Buthidae, y 30 -44% en las familias Luridae, Scorpionidae y Euscorpidae, del total de los ESTs clasificados como componentes proteicos del veneno. Además de las toxinas previamente caracterizadas (fundamentalmente neurotoxinas y otras toxinas de canales iónicos), otros potenciales componentes del veneno como LVPs (por su nombre en inglés "lipolysis activating factors"), fosfolipasas A2 (PLA2), péptidos antimicrobianos (AMP), metaloproteasas, proteínas salivales secretorias de garrapata, precursores de péptidos citolíticos y otras proteínas ricas en cisteínas con péptidos de secreción, se identificaron a nivel transcripcional.

A diferencia de los estudios antes mencionados, Morgenstern y colaboradores analizaron ESTs provenientes de glándulas en estado de mantenimiento (Morgenstern *et al.* 2011). En este caso, la librería de cDNA se construyó a partir de un sólo individuo de la especie *Hottentota judaicus* perteneciente a la familia Buthidae, el cual se mantuvo sin alimento por 14 días para evitar la estimulación de la glándula en la producción de veneno y no fue ordeñado previo a la remoción del telson. Comparado con otras especies, el perfil transcripcional mantuvo un nivel de secuencias similares a toxinas bajo (incluso menor que en especies no buthides,) correspondiente al 24% del total de ESTs, y mostraba proporciones atípicas de las diferentes familias de toxinas. En efecto, se observó que la abundancia de toxinas modificadoras de canales de sodio era relativamente baja, mientras que las toxinas bloqueadoras de canales de potasio eran las más abundantes, mostrando un patrón de expresión poco frecuente en la familia Buthidae. Además, la proporción de las clases más importantes de toxinas de sodio mostró que las tipo α estaban claramente sub-representadas con una relación $\alpha:\beta$ de 24:76 (comparado con una relación 40:20 en *B. occitanus*).

Tabla 4. Perfiles transcripcionales de glándulas de veneno de algunas especies de alacrán

Species	Family	Sequences	%EST	Reference
* <i>Hadrurus gertschi</i>	Luridae	147 ESTs 68 clusters	31% CV 19% PV 50% N/H	Schwartz <i>et al.</i> 2007
* <i>Centruroides noxius</i> Hoffmann	Buthidae	126 ESTs	55% CV 30% PC 15% N/H	datos no publicados; Carreño Campos, 2007
* <i>Buthus occitanus</i> <i>israelis</i>	Buthidae	450 ESTs	78% CV - -	Kozminsky-Atias <i>et al.</i> 2008
* <i>Tityus discrepans</i>	Buthidae	112 ESTs 51 clusters	50% CV 13% PC 37% N/H	D'Suze <i>et al.</i> 2009
* <i>Scorpiops jendeki</i>	Euscorpidae	871 ESTs 293 clusters	40% CV 30% PC 24% N/H	Ma <i>et al.</i> 2009
* <i>Lychas mucronatus</i>	Buthidae	738 ESTs 380 clusters	55% CV 22% PC 23% N/H	Ruiming <i>et al.</i> 2010
* <i>Heterometrus</i> <i>petersii</i>	Scorpionidae	486 ESTs 184 clusters	68% CV 20% PC 12% N/H	Ma <i>et al.</i> 2010
** <i>Hottentota judaicus</i>	Buthidae	537 ESTs 283 clusters	24% CV 39% PC 37% N/H	Morgenstern <i>et al.</i> 2011

CV, componente del veneno; PC, procesos celulares; N/H, no determinado/hipotético. *Glándula activa o en regeneración. ** Glándula en mantenimiento o reposo.

Por otra parte, se observó que una proporción significativa de los transcritos similares a toxinas (NaTx en particular) tenían características que muy probablemente impedían el procesamiento correcto de estos durante la traducción en la glándula. Por ejemplo, se encontraron ESTs similares a la toxina km-BUTX-Hj1c sin la secuencia correspondiente al péptido señal que, una vez comparados con la organización genómica de esta toxina, parecían haber perdido el exón 1 que incluye además una parte de la región 5'UTR; en otros casos se había perdido el codón de paro, o los ESTs mostraban una acumulación anormal de mutaciones.

Sorprendentemente, la toxina BjxtrIT que es considerada como el componente mayoritario del veneno de este alacrán (~1% del peso seco del veneno crudo) no se detectó a nivel transcripcional. En conjunto, estas observaciones indican que el transcriptoma no refleja directamente la composición del veneno, y más bien parece ser un indicador de los niveles de transcritos necesarios para mantener la glándula en un estado basal.

Algunos de los procesos celulares comúnmente representados en estas librerías, corresponden a transcripción, metabolismo de proteínas y transporte. Cabe mencionar que en cada uno de los reportes antes mencionados, se obtuvieron ESTs que no tienen similitud significativa en las bases de datos. Esto puede implicar que existen transcritos cuya función no se ha descrito, y que podrían representar nuevas familias de genes específicas del alacrán. En conjunto, esta información es sumamente interesante, ya que a diferencia de los perfiles proteómicos realizados previamente, ofrecen una visión integral de los procesos biológicos que ocurren en la glándula de veneno.

Hasta este momento, el uso de estrategias de secuenciación masiva, en particular de la pirosecuenciación, ha sido empleada únicamente en el estudio del transcriptoma del alacrán africano *Pandinus imperator* (Roeding *et al.* 2009). En este estudio, se obtuvieron alrededor de 429,000 lecturas con una longitud promedio de 223 ± 50 nt, que fueron ensambladas en 8334 contigs (Tabla 5) con el programa MIRA (Fig. 6). Haciendo una búsqueda por homología de estas secuencias sobre las bases de datos UniProt, FlyBase y de otros quelicerados, se observó que aproximadamente 72% de las secuencias ensambladas tienen identidad significativa con al menos una secuencia de las bases de datos utilizadas, tomando como único criterio de corte un valor de expectación de $1e-04$. A diferencia de los trabajos antes descritos, este estudio no busca identificar familias de toxinas y se enfoca en un análisis filogenómico con familias multigénicas, para lo cual se usaron 149 genes ortólogos de 67 taxa diferentes, teniendo como objetivo entender la relación entre artrópodos incluyendo los datos de *Pandinus imperator*. Si bien ofrece un punto de referencia interesante por la

diversidad de secuencias generadas, no es posible discernir entre los ESTs glándula específicos y las secuencias que se expresan de manera indistinta en el organismo dado que las librerías de cDNA se construyeron a partir de transcritos del cuerpo completo del alacrán incluyendo el telson.

Tabla 5. Resultados del ensamblado del transcriptoma de *P. imperator*

reads	contigs	singlets	debris
428, 844	8, 334	26, 147	70, 082
223 ± 50pb	422 ± 50pb	226 pb	256 pb

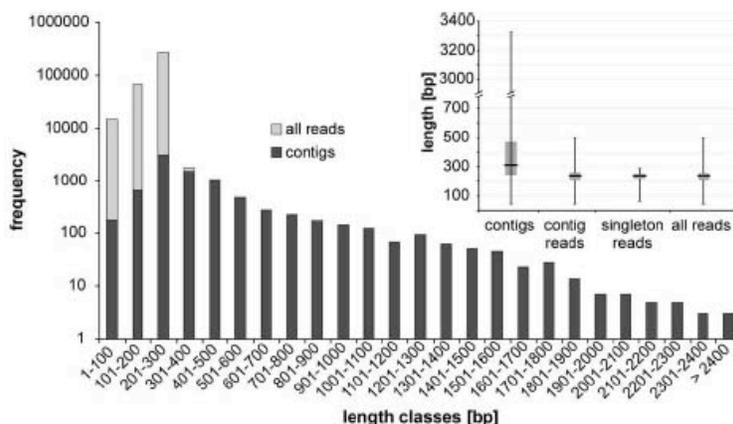


Figura 6. Frecuencia de los contigs, lecturas y *singlets* de *P. imperator* en rangos de longitud

Por todo lo anterior, es interesante plantear la posibilidad de recurrir a metodologías como la pirosecuenciación, para lograr un estudio más completo del universo transcripcional de los alacranes, con un especial énfasis en las glándulas de veneno. Dado que la pirosecuenciación es hoy en día una herramienta poderosa capaz de generar más de 500,000 secuencias de calidad con longitudes de entre 100 y 450 nucleótidos (Margulies *et al.* 2005; Huse *et al.* 2007), permite no sólo cubrir la diversidad transcripcional de los órganos productores de veneno (caracterización cualitativa), si no que nos da la oportunidad de hacer un estudio cuantitativo en términos de niveles de expresión génica. Un estudio de esta

naturaleza permitirá además generar información que sentará las bases para estudios moleculares posteriores a gran escala.

3 Hipótesis

Existen diferencias a nivel transcripcional entre el estado de reposo y el estado activo de la glándula de veneno de alacrán, cuantificables en términos de abundancia y diversidad de secuencias expresadas que pueden identificarse por pirosecuenciación.

4 Objetivo general

Identificar las diferencias entre los perfiles transcripcionales de la glándula de veneno del alacrán *C. noxius*, comparando la diversidad y abundancia de secuencias expresadas en estados de reposo y actividad.

5 Objetivos particulares

- Obtener un ensamblado *de novo* confiable de las secuencias obtenidas con las lecturas de pirosecuenciación.
- Identificar las secuencias que se expresan específicamente en el cuerpo del alacrán, en el telson o de manera indistinta en el organismo.
- Proponer una estrategia de anotación para asignar función a las secuencias ensambladas.
- Identificar las familias de toxinas que se expresan en la glándula de veneno.
- Comparar con pruebas estadísticas la abundancia transcripcional de las secuencias glándula específicas antes y después de la ordeña.

6 Metodología

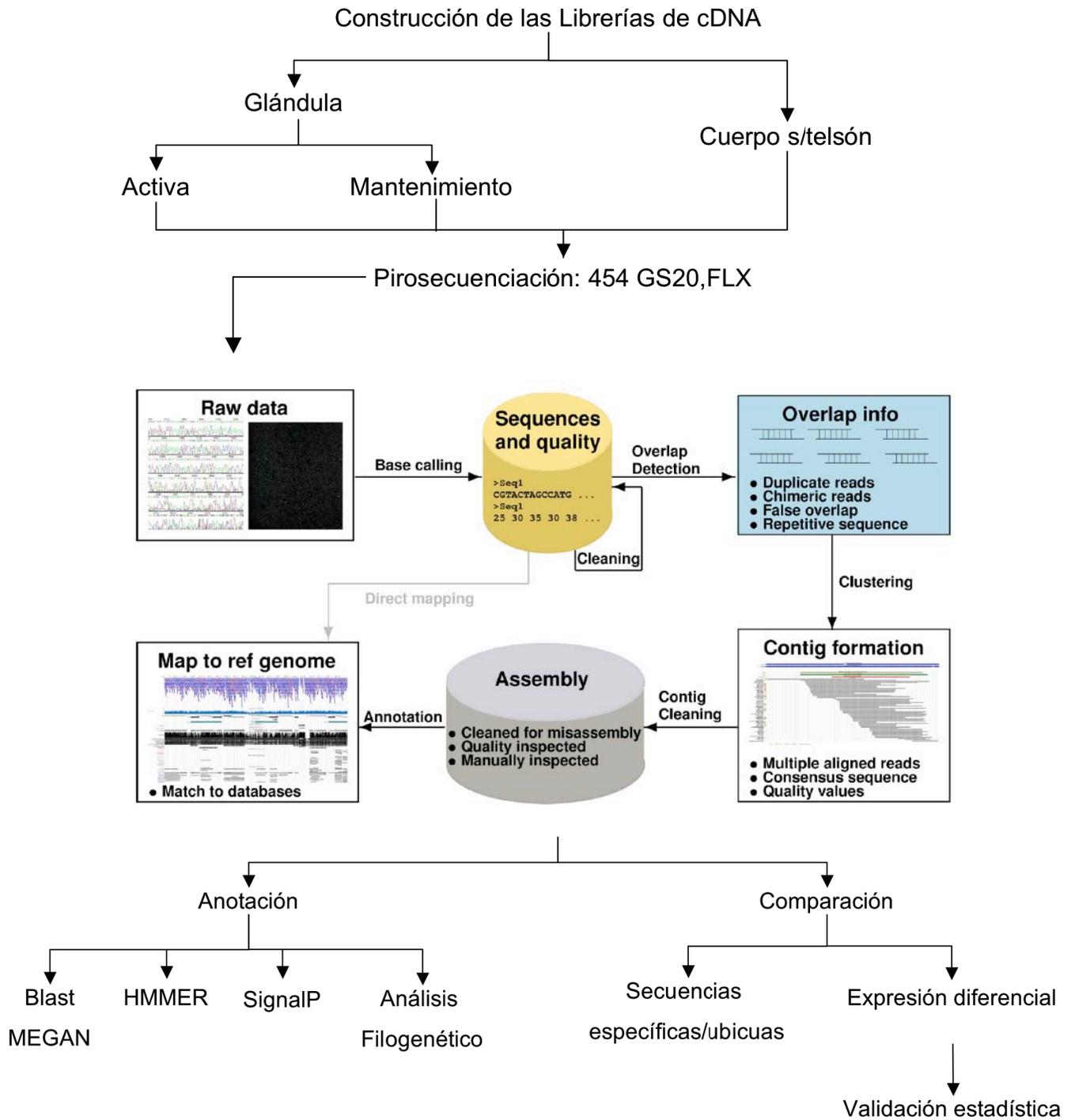


Figura 7. Diagrama de flujo de la metodología

6.1 Preparación de las librerías de cDNA

Se construyeron 9 librerías de cDNA de la especie *Centruroides noxius* Hoffmann, 1932, a partir de RNA extraído del cuerpo de un alacrán sin telson (3 librerías) y del telson de 20 individuos en dos estados diferentes:

- Activo o de regeneración: el telson se removió 5 días después de extraer el veneno por estimulación eléctrica (3 librerías)
- Pasivo o de mantenimiento: sin extracción de veneno previa a la remoción del telson (3 librerías)

El RNA total se extrajo con TRIZOL (Invitrogen). Para la síntesis del cDNA se utilizó el paquete comercial Message Amp-II (Ambion) siguiendo el protocolo recomendado por el proveedor. Para la síntesis de la primera cadena de cDNA se utilizaron oligos T7 (dT). Después de una segunda reacción de síntesis, 5-10 ng del cDNA de doble cadena se amplificaron por transcripción *in vitro*; el RNA anti-sentido (aRNA) resultante se purificó usando columnas Qiagen RNAeasy (Qiagen) para ser usado como templado en una segunda ronda de síntesis de cDNA. En este paso se usaron nonámeros al azar (Amersham) para la síntesis de la primera cadena. El cDNA resultante se purificó con el paquete comercial DNA Clear (Ambion) y posteriormente se nebulizó para obtener fragmentos de 200-700 pb previo a la secuenciación.

6.2 Pirosecuenciación por GS20-454, GS-FLX y FLX-Titanium

Las muestras de cDNA fueron reparadas en los extremos y ligadas a adaptadores. El enriquecimiento en las perlas de estreptavidina, la desnaturalización del DNA y el PCR en emulsión se hicieron de acuerdo al protocolo descrito por Margulies y colaboradores (Margulies *et al.* 2005). Tanto la preparación de las librerías como la secuenciación con los sistemas de 454

Roche se llevaron a cabo en el Laboratorio Nacional de Genómica para la Biodiversidad (LANGEBIO), Cinvestav Unidad Irapuato, México.

6.3 Ensamblado *de novo* de las lecturas de pirosecuenciación

Se hizo un ensamblado global que incluyó todas las lecturas obtenidas en las corridas de pirosecuenciación de la glándula activa, en reposo y del cuerpo sin el telson. Para ello, se utilizó el ensamblador Newbler 2.5 en modo de cDNA con los parámetros preestablecidos en el programa.

Adicionalmente, las lecturas clasificadas como “*singlets*” en el ensamblado de Newbler, se re-ensamblaron con MIRA en modo cDNA, tomando como longitud mínima 50 pb y por lo menos 2 lecturas en cada contig.

6.4 Análisis cualitativo de la diversidad de transcritos

6.4.1 Búsquedas por homología

Las secuencias ensambladas y los *singlets* se compararon con Blastx y Blastp (Altschul *et al.* 1990) en la bases de datos NCBI-NR, FlyBase y la colección de toxinas depositada en ToxProt (http://www.expasy.ch/sprot/tox-prot/tox-prot_stat.html). Los criterios de corte considerados para filtrar alineamientos significativos con ambas estrategias de Blast fueron:

Valor de expectación: 1e-04

Porcentaje de identidad: 30%

Porcentaje de alineamiento del contig: 30%

6.4.2 Búsqueda de dominios de proteínas

Se utilizó el programa HMMER (Eddy 1998), basado en búsquedas con modelos ocultos de Markov sobre la base de datos de Pfam para identificar dominios de proteínas. Se tomó un corte de valor de expectación de 0,1.

6.4.3 Búsqueda de péptidos señal

Las secuencias ensambladas fueron se analizaron en el servidor SIGNALP (Emanuelsson *et al.* 2007), el cual predice la presencia y ubicación de los sitios de corte de péptidos señal en las secuencias de amino ácidos tanto en eucariotes como en procariotes. Este servidor incorpora dos algoritmos diferentes, uno basado en redes neurales y el otro en modelos ocultos de Markov. Las secuencias con expresión específica en el telson que mostraron tener un péptido señal predicho con ambas estrategias de búsqueda fueron consideradas como posibles componentes del veneno.

6.4.4 Análisis filogenético

Las secuencias ensambladas similares a toxinas se alinearon por familias con Clustalw (Larkin *et al.* 2007) y se sometieron a un análisis filogenético por Neighbor Joining con 100 pseudoréplicas de bootstrap usando la paquetería Phylip (Felsenstein 2005). Las topologías consenso fueron usadas para clasificar nuevas familias de toxinas en esta especie de alacrán.

6.4.5 Clasificación taxonómica y mapeo funcional

Se utilizó el programa de análisis de datos metagenómicos MEGAN (Huson *et al.* 2007; Mitra *et al.* 2011) para obtener un perfil general de las especies representadas en las secuencias ensambladas con similitud significativa en NCBI-NR. MEGAN utiliza un algoritmo de búsqueda del ancestro común mas cercano, denominado “LCA-assignment algorithm” (LCA = lowest common ancestor) para agrupar las secuencias de entrada en la jerarquía taxonómica de NCBI, que en este momento comprende mas de 670,000 taxa. Cuando una secuencia se alinea de manera específica sobre un taxón único, se asigna directamente en ese nodo de la jerarquía. Entre menos específico sea el alineamiento sobre varios taxa, la secuencia será

posicionada en nodos más altos. Por ejemplo, si una secuencia alinea sobre los taxa *a* y *b* siendo *a* un ancestro de *b*, el alineamiento sobre el ancestro *a* será descartado y sólo se mantendrá el taxón *b* en el archivo de salida. Para lograr esta clasificación, MEGAN requiere umbrales de corte tanto del “bit score” como del número de lecturas que alinean sobre un taxón; por defecto, se requieren al menos 5 secuencias alineadas sobre un taxón para declararlo como representado en el conjunto de datos que se están analizando. Cuando el número de secuencias en un taxón es menor al umbral de corte, las secuencias clasificadas en esta posición de la jerarquía suben de nodo, hasta que alguno cumpla el mínimo de secuencias para ser considerado.

MEGAN fue alimentado con los archivos de salida de Blastp y Blastx de las secuencias ensambladas y de las lecturas clasificadas como *singlets*.

6.5 Análisis cuantitativo de la abundancia de transcritos

Se hicieron comparaciones pareadas de ambos estados de glándula diferenciando el sistema de 454 utilizado (GS20, FLX y FLX-Titanium) para evaluar la reproducibilidad de los experimentos de secuenciación. Para ello fue necesario construir una matriz global donde cada columna correspondía al número de lecturas por corrida de pirosecuenciación que conforman cada transcrito ensamblado como se describe a continuación:

ID	isogrupo	GS20-TA	GS20-TM	FLX-TA	FLX-TM	FLXT-TA	FLXT-TM
1	isogrupo01	a	b	c	d	e	f

a-f: número de lecturas por corrida. TA: telson activo; TM: telson en mantenimiento; FLXT: FLX-Titanium

Esta matriz se utilizó como archivo de entrada para la prueba de Fisher (Triola 2004), calculada con la paquetería de análisis estadístico R. De igual forma se hizo una comparación global tomando como matriz de entrada la sumatoria de las lecturas por transcrito en las dos condiciones de glándula, como se muestra a continuación:

ID	isogrupo	TA	TM
1	isogrupo01	a + c + e	b + d + f



Los valores p obtenidos con el estadístico de Fisher se sometieron a una segunda evaluación para calcular el valor Q (Storey 2004), usando el módulo QVALUE de la paquetería R. Para ambos estadísticos se tomó un valor $\alpha = 0,05$.

6.6 Procesamiento de archivos

A lo largo del desarrollo de la metodología descrita, fue necesario diseñar una serie de programas en PERL para el manejo y procesamiento de los datos.

7 Resultados y Discusión

7.1 Librerías de cDNA

Se construyeron tres librerías de cDNA a partir de RNA extraído del cuerpo sin telson y del telson con glándulas de veneno en estados activo y de mantenimiento de alacranes de la especie *Centruroides noxius*. Cada una de ellas se secuenció de manera independiente, logrando un total de 1,4 millones de lecturas con longitudes promedio de 100 nucleótidos para las librerías de glándula (sistema GS20), y 240 para la librería correspondiente al cuerpo del alacrán (sistema GS FLX). Para una segunda ronda de secuenciación se construyeron tres librerías más, bajo los criterios antes mencionados. Se seleccionaron RNAs de longitud menor a la recomendada para el sistema FLX Titanium con la finalidad de reducir la pérdida de transcritos cortos que pudieran expresarse en la glándula. Debido a la selección de tamaño de las secuencias, no fue posible alcanzar la longitud óptima que ofrece el sistema 454 Titanium y el tamaño de las lecturas obtenidas fue similar al que arroja el sistema GS20 en el caso de los transcritos de glándula, y de tamaños híbridos entre GS20 y GS FLX para los transcritos de cuerpo. Finalmente, se hizo una tercera corrida de pirosecuenciación siguiendo el protocolo recomendado por el proveedor para alcanzar un rendimiento óptimo del sistema FLX Titanium y, en este caso, las lecturas alcanzaron longitudes superiores a 300pb (Tabla 6). Es importante resaltar que los sistemas GS20, GS FLX y FLX Titanium son actualizaciones y mejoras en los protocolos de pirosecuenciación que de manera sucesiva incrementaron la cantidad y la longitud de las lecturas. Por ello, el uso de estos tres sistemas durante el proceso de secuenciación del transcriptoma de *C. noxius* responde a una evolución temporal de la plataforma de 454 ROCHE.

Si bien el protocolo mismo de preparación de las librerías para la corrida de pirosecuenciación con FLX Titanium pudo haber eliminado transcritos cortos, la longitud de las lecturas obtenidas resultó ser un ancla de secuencia que contribuyó fundamentalmente a ensamblar transcritos correspondientes a

procesos celulares del alacrán, tanto telson específicos como de expresión mixta. El rendimiento de este experimento permitió duplicar el número de lecturas y por tanto, la cobertura del transcriptoma. En total, el análisis incluyó más de tres millones de lecturas de longitudes variables (Tabla 6) que, habiendo sido generadas en tres experimentos de secuenciación independientes, permitieron no sólo hacer un ensamblado global, sino análisis estadísticos confiables del número de transcritos con diferencias de expresión significativos. Estos análisis se describirán en las secciones posteriores.

Tabla 6. Número de lecturas por corrida de pirosecuenciación

Librerías	Lecturas Totales	Corrida 454	Longitud (pb)	Sistema 454
Mantenimiento	1 249 489	665 311	100	GS20
		124 949	134	FLX
		459 229	352	Titanium
Actividad	981 028	404 812	99	GS20
		92 507	95	FLX
		483 709	326	Titanium
Cuerpo s/telson	777 532	362 272	240	FLX
		37 568	173	FLX
		377 692	340	Titanium
Total	3 008 049			

7.2 Ensamblado

Para ensamblar el transcriptoma de *Centruroides noxius* fue necesario evaluar diferentes programas y algoritmos. Los ensambladores Newbler y MIRA se eligieron por ser programas robustos en el ensamblado de datos de 454 (Kumar y Blaxter, 2010). Los resultados se discuten a continuación.

7.2.1 Newbler cDNA V2.3 y 2.5

Las versiones 2.3 (liberada en octubre del 2009) y 2.5 (liberada en diciembre del 2010) de Newbler cuentan con un modo de ensamblado particular para ESTs. Además de la clasificación tradicional de las lecturas (ensambladas,

parcialmente ensambladas, *singlets*, outliers y repetidas), bajo la condición de cDNA se definen nuevos conceptos de secuencias, los “isotigs” y los “isogrupos”. Cada isotig está compuesto por varios contigs relacionados entre ellos, y puede considerarse como análogo a un transcrito individual. Aquellos isotigs que en conjunto formen un mismo isogrupo pueden pensarse como variantes de *splicing* de un gen particular, donde los contigs se consideran como exones potenciales. Por ejemplo, un isogrupo compuesto por 12 contigs y 16 isotigs, teóricamente representa un gen con 12 exones y 16 variantes de *splicing* posibles; el sentido biológico de esta extrapolación debe ser cuidadosamente evaluado. Las conexiones entre los contigs de un mismo isogrupo representan lecturas que divergen hacia dos o más contigs diferentes, o por un pico de profundidad en una zona determinada (Fig. 8). Estos picos se definen siguiendo las siguientes reglas: la profundidad del alineamiento tiene al menos 10 lecturas; un mínimo de 20% de las lecturas alineadas deben tener orientación opuesta relativa a la orientación más abundante de las lecturas alineadas hasta ese punto; los picos de profundidad deben tener más de 10 bases de separación entre ellos.

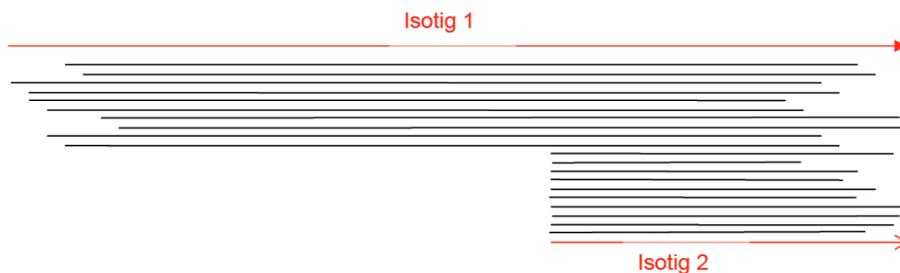


Figura 8. Detección de picos de profundidad

7.2.2 MIRA EST

Tomando como referencia el estudio transcriptómico del alacrán *P. imperator* (Roeding *et al.* 2009), donde las lecturas obtenidas por 454-FLX se ensamblaron con MIRA-EST, decidimos probar dicho ensamblador en el estudio de *C. noxius*. Inicialmente, MIRA se diseñó específicamente para ensamblar

secuencias de transcritos usando un algoritmo glotón, como se describió previamente. Cuenta con dos formas de trabajo diferentes, una para secuencias de genomas completos y otra para cDNA (Chevreux *et al.* 2004). Adicionalmente pueden activarse parámetros de búsqueda de polimorfismos en ambas formas de trabajo. Una de las principales desventajas de este ensamblador es la clasificación de lecturas “debris”, ya que a diferencia de Newbler, no podemos discernir entre outliers, *singlets* y secuencias repetidas. A pesar de que MIRA reporta un número determinado de lecturas denominadas *singlets* en los archivos de salida, la cantidad de secuencias en esta categoría es sumamente pequeña considerando el número de lecturas que se usan en los archivos de entrada y el número de secuencias debris.

7.2.3 Newbler vs MIRA

Previo a la última ronda de pirosecuenciación del transcriptoma de *C. noxius*, se evaluaron ambos ensambladores con el fin de elegir el programa más eficiente y confiable bajo criterios como aprovechamiento de las lecturas, longitud de las secuencias ensambladas, proporción de *singlets* y de secuencias repetidas. En ambos casos se eliminaron las secuencias de baja calidad o de longitud inferior a 50 nucleótidos. Además de este primer filtro de calidad, fue necesario hacer ensamblados parciales de cada una de las corridas de 454, debido al alto número de secuencias en cada una de ellas, que en conjunto representan un costo computacional muy elevado. En cada ensamblado parcial, se eliminaron las secuencias clasificadas por Newbler como repetidos, o debris por MIRA, ya que teóricamente no aportan información adicional en ensamblados posteriores. Con las lecturas filtradas, se hicieron ensamblados globales usando los parámetros preestablecidos en cada programa en modo de genoma y cDNA. Adicionalmente, en el caso de MIRA los valores de identidad de los alineamientos pareados así como la longitud mínima de empalme, se igualaron a los valores estándar de Newbler con el fin de tener ensamblados bajo los mismos criterios de corte y hacer una comparación más objetiva.

De estos ensamblados podemos concluir lo siguiente: el número de lecturas clasificadas por MIRA como “debris” que fueron eliminadas, varía mucho de una condición a otra, teniendo en los ensamblados globales desde 47% hasta 72% de lecturas totales aprovechadas (Tabla 7). Por otra parte, el número de contigs generados en modo de EST duplica la cifra obtenida en modo de genoma bajo los mismos parámetros de corte. Es notable que Newbler v2.3 aprovecha de manera mucho más eficiente las lecturas de entrada, incluso después de los ensamblados parciales, pues se logró ensamblar 98% de estas.

Tabla 7. Datos cuantitativos de los cuatro ensamblados con Newbler y MIRA

	Newbler 2.3		MIRA (accurate)			
	Genoma	cDNA	Genoma	EST	Genoma Default	EST Default
reads Filtrados	816 328	1 404 274		677 562		1 025 092
% filtrados	57	98		47		72
reads ensamblados	467 171	970 646	520 413	632 721	855 061	963 234
parciales	129 477	135 750	-	-	-	-
singlets	100 070	264 406	-	-	-	-
outlier / debris	23 572	14 558	156 538	43 973	169 989	61 799
repeats	96 035	18 914	-	-	-	-
contigs	23 991	15 659	20 291	46 848	34 325	69 155
logitud promedio (pb)	298	382	302	244	312	268
contigs > 500pb	3 392	3 590	2 717	3 864	3 129	987
mayor longitud (pb)	2 692	2 984	3 084	3 037	3 236	3 865

Por otra parte, se hizo una comparación de la proporción de contigs o isotigs en diferentes rangos de longitud (frecuencia normalizada respecto al total de contigs por ensamblado). De la figura 9 es evidente que el ensamblado con Newbler v2.3 cDNA mantiene una mayor proporción de isotigs de longitud superior a 500pb mientras que MIRA en modo de EST arroja una mayor proporción de secuencias fragmentadas de longitud entre 100 y 300pb. Si bien la tabla 7 indica que MIRA EST tiene más contigs largos (en números absolutos), e incluso la

secuencia ensamblada de mayor longitud con este algoritmo supera la mejor longitud alcanzada por Newbler, el número de contigs y el promedio del tamaño de las secuencias ensambladas con MIRA hacen evidente que el ensamblado global se mantiene altamente fragmentado.

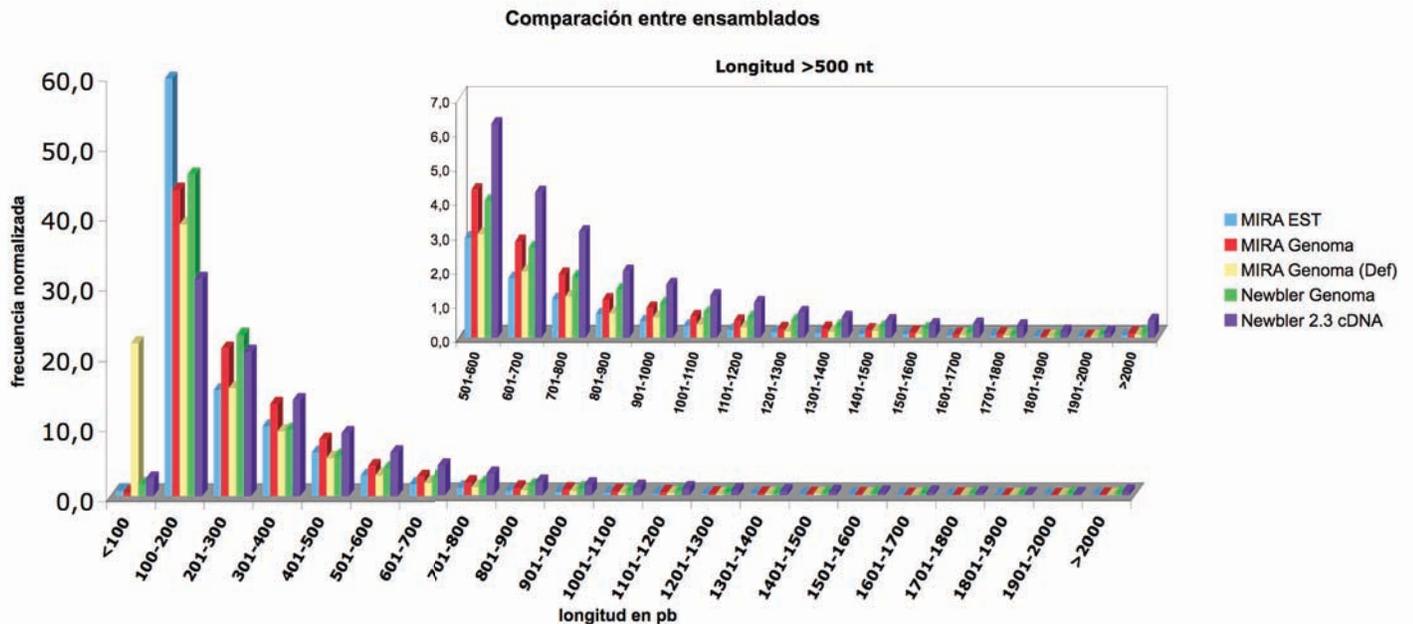


Figura 9. Comparación de la longitud de los contigs/isotigs obtenidos con Newbler y MIRA. En la parte superior se observan aquellos con longitud superior a 500pb. La frecuencia fue normalizada respecto al total de contigs/isotigs de cada ensamblado, para tener una comparación mas objetiva.

Dado que bajo los criterios de aprovechamiento de lecturas, número de contigs/isotigs ensamblados y longitud promedio de las secuencias, Newbler muestra claras ventajas respecto a MIRA, se eligió este programa para el ensamblado *de novo* de todas las lecturas de *C. noxius* obtenidas en los tres experimentos de pirosecuenciación. Los resultados se muestran en la siguiente sección.

7.2.4 Ensamblado global

Longitud del ensamblado

El ensamblado global con Newbler v2.5 cDNA arrojó 26 672 isotigs (aproximadamente 19 Mb ensambladas), agrupados en 18 979 isogrupos (Tabla 8), que en total representan más del 80% de las lecturas aprovechadas. Tomando como referencia el número de genes de *Drosophila melanogaster* (aproximadamente 15 000, FlyBase) y de la garrapata *Ixodes scapularis* (24 925, VectorBase), un arácnido filogenéticamente más cercano al alacrán, podemos sugerir que el número de isogrupos pensados como genes únicos, es congruente con el rango de secuencias de los artrópodos antes mencionadas.

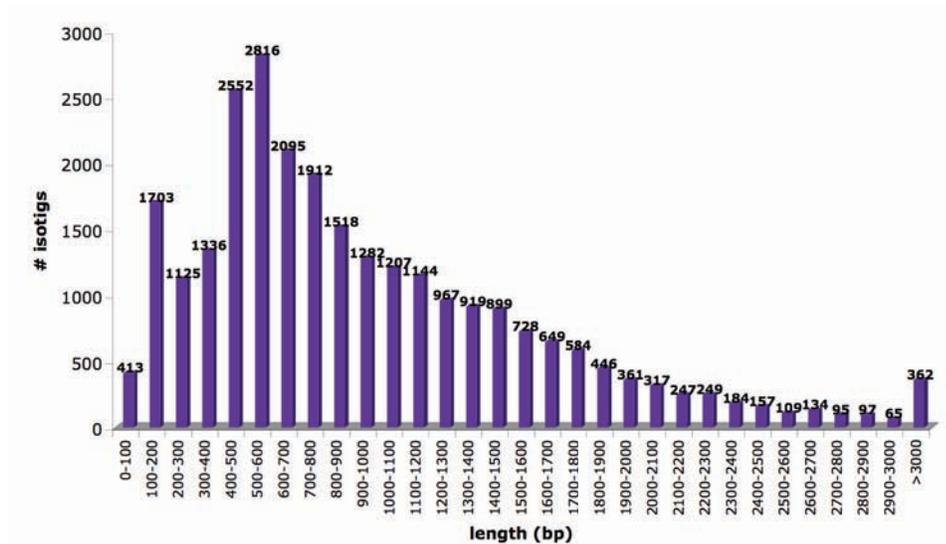
Tabla 8. Resultados del ensamblado global con Newbler v2.5 cDNA

Reads			
numberAssembled	2025125	numberSingleton	424134
numberPartial	452407	numberOutlier	95909
numberRepeat	3894	numberTooShort	47638
Total	2481426	Total	567681
Isotigs		Isogroups	
numberOfIsotigs	26672	numberOfIsogroups	18979
avgContigCnt	2,1	avgContigCnt	2,5
largestContigCnt	16	largestContigCnt	15553
numberWithOneContig	16257	numberWithOneContig	15385

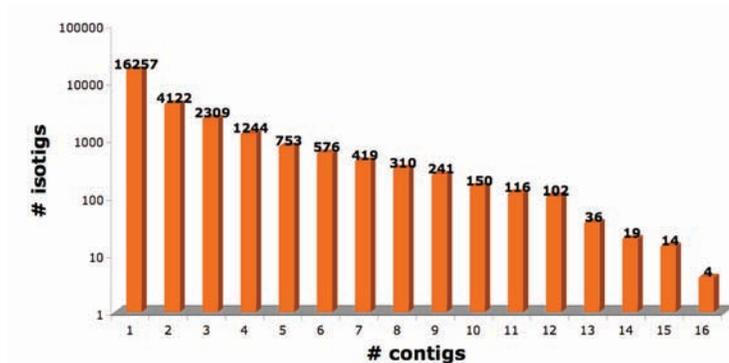
La longitud promedio de los isotigs fue de 950 nucleótidos, y se obtuvieron 19 543 isotigs superiores a 500 pb (Fig. 10A). Por otra parte, aprovechando la clasificación en isotigs e isogrupos generada por Newbler, podemos ver que 39% de las secuencias ensambladas está conformada por más de 2 contigs o exones (Fig. 10B) y únicamente 17% de los isogrupos, es decir, de genes únicos, tiene más de una variante de *splicing* (Fig. 10C). Estudios recientes han mostrado que el *splicing* alternativo es un proceso recurrente en organismos eucariotes; se ha estimado que entre 70 y 95% de los genes humanos producen isoformas alternativas de mRNA (Wang *et al.* 2008), y 46% de los genes de *Drosophila*

melanogaster muestran *splicing* alternativo durante el desarrollo (Hansen *et al.* 2009). En comparación con estos organismos modelo, es claro que la proporción de genes de alacrán con variantes de *splicing* es muy reducido, sin embargo, para validar esta extrapolación de los resultados del ensamblado será necesario generar información genómica que permita analizar la complejidad en cuanto al contenido de intrones y exones de los genes de este organismo.

A.



B.



C.

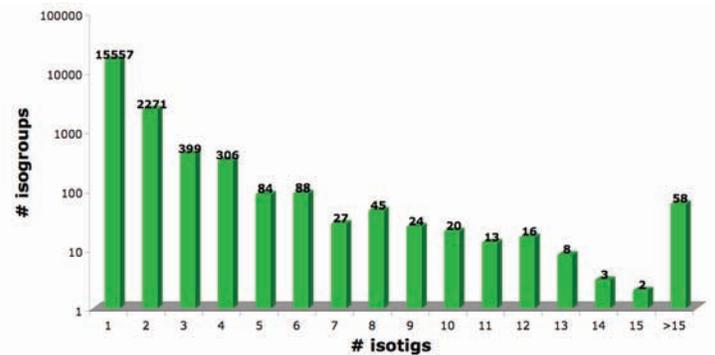


Figura 10. Características generales del ensamblado. A. Longitud de los isotigs ensamblados. B. Número de exones por variante de *splicing* (contigs por isotig). C. Número de variantes de *splicing* por transcrito único (isotigs por isogrupo)

Puesto que el ensamblado global muestra un alto contenido de *singlets*, estos fueron re-ensamblados con MIRA. Se obtuvieron 53, 218 contigs con una N50 = 312 pb, lo cual indica que no hubo una mejora significativa en cuanto al aprovechamiento de estas lecturas y que se trata de un ensamblado muy fragmentado y de baja calidad. Por ello, sólo consideraremos las secuencias ensambladas con Newbler en los análisis que se describen en las siguientes secciones.

Contigs de expresión glándula/cuerpo específica o ubicua

El tipo de expresión de los transcritos (glándula/cuerpo específica o ubicua) se determinó calculando el número de lecturas por librería que conforman cada isogrupo. Se identificaron 4, 523 isogrupos glándula específicos, 669 cuerpo específicos y 13, 787 de expresión ubicua (Fig. 11).

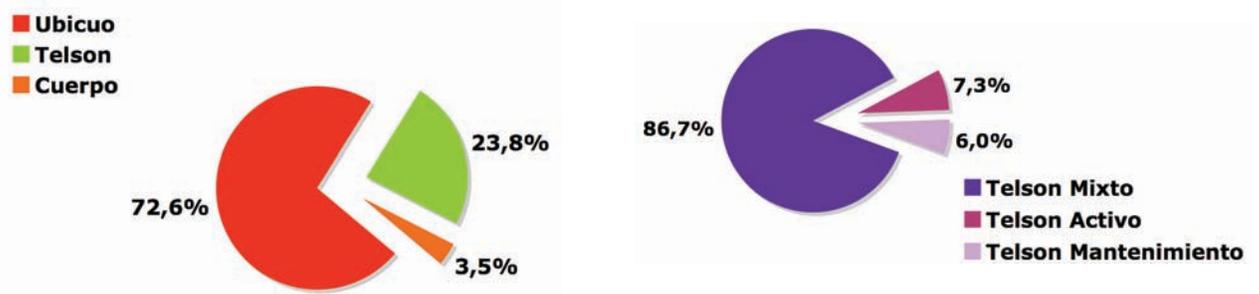


Figura 11. Tipo de expresión de los transcritos. Porcentaje de transcritos de expresión específica (glándula o cuerpo) y ubicua.

Más del 70% de los isotigs muestran expresión ubicua en el organismo, y llama la atención la disparidad en los porcentajes de isogrupos que tienen expresión específica. Esto no implica que cerca del 24% de los genes de alacrán se expresen exclusivamente en glándula, sino que este conjunto de isogrupos representan por una parte, transcritos fragmentados que no pudieron ensamblarse dado que las secuencias de glándula están enriquecidas con lecturas de

longitudes cortas, resultantes de la secuenciación con el sistema GS20 (Tabla 6). Por otro lado, dado que el 90% de las secuencias glándula específicas muestran marcos de lectura abiertos de más de 30pb de longitud, es posible sugerir que algunas de estas secuencias representen transcritos cortos que pudieran jugar un papel importante en procesos regulatorios de la glándula de veneno.

Sin embargo, también es importante considerar que dentro del conjunto de transcritos de expresión ubicua, podríamos encontrar secuencias que no llegan a ser traducidas en el cuerpo ya que su función está restringida a la glándula. En efecto, sabemos que las polimerasas de RNA transcriben regiones del genoma de manera no específica (Dinger *et al.* 2009; Jacquier 2009), por lo que cabe suponer que habrá genes que si bien funcionan exclusivamente en la glándula (como el caso de las toxinas), llegan a transcribirse en otros tejidos del cuerpo del alacrán y por tanto, están representados en bajos niveles en las librerías construidas a partir de RNA de cuerpo. Esta observación se vuelve muy importante en el paso de búsqueda de función de las secuencias ensambladas, en particular para determinar la diversidad de toxinas que puede existir en esta especie de alacrán. Bajo esta perspectiva, la búsqueda de transcritos similares a toxinas se hizo en el conjunto de isogrupos cuya composición incluye hasta 10% de lecturas provenientes de las librerías de cuerpo, como se discutirá más adelante.

7.3 Corridas Acumuladas

Podemos considerar diferentes parámetros como indicadores de la cobertura alcanzada de un genoma o transcriptoma en proceso de secuenciación: número de secuencias ensambladas, número de *singlets* y de secuencias repetidas. Para tener un estimado de la cobertura del transcriptoma de *C. noxius* alcanzada hasta el momento, se hicieron ensamblados acumulados de las corridas de pirosecuenciación, donde la corrida 1 (run1) comprende las lecturas obtenidas con el sistema GS20 de glándula, la corrida 2 (run2) el acumulado GS20 + FLX, la corrida 3 (run3) el acumulado GS20 + FLX + corrida de longitud híbrida

entre GS20 y FLX y la corrida 4 (run4) el acumulado GS20 + FLX + corrida de longitud híbrida entre GS20 y FLX + FLX-Titanium.

Teóricamente, se espera que al llegar a la cobertura máxima, el número acumulado de secuencias ensambladas debe alcanzar un comportamiento asintótico. En el caso del transcriptoma de *C. noxius*, observamos que el número de isotigs e isogrupos ensamblados ha aumentado a lo largo del proceso de pirosecuenciación, en particular después de agregar las lecturas obtenidas con la ultima corrida de 454 (Fig. 12). Esto puede deberse a que el protocolo de preparación de las librerías del sistema Titanium discrimina fragmentos de RNA de tamaños cortos y favorece la retención de fragmentos largos para la amplificación en emulsión, y por tanto, efectivamente se están agregando transcritos nuevos al ensamblado global que antes no se consideraban por los requerimientos de los sistemas GS20 y FLX.

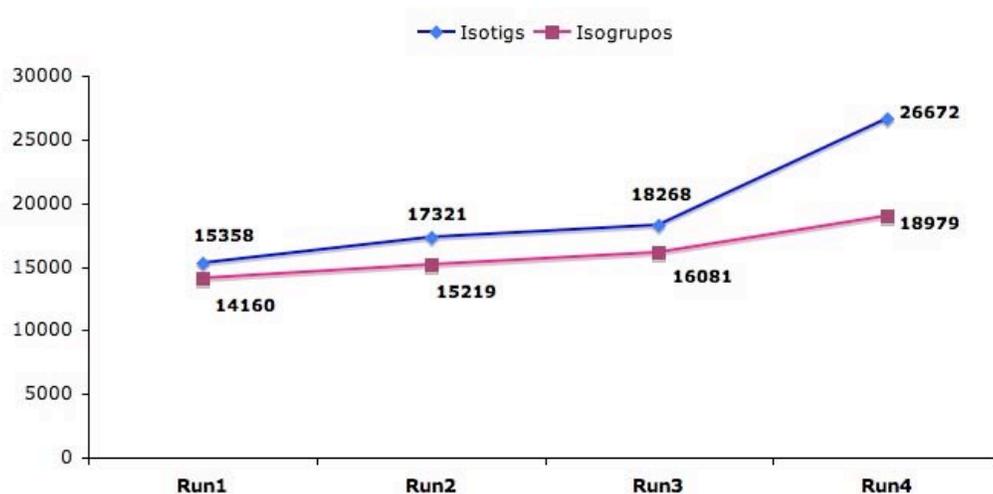


Figura 12. Número acumulado de secuencias ensambladas

Por otra parte, dos indicadores que deben considerarse, son el número de lecturas clasificadas como *singlets* y secuencias repetidas. En un escenario de cobertura máxima, se espera un número bajo de *singlets*, puesto que la mayor parte de las lecturas debe ser ensamblada en un contig. En cuanto a las

secuencias repetidas, podríamos esperar un incremento conforme se agregan lecturas al ensamblado, lo cual hablaría de una mayor profundidad de los transcritos secuenciados. Al igual que en el caso de los isotigs acumulados, el transcriptoma de *C. noxius* ha mostrado un incremento en el número de *singlets* y, si bien las secuencias repetidas han aumentado a lo largo del proceso de secuenciación, estas comprenden una proporción muy pequeña del total de las lecturas obtenidas y por tanto indican que la profundidad de los transcritos es baja (Fig. 13). En conjunto, estos tres indicadores muestran que es necesario continuar con el protocolo de secuenciación, dado que no hemos llegado a cubrir el transcriptoma de manera óptima.

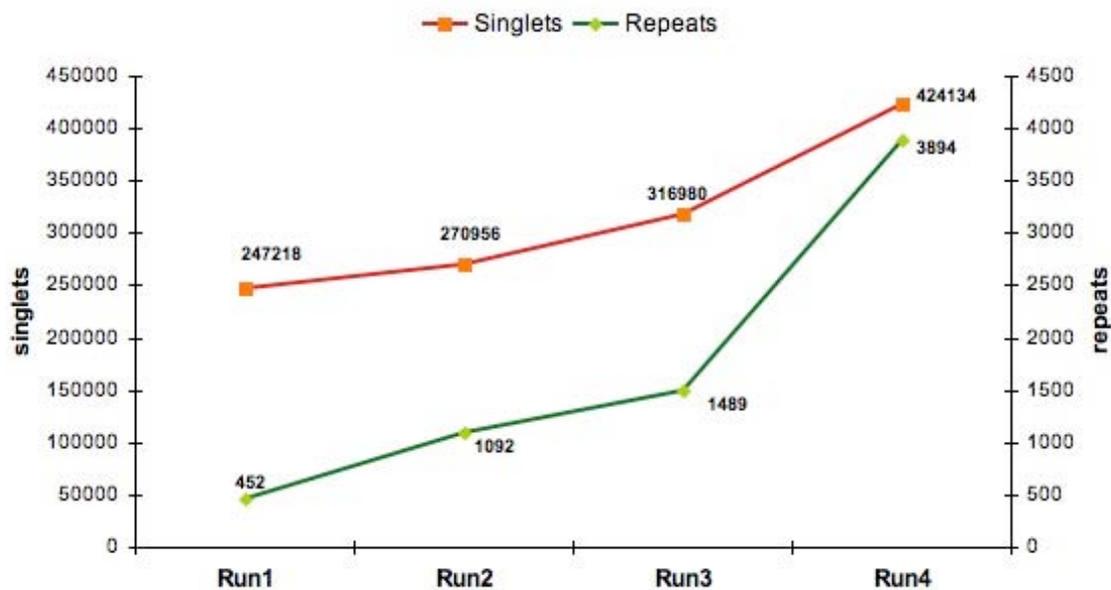


Figura 13. *Singlets* y secuencias repetidas acumuladas

7.4 Análisis cualitativo de los transcritos de *Centruroides noxius*

7.4.1 Clasificación Taxonómica

Dado que la extracción de RNA total se hizo a partir de un macerado del cuerpo del alacrán, es posible pensar que el universo de lecturas generadas por 454 incluya algunas secuencias provenientes de organismos comensales o

parásitos (de tracto digestivo o respiratorio, por ejemplo). Por ello, conocer la proporción de secuencias no específicas de alacrán se vuelve un criterio de control importante para validar el grado de pureza en cuanto al contenido de secuencias *bona fide* de artrópodos que se reportan en el transcriptoma. Para discriminar entre las secuencias artrópodo específicas y no específicas, usamos el programa MEGAN de análisis de metagenomas (Huse *et al.* 2007). MEGAN asigna cada secuencia al taxón más cercano en función de los hits de Blast en un e-value determinado. El resultado de este análisis puede verse en la figura 14. A pesar de que se observan secuencias de origen viral, bacteriano, fúngico y vegetal, estas no fueron consideradas en el ensamblado pues se mantuvieron clasificadas como *singlets*. Si bien existen pocos reportes que describen la flora microbiana de los artrópodos, se ha observado que en insectos como *Bactrocera dorsalis* las Gamaproteobacterias, Actinobacterias y Firmicutas son predominantes en el tracto digestivo (Wang *et al.* 2011), mientras que en diferentes especies de garrapatas (*Ixodes ricinus*, *Dermacentor reticulatus* y *Haemaphysalis concinna*) se ha visto que más del 80% de las secuencias ribosomales de bacterias aisladas del tracto digestivo corresponden a bacterias Gram+, en particular *Bacillus* y *Paenibacillus* (Rudolf *et al.* 2009). Por otra parte, en los colmillos y en las glándulas de veneno de *Loxosceles laeta* se han aislado especies de *Clostridium*, las cuales se piensa son inoculadas junto con el veneno y podrían exacerbar el daño dermonecrótico (Catalán *et al.* 2010). En el caso de *C. noxius*, se observó que las Actinobacterias, Flavobacterias, Firmicutas (*Bacillus* y *Clostridium*) y Proteobacterias son los géneros bacterianos más representados entre las lecturas del cuerpo y glándula analizadas (Fig. suplementaria 1); sin embargo, la presencia de estas bacterias, así como el tejido de origen, deberán ser validados con análisis posteriores de secuencias ribosomales.

Es evidente que la clasificación de las secuencias ensambladas en la jerarquía taxonómica está sesgada a las especies mejor representadas en las bases de datos actuales, sin embargo es claro el enriquecimiento de transcritos de origen eucariote, fundamentalmente hacia cordados y artrópodos. Aproximadamente 48% de las secuencias con homología en NCBI-NR se

agruparon en el clado de los artrópodos, 44% de las cuales (21% del total de secuencias con homología) pertenecen específicamente a quelicerados. En la figura 14B es evidente que dentro de la familia Buthidae predominan secuencias telson específicas, entre las que se encuentran toxinas y proteínas secretorias.

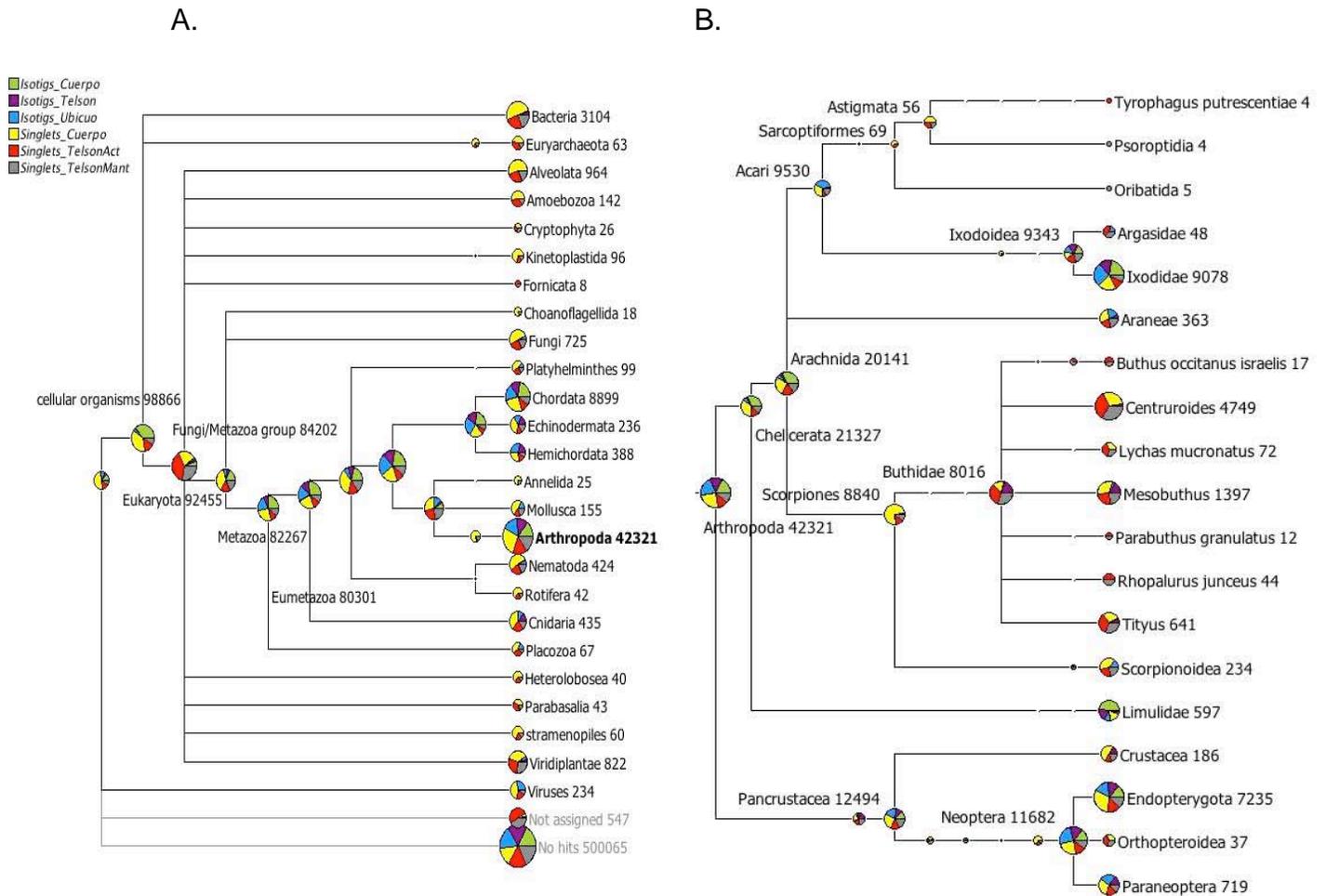


Figura 14. Clasificación taxonómica de los transcritos de *C. noxius*. A. Representación general de los diferentes taxa a nivel de isotigs y *singlets*. B. Acercamiento del nodo de los artrópodos.

7.4.2 Búsquedas por homología

Se planteó una estrategia de anotación basada en la comparación con alineamientos locales (Blast) de los isotigs del ensamblado global obtenido con Newbler v2.5 cDNA sobre tres diferentes bases de datos, que en conjunto permiten ampliar el universo de búsqueda. Estas son NCBI-NR, la colección de

péptidos de *Drosophila melanogaster* (FlyBase) y la base de datos ToxProt, que contiene todas las toxinas caracterizadas hasta el momento. Los resultados se filtraron en función del valor esperado ($1e-04$), porcentaje de identidad (30%) y cobertura (30% del isotig alineado). Con estos criterios, cerca de 50% de los isotigs tuvieron resultados significativos en al menos una de las bases de datos utilizadas (Fig. 15).

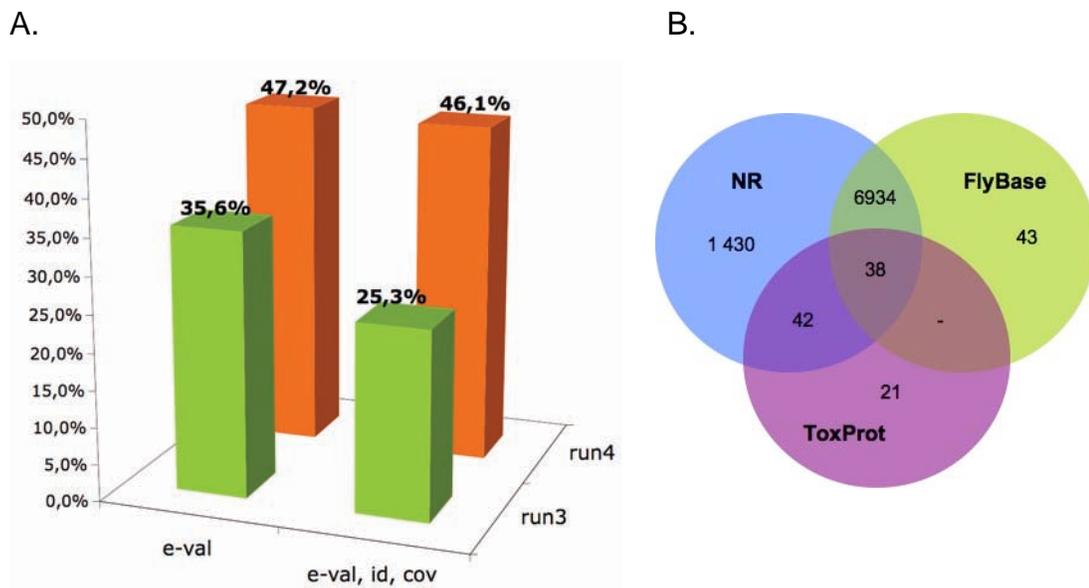


Figura 15. Secuencias ensambladas con homología en bases de datos. A. Comparación de la proporción de isotigs con alineamientos significativos en NCBI-NR en los ensamblados 3 (sin incluir lecturas FLX-Titanium) y 4 (considerando FLX-Titanium). B. Isogrupos de la corrida 4 con alineamientos significativos en las bases de datos utilizadas para las búsquedas por homología.

Dado que en este momento contamos con una amplia colección de secuencias de transcritos de *Ixodes scapularis*, una garrapata filogenéticamente cercana al alacrán, es posible aprovechar estas secuencias como punto de referencia en cuanto a la proporción de ESTs de *C. noxius* que podemos esperar con similitud significativa a las secuencias de *D. melanogaster*. Para ello se comparó el conjunto de 24 925 transcritos ensamblados de *I. scapularis* con la colección de transcritos y péptidos de *D. melanogaster*, a nivel de nucleótidos y de amino ácidos (Tabla 9).

Tabla 9. Comparación pareada entre transcriptomas. Los transcriptomas de *I. scapularis*, *C. noxius* y *D. melanogaster* fueron comparados con dos estrategias de Blast (nucleótidos y proteínas)

	<i>Ixodes scapularis</i>		<i>Centruroides noxius</i>	
	nucleótidos	amino ácidos	nucleótidos	amino ácidos
<i>D. melanogaster</i>	35%	48%	5,4%	37%

Tomando en cuenta que el tiempo de divergencia entre los artrópodos y los insectos ha sido calculado en más de 700 millones de años (Pisani *et al.* 2004), podemos esperar poca conservación de las secuencias de nucleótidos, la cual se refleja en el bajo porcentaje de transcritos con alineamientos significativos de Blastn. A diferencia del caso de *I. scapularis*, la comparación a nivel de nucleótidos y amino ácidos entre los isotigs ensamblados de *C. noxius* y *D. melanogaster* revela porcentajes mucho más bajos (Tabla 9). Esto puede tener diferentes causas. En primer lugar es posible pensar que efectivamente el tiempo de divergencia entre los insectos y los escorpiones ha sido lo suficientemente largo para rastrear exhaustivamente los transcritos de un organismo en el otro. Por otra parte, desconocemos la proporción de secuencias genómicas (intergénicas no codificantes) transcritas de manera inespecífica que están representadas en las lecturas de 454 (Dinger *et al.* 2009) que difícilmente estarían presentes en las secuencias de la mosca. También es posible pensar que existen familias de transcritos específicos de alacrán que no han sido descubiertos, y por tanto, no tienen homología con otras secuencias de *D. melanogaster*.

7.4.3 Genes eucariotes esenciales

Además de la validación taxonómica de las secuencias ensambladas de alacrán, una forma de corroborar la calidad de los isotigs obtenidos, consiste en buscar genes eucariotes esenciales en el transcriptoma de *C. noxius*.

Tabla 10. Genes eucariotes esenciales de copia única identificados en *C. noxius* (cobertura > 70%)

Genes esenciales eucariotes
DNA-directed RNA polymerases I, II, and III 14.5 kDa polypeptide 27 kDa polypeptide 8.3 kDa polypeptide
Eukaryotic peptide chain release factor subunit 1
Eukaryotic translation initiation factor 5A
GTP-binding and nucleic acid-binding protein YchF
GTPase and tRNA-U34 5-formylation enzyme TrmE
Glutamyl-tRNA(Gln) amidotransferase subunit A (EC 6.3.5.7)
Glycyl-tRNA synthetase (EC 6.1.1.14) @ Glycyl-tRNA synthetase (EC 6.1.1.14), mitochondrial
HBS1 protein
LSU ribosomal proteins
Leucyl-tRNA synthetase (EC 6.1.1.4)
Lysyl-tRNA synthetase (class II) (EC 6.1.1.6) Lysyl-tRNA synthetase (class II) (EC 6.1.1.6), mitochondrial
Phenylalanyl-tRNA synthetase alfa chain (EC 6.1.1.20) mitochondrial
RNA polymerase III transcription initiation factor (TFIIIC) 95 kDa subunit
Ribonucleases P/MRP protein subunit POP7 (EC 3.1.26.5)
SSU ribosomal proteins
Seryl-tRNA synthetase (EC 6.1.1.11)
TATA-box binding protein
Threonyl-tRNA synthetase (EC 6.1.1.3)
Transcription initiation factor IIB IIE beta subunit IIF beta subunit IIH cyclin-dependent kinase 7 IIH p34 subunit IIH p44 subunit IIIA IIIB 70 kDa subunit
Translation elongation factor 1 alfa subunit beta subunit gamma subunit
Translation elongation factor Tu
Tryptophanyl-tRNA synthetase (EC 6.1.1.2)
Tyrosyl-tRNA synthetase (EC 6.1.1.1)
proteasome regulatory subunit Rpn10
proteasome subunit alfa1 beta1 (EC 3.4.25.1)
ubiquitin / LSU ribosomal protein L40e
ubiquitin / SSU ribosomal protein S27Ae
ubiquitin-like protein fubi / SSU ribosomal protein S30e

Para ello, se utilizó una base de datos con proteínas esenciales de copia única obtenidas de organismos modelo secuenciados (la base de datos fue proporcionada por el Dr. Robert Edwards, San Diego State University, EUA. <http://edwards.sdsu.edu/labsite/>), que comprende mamíferos (*Homo sapiens*, *Pan troglodytes*, *Mus musculus*, *Rattus norvegicus*, *Canis familiaris*), aves (*Gallus gallus*), insectos (*Apis mellifera*, *Drosophila melanogaster*, *Anopheles gambiae*), plantas (*Arabidopsis thaliana*) peces (*Danio rerio*, *Tetraodon nigroviridis*, *Fugu rubripes*) y levadura (*Saccharomyces cerevisiae*). Las proteínas identificadas con una cobertura superior a 70% se muestran en la tabla 10 y en la tabla

suplementaria 1. Entre estas proteínas esenciales, destacan las tRNA sintetisas, las proteínas ribosomales y las proteínas involucradas en procesos de transcripción y traducción. Es importante resaltar que no se encontraron RNAs ribosomales en el ensamblado ni tampoco a nivel de *singlets*.

7.4.4 Mapeo de las secuencias en las rutas metabólicas de KEGG

Para lograr una visión global de los procesos celulares representadas en el transcriptoma de *C. noxius*, se hizo un mapeo de las secuencias ensambladas sobre las redes de interacciones moleculares descritas en la “Kyoto Encyclopedia of Genes and Genomes” (KEGG) con el programa de análisis de metagenomas MEGAN (Mitra *et al.* 2011). A partir de esta comparación, fue posible hacer inferencias interesantes.

A nivel metabólico, el transcriptoma del alacrán cubre, si bien no de manera óptima, todas las redes de interacciones descritas en eucariotes (Fig. suplementaria 2). Estas se ven particularmente representadas en los isotigs de expresión ubicua, cuerpo específicas y a nivel de *singlets*. Los mecanismos de procesamiento de la información genética están claramente representados, lo cual es congruente con los resultados de búsqueda de genes esenciales descritos previamente. En efecto, se encontraron proteínas ribosomales, factores de inicio y término de la transcripción y la traducción, tRNA sintetisas, subunidades del proteasoma, entre otros, contenidos en la tabla 10 y agrupados en la figura 16. Dentro de la categoría de motilidad celular se deriva una observación importante: únicamente el proceso de regulación del citoesqueleto de actina está enriquecido con secuencias ensambladas de alacrán, mientras que los mecanismos de motilidad bacteriana (quimiotaxis y ensamblado del flagelo) se ven reflejados a nivel de *singlets* con números de lecturas despreciables.

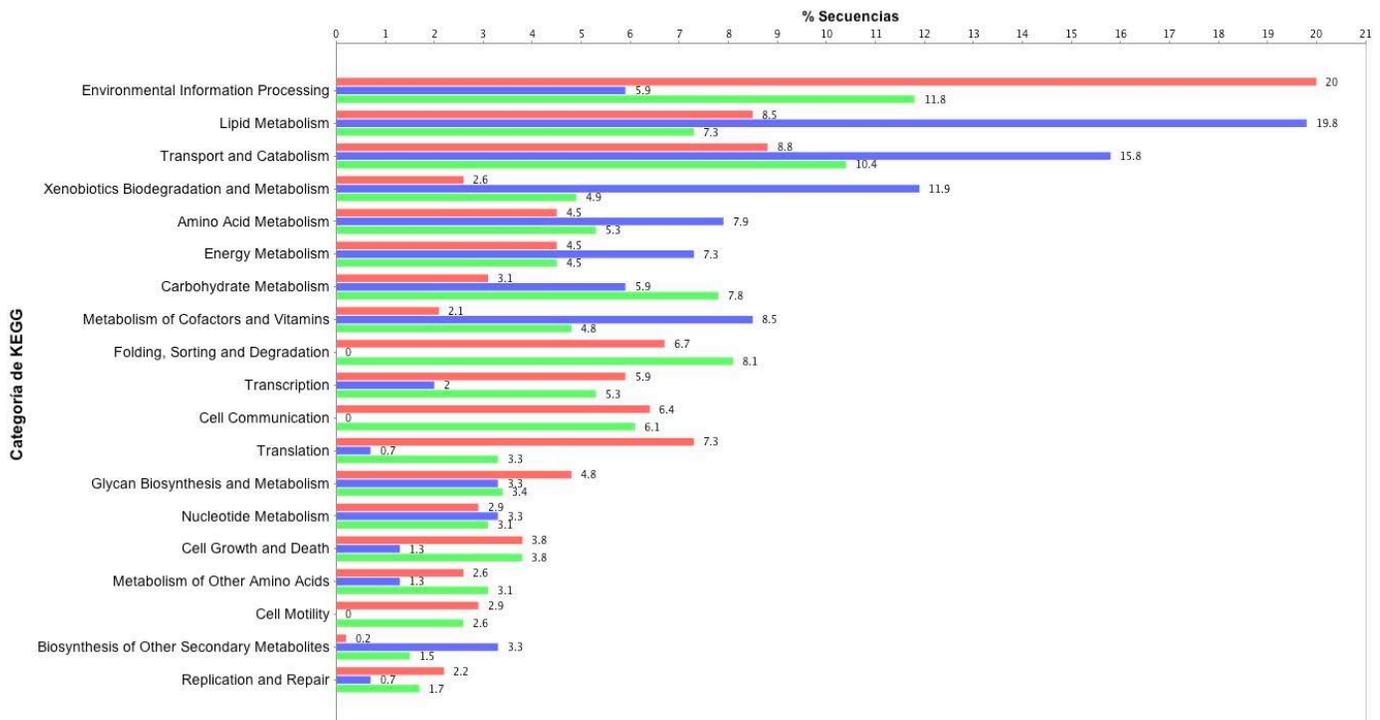


Figura 16. Procesos metabólicos, celulares y de manejo de la información genética representados en el transcriptoma de *C. noxius*. En rojo se muestran las secuencias glándula específicas; en azul las secuencias cuerpo específicas y en verde las secuencias de expresión ubicua.

7.4.5 Toxinas

A lo largo de varios años de investigación, se ha logrado caracterizar una amplia gama de toxinas moduladoras de la función de canales de sodio y bloqueadoras de canales de potasio en el veneno del alacrán mexicano *C. noxius*, que han mostrado ser herramientas fundamentales tanto en el desarrollo de antivenenos como en el estudio del funcionamiento de los canales iónicos. Para complementar la información que se tiene a nivel proteómico en cuanto a la diversidad de toxinas en esta especie alacrán, se analizaron los transcritos telson específicos con una batería de programas y algoritmos como se describe a continuación.

La estrategia de búsqueda de toxinas incluyó el uso de Blastp de todos los marcos de lectura abiertos en los isotigs sobre la base de datos ToxProt, que contiene todas las secuencias similares a toxinas conocidas hasta el momento. Se usó también el programa SignalP para identificar péptidos señal necesarios para la secreción de los componentes del veneno. Posteriormente, los isotigs se agruparon y alinearon con el programa Clustalw por familias de toxinas; en algunos casos se construyeron árboles filogenéticos por Neighbor-Joining con la paquetería PHYLIP para lograr una clasificación más detallada de las secuencias. Los porcentajes de identidad y de cobertura de las secuencias similares a toxinas se muestran en la tabla suplementaria 2.

Tabla 11. Familias de toxinas identificadas en el transcriptoma de *C. noxius*

Tipo de Toxina		# Isogrupos	# Reads	% identidad
Modificadoras de canales iónicos	Canales de Sodio	27	7150	44-100%
	Canales de Potasio	15	9459	60-100%
	Canales de Calcio	2	801	45-70%
	Otras LVP1	4	2845	38-45%
Zinc Metaloproteasas	Antareasa	7	5050	40-60%
	Astacin-like			
	VMP			
Fosfolipasas	PLA2	1	5	50%
Inhibidores de Proteasas	Kunitz-like	4	606	50-70%
Serin-Proteasas	Ser-Proteasas secretadas en glándula salivales	3	282	50%
Lipasa	Dominio de colipasa	1	230	60%
Antimicrobianos	Porina	1	6	60-78%
Otros componentes del veneno	Venom isulin-like growth factor binding protein	2	2012	30-40%
	Hyaluronidasa	1	15	70%
	Alergeno	1	26	66%
	Toxin-like peptide	1	1262	60%
	Neurotoxinas	2	552	78%
TOTAL		72	30301	

Con esta estrategia, se logró identificar 72 isogrupos similares a toxinas de familias diversas, las cuales se muestran en la tabla 11. Estos isogrupos

comprenden únicamente el 0,4% de los transcritos ensamblados. Es importante aclarar que se utilizó un filtro adicional respecto a la composición de las secuencias que se clasificaron como toxinas. Como se discutió en la sección 7.2.4, tomando en cuenta que los loci de toxinas pueden ser transcritos en otros tejidos del alacrán de manera inespecífica (aunque su función está restringida a la glándula de veneno), se tomaron diferentes puntos de corte en cuanto al contenido máximo en los isogrupos de lecturas provenientes de las librerías construidas a partir de RNA extraído del cuerpo del alacrán que podía incluirse para buscar transcritos que pudiesen ser secretados en el veneno de *C. noxius*. Inicialmente se hizo una búsqueda de toxinas en todos los isogrupos sin discriminar el porcentaje de lecturas de cuerpo; en este caso, el número de posibles componentes del veneno era muy elevado pues incluía enzimas como proteasas y fosfodiesterasas que en realidad corresponden a enzimas presentes en el cuerpo del alacrán, sin señales de secreción ni lecturas de glándula. Con estas observaciones, se evaluó el contenido de toxinas en aquellos isogrupos con un contenido de entre 0-15% de lecturas del cuerpo del alacrán. Mientras que con porcentajes inferiores al 10% no era posible rastrear algunos componentes *bona fide* del veneno identificados en el análisis previo, al considerar porcentajes entre 10 y 15% se incrementaba de manera importante el número de falsos positivos. Por esta razón, se consideraron secuencias con hasta 10% de su composición total con lecturas provenientes del cuerpo del alacrán y, si bien este punto de corte puede mantener algunos isogrupos erróneamente clasificados como componentes del veneno, la evaluación individual de las secuencias así como la búsqueda de péptidos señal y dominios (motivos de cisteínas, dominios Kunitz, entre otros), eliminaron aquellos isotigs con similitud a toxinas poco significativa.

7.4.5.1 Toxinas CAP

Las proteínas CAP [(*Cystein-rich secretory proteins* (CRISP), *Antigen 5* (Ag5), *Pathogenesis-related* (PR-1)] se encuentran distribuidas en un gran número de organismos procariones y eucariotes no vertebrados. Generalmente los miembros de esta superfamilia de proteínas se secretan y comparten similitudes

funcionales dadas por la conservación del dominio CAP; sin embargo, variaciones estructurales fuera de este dominio alteran la especificidad a diferentes blancos de acción y por tanto, alteran las consecuencias biológicas que estas tienen (revisado por Fry BG *et al.* 2009). Los alérgenos identificados en los venenos de insectos y de alacranes (alérgeno 5 de *Tityus serrulatus*) son proteínas secretadas ricas en cisteínas, por lo que se clasifican dentro de la familia CRISP. Se identificó un isotig conformado por lecturas provenientes del telson en estados activo y de mantenimiento, 66% idéntico al alérgeno 5 del alacrán *Tityus serrulatus* (Fig. 17). Aunque no existen reportes claros sobre la presencia de alérgenos en los venenos de alacranes puesto que estos han sido asociados a los venenos de insectos como avispa, abejas y hormigas, este resultado hace posible pensar que los venenos de alacranes de la familia *Buthidae* contienen proteínas potencialmente causantes de reacciones alérgicas en casos de picadura. Será interesante confirmar la respuesta inmune que los alérgenos de alacrán puedan despertar.

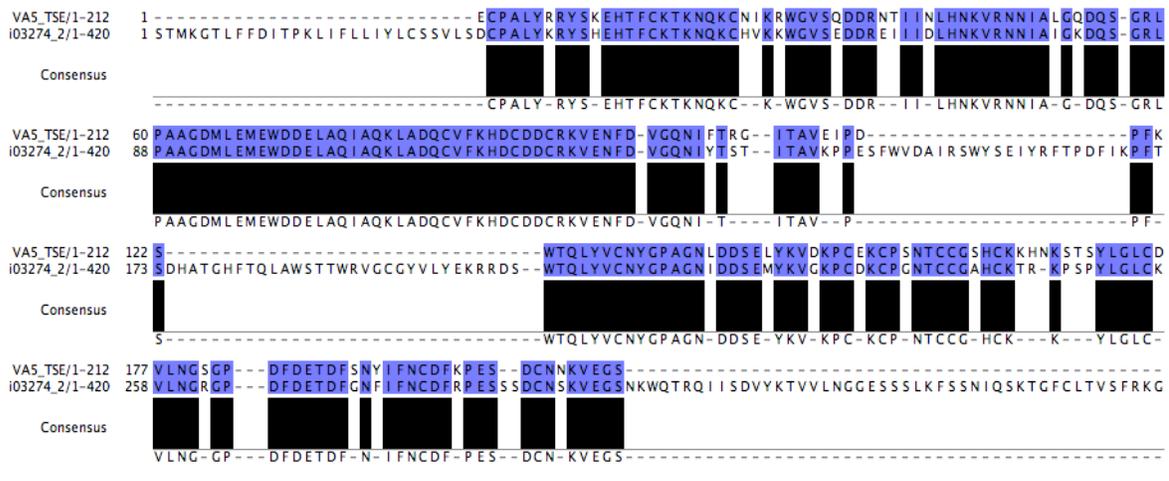


Figura 17. Alineamiento del alérgeno 5 de *T. serrulatus* con una secuencia de *C. noxius*

7.4.5.2 Péptidos con actividad antimicrobiana

Existen múltiples ejemplos de péptidos con actividad antimicrobiana caracterizados en venenos de diferentes especies de alacranes. La expresión en sistemas heterólogos y la síntesis química de dichas proteínas es de suma importancia para contender con bacterias resistentes a fármacos tradicionalmente utilizados en el tratamiento de infecciones. Algunos de estos ejemplos son la vejovina aislada del veneno del alacrán mexicano *Vaejovis mexicanus* (Hernández-Aponte *et al.* 2011); la parabutoporina de *Parabuthus schlechteri*, la escorpina aislada en especies como *Pandinus imperator* (Conde *et al.* 2000) y *Tytius costatus* (Diego-García *et al.* 2007), las bactridinas de *Tityus discrepans* (Díaz *et al.* 2009) y la mucroporina en *Lychas mucronatus* (Dai *et al.* 2008), entre otros.

La parabutoporina (PBPO) es un péptido de 45 amino ácidos que se caracterizó recientemente en el veneno del alacrán africano *Parabuthus schlechteri* (Remijsen *et al.* 2010). Se ha observado que tiene efectos antimicrobianos en bacterias Gram+ y Gram-, así como efectos antifúngicos en concentraciones micromolares. Por otra parte, se ha propuesto que la PBPO interfiere directamente con funciones celulares del sistema inmune innato en humanos, ya que puede actuar como quimioattractor de neutrófilos y retardar su apoptosis. Los *singlets* similares a la PBPO se muestran en a figura 18.

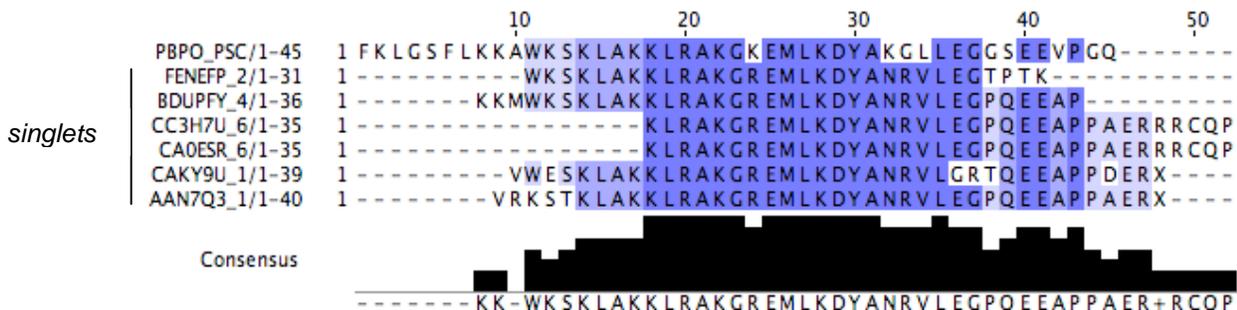


Figura 18. Alineamiento de la PBPO y 6 *singlets* de *C. noxius*

7.4.5.3 Toxinas que afectan canales iónicos

Los componentes mejor caracterizados en los venenos de los alacranes son los péptidos que reconocen canales iónicos (Possani *et al.* 2000) y receptores en membranas excitables. Estas toxinas se han clasificado en función de la especificidad de las especies a las que pueden afectar (mamíferos, insectos, crustáceos), de sus receptores blanco (sodio, potasio), la longitud de las secuencias (largas o cortas), el mecanismo de acción y sitios de unión (α y β).

Toxinas moduladoras de la función de canales de sodio

Una de las características que distingue a los venenos de las especies de la familia Buthidae es la abundancia de toxinas modificadoras de canales de sodio, puesto que estas son prácticamente inexistentes en otras familias de alacranes (por Rodríguez de la Vega *et al.* 2005). Las toxinas que modulan la actividad de los canales de sodio, pueden constituir hasta el 10% del contenido proteico de algunos venenos. Son péptidos largos de entre 60 y 76 amino ácidos, estabilizadas por cuatro puentes disulfuro, tres de los cuales están conservados en todas las familias de alacranes. Como se describió en la sección 1.1, las toxinas que afectan canales de sodio pueden dividirse en dos grupos principales: las toxinas tipo α , que se unen al sitio 3 de los canales de sodio de manera independiente de voltaje, y las tipo β , que se unen al sitio 4. En función de sus diferencias farmacológicas, las toxinas α se dividen en tres subgrupos: las toxinas α clásicas, las cuales son particularmente tóxicas en mamíferos; las similares al tipo α , que pueden actuar tanto en insectos como en mamíferos, y las α específicas de insectos (Bosmans & Tytgat, 2007). Las toxinas de *C. noxius* específicas para los mamíferos son la Cn2-4 y Cn6-9, mientras que las específicas para los insectos y los crustáceos son la Cn1, 5, 10-12.

Con la estrategia de búsqueda planteada en este trabajo, se identificaron 27 isogrupos similares a toxinas modificadoras de canales de sodio. La topología resultante, que incluye toxinas conocidas de *C. noxius*, LVPs, y otras neurotoxinas

de especies como *Lychas mucronatus*, *Mesobuthus martensii*, *Centruroides suffusus suffusus*, *Buthus granulatus* y *Hottentota judaicus*, se muestra en la figura 19.

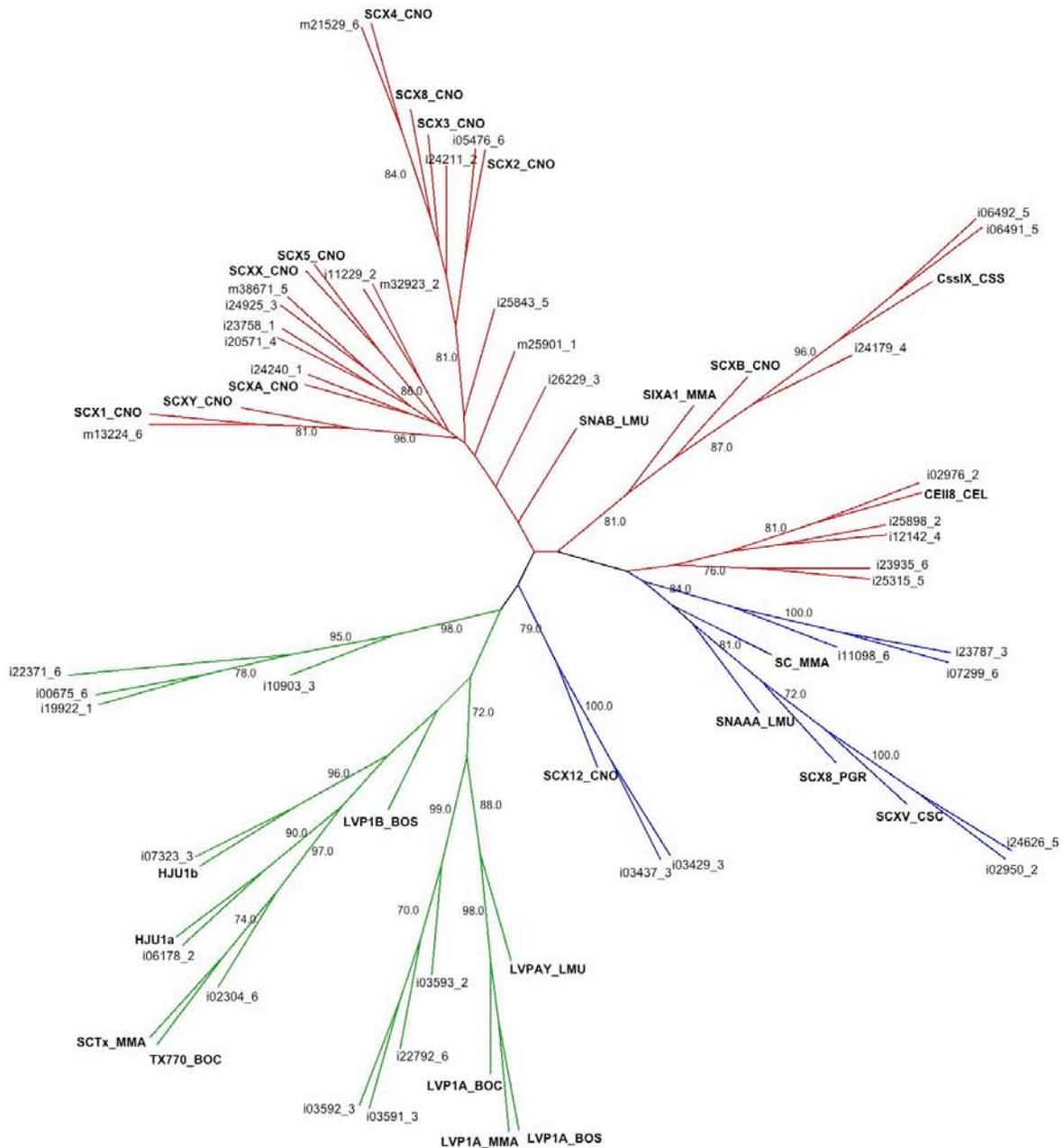


Figura 19. Filogenia de las secuencias similares a toxinas modificadoras de canales de sodio. La topología consenso fue obtenida por NJ con 100 pseudoréplicas de bootstrap. Azul: α -N α T ξ ; $\rho\phi\phi$: β -NaTx; verde: LVPs.

A partir de este árbol se pueden hacer observaciones interesantes. A diferencia de los datos bioquímicos que muestran que sólo la toxina Cn12 pertenece claramente a las α -toxinas, encontramos, por lo menos a nivel transcripcional, otras secuencias que se agrupan en esta sub-familia. Por otro lado, existen otros isotigs similares a β neurotoxinas de especies como *Centruroides suffusus*, *Mesobuthus martensii* y *Lychas mucronatus*, que no habían sido descritas en esta especie. Entre estas neurotoxinas, destacan el isotig06491 que es 94% idéntico al péptido CsslX recientemente caracterizado en el veneno de *Centruroides suffusus* (Espino-Solis *et al.* 2011); el isotig02976, 91% idéntico a la β -toxina Cell8 de *Centruroides elegans*; y los isotigs 06178 y 07323 similares a las toxinas Hj1a y Hj1b de *Hottentotta judaicus*.

Otro clado interesante, comprende secuencias de LVPs, por su nombre en inglés “lipolysis activating factor”. La primera toxina identificada dentro de esta familia fue aislada del veneno del alacrán *Buthus occitanus tunetanus* (Soudani *et al.* 2005). Esta proteína resultó ser homóloga a toxinas modificadoras de canales de sodio, pero a diferencia de estas, las LVPs mantienen una estructura heterodimérica formando un puente disulfuro entre la cadena α (de 69 residuos de longitud) y la cadena β (de 73 residuos de longitud). Ambas cadenas tienen 7 cisteínas, una de las cuales les permite formar el puente que las une. Las LVPs inducen una respuesta lipolítica en células adiposas a través de una vía mediada por adrenoreceptores tipo β . Esta familia de proteínas fue también identificada como uno de los componente más abundantes en el veneno del alacrán *L. mucronatus*, conformando entre 3 y 17% del total de péptidos similares a toxinas (Ruiming *et al.* 2010). Como se puede ver en la topología de la figura 20, la familia de LVPs está representada en el alacrán *C. noxius*, por lo menos a nivel transcripcional. Los isotigs similares a estas proteínas, mantienen aproximadamente 38-45% de identidad (Fig. 20).

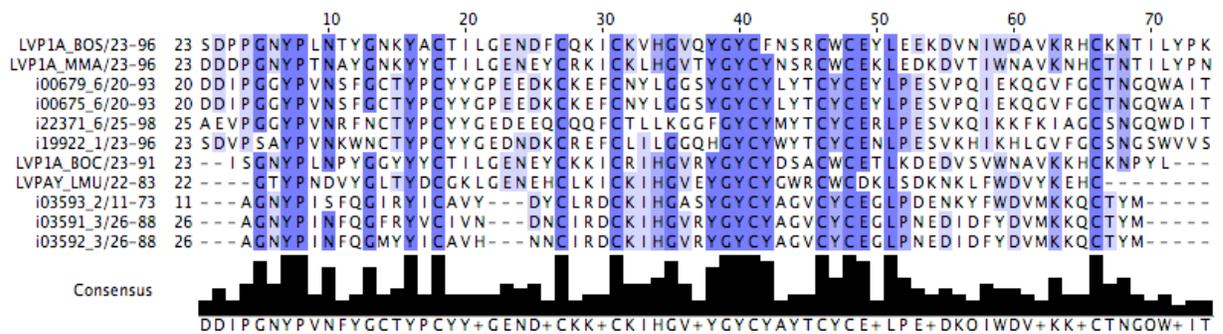


Figura 20. Alineamiento de las secuencias similares a LVP de *L. mucronatus* y *B. occitanus*

Las β -toxinas previamente descritas en *C. noxius* se muestran en la figura 21; otras secuencias similares a toxinas de las subfamilias α y β se muestran en la figura 22 (el alineamiento comprende únicamente las secuencias correspondientes a los péptidos maduros).

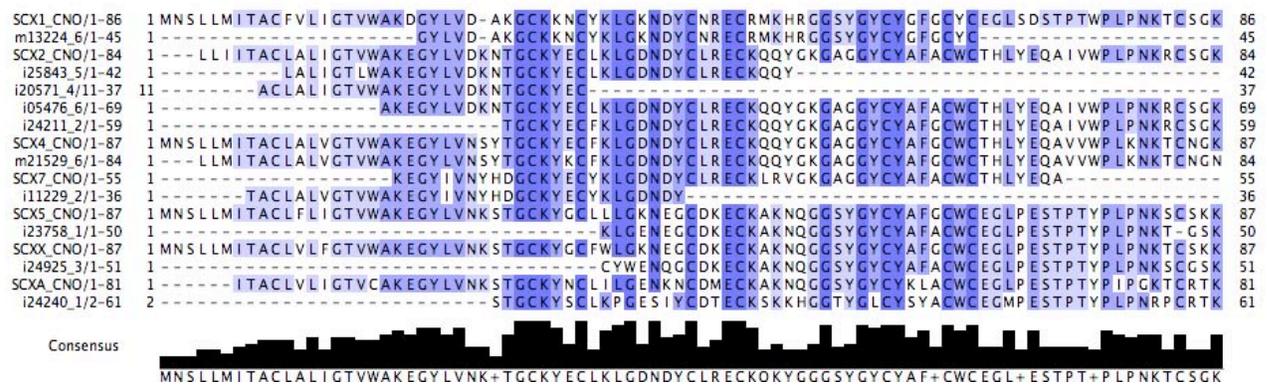


Figura 21. Alineamiento de los isotigs correspondientes a las β -toxinas de *C. noxius*

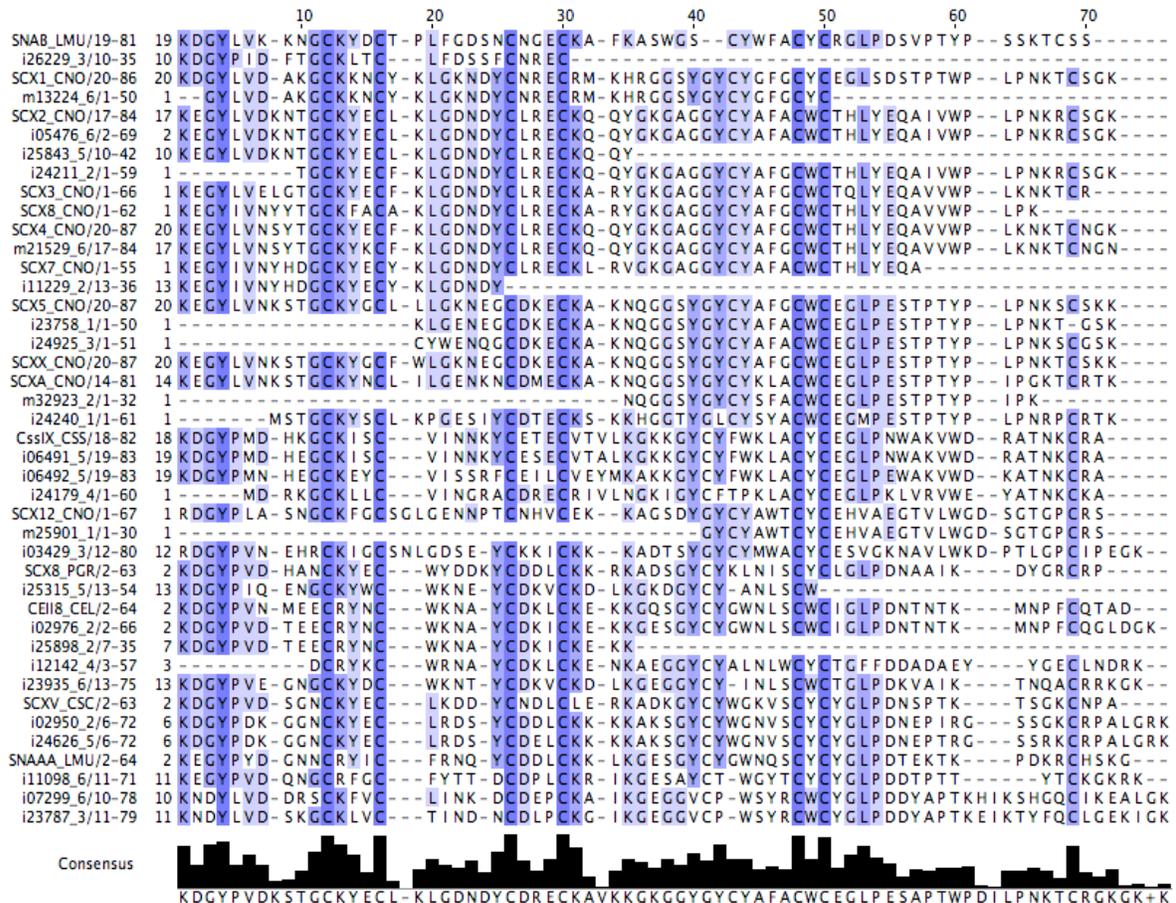


Figura 22. Alineamiento de las α y β toxinas identificadas a nivel transcripcional en *C. noxius*

Toxinas bloqueadoras de canales de potasio

Las toxinas que se unen a los canales de potasio son péptidos cortos de entre 22 y 40 amino ácidos, estabilizados por tres puentes disulfuro. Estas toxinas se encuentran en bajas cantidades en los venenos de alacranes de la familia Buthidae, representando hasta 0,1% del veneno crudo (Rodríguez de la Vega *et al.* 2004). En el caso particular de *C. noxius*, estas comprenden la slotoxina (α -1.11), la noxiustoxina (α -2.1 y α -2.4), la cobatoxina (α -10.1 y 10.2) y las toxinas Erg (γ -1.1, 3.1, 4.13, 4.11 y 4.2).

Se identificaron 15 isogrupos similares a toxinas bloqueadoras de canales de potasio que mantienen porcentajes de identidad entre 60 y 100% con respecto

Un análisis detallado de las secuencias tipo α reveló la presencia de transcritos similares a algunas toxinas de otras especies de la familia Buthidae (Fig. 24A). En efecto, los isotigs 00015 y 01725 mostraron 43 y 59% de identidad con la toxina KA171 de alacrán *M. martensii*; los isotigs 00588 y 00559 mantienen 92% de identidad con la toxina KAX2D de *C. suffusus* y el isotig 11041, 40% y 36% de identidad con las toxinas KAX45 de *T. costatus* y KA163 de *M. martensii*, respectivamente. Por otra parte, no fue posible rastrear todas las toxinas tipo Erg conocidas a nivel bioquímico (Pardo-López *et al.* 2002); sólo las Erg1 (KGX11) y 2 (KGX31) estuvieron representadas por isotigs con 100% de identidad (Fig. 24B).

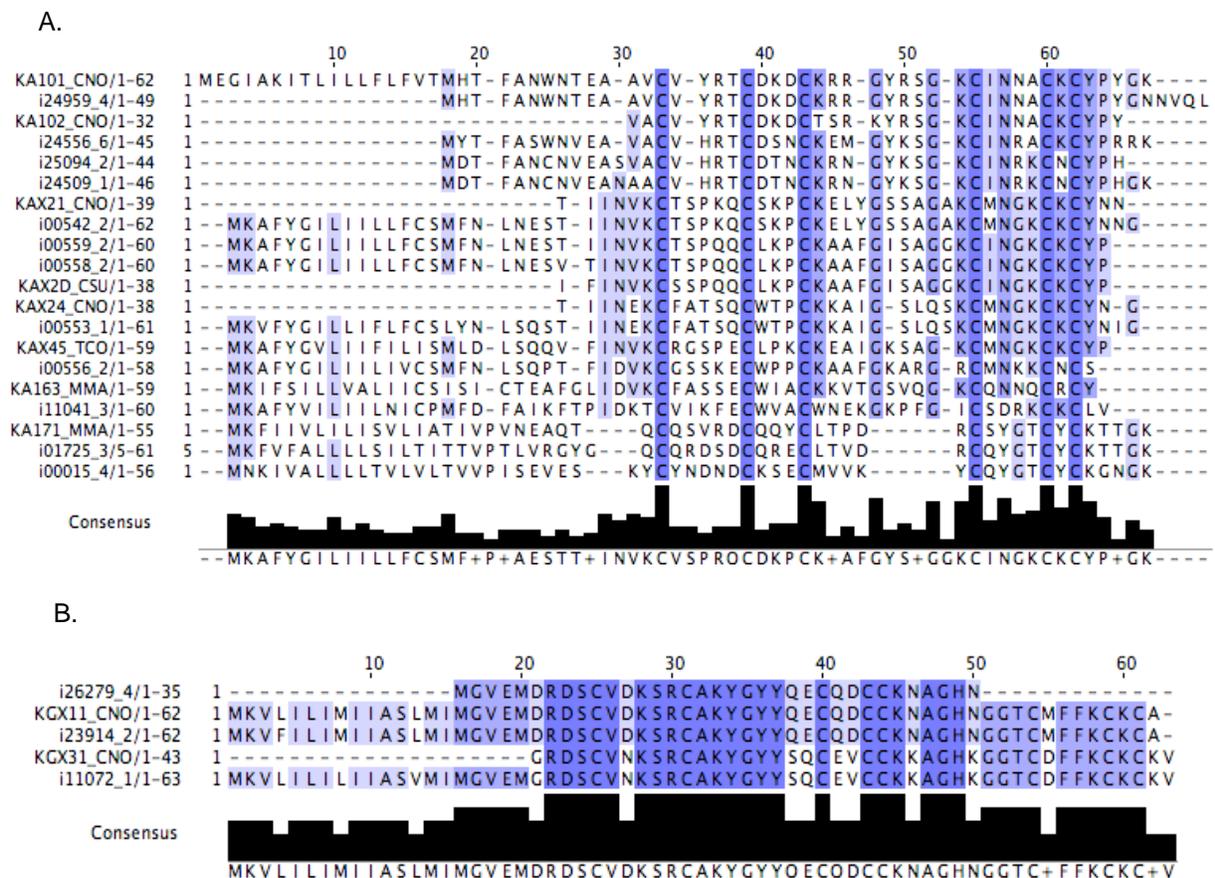


Figura 24. Toxinas bloqueadoras de canales de potasio. A. Toxinas tipo α . B. Toxinas gama Erg1 (KGX11) y Erg2 (KGX31)

Si bien las toxinas de potasio tipo β parecen estar representadas de manera ubicua en los alacranes del genero *Tityus* (*T. costatus*, *T. discrepans* y *T. trivittatus*) dentro de la familia Buthidae y en algunas especies de la familia Luridae, poco se sabe acerca de su función biológica; por ello, algunos reportes las consideran péptidos huérfanos (Diego-García *et al.* 2007). Se caracterizan por tener un extremo amino muy largo en el que se distingue un dominio rico en cisteínas. Puesto que *C. noxius* es una especie que pertenece a la familia Buthidae, no es sorprendente encontrar secuencias similares a toxinas β , como las que se muestran en el alineamiento de la figura 25. Dentro de esta familia, se identificó una secuencia (isotig23861) 74% idéntica a la toxina KBX1_TITCO (scorpine-like), un péptido con actividad antimicrobiana y bloqueadora de canales de potasio previamente identificado en las especies *Pandinus imperator* (Conde *et al.* 2000) y *Tityus costatus* (Diego-García *et al.* 2008), además de dos isotigs (19465 y 25292) que mantienen porcentajes elevados de identidad con la toxina KIK2 del alacrán *T. discrepans* (72 y 81%, respectivamente).

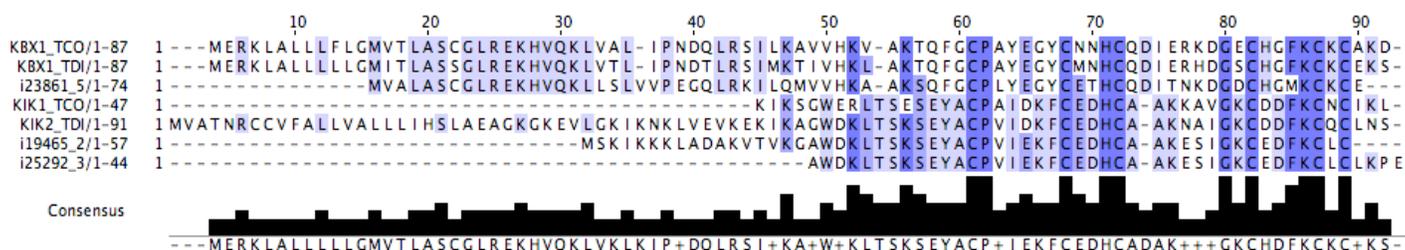


Figura 25. Alineamiento de los isotigs similares a toxinas de potasio tipo β .

A pesar de que no fue posible recuperar las secuencias de amino ácidos completas, se observan motivos y patrones de cisteínas conservados a lo largo del alineamiento. Con estos resultados, se puede concluir que las toxinas de potasio tipo β se encuentran expresadas de manera ubicua a lo largo de la familia Buthidae, incluyendo además del género *Tityus*, al género *Centruroides*.

Tanto en el caso de las toxinas que modulan la función de canales de sodio como en las que bloquean canales de potasio, no fue posible rastrear los transcritos de las toxinas caracterizadas a nivel proteico en el veneno de *C. noxius*. Tratándose de péptidos con estructuras muy estables y ricas en puentes disulfuro, es posible sugerir que no se requiera que los genes que los codifican muestren niveles de expresión elevados para mantener concentraciones de toxinas basales. Al tener una baja representación a nivel transcripcional, se vuelve poco probable que sean secuencias detectables por el sistema de pirosecuenciación. Adicionalmente, en algunas especies de alacranes como *Hottentotta judaicus* (Morgenstern *et al.* 2011) y *Hadrurus gertschi* (datos no publicados) las toxinas mayoritarias del veneno no se han sido detectado como transcritos, incluso haciendo búsquedas específicas con oligonucleótidos diseñados a partir de las secuencias de amino ácidos correspondientes, lo cual es congruente con los resultados descritos hasta el momento en *Centruroides noxius*.

Toxinas bloqueadoras de canales de calcio

Igualmente interesante fue observar isotigs similares a toxinas bloqueadoras de canales de calcio, las cuales no habían sido caracterizadas en alacranes de esta especie (Fig. 26). En particular, el isotig11091 es 46% idéntico a la toxina SCX8 del alacrán *B. occitanus*, cuya función (deducida por similitud) es inhibir receptores de rianodina. El isotig04299 es 67% idéntico a la toxina Hj1a del alacrán *H. judaicus*; si bien la función de este péptido no se conoce, siendo 73% idéntico a la SCX8_BUTOS podemos suponer que podría tener especificidad por algunos canales de calcio. Contrario a algunos reportes donde se ha descrito que los péptidos señal y propéptidos de las toxinas de alacrán se mantienen altamente conservados (Kozminsky-Atias *et al.* 2008), en este ejemplo se observa una divergencia importante en las secuencias correspondientes al péptido señal y una mejor conservación de los residuos del péptido maduro.

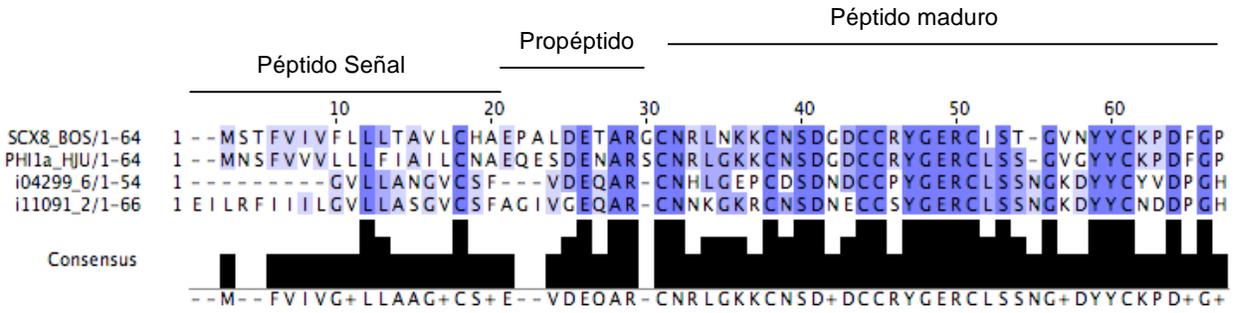


Figura 26. Isotigs similares a toxinas bloqueadoras de canales de calcio

7.4.5.4 Proteínas con dominio Kunitz

Las toxinas con dominios Kunitz se han encontrado en los venenos de reptiles, anémonas marinas, avispa, conos marinos y algunas especies de arácnidos. En el caso de las serpientes, estas proteínas inhiben factores de coagulación como la plasmina y la trombina, mientras que otras muestran actividades neurotóxicas, usando como blancos de acción canales de calcio y potasio. Las proteínas tipo Kunitz son componentes mayoritarios de la secreciones hematófagas de garrapatas y otros insectos, inhibiendo la función del factor Xa (Fry *et al.* 2009). En el caso de *Centruroides noxius*, se identificaron cuatro isogrupos similares a péptidos inhibidores de serin proteasas con dominios tipo Kunitz (Fig. 27) reportados en la anémona *Anemonia sulcata* (KC2_ANESU) y en el alacrán *Lychas mucronatus* (VP9_LYCMC). Al igual que en las toxinas que afectan canales de calcio, se observa una mayor divergencia en el péptido señal mientras que los péptidos maduros que contienen el dominio Kunitz se mantienen altamente conservados.

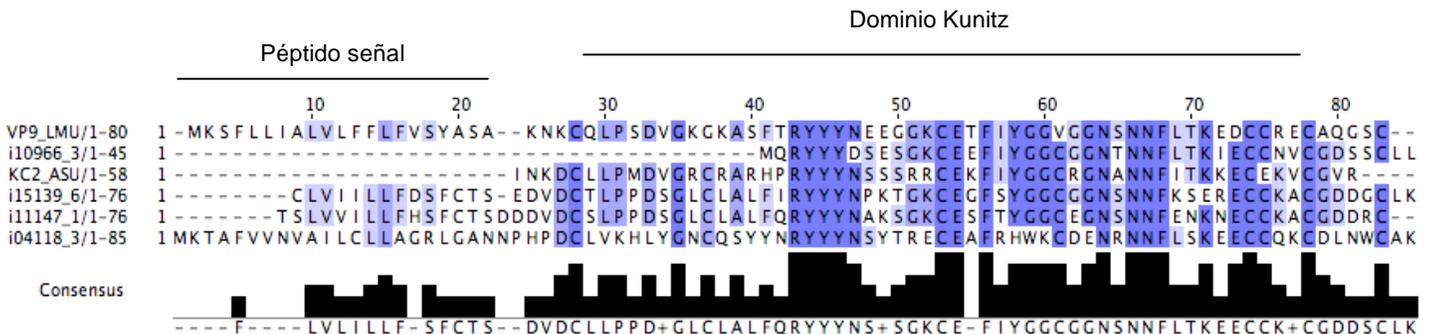


Figura 27. Alineamiento de isotigs con inhibidores de serin proteasas con dominios tipo Kunitz

7.4.5.5 Fosfolipasas

Las fosfolipasas A2 (PLA2) conforman una familia de proteínas secretorias y enzimas citosólicas que se han reclutado de manera convergente en insectos, cefalópodos, arácnidos y reptiles. Pueden actuar como potentes neurotoxinas presinápticas, pues la hidrólisis de lípidos de membrana de células nerviosas altera la conformación de la bicapa y por tanto, bloquea la liberación de neurotransmisores. Estudios previos han mostrado la presencia de PLA2 en los venenos de alacranes de la familia Luridae, como *Anuroctonus phaiodactylus* (Valdez-Cruz *et al.* 2007), *Pandinus imperator* y *Heterometrus fulvipes* (Hariprasad *et al.* 2007), sin embargo, estas proteínas no se han reportado en alacranes de la familia Buthidae.

Se identificó una secuencia 57% idéntica a PLA2 secretorias de insectos (*Apis dorsala* y *Apis mellifera*). Aunque no fue posible recuperar la secuencia de amino ácidos completa (Fig. 28), será interesante validar experimentalmente la presencia de esta proteína en el veneno de *C. noxius*.

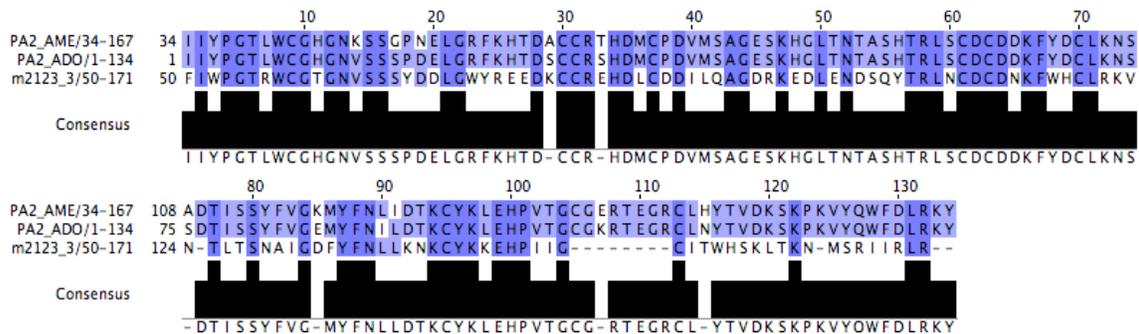


Figura 28. Alineamiento del isotig similar a PLA2 secretorias de insectos

7.4.5.6 Metaloproteasas

Las proteasas, en particular las serin y metaloproteasas se han estudiado ampliamente en los venenos de serpientes, pues son responsables de los efectos

locales y sistémicos durante los envenenamientos por mordeduras de estos reptiles. En general, estas proteasas son capaces de promover la hidrólisis de proteínas como fibrinógenos y fibrinas, lo cual ocasiona problemas de coagulación y fibrinólisis (Costa *et al.* 2010). Recientemente se ha observado la presencia de metaloproteasas y serin-proteasas en venenos de otros organismos, como alacranes, arañas y garrapatas. En este estudio se encontraron siete isogrupos similares a metaloproteasas caracterizadas en arácnidos; la topología consenso se muestra en la figura 29. Se observan tres clados diferentes; uno de ellos incluye secuencias similares a la antareasa reportada en *T. serrulatus*; en otro clado vemos un isotigs similar a la metaloproteasa-1 reportada en *M. eupens* y el tercero comprende un isotig similar a una VMPA proveniente del veneno de *Latrodectus hesperus*.

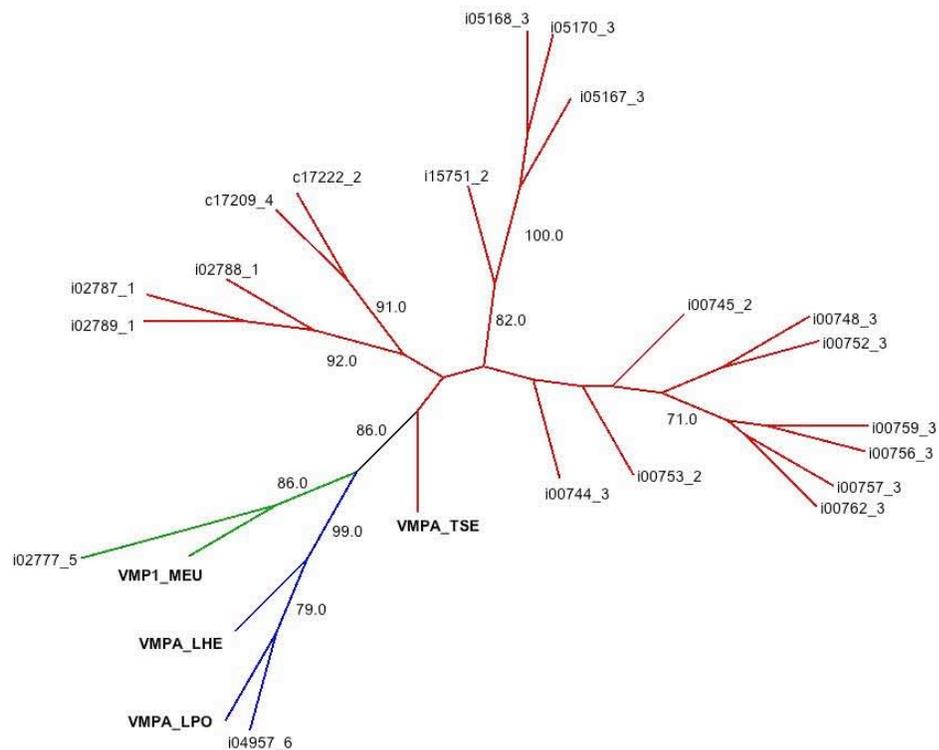


Figura 29. Filogenia de las secuencias similares a metaloproteasas de venenos de arácnidos. La topología consenso fue obtenida por NJ con 100 pseudoréplicas de bootstrap. Rojo: isotigs similares a antareasa; verde: isotig similar a la VMP1; azul: isotigs similares a la astacina arañas.

Se identificó un isogrupo con similitud a la metaloproteasa-1 del alacrán *Mesobuthus eupeus*, la cual no ha sido estudiada funcionalmente hasta el momento. Las secuencias de ambas especies mantienen 50% de identidad con una cobertura de 100% (Fig. 30).

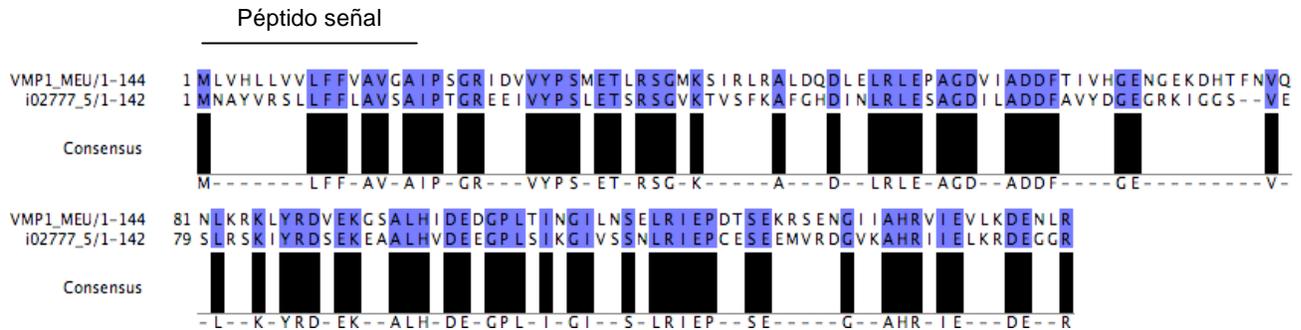


Figura 30. Alineamiento de una secuencia de *C. noxius* similar a la VMP1 del alacrán *M. eupeus*.

Algunos estudios clínicos han reportado que los venenos de alacranes pueden inducir pancreatitis aguda en un evento de envenenamiento. Recientemente se identificó una fracción purificada del veneno del alacrán brasileño *Tityus serrulatus* con capacidad de cortar proteínas SNARE en tejido pancreático exócrino. Por medio de experimentos de secuenciación directa de los péptidos de esta fracción, se mostró la presencia de una metaloproteasa dependiente de zinc denominada antareasa (Fletcher *et al.* 2009). Esta enzima corta de manera específica las v-SNAREs en regiones cercanas a la sitio de anclaje a la membrana, y por tanto impide que las reacciones de fusión se lleven a cabo de manera normal. Puesto que las proteínas SNARE son críticas para el transporte vesicular selectivo entre compartimentos celulares, las modificaciones o daños a estas proteínas tiene consecuencias irreversibles para las células. Por ello, las toxinas que pudieran atacar o interferir con esta maquinaria de transporte, tienen efectos significativos en procesos celulares. En particular, cinco isogrupos tienen porcentajes de identidad >50% con la antareasa de *T. serrulatus* (Fig. 31).

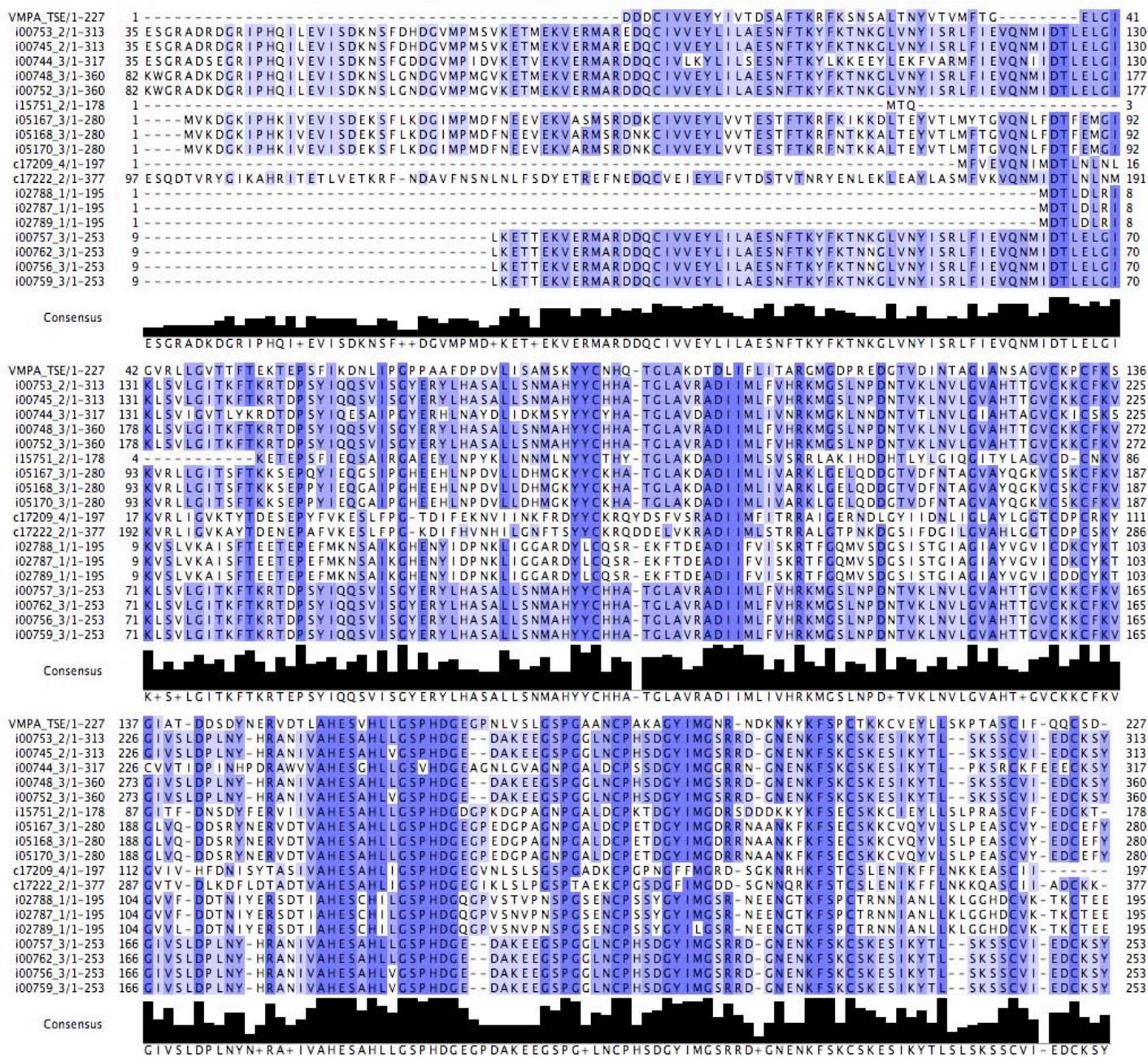


Figura 31. Alineamiento de los isotopos similares a la secuencia de antipain de *T. serrulatus*.

Se encontró otro isotipo similar a una zinc metaloproteasa (Fig. 32), la astacina, identificada en el veneno de arañas de los géneros *Loxosceles* y *Latrodectus*. Esta proteína provoca el despegado de células endoteliales en cultivos celulares, así como la degradación de fibronectina y fibrinógeno *in vitro* (Trevisan-Silva *et al.* 2010). Si bien la función de la astacina en el veneno de estas

arañas no ha sido estudiado a detalle, es probable que facilite la difusión de otras toxinas después de la mordedura. Alternativamente, podría estar relacionada con el procesamiento proteolítico de algunos componentes del veneno, o funcionar como ayuda digestiva extra-oral de las presas de estos arácnidos. La predicción con SignalP muestra que en *C. noxius* el péptido señal tiene una mayor longitud respecto a los péptidos de arañas.

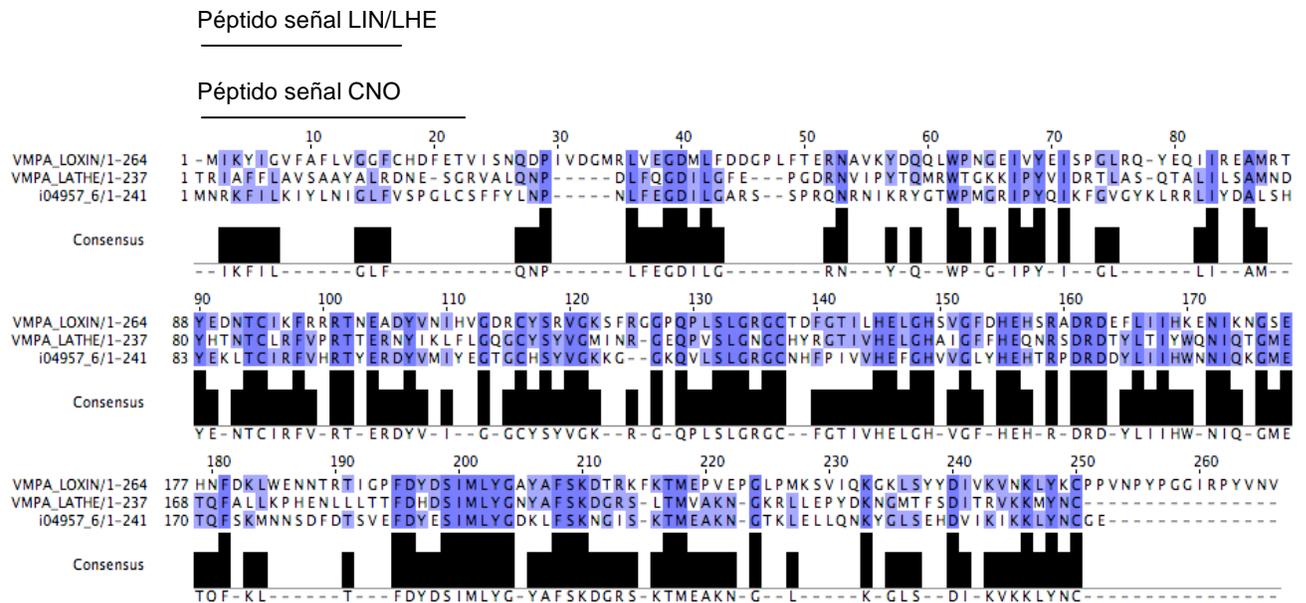


Figura 32. Alineamiento de una secuencia de *C. noxius* similar a astacina de arácnidos

7.4.5.7 Otros componentes del veneno

- Hialuronidasa

La hialuronidasa hidroliza al ácido hialurónico, el cual está ampliamente distribuido en la matriz extracelular. Esta proteína está presente de manera ubicua en el veneno de muchas especies, como reptiles, alacranes, abejas, avispas y peces, actúa como factor de difusión y facilita la permeabilidad de los tejidos y por tanto la dispersión de toxinas y otros componentes de los venenos. La primera secuencia completa reportada en alacrán, corresponde a la BmHYA1 proveniente de la especie *Buthus martensii* (Feng *et al.* 2010). Se encontró un isotig similar a

hialuronidasa dentro de las secuencias de glándula de *C. noxius*, sin embargo, la cobertura de la BmHYA1 es únicamente de 24% con un porcentaje de identidad de 80% (Fig. 33).

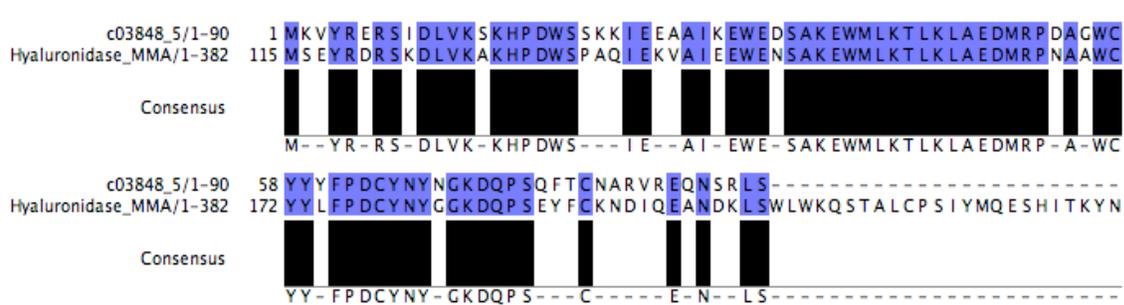


Figura 33. Alineamiento de una secuencia de *C. noxius* similar a la hialuronidasa del veneno de *M. martensii*

- Lipasa

Se encontró un isotig 60% idéntico a la proteína 164 del veneno del alacrán *L. mucronatus* (Ruiming *et al.* 2010). Si bien la función de esta toxina no se ha estudiado y sólo se ha descrito a nivel transcripcional, se identificó un dominio de colipasa de 50 residuos (de la posición 25 a 75), así como un péptido señal que cubre los primeros 20 residuos de ambas secuencias (Fig. 34).

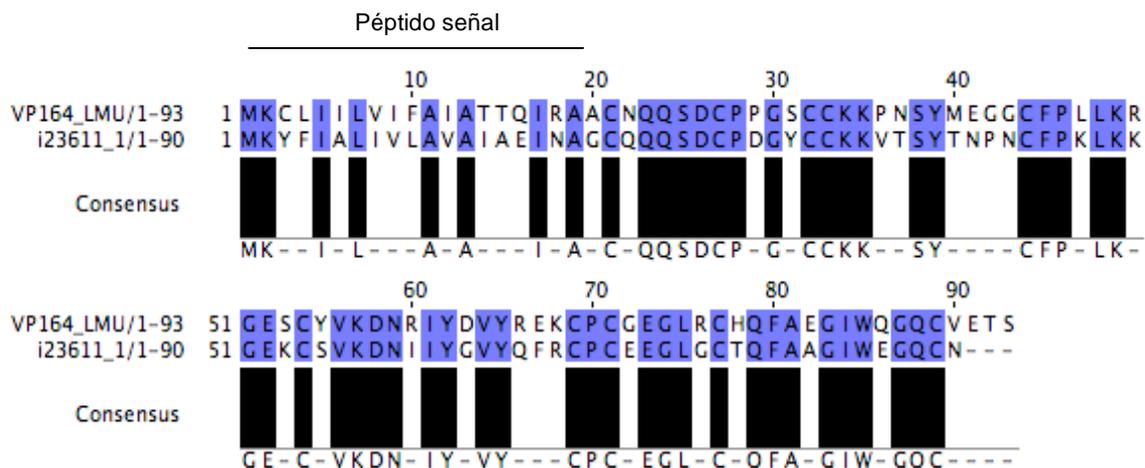
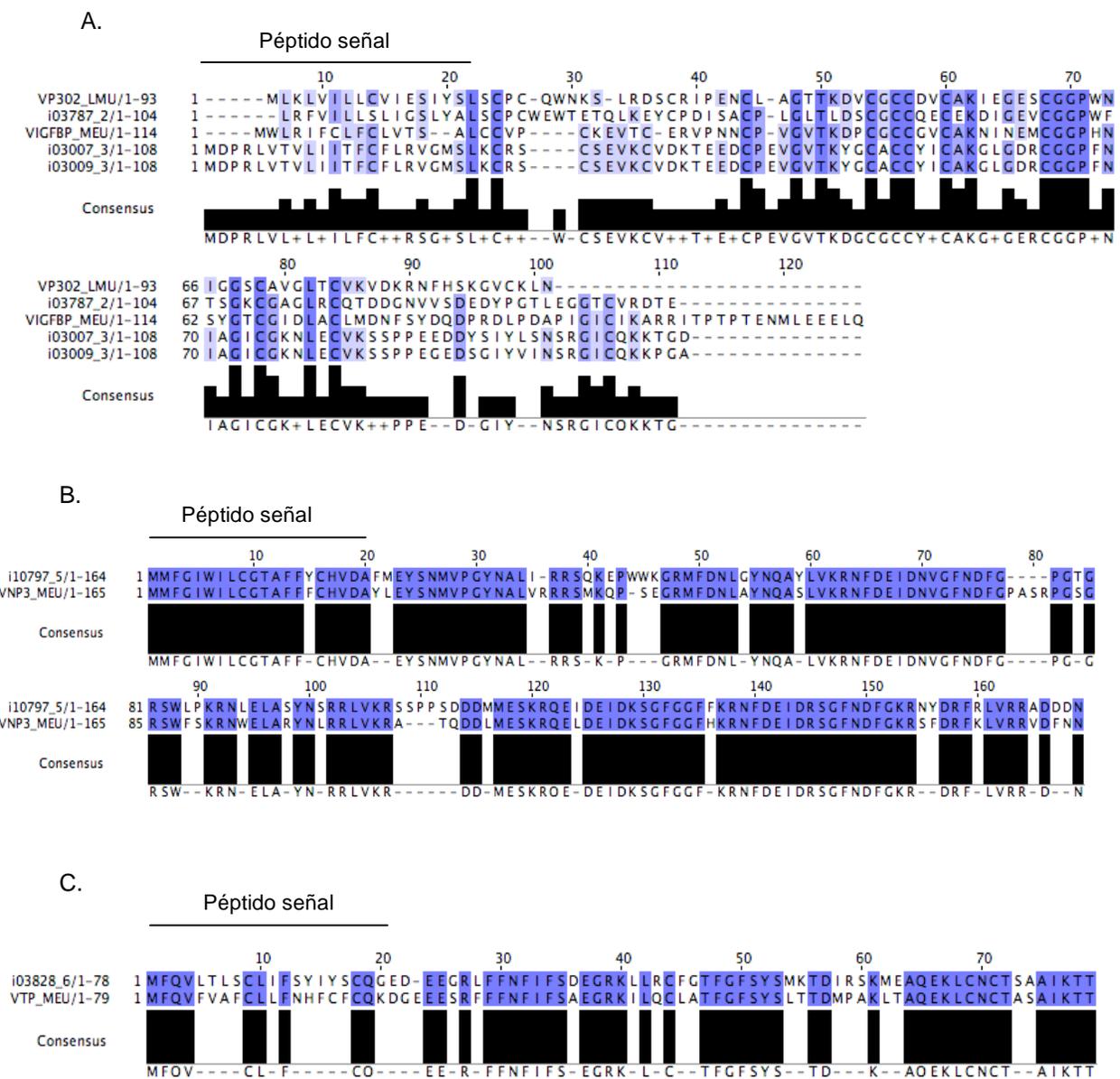


Figura 34. Alineamiento de un isotig similar a la proteína VP164 del veneno de *L. mucronatus*.

- Otros neuropéptidos

Se identificaron cinco isogrupos similares a péptidos encontrados en los venenos de las especies *L. mucronatus* (venom peptide VP302), *M. eupeus* (MeVIGFBP-1: *venom insulin growth factor binding protein*; VTP: *venom toxin-like peptide*; VNP3: *venom neuropeptide 3*) y *R. junceus* (VTX: *venom toxin-like peptide*); sin embargo, la función de estas toxinas no se ha descrito hasta el momento (Fig. 35).



transcripcional, deberán ser validadas experimentalmente. Puesto que en algunos de los ejemplos descritos se observan diferencias importantes a nivel de los péptidos señal con respecto a los péptidos encontrados en otras especies de alacranes, es probable que algunos de estos transcritos no puedan ser procesados durante la traducción para la secreción en el veneno y se degraden en la glándula. Con estas observaciones, es claro que el perfil transcripcional de la glándula de veneno no refleja directamente la composición proteica del veneno del alacrán. No podemos perder de vista que quizás algunas proteínas glándula específicas reportadas como potenciales toxinas (ej. metaloproteasas, inhibidores de proteasas, VIGFVP-1), estén involucradas en el funcionamiento de la glándula y en el procesamiento de otros componentes del veneno, y no necesariamente tengan *per se* efectos tóxicos. En este sentido, será igualmente interesante localizar los *loci* correspondientes a estas familias de proteínas para analizar su organización, posibles rearrreglos y dinámicas genómicas que han dado origen a la producción del veneno con una composición proteica determinada. Desde una perspectiva evolutiva, es evidente que la expresión específica de las toxinas descritas en este trabajo, han generado una herramienta de predación y de defensa, con blancos de acción en especies filogenéticamente lejanas como insectos y mamíferos.

7.5 Análisis cuantitativo de los niveles de expresión en la glándula de veneno en estados de actividad y mantenimiento

El protocolo de pirosecuenciación permite hacer inferencias respecto a la representación de determinados transcritos en un conjunto de datos. Dado que las muestras de cDNA no requieren ser amplificadas para la construcción de las librerías (Margulies *et al.* 2005), se espera que el número de lecturas de un transcrito refleje la abundancia de éste en una condición en particular. Esta ventaja de la plataforma de 454 se aprovechó para determinar la proporción de transcritos que se expresan diferencialmente en la glándula cuando se encuentra en estado de reposo o cuando está comprometida en la regeneración del veneno.

Para este propósito, se usaron dos pruebas estadísticas que se describen a continuación.

7.5.1 Prueba exacta de Fisher

La prueba exacta de Fisher (Triola 2004) es una prueba de significancia estadística usada en el análisis de tablas de contingencia para calcular la desviación con respecto a una hipótesis nula. El valor p derivado de esta prueba, corresponde a la probabilidad de obtener un valor del estadístico de prueba que sea al menos tan extremo como el que representa a los datos muestrales, suponiendo que la hipótesis nula es verdadera. El nivel de significancia (α) de la prueba denota la probabilidad de que el estadístico de prueba caiga en la región crítica cuando la hipótesis nula es verdadera. Si el estadístico de prueba cae en la región crítica, se debe rechazar la hipótesis nula, de modo que α es la probabilidad de cometer el error de rechazar la hipótesis nula cuando es verdadera. Lo anterior implica que la prueba de Fisher estima la tasa de falsos positivos, que se puede representar de la siguiente manera:

$$\text{Tasa de falsos positivos} \approx \# \text{ falsos positivos} / \# \text{ pruebas nulas verdaderas}$$

7.5.2 Estadístico Q

A diferencia de la prueba de Fisher, el estadístico Q controla la proporción esperada de hipótesis nulas incorrectamente rechazadas (Storey et al. 2004), lo cual se conoce como "False Discovery Rate" (FDR) y se representa de la siguiente forma:

$$\text{False discovery rate} \approx \# \text{ falsos positivos} / \# \text{ pruebas significativas}$$

Por ejemplo, si se acepta una tasa de falsos positivos de 5% con la prueba exacta de Fisher, quiere decir que en 5% de los casos, aceptaremos una hipótesis nula como significativa. Por otra parte, un FDR de 5% implica que entre las pruebas que se consideraron significativas con la prueba exacta de Fisher, 5%

serán falsos positivos. Si tenemos 100 pruebas significativas, esto resulta en 5 falsos positivos.

7.5.3 Comparación entre estados de glándula

Teniendo experimentos independientes de secuenciación con 3 sistemas diferentes de 454, se pudo evaluar la reproducibilidad de las corridas en cuanto a los niveles de transcripción detectados en cada una de ellas. Para este propósito, se calculó el número de secuencias con diferencias de expresión en glándula activa y en reposo por corrida de 454 con la prueba exacta de Fisher, tomando un valor α de 0,05 (H_0 = la abundancia transcripcional de un gen no tiene diferencias significativas entre ambas condiciones de estudio de la glándula; H_1 = el gen se expresa de manera preferencial en una condición de la glándula de veneno). Los valores p obtenidos fueron sometidos a una segunda evaluación con el estadístico Q, tomando igualmente un valor α de 0,05 (Fig. 36).

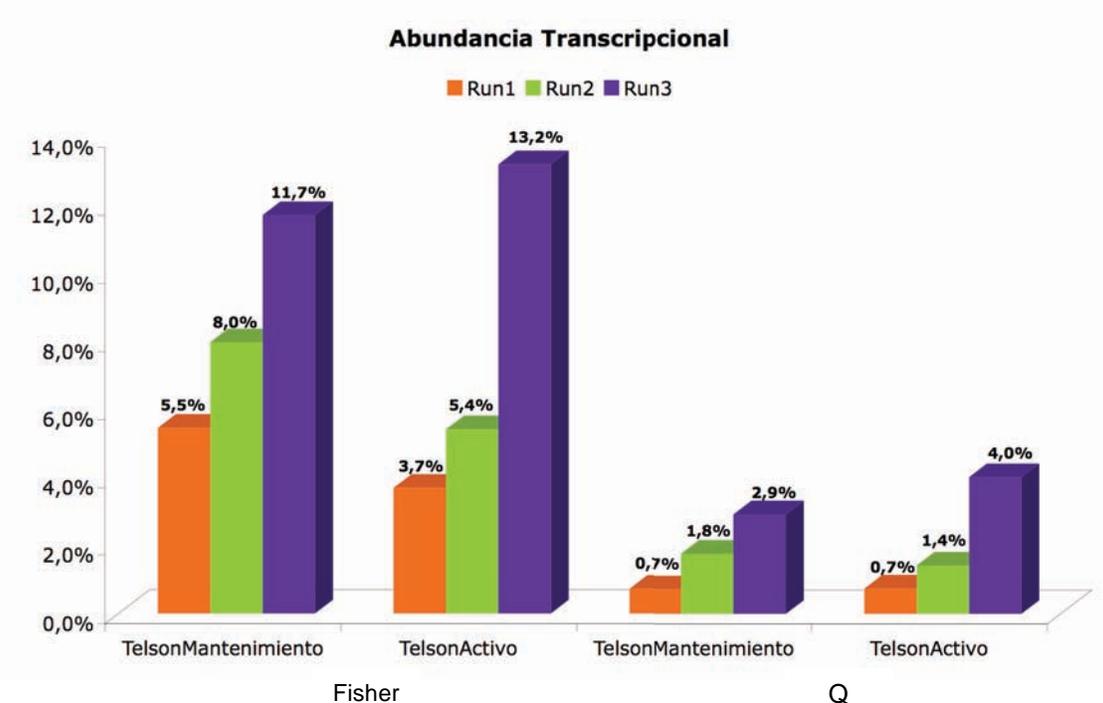
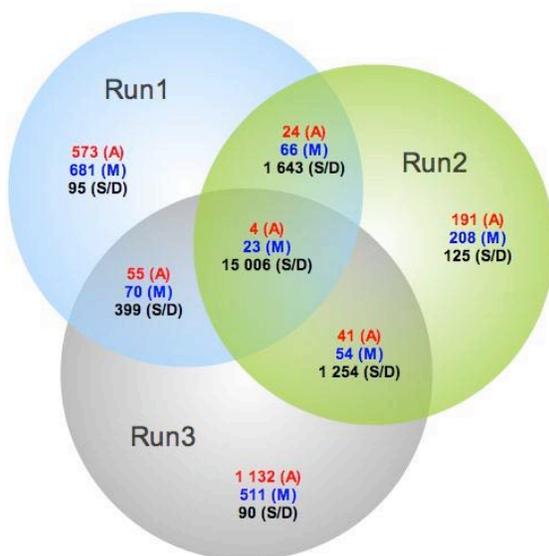


Figura 36. Porcentaje de isogrupos con mayor expresión en glándula activa o en mantenimiento. Estas proporciones fueron calculadas con la prueba exacta de Fisher y el estadístico Q ($\alpha = 0,05$).

En primer lugar, se puede destacar que los experimentos 1 y 2 de pirosecuenciación muestran un mayor porcentaje de secuencias expresadas en la glándula en estado de reposo. Si bien podríamos esperar que la glándula mostrara una mayor actividad transcripcional durante el periodo de regeneración del veneno perdido en la ordeña, ésta no se ve directamente reflejada en las corridas 1 y 2. Por otro lado, es sorprendente que en el experimento 3 cerca del 25% de los isogrupos muestra diferencias de abundancia transcripcional en glándula bajo el estadístico de Fisher y 7% bajo el estadístico Q.

Como se observa en la figura 37, la intersección de los tres experimentos es mínima, pues únicamente cuatro isogrupos son consistentemente expresados de manera preferencial en glándula en estado de reposo validado por el estadístico Q. De esta comparación se vuelve evidente que la reproducibilidad de las corridas es baja y puede deberse a una cobertura parcial y heterogénea del transcriptoma de este organismo tomando las corridas de 454 por separado.

A.



B.

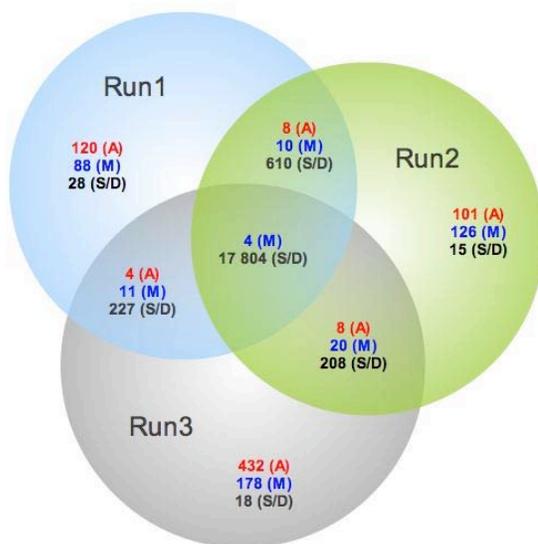


Figura 37. Evaluación de los isogrupos con expresión diferencial consistente en los experimentos de pirosecuenciación. A. Intersección de los isogrupos validados con el estadístico de Fisher. B. Intersección de los isogrupos validados con el estadístico Q. (A): glándula activa; (M): glándula en mantenimiento; (S/D): sin diferencia de expresión.

Basados en las observaciones anteriores, es claro que no podemos asumir homogeneidad en la cobertura de los transcritos en los experimentos de secuenciación de manera independiente. Por ello, se hizo una evaluación de las diferencias de abundancia transcripcional tomando en cuenta las tres corridas de 454 de manera conjunta. De este análisis (Fig. 38), 599 isogrupos mostraron mayor expresión en la glándula activa y 400 en la glándula en estado de reposo, que corresponden al 3% y al 2% del total de isogrupos ensamblados, respectivamente.

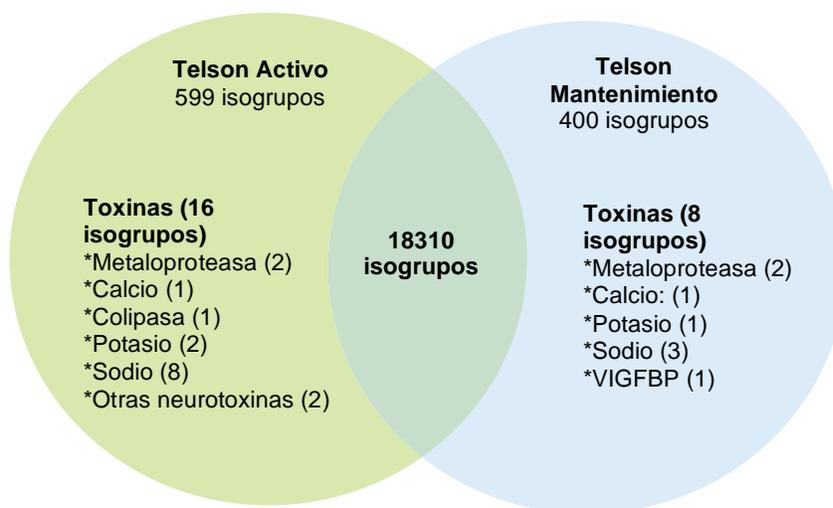


Figura 38. Isogrupos con diferencias de abundancia transcripcional validadas con el estadístico Q entre estados de glándula.

Entre estos transcritos, destacan secuencias similares a toxinas, como se muestra en la figura 39. Estas observaciones son congruentes con los perfiles transcripcionales que se han estudiado hasta el momento en otras especies de alacranes. En efecto, se ha observado de manera recurrente que las librerías de cDNA construidas a partir de RNA extraído de glándulas en proceso de regeneración del veneno (posterior a la ordeña por estimulación eléctrica), están enriquecidas con secuencias similares a toxinas, mientras que al estudiar

glándulas en reposo, esta proporción disminuye de manera importante y sólo se observan transcritos que permiten mantener un contenido basal de toxinas. Dentro de los isogrupos similares a NaTx que se expresan de manera diferencial entre estados de glándula, es importante notar que ninguno de ellos corresponde a la toxina Cn2 (isotigs 05476, 20571 y 24211; Fig. 23), que es la toxina de mayor toxicidad para mamíferos. Tomando en cuenta que los mamíferos no son los principales predadores y tampoco una presa natural de los alacranes, no es sorprendente que la producción de esta toxina a nivel transcripcional no sea prioritaria durante la regeneración del veneno. Por otra parte, dos de los cinco isogrupos similares a la antareasa de *T. serrulatus* están más representados en el estado activo y un tercero en el estado de mantenimiento de la glándula.

Cabe plantear la posibilidad de que algunos de estos transcritos sean expresados sin llegar a procesarse para ser posteriormente secretados en el veneno, como se observó con otras secuencias de toxinas en el perfil transcripcional del alacrán *H. judaicus* (Morgestern *et al.* 2011), lo cual implica que los *loci* de toxinas son regiones del genoma activamente transcritos, aunque no necesariamente los productos de este proceso sean funcionales a nivel de proteínas.

Adicionalmente, los isogrupos que mostraron diferencias de expresión validada por el estadístico Q fueron mapeados sobre las rutas metabólicas de KEGG. La distribución de los procesos celulares se muestra en la figura 39. De esta clasificación funcional, destacan por la proporción de secuencias de expresión en glándula activa, las categorías correspondientes al metabolismo de carbohidratos, de lípidos y de amino ácidos. Estas observaciones son congruentes con datos referentes al costo metabólico asociado a la producción de veneno, los cuales mostraron a través de mediciones de consumo de oxígeno en el alacrán *Parabuthus transvaalicus*, que hay un incremento de casi 40% en el consumo de oxígeno durante las 72 horas posteriores a la extracción del veneno (Nisani *et al.* 2007).

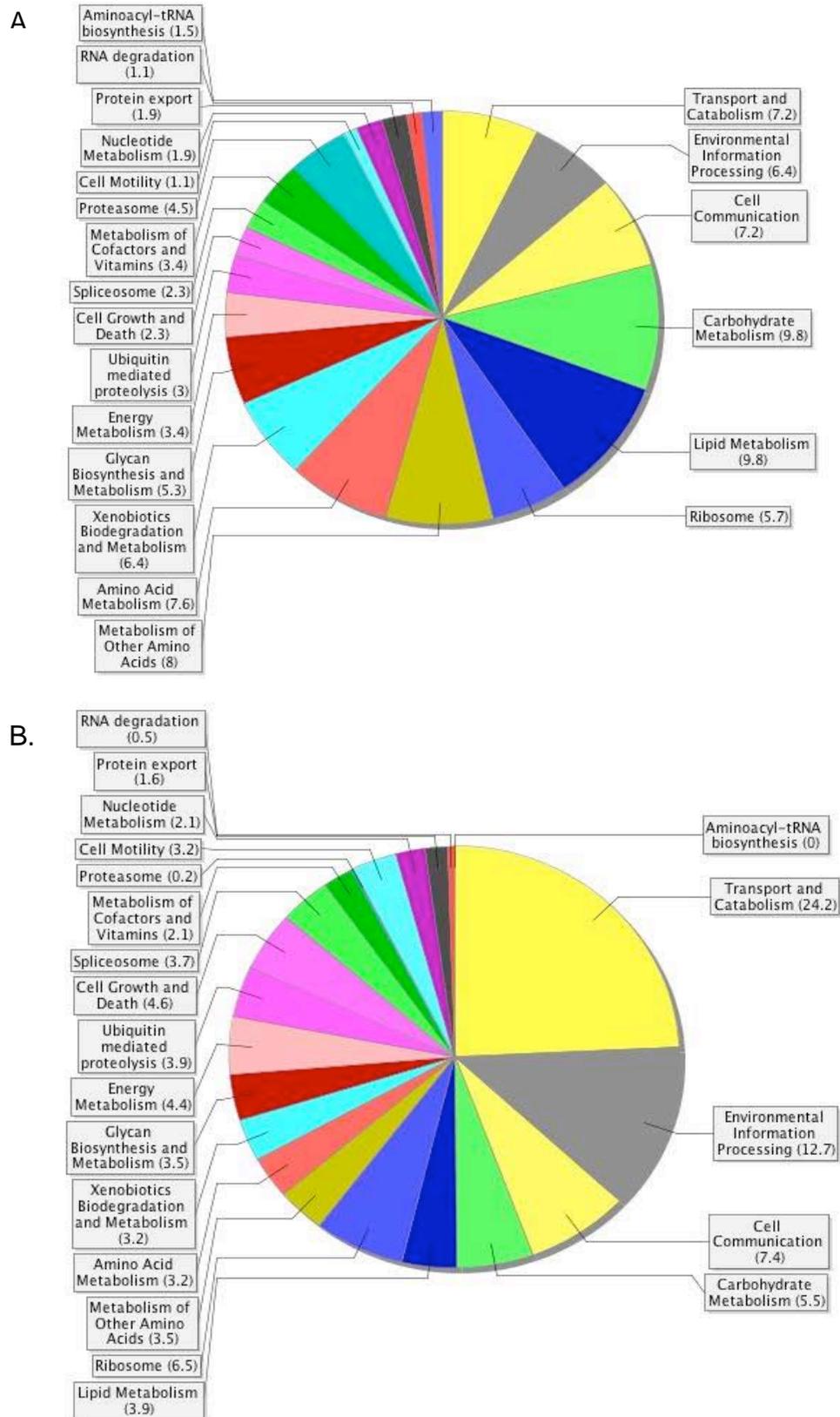


Figura 39. Categorías funcionales representadas de manera diferencial entre estados de glándula. Los valores entre paréntesis representan el porcentaje de secuencias con alineamientos significativos en NR-NCBI. A. Telson activo. B. Telson en reposo o mantenimiento.

También se observa una mayor proporción de componentes del proteasoma transcritos preferencialmente en el telson en estado activo, lo cual puede reflejar que la glándula requiere un mayor control de calidad de las proteínas que se están sintetizando para regresar a su estado basal. Considerando que la extracción eléctrica del veneno es un estrés externo que afecta directamente la glándula, no es sorprendente encontrar estos componentes altamente representadas a nivel transcripcional.

El conjunto de secuencias comprendidas en las categorías de metabolismo de amino ácidos y de biosíntesis de aminoacil-tRNAs obtenidas en estado activo, sugiere que la glándula mantiene una actividad traduccional elevada, probablemente relacionada con la regeneración del veneno perdido durante la ordeña. En estado de mantenimiento o reposo, se observa una mayor representación de transcritos relacionados al procesamiento de estímulos externos (información ambiental), que incluyen proteínas de transporte de membrana y mecanismos de transducción de señales. A diferencia del estado activo, en reposo parece haber más eventos de endo/exocitosis, actividad de peroxisomas y lisosomas, reflejados en el alto porcentaje de secuencias involucradas en catabolismo y transporte.

Sin embargo, cabe destacar que la mayor proporción de genes expresados de manera diferencial entre estados de la glándula (cerca del 80% en ambos casos), no tienen homología con secuencias depositadas en las bases de datos o corresponden a proteínas hipotéticas. Por esta razón, es posible plantear que las diferencias funcionales y de regulación que ocurren durante la regeneración del veneno o que permiten mantener a la glándula en un estado basal, están dadas por genes de función hasta el momento desconocida que deberán ser cuidadosamente analizados en estudios posteriores.

8 Conclusiones y perspectivas

Como resultado de este estudio, se logró obtener un ensamblado robusto del transcriptoma de *Centruroides noxius* tanto a nivel cuantitativo (longitud de los isotigs, aprovechamiento de las lecturas) como cualitativo (identificación genes esenciales completos, toxinas reportadas). Sin embargo, es claro que no se pudo alcanzar una cobertura óptima, según refleja el alto número de *singlets*, las pocas secuencias repetidas y el mapeo funcional.

El diseño de los experimentos de pirosecuenciación permitió evaluar de manera diferencial la gama de ESTs presentes en la glándula de veneno y en el cuerpo del alacrán. En este sentido, la clasificación taxonómica de las secuencias ensambladas y la identificación de genes eucariotes esenciales revelaron que el ensamblado está constituido por secuencias *bona fide* de un artrópodo, y nos permiten descartar eventos de contaminación dentro de las secuencias reportadas como transcritos de alacrán. Así mismo, el análisis de las secuencias similares a toxinas mostró que existe una amplia gama de péptidos y familias de toxinas nunca antes identificados en el veneno de *C. noxius*. Esta observación confirma que a lo largo de la historia evolutiva de estos arácnidos, han ocurrido diferentes eventos de duplicación de genes que han permitido generar una combinatoria de proteínas que los alacranes utilizan eficientemente con fines de depredación y defensa.

El análisis estadístico global mostró que efectivamente, existen genes que se expresan de manera preferencial después de la extracción de veneno, entre los que destacan secuencias similares a toxinas y genes involucrados en diferentes rutas metabólicas.

De manera general, podemos concluir que los resultados obtenidos en este trabajo representan el primer estudio a gran escala de las secuencias expresadas

en el alacrán *Centruroides noxius* que, a diferencia de los perfiles transcripcionales de alacranes hasta el momento reportados, buscó describir de manera integral el universo de ESTs presentes en un organismo de enorme relevancia médica y evolutiva.

Estudios posteriores deberán incrementar la cobertura del transcriptoma del alacrán, utilizando protocolos de secuenciación masiva que complementen los datos de pirosecuenciación con los que contamos en este momento. Es posible plantear el uso de otras plataformas como Illumina o Solid, que ayudarán a ensamblar las secuencias de los transcritos obtenidos de manera parcial en este estudio, y permitirán hacer validaciones estadísticas más robustas respecto a los genes expresados diferencialmente en la glándula de veneno.

Puesto que no fue posible rastrear las secuencias de amino ácidos completas de algunas toxinas, se propone que los transcritos ensamblados en este estudio puedan usarse como anclas para el diseño de oligonucleótidos específicos y hacer búsquedas dirigidas de las toxinas completas de interés. En este sentido, se deberá también evaluar la presencia de las toxinas encontradas a nivel transcripcional en el proteoma de la glándula de veneno. La validación funcional de estas, será fundamental para su aplicación biotecnológica y como modelos en estudios de estructura-función de proteínas. Por otro lado, las diferencias en niveles de expresión deberán analizarse con metodologías de biología molecular, como PCR en tiempo real, y correlacionarse a nivel proteómico.

Este estudio sienta las bases para hacer búsquedas de familias de genes específicas de alacrán, que puedan ser de interés para entender procesos celulares complejos, o que puedan tener aplicaciones biotecnológicas.

9 Bibliografía

1. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol.* 215: 403-10.
2. Bosmans F, Tytgat J. 2007. Voltage-gated sodium channel modulation by scorpion α -toxins. *Toxicon.* 49: 142–58.
3. Boyer LV, Theodorou AA, Berg RA, Mallie J, Chávez-Méndez A, García-Ubbelohde W, Hardiman S, Alagón A. 2009. Antivenom for critically ill children with neurotoxicity from scorpion stings. *N Engl J Med.* 360: 2090-8.
4. Carreño Campos C. 2007. Caracterización molecular de un banco de cDNA de la glándula venenosa del alacrán mexicano *Centruroides noxius* Hoffman (Tesis de Licenciatura – UAEM).
5. Catalán A, Espoz MC, Cortés W, Sagua H, González J, Araya JE. 2010. Tetracycline and penicillin resistant *Clostridium perfringens* isolated from the fangs and venom glands of *Loxosceles laeta*: its implications in loxoscelism treatment. *Toxicon.* 56:890-6.
6. Cestèle S, Catterall WA. 2000. Molecular mechanisms of neurotoxin action on voltage-gated sodium channels. *Biochimie.* 82:883-92.
7. Chevreux B, Pfisterer T, Drescher B, Driesel AJ, Müller WE, Wetter T, Suhai S. 2004. Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. *Genome Res.* 14: 1147-59.
8. Chuang RS, Jaffe H, Cribbs L, Perez-Reyes E, Swartz KJ. 1998 Inhibition of

- T-type voltage-gated calcium channels by a new scorpion toxin. *Nat Neurosci.* 1:668-74.
9. Conde R, Zamudio FZ, Rodríguez MH, Possani LD. 2000. Scorpine, an anti-malaria and anti-bacterial agent purified from scorpion venom. *FEBS Lett.* 471:165-8.
 10. Conde R, Zamudio FZ, Becerril B, Possani LD. 1999. Phospholipin, a novel heterodimeric phospholipase A2 from *Pandinus imperator* scorpion venom. *FEBS Lett.* 460: 447-50.
 11. Corona M, Coronas FV, Merino E, Becerril B, Gutiérrez R, Rebolledo-Antunez S, García DE, Possani LD. 2003. A novel class of peptide found in scorpion venom with neurodepressant effects in peripheral and central nervous system of the rat. *Biochim Biophys Acta.* 1649:58-67
 12. Corona M, Gurrola GB, Merino E, Cassulini RR, Valdez-Cruz NA, García B, Ramírez-Domínguez ME, Coronas FI, Zamudio FZ, Wanke E, Possani LD. 2002. A large number of novel Ergtoxin-like genes and ERG K⁺-channels blocking peptides from scorpions of the genus *Centruroides*. *FEBS Lett.* 532: 121-6.
 13. Corzo G, Escoubas P, Villegas E, Barnham KJ, He W, Norton RS, Nakajima T. 2001. Characterization of unique amphipathic antimicrobial peptides from venom of the scorpion *Pandinus imperator*. *Biochem J.* 359:35-45.
 14. Costa Jde O, Fonseca KC, Garrote-Filho MS, Cunha CC, de Freitas MV, Silva HS, Araújo RB, Penha-Silva N, de Oliveira F. 2010. Structural and functional comparison of proteolytic enzymes from plant latex and snake venoms. *Biochimie.* 92:1760-5.
 15. Dai C, Ma Y, Zhao Z, Zhao R, Wang Q, Wu Y, Cao Z, Li W. 2008.

Mucroporin, the first cationic host defense peptide from the venom of *Lychas mucronatus*. *Antimicrob Agents Chemother.* 52: 3967-72.

16. DeBin JA, Maggio JE, Strichartz GR. 1993. Purification and characterization of chlorotoxin, a chloride channel ligand from the venom of the scorpion. *Am J Physiol.* 264:C361-9.
17. Díaz P, D'Suze G, Salazar V, Sevcik C, Shannon JD, Sherman NE, Fox JW. 2009. Antibacterial activity of six novel peptides from *Tityus discrepans* scorpion venom. A fluorescent probe study of microbial membrane Na⁺ permeability changes. *Toxicon.* 54: 802-17.
18. Diego-García E, Abdel-Mottaleb Y, Schwartz E, Rodríguez de la Vega RC, Tytgat J, Possani LD. 2008. Cytolytic and K channel blocking activities of β -KTx and scorpine-like peptides purified from scorpion venoms. *Cell Mol Life Sci.* 65: 187-200.
19. Diego-García E, Schwartz EF, D'Suze G, González SA, Batista CV, García BI, de la Vega RC, Possani LD. 2007. Wide phylogenetic distribution of Scorpine and long-chain β -KTx-like peptides in scorpion venoms: identification of "orphan" components. *Peptides.* 28: 31-7.
20. Dinger ME, Amaral PP, Mercer TR, Mattick JS. 2009. Pervasive transcription of the eukaryotic genome: functional indices and conceptual implications. *Brief Funct Genomic Proteomic.* 8: 407-23.
21. D'Suze G, Schwartz EF, García-Gómez BI, Sevcik C, Possani LD. 2009. Molecular cloning and nucleotide sequence analysis of genes from a cDNA library of the scorpion *Tityus discrepans*. *Biochimie.* 91: 1010-9.
22. Eddy SR. 1998. Profile hidden Markov models. *Bioinformatics.* 14:755-63.

23. Emanuelsson O, Brunak S, von Heijne G, Nielsen H. 2007. Locating proteins in the cell using TargetP, SignalP and related tools. *Nat Protoc.* 2: 953-71.
24. Espino-Solis GP, Estrada G, Olamendi-Portugal T, Villegas E, Zamudio F, Cestèle S, Possani LD, Corzo G. 2011. Isolation and molecular cloning of beta-neurotoxins from the venom of the scorpion *Centruroides suffusus suffusus*. *Toxicon.* 57: 739-46.
25. Felsenstein, J. 2005. PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.
26. Feng L, Gao R, Meng J, Gopalakrishnakone P. 2010. Cloning and molecular characterization of BmHYA1, a novel hyaluronidase from the venom of Chinese red scorpion *Buthus martensi* Karsch. *Toxicon.* 56: 474-9.
27. Fletcher PL Jr, Fletcher MD, Weninger K, Anderson TE, Martin BM. 2010. Vesicle-associated membrane protein (VAMP) cleavage by a new metalloprotease from the Brazilian scorpion *Tityus serrulatus*. *J Biol Chem.* 285: 7405-16.
28. Fry BG, Roelants K, Champagne DE, Scheib H, Tyndall JDA, King GF, Nevalainen TJ, Norman JA, Lewis RJ, Norton RS, Renjifo C, Rodríguez de la Vega RC. 2009. The Toxicogenomic Multiverse: Convergent Recruitment of Proteins Into Animal Venoms. *Annu Rev Genomics Hum Genet.* 10: 483-511.
29. García C, Calderón-Aranda ES, Anguiano GA, Becerril B, Possani LD. 2003. Analysis of the immune response induced by a scorpion venom sub-

- fraction, a pure peptide and a recombinant peptide, against toxin Cn2 of *Centruroides noxius* Hoffmann. *Toxicon*. 41: 417-27.
30. García ML, Gao Y, McManus OB, Kaczorowski GJ. 2001. Potassium channels: from scorpion venoms to high-resolution structure. *Toxicon*. 39(6):739-48.
31. García-Valdés J, Zamudio FZ, Toro L, Possani LD. 2001. Slotoxin, α KTx1.11, a new scorpion peptide blocker of MaxiK channels that differentiates between alfa and alfa+beta (beta1 or beta4) complexes. *FEBS Lett*. 505: 369-73.
32. Gordon D, Gurevitz M. 2003. The selectivity of scorpion alpha-toxins for sodium channel subtypes is determined by subtle variations at the interacting surface. *Toxicon*.41:125-8.
33. Guan R, Wang CG, Wang M, Wang DC. 2001. A depressant insect toxin with a novel analgesic effect from scorpion *Buthus martensii* Karsch. *Biochim Biophys Acta*. 1549:9-18.
34. Hansen KD, Lareau LF, Blanchette M, Green RE, Meng Q, Rehwinkel J, Gallusser FL, Izaurralde E, Rio DC, Dudoit S, Brenner SE. 2009. Genome-wide identification of alternative splice forms down-regulated by nonsense-mediated mRNA decay in *Drosophila*. *PLoS Genet*. 5:e1000525.
35. Hariprasad G, Singh B, Das U, Ethayathulla AS, Kaur P, Singh TP, Srinivasan A. 2007. Cloning, sequence analysis and homology modeling of a novel phospholipase A2 from *Heterometrus fulvipes* (Indian black scorpion). *DNA Seq*. 18: 242-46.
36. Hernández-Aponte CA, Silva-Sanchez J, Quintero-Hernández V, Rodríguez-Romero A, Balderas C, Possani LD, Gurrola GB. 2011. Vejovine,

- a new antibiotic from the scorpion venom of *Vaejovis mexicanus*. *Toxicon*. 57: 84-92.
37. Hoffmann A. 2003. *El Maravilloso Mundo de los Arácnidos*, 3^a ed. México, FCE-SEP-CONACyT. pp 33-46.
38. Huse SM, Huber JA, Morrison HG, Sogin ML, MarkWelch D. 2007. Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biol*. 8: R143.
39. Huson DH, Auch AF, Qi J, Schuster SC. 2007. MEGAN analysis of metagenomic data. *Genome Res*. 17: 377-86.
40. Jacquier A. 2009. The complex eukaryotic transcriptome: unexpected pervasive transcription and novel small RNAs. *Nat Rev Genet*. 10: 833-44.
41. Jeyaprakash A, Hoy MA. 2009. First divergence time estimate of spiders, scorpions, mites and ticks (subphylum: Chelicerata) inferred from mitochondrial phylogeny. *Exp Appl Acarol*. 47: 1-18.
42. Kozminsky-Atias A, Bar-Shalom A, Mishmar D, Zilberberg N. 2008. Assembling an arsenal, the scorpion way. *BMC Evol Biol*. 8: 333.
43. Kumar S, Blaxter ML. 2010. Comparing *de novo* assemblers for 454 transcriptome data. *BMC Genomics*. 11: 571.
44. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ and Higgins DG. 2007. ClustalW and ClustalX version 2. *Bioinformatics*. 23: 2947-48.
45. Li S, Ma Y, Jang S, Wu Y, Liu H, Cao Z, Li W. 2009. A HindIII BAC library construction of *Mesobuthus martensii* Karsch (Scorpiones: Buthidae): an

- important genetic resource for comparative genomics and phylogenetic analysis. *Genes Genet Syst.* 84: 417-24.
46. Ma Y, Zhao R, He Y, Li S, Liu J, Wu Y, Cao Z, Li W. 2009. Transcriptome analysis of the venom gland of the scorpion *Scorpiops jendeki*: implication for the evolution of the scorpion venom arsenal. *BMC Genomics.* 10: 290.
47. Ma Y, Zhao Y, Zhao R, Zhang W, He Y, Wu Y, Cao Z, Guo L, Li W. 2010. Molecular diversity of toxic components from the scorpion *Heterometrus petersii* venom revealed by proteomic and transcriptome analysis. *Proteomics.* 10: 2471-85.
48. Margulies M, Egholm M, Altman WE, Attiya S, *et al.* 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature.* 437: 376-80.
49. Metzker ML. Sequencing Technologies – the next generation. 2009. *Nat Rev Genet.* 11: 31-46.
50. Millar JR, Koren S, Sutton G. 2010. Assembly algorithms for next-generation sequencing data. *Genomics.* 95: 315-27.
51. Mitra S, Rupek P, Richter DC, Urich T, Gilbert JA, Meyer F, Wilke A, Huson DH. 2011. Functional analysis of metagenomes and metatranscriptomes using SEED and KEGG. *BMC Bioinformatics.* 2 Suppl 1:S21.
52. Morgenstern D, Rohde BH, King GF, Tal T, Sher D, Zlotkin E. 2011. The tale of a resting gland: transcriptome of a replete venom gland from the scorpion *Hottentotta judaicus*. *Toxicon.* 57: 695-703.
53. Nieto AR, Gurrola GB, Vaca L, Possani LD. 1996. Noxiustoxin 2, a novel K⁺ channel blocking peptide from the venom of the scorpion *Centruroides noxius* Hoffmann. *Toxicon.* 34: 913-22.

54. Nisani Z, Dunbara S, Hayesa WK. 2007 Cost of venom regeneration in *Parabuthus transvaalicus* (Arachnida: Buthidae). *Comp Biochem Physiol A Mol Integr Physiol.* 147: 509-13.
55. Olamendi-Portugal T, García BI, López-González I, Van Der Walt J, Dyason K, Ulens C, Tytgat J, Felix R, Darszon A, Possani LD. 2002. Two new scorpion toxins that target voltage-gated Ca²⁺ and Na⁺ channels. *Biochem Biophys Res Commun.* 299: 562–68
56. Pardo-López L, García-Valdés J, Gurrola GB, Robertson GA, Possani LD. 2002. Mapping the receptor site for ergotoxin, a specific blocker of ERG channels. *FEBS Lett.* 510:45-9.
57. Pisani D, Poling LL, Lyons-Weiler M, Hedges SB. 2004. The colonization of land by animals: molecular phylogeny and divergence times among arthropods. *BMC Biol.* 2:1.
58. Possani LD, Becerril B, Delepierre M, Tytgat J. 1999. Scorpion toxins specific for Na⁺-channels. *Eur J Biochem.* 264:287-300.
59. Possani LD, Merino E, Corona M, Bolivar F, Becerril B. 2000. Peptides and genes coding for scorpion toxins that affect ion-channels. *Biochimie.* 82: 861-8.
60. Ramírez-Dominguez ME, Olamendi-Portugal T, García U, García C, Arechiga H, Possani LD. 2002. Cn11, the first example of a scorpion toxin that is a true blocker of Na⁽⁺⁾ currents in crayfish neurons. *J Exp Biol.* 205: 869-76.
61. Remijsen Q, Verdonck F, Willems J. 2010. Parabutoporin, a cationic amphipathic peptide from scorpion venom: much more than an antibiotic. *Toxicon.* 55: 180-5.

62. del Río-Portilla F, Hernández-Marín E, Pimienta G, Coronas FV, Zamudio FZ, Rodríguez de la Vega RC, Wanke E, Possani LD. 2004. NMR solution structure of Cn12, a novel peptide from the Mexican scorpion *Centruroides noxius* with a typical beta-toxin sequence but with alfa-like physiological activity. *Eur J Biochem.* 271: 2504-16.
63. Rodríguez de la Vega RC, Possani LD. 2004. Current views on scorpion toxins specific for K⁺-channels. *Toxicon*, 43: 865-75.
64. Rodríguez de la Vega RC, Possani LD. 2005. Overview of scorpion toxins specific for Na⁺ channels and related peptides: biodiversity, structure–function relationships and evolution. *Toxicon*, 46: 831–44.
65. Rodríguez de la Vega RC, Schwartz EF, Possani LD. 2010. Mining on scorpion venom biodiversity. *Toxicon.* 56: 1155-61.
66. Roeding F, Borner J, Kube M, Klages S, Reinhardt R, Burmester T. 2009. A 454 sequencing approach for large scale phylogenomic analysis of the common emperor scorpion (*Pandinus imperator*). *Mol Phylogenet Evol.* 53: 826-34.
67. Rudolf I, Mendel J, Sikutová S, Svec P, Masáříková J, Nováková D, Bunková L, Sedláček I, Hubálek Z. 2009. 16S rRNA gene-based identification of cultured bacterial flora from host-seeking *Ixodes ricinus*, *Dermacentor reticulatus* and *Haemaphysalis concinna* ticks, vectors of vertebrate pathogens. *Folia Microbiol (Praha).* 54:419-28.
68. Ruiming Z, Yibao M, Yawen H, Zhiyong D, Yingliang W, Zhijian C, Wenxin L. 2010. Comparative venom gland transcriptome analysis of the scorpion *Lychas mucronatus* reveals intraspecific toxic gene diversity and new venomous components. *BMC Genomics.* 11: 452.

69. Schneider MC, Zacaro AA, Pinto-Da-Rocha R, Candido DM, Cella DM. 2009. A comparative cytogenetic analysis of 2 Bothriuridae species and overview of the chromosome data of Scorpiones. *J Hered.* 100: 545-55.
70. Schneider MC, Cella DM. 2010. Karyotype conservation in 2 populations of the parthenogenetic scorpion *Tityus serrulatus* (Buthidae): rDNA and its associated heterochromatin are concentrated on only one chromosome. *J Hered.* 101: 491-6.
71. Schwartz E, Diego-García E, Rodríguez de la Vega RC, Possani LD. 2007. Transcriptome analysis of the venom gland of the Mexican scorpion *Hadrurus gertschi* (Arachnida: Scorpiones). *BMC Genomics.* 8: 119.
72. Soudani N, Gharbi-Chihi J, Srairi-Abid N, Yazidi CM, Planells R, Margotat A, Torresani J, El Ayeub M. 2005. Isolation and molecular characterization of LVP1 lipolysis activating peptide from scorpion *Buthus occitanus tunetanus*. *Biochim Biophys Acta.* 1747: 47-56.
73. Storey JD, Taylor JE, and Siegmund D. 2004. Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates: A unified approach. *J Roy Statistical Society.* 66: 187-205.
74. Torres-Larios A, Gurrola GB, Zamudio FZ, Possani LD. 2000. Hadrurin, a new antimicrobial peptide from the venom of the scorpion *Hadrurus aztecus*. *Eur J Biochem.* 267:5023-31.
75. Trevisan-Silva D, Gremski LH, Chaim OM, da Silveira RB, Meissner GO, Mangili OC, Barbaro KC, Gremski W, Veiga SS, Senff-Ribeiro A. 2010. Astacin-like metalloproteases are a gene family of toxins present in the venom of different species of the brown spider (genus *Loxosceles*). *Biochimie.* 92: 21-32.

76. Triola MF. 2004. Estadística. 9a ed. Pearson Education. México. Pearson Educación. pp. 366-383.
77. Tytgat J, Chandy KG, García ML, Gutman GA, Martin-Eauclaire MF, van der Walt JJ, Possani LD. 1999. A unified nomenclature for short-chain peptides isolated from scorpion venoms: alpha-KTx molecular subfamilies. *Trends Pharmacol Sci.* 20:444-7.
78. Valdez-Cruz NA, Segovia L, Corona M, Possani LD. 2007. Sequence analysis and phylogenetic relationship of genes encoding heterodimeric phospholipases A2 from the venom of the scorpion *Anuroctonus phaiodactylus*. *Gene.* 396: 149-58.
79. Valdivia HH, Possani LD. 1998. Peptide toxins as probes of ryanodine receptor structure and function. *Trends Cardiovasc Med.* 8:111-8.
80. Wang H, Jin L, Zhang H. 2011. Comparison of the diversity of the bacterial communities in the intestinal tract of adult *Bactrocera dorsalis* from three different populations. *J Appl Microbiol.* 110:1390-401.
81. Wang ET, Sandberg R, Luo S, Khrebtkova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB. 2008. Alternative isoform regulation in human tissue transcriptomes. *Nature.* 456:470-6.
82. Zamudio FZ, Conde R, Arévalo C, Becerril B, Martin BM, Valdivia HH, Possani LD. 1997. The mechanism of inhibition of ryanodine receptor channels by imperatoxin I, a heterodimeric protein from the scorpion *Pandinus imperator*. *J Biol Chem.* 272:11886-94.

10 Material Suplementario

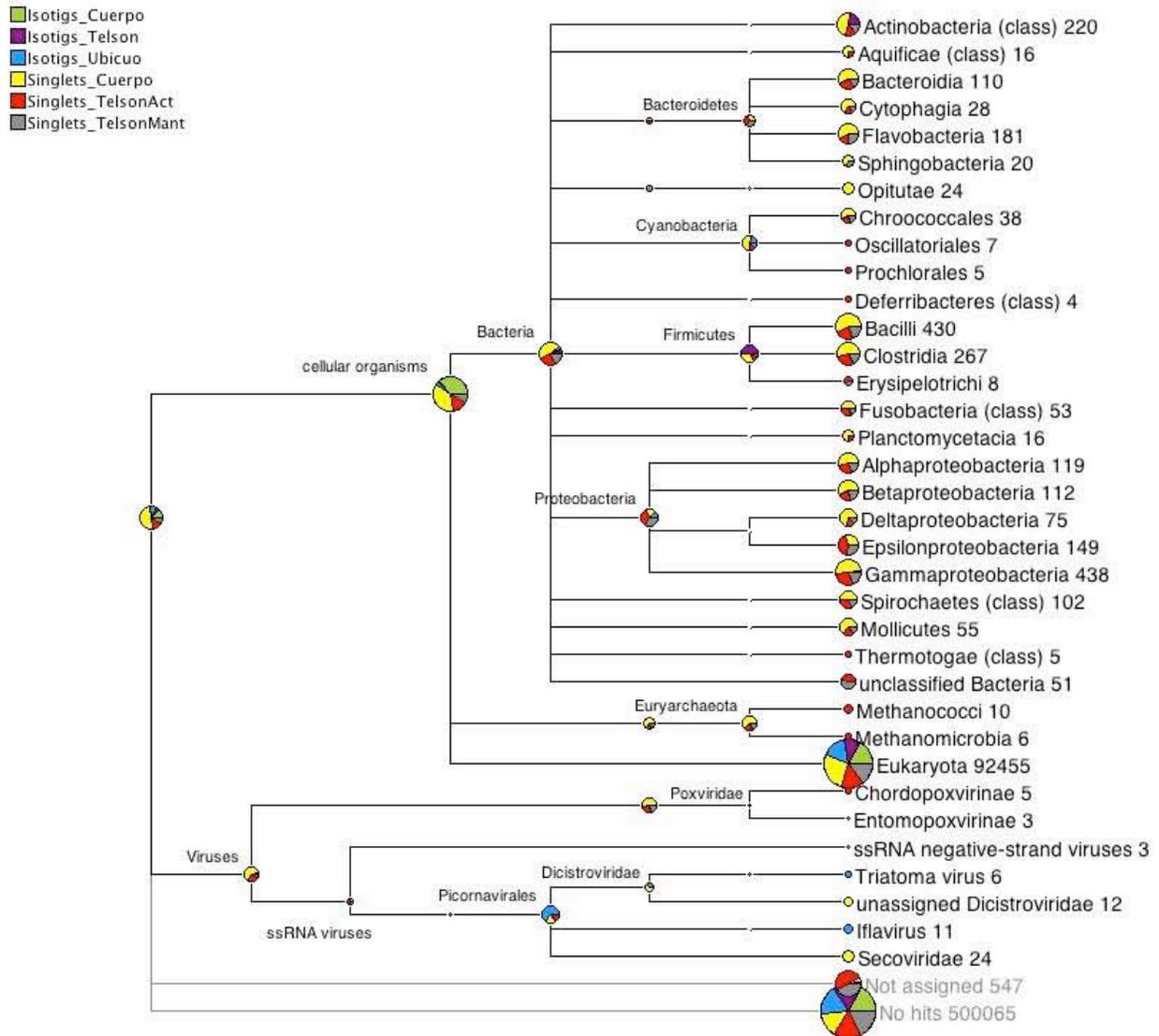


Figura suplementaria 1. Géneros bacterianos representados en las secuencias ensambladas y los *singlets* de *C. noxius*

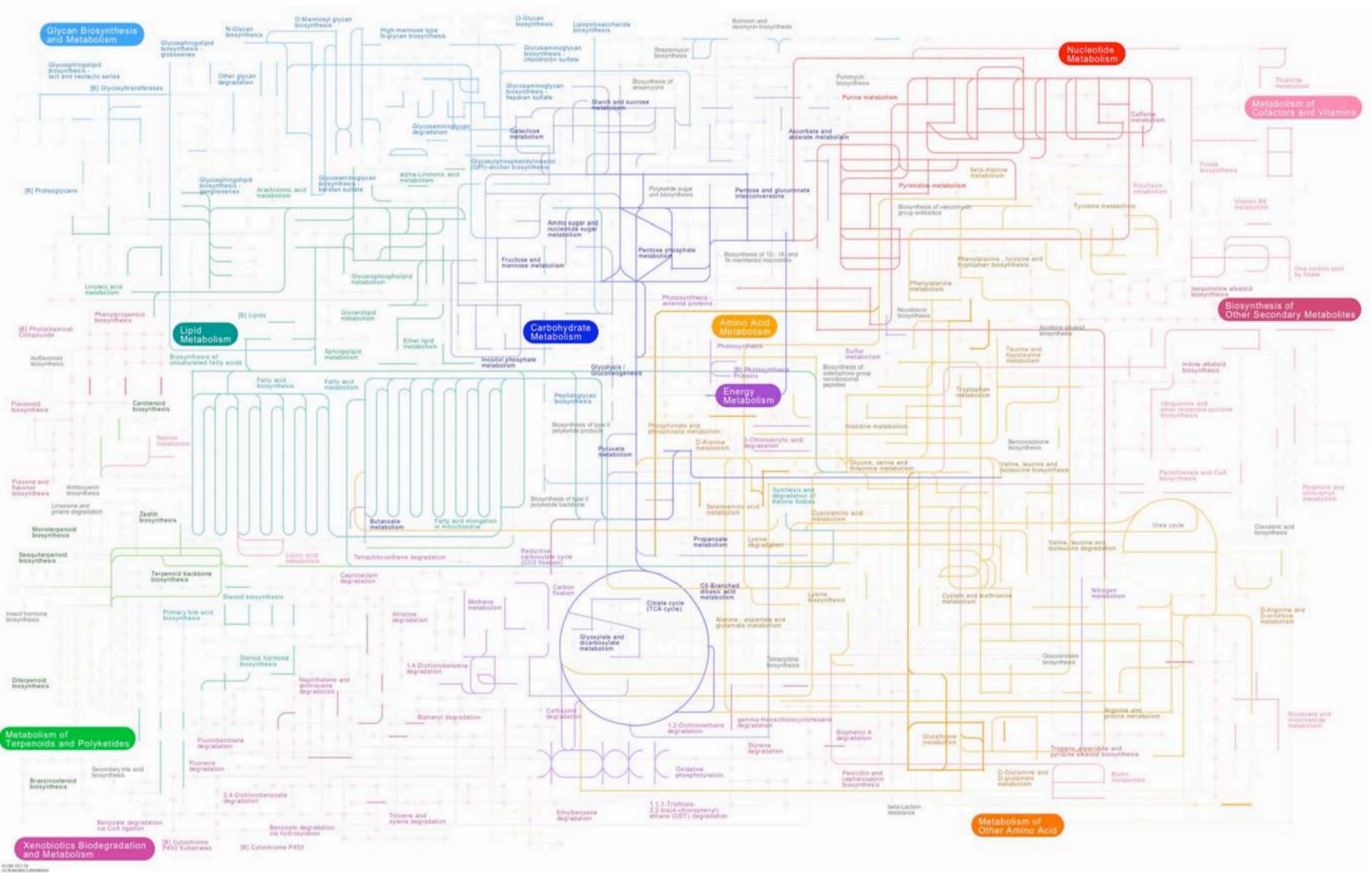


Figura suplementaria 2. Redes metabólicas representadas en el transcriptoma de *C. noxius*. Las líneas oscuras indican pasos enzimáticos cubiertos; las líneas claras pasos enzimáticos faltantes

Tabla suplementaria 1. Genes eucariotes esenciales de copia única identificados en el transcriptoma de *C. noxius*

Gen
Arginyl-tRNA synthetase (EC 6.1.1.19)
Asparaginyl-tRNA synthetase (EC 6.1.1.22)
Asparaginyl-tRNA synthetase (EC 6.1.1.22), mitochondrial
Aspartyl-tRNA synthetase (EC 6.1.1.12)
COG0536: GTP-binding protein Obg
DNA-directed RNA polymerase II 13.2 kDa polypeptide (EC 2.7.7.6)
DNA-directed RNA polymerase II 13.3 kDa polypeptide (EC 2.7.7.6)
DNA-directed RNA polymerase II 19 kDa polypeptide (EC 2.7.7.6)
DNA-directed RNA polymerase III 12.5 kDa polypeptide (EC 2.7.7.6)
DNA-directed RNA polymerases I, II, and III 14.5 kDa polypeptide (EC 2.7.7.6)
DNA-directed RNA polymerases I, II, and III 27 kDa polypeptide (EC 2.7.7.6)
DNA-directed RNA polymerases I, II, and III 8.3 kDa polypeptide (EC 2.7.7.6)
Eukaryotic peptide chain release factor subunit 1
Eukaryotic translation initiation factor 5A
GTP-binding and nucleic acid-binding protein YchF
GTPase and tRNA-U34 5-formylation enzyme TrmE
Glutamyl-tRNA(Gln) amidotransferase subunit A (EC 6.3.5.7)
Glycyl-tRNA synthetase (EC 6.1.1.14) @ Glycyl-tRNA synthetase (EC 6.1.1.14), mitochondrial
HBS1 protein
LSU ribosomal protein L10Ae (L1p)
LSU ribosomal protein L10e (L16p)
LSU ribosomal protein L11e (L5p)
LSU ribosomal protein L11p (L12e), mitochondrial
LSU ribosomal protein L12e (L11p)
LSU ribosomal protein L13Ae (L13p)
LSU ribosomal protein L13e
LSU ribosomal protein L14e
LSU ribosomal protein L15e
LSU ribosomal protein L16p (L10e), mitochondrial
LSU ribosomal protein L17e (L22p)
LSU ribosomal protein L18Ae
LSU ribosomal protein L18e
LSU ribosomal protein L1e (L4p)
LSU ribosomal protein L20p, mitochondrial
LSU ribosomal protein L21e
LSU ribosomal protein L22e
LSU ribosomal protein L22p (L17e), mitochondrial
LSU ribosomal protein L23Ae (L23p)
LSU ribosomal protein L23p (L23Ae), mitochondrial
LSU ribosomal protein L24e
LSU ribosomal protein L24p (L26e), mitochondrial
LSU ribosomal protein L26e (L24p)

LSU ribosomal protein L27Ae (L15p)
LSU ribosomal protein L27e
LSU ribosomal protein L28e
LSU ribosomal protein L29e
LSU ribosomal protein L30e
LSU ribosomal protein L31e
LSU ribosomal protein L32e
LSU ribosomal protein L33p, mitochondrial
LSU ribosomal protein L35Ae
LSU ribosomal protein L36e
LSU ribosomal protein L36p, mitochondrial
LSU ribosomal protein L37Ae
LSU ribosomal protein L38e
LSU ribosomal protein L39e
LSU ribosomal protein L39mt, mitochondrial
LSU ribosomal protein L3e (L3p)
LSU ribosomal protein L3p (L3e), mitochondrial
LSU ribosomal protein L40mt, mitochondrial
LSU ribosomal protein L44e
LSU ribosomal protein L44mt, mitochondrial
LSU ribosomal protein L46mt, mitochondrial
LSU ribosomal protein L47mt, mitochondrial
LSU ribosomal protein L48mt, mitochondrial
LSU ribosomal protein L49mt, mitochondrial
LSU ribosomal protein L4p (L1e), mitochondrial
LSU ribosomal protein L53mt, mitochondrial
LSU ribosomal protein L5e (L18p)
LSU ribosomal protein L6e
LSU ribosomal protein L7/L12 (L23e), mitochondrial
LSU ribosomal protein L7Ae
LSU ribosomal protein L7e (L30p)
LSU ribosomal protein L8e (L2p)
LSU ribosomal protein L9e (L6p)
LSU ribosomal protein L9p, mitochondrial
LSU ribosomal protein P0 (L10p)
LSU ribosomal protein P1 (L7/L12)
Leucyl-tRNA synthetase (EC 6.1.1.4)
Lysyl-tRNA synthetase (class II) (EC 6.1.1.6) @ Lysyl-tRNA synthetase (class II) (EC 6.1.1.6), mitochondrial
Phenylalanyl-tRNA synthetase alpha chain (EC 6.1.1.20)
Phenylalanyl-tRNA synthetase alpha chain (EC 6.1.1.20), mitochondrial
RNA polymerase III transcription initiation factor (TFIIIC) 95 kDa subunit
Ribonucleases P/MRP protein subunit POP7 (EC 3.1.26.5)
SSU ribosomal protein MRP10, mitochondrial
SSU ribosomal protein S10e
SSU ribosomal protein S10p (S20e), mitochondrial
SSU ribosomal protein S11e (S17p)

SSU ribosomal protein S11p (S14e), mitochondrial
SSU ribosomal protein S12e
SSU ribosomal protein S12p (S23e), mitochondrial
SSU ribosomal protein S14e (S11p)
SSU ribosomal protein S14p (S29e), mitochondrial
SSU ribosomal protein S15Ae (S8p)
SSU ribosomal protein S15e (S19p)
SSU ribosomal protein S16p, mitochondrial
SSU ribosomal protein S17p (S11e), mitochondrial
SSU ribosomal protein S18e (S13p)
SSU ribosomal protein S19e
SSU ribosomal protein S21e
SSU ribosomal protein S22mt, mitochondrial
SSU ribosomal protein S23e (S12p)
SSU ribosomal protein S23mt, mitochondrial
SSU ribosomal protein S24e
SSU ribosomal protein S24mt, mitochondrial
SSU ribosomal protein S26e
SSU ribosomal protein S27e
SSU ribosomal protein S28e
SSU ribosomal protein S28mt, mitochondrial
SSU ribosomal protein S29e (S14p)
SSU ribosomal protein S29mt, mitochondrial
SSU ribosomal protein S33mt, mitochondrial
SSU ribosomal protein S34mt, mitochondrial
SSU ribosomal protein S3e (S3p)
SSU ribosomal protein S4e
SSU ribosomal protein S5p (S2e), mitochondrial
SSU ribosomal protein S6e
SSU ribosomal protein S6p, mitochondrial
SSU ribosomal protein S7e
SSU ribosomal protein S7p (S5e), mitochondrial
SSU ribosomal protein S8e
SSU ribosomal protein S9e (S4p)
SSU ribosomal protein S9p (S16e), mitochondrial
SSU ribosomal protein SAe (S2p)
Seryl-tRNA synthetase (EC 6.1.1.11)
TATA-box binding protein
Threonyl-tRNA synthetase (EC 6.1.1.3)
Transcription initiation factor IIB
Transcription initiation factor IIE beta subunit
Transcription initiation factor IIF beta subunit
Transcription initiation factor IIH cyclin-dependent kinase 7
Transcription initiation factor IIH p34 subunit
Transcription initiation factor IIH p44 subunit
Transcription initiation factor IIIA
Transcription initiation factor IIIB 70 kDa subunit

Translation elongation factor 1 alpha subunit
Translation elongation factor 1 beta subunit
Translation elongation factor 1 gamma subunit
Translation elongation factor Tu
Tryptophanyl-tRNA synthetase (EC 6.1.1.2)
Tyrosyl-tRNA synthetase (EC 6.1.1.1)
Tyrosyl-tRNA synthetase (EC 6.1.1.1) ## cluster 3
proteasome regulatory subunit Rpn10
proteasome regulatory subunit Rpn11
proteasome regulatory subunit Rpn12
proteasome regulatory subunit Rpn3
proteasome regulatory subunit Rpn5
proteasome regulatory subunit Rpn6
proteasome regulatory subunit Rpn7
proteasome regulatory subunit Rpn8
proteasome regulatory subunit Rpn9
proteasome regulatory subunit Rpt1
proteasome regulatory subunit Rpt2
proteasome regulatory subunit Rpt4
proteasome regulatory subunit Rpt6
proteasome regulatory subunit S5b
proteasome regulatory subunit p28
proteasome regulatory subunit p29
proteasome subunit alpha1 (EC 3.4.25.1)
proteasome subunit alpha2 (EC 3.4.25.1)
proteasome subunit alpha3 (EC 3.4.25.1)
proteasome subunit alpha4 (EC 3.4.25.1)
proteasome subunit alpha5 (EC 3.4.25.1)
proteasome subunit alpha6 (EC 3.4.25.1)
proteasome subunit alpha7 (EC 3.4.25.1)
proteasome subunit beta1 (EC 3.4.25.1)
proteasome subunit beta2 (EC 3.4.25.1)
proteasome subunit beta3 (EC 3.4.25.1)
proteasome subunit beta4 (EC 3.4.25.1)
proteasome subunit beta6 (EC 3.4.25.1)
proteasome subunit beta7 (EC 3.4.25.1)
ubiquitin / LSU ribosomal protein L40e
ubiquitin / SSU ribosomal protein S27Ae
ubiquitin-like protein fubi / SSU ribosomal protein S30e

Tabla suplementaria 2. Isogrupos similares a toxinas

Isogrupo	Tipo de Toxina	Toxina	%Identidad	%Cobertura ORF	%Cobertura Toxina
isogroup00005	α KTx	KA171_MMA	43	100	100
isogroup00035	α KTx	KAX21_CNO	100	63	100
isogroup00095	α KTx	KA171_MMA	77	49	55
isogroup03436	α KTx	KA163_MMA	36	97	98
isogroup16816	α KTx	KA101_CNO	70	65	48
isogroup16863	α KTx	KA102_CNO	71	69	97
isogroup17266	α KTx	KA101_CNO	100	90	71
isogroup17401	α KTx	KA101_CNO	68	100	71
isogroup11772	β KTx	KIK2_TDI	72	32	63
isogroup16053	β KTx	KBX1_TCO	48	76	36
isogroup16168	β KTx	KBX1_TCO	74	99	84
isogroup17599	β KTx	KIK2_TDI	81	98	47
isogroup03452	γ KTx	KGX31_CNO	100	68	100
isogroup16221	γ KTx	KGX11_CNO	98	100	100
isogroup18586	γ KTx	KGX11_CNO	100	100	56
isogroup00487	CaTx	SCXC1_MMA	58	80	67
isogroup03464	CaTx	SCXC1_MMA	44	92	95
isogroup15918	Lipasa	VP164_LMU	60	100	97
isogroup00042	LVP-NaTx	LVP1A_BOS	32	54	54
isogroup00331	LVP-NaTx	LVP1A_BOC	41	100	97
isogroup01026	LVP-NaTx	LVP1A_MMA	31	71	71
isogroup15099	LVP-NaTx	LVP1A_BOC	39	100	98
isogroup00230	NaTx	SCXV_CSC	70	89	102
isogroup00311	NaTx	SCX12_CNO	54	81	97
isogroup00798	NaTx	SCX2_CNO	100	100	82
isogroup01026	NaTx	HJ1A_HJU	55	77	100
isogroup01137	NaTx	CSSIX_CSS	94	100	100
isogroup01524	NaTx	SCXB_BOC	50	67	80
isogroup01533	NaTx	HJ1B_HJU	53	100	67
isogroup03359	NaTx	SIXA1_MMA	41	90	96
isogroup03547	NaTx	SCX4_CLL	97	100	41
isogroup12229	NaTx	SCNA1_TDI	41	72	91
isogroup12878	NaTx	SCX2_CNO	100	73	32
isogroup14678	NaTx	TX11_ACR	34	65	88
isogroup16065	NaTx	SCX2_CSC	94	54	57
isogroup16094	NaTx	SCX5_CSC	51	62	83
isogroup16486	NaTx	SCXB_CNO	53	97	92
isogroup16518	NaTx	SCX2_CNO	97	100	70
isogroup16547	NaTx	SCX2_CSC	70	98	69
isogroup16933	NaTx	SCXV_CSC	69	89	100
isogroup17232	NaTx	SCX3_CSC	96	74	55

isogroup18150	NaTx	SCX2_CNO	98	100	50
isogroup18205	NaTx	CEII8_CEL	90	86	47
isogroup18536	NaTx	SCAS_MMA	50	103	42
m25901	NaTx	SCX12_CNO	97	100	45
m2123	PLA2	PLA2_AME	50	100	80
isogroup00380	Tx	VTx_MEU	62	100	100
isogroup01539	Tx	VTx_RJU	32	95	100
isogroup03301	Tx	VNP3_MMA	78	100	100
contig03848	VC	BmHYA1_MMA	80	100	24
isogroup00241	VC	VIGFBP_MEU	37	100	89
isogroup00368	VC	VP302_LMU	44	90	100
isogroup03274	VC	VA5_TSE	66	90	74
<i>singlets</i>	VC	PBPO_PSC	78	100	80
isogroup00032	VMP	VMPA_TSE	41	96	84
isogroup00046	VMP	VMPA_TSE	46	74	100
isogroup00210	VMP	VMP1_MEU	50	100	100
isogroup00213	VMP	VMPA_TSE	42	99	85
isogroup00660	VMP	VMPA_LIN	45	9	81
isogroup00717	VMP	VMPA_TSE	59	84	104
isogroup08058	VMP	VMPA_TSE	53	98	77
isogroup00690	VSP	VSPA_TGA	39	87	40
isogroup02446	VSP	VSPA_TGA	35	72	58
isogroup05743	VSP	VSP2_ABI	31	66	100
isogroup00435	VSPI	KC2_ASU	44	61	93
isogroup03392	VSPI	KC2_ASU	68	84	66
isogroup03499	VSPI	KC2_ASU	60	70	91
isogroup07446	VSPI	KC2_ASU	55	72	95

KTx: toxinas bloqueadoras de canales de potasio; CaTx: toxinas bloqueadoras de canales de calcio; NaTx: toxinas moduladoras de la función de canales de sodio; PLA2: fosfolipasa A2; Tx: otras neurotoxinas; VC: componente del veneno; VMP: metaloproteasa; VSP: serin proteasa; VSPI: inhibidor de serin proteasas