Divergencia, hibridación y un modelo para el estudio de la especiación en mamíferos

Mag. Matías Feijoo

Orientador: Dr. Enrique P. Lessa

Co-orientador: Dr. Robert J. Baker

Tesis de Doctorado en Ciencias Biológicas

PEDECIBA

ÍNDICE

Resumen Pág. 4

INTRODUCCIÓN Pág. 6

1. Datos a escala genómica Pág. 7
2. Ecologia de la especiación, selección natural y abordajes de estudio Pág. 8
3. Taxa seleccionados para este estudio Pág. 11
3.1 *Complejo Uroderma bilobatum* Pág. 11
3.2 *Conflictos filogenéticos dentro de Eutheria* Pág. 12

OBJETIVOS Pág. 15

4. Justificación Pág. 16
5. Objetivo general Pág. 16
6. Objetivos específicos Pág. 17

MÉTODOS Pág. 18

7. Manejo de datos genómicos Pág. 19
7.1 *Edición, ensamblado y anotación del transcriptoma* Pág. 19
8. Análisis de datos Pág. 20
8.1 *Selección positiva episódica* Pág. 21
8.2 *Métodos para reconstruir el árbol de las especies* Pág. 23

RESULTADOS Pág. 31

9. Artículo I Pág. 32
9.1 Resumen artículo I Pág. 32
10. Artículo II Pág. 35
10.1 Resumen artículo II Pág. 35

CONCLUSIONES Y PERSPECTIVAS Pág. 39

11. Conclusiones y perspectivas Pág. 40

BIBLIOGRAFÍA Pág. 43

12. Referencias Pág. 44

ANEXOS Pág. 52

13. Anexos Pág. 53
10.1 Anexo I Pág. 53
10.2 Anexo II Pág. 54

Resumen

Las nuevas tecnologías de secuenciación masiva han permitido avances significativos en diversas áreas de investigación, incluyendo el estudio de la biología evolutiva de las especies. En este sentido, esta tesis presenta la aplicación de estas herramientas para aportar al conocimiento de los procesos de diversificación en mamíferos a dos niveles, abordados en sendos manuscritos. El primero de estos manuscritos analiza los genes blanco de la selección positiva a lo largo de la especiación de los murciélagos, tanto a nivel de linajes ancestrales como en la divergencia de dos especies incipientes, como lo son las subespecies de *Uroderma bilobatum* ("murciélago de campamento"), pertenecientes a la familia Phyllostomidae. Esta familia presenta una gran diversidad (>140 especies) en la región Neotropical, e incluye linajes especializados para muy diversos regímenes de alimentación (insectívoros, nectarívoros, frugívoros, sanguívoros, carnívoros, omnívoros, palinívoros). La especialización trófica se refleja en la anatomía y función de las glándulas salivares submandibulares. Para avanzar en la comprensión de la divergencia adaptativa temprana en el género *Uroderma*, se obtuvieron transcriptomas de glándulas submandibulares en dos subespecies de *U. bilobatum* que se encuentran en una fase temprana de la especiación. Estos datos, combinados con otras secuencias disponibles, permitieron examinar la acción de la selección positiva en la divergencia de múltiples genes en estos y otros murciélagos. Los principales genes que han evolucionado bajo selección positiva se asocian a funciones del sistema inmune y el metabolismo lipídico y tienen como implicancia biológica, su asociación directa con la respuesta a patógenos, sistemas de refugios, dieta y los requerimientos energéticos durante el vuelo. El segundo manuscrito presenta el estudio de las relaciones filogenéticas dentro de los mamíferos euterios, combinando información genómica y transcriptómica, y utilizando métodos de análisis que combinan árboles de genes y árboles de especies. Dicho manuscrito examina las relaciones filogenéticas de los Chiroptera, para las cuales hay hipótesis alternativas con cierto apoyo científico. Sin embargo, el estudio de las relaciones no se limitó a los Chiroptera sino que se extendió a otros grupos (Dermoptera, Pholidota y Arctoidea) apoyándose en la suficiencia de datos disponibles en bases de datos públicas. Al mismo tiempo, este manuscrito procura superar las limitaciones que imponen el uso exclusivo de datos genómicos en el número de taxones representados en los análisis. Para ello, se propone un protocolo para obtener una mucho mayor representación taxonómica incorporando datos transcriptómicos. Con ello se

mantiene un enfoque multigénico, aunque a una escala mucho menor que la del genoma. Con detalle en el grupo de los quirópteros, los resultados obtenidos sugieren la parafilia del grupo de los murciélagos que ecolocalizan. En cuanto a los otros grupos analizados, se apoya la ubicación de Dermoptera como grupo hermano de los primates, la ubicación de Pholidota como grupo hermano de los carnívoros y dentro de Arctoidea, se ordenarían de forma tal que los pinnípedos serían el grupo hermano del conjunto Ursidae-Musteloidea. Más allá de los resultados presentados en la tesis, la misma representa la base para futuros estudios en el complejo *U. bilobatum* así como para el estudio de la familia de los filostómidos, la cual presenta una amplia diversidad biológica a nivel de especies, características morfológicas, etológicas y de dieta, entre otras.

INTRODUCCIÓN

1. Datos a escala genómica

Grandes avances tecnológicos en la última década han permitido la obtención de datos masivos de secuencias (HTS: *high throughput sequencing*), generando en algunos casos hasta 1.8 tera bases totales (http://www.illumina.com/systems/sequencing.html) durante el proceso de secuenciación. Esto ha redimensionado el foco de atención de diversas áreas de estudios a una escala mayor, permitiendo el abordaje de varios temas a escala genómica. En sus inicios, el uso de estas tecnologías de secuenciación se centralizó en la generación de genomas completos de organismos modelos en investigación y de interés comercial. Ejemplo de esto han sido los trabajos generados en procariotas, donde ha existido un sesgo en los grupos secuenciados (Zhi et al. 2012; Hugenholtz 2002) con especial énfasis en aquellos grupos que contienen patógenos de humanos (Brown 2001) u organismos de interés biotecnológico (Zhi et al. 2012). También son ejemplo de este sesgo los trabajos en las subfamilias de gramíneas (Vogel et al. 2010), Ehrhartoideae (arroz) y Panicoideae (maíz, sorgo, caña de azúcar y mijos), los trabajos en la levadura *Saccharomyces cerevisiae* (e.g. Duina et al. 2014), en *Droshophilla* (e.g. Russell 2012) que es el modelo biológico clásico para el estudio del genoma eucariota, así como también, los trabajos que involucran el genoma humano y sus especies cercanas con una perspectiva biomédica (e.g. Collins et al. 2003).

A medida que el desarrollo tecnológico avanza, se ajustan el funcionamiento y la eficiencia, y se desarrollan nuevas variantes en los protocolos utilizados, se evidencia una tendencia a la reducción rápida de los costos de las mismas y se estabilizan las técnicas para su aplicación en la secuenciación a escala genómica (Bahassi y Stambrook 2014; Koboldt et al. 2013; Mardis 2013). Con estos avances, diversos proyectos de gran escala genómica han sido planteados, incluyendo en particular, los proyecto para secuenciar mil genomas humanos, mil de *Drosophila melanogaster*, y mil de diversas especies de plantas. Sin embargo, en este escenario los estudios en especies no-modelos y el acceso a este tipo de datos en múltiples organismos también se ven favorecidos (Ellegren 2014; Faino et al. 2014), permitiendo plantear diversos proyectos a nivel mundial que prevén llegar a secuenciar mil genomas de mamíferos y más aún, llegar a diez mil genomas de especies de vertebrados.

Dentro de los mamíferos y como ejemplo emblema de una especie no-modelo para la cual se secuenció y ensambló su genoma con datos de HTS, encontramos al oso Panda (*Ailuropoda melanoleuca*) (Li et al. 2009). A pesar de estos antecedentes, es aún poco probable pensar en generar los genomas completos para las especies que se deseen estudiar y menos aún si se

pretende tener una cobertura a escala poblacional. Por esto, la mayoría de los trabajos se han volcado a una alternativa de menor costo y que también forma parte de las nuevas tecnologías. Esta alternativa involucra el estudio de la expresión génica en distintos tejidos, órganos y condiciones como aproximación al entendimiento del funcionamiento de los mismos, en lo que define el área de la transcriptómica (Rudd 2003; Dong et al. 2005; Mortazavi 2008). Este tipo de estudios permite centralizarse en la fracción codificante del genoma y su análisis, involucrando esto último, no solo el estudio a nivel de los valores de expresión sino también indagar en las características y diferencias a nivel de las secuencias nucleotídicas.

En este sentido, estos abordajes genómicos no solo presentan desafíos a nivel de las tecnologías de secuenciación si no que también a nivel del análisis de los datos generados. Esto se debe a la magnitud de información disponible (entre 30Gb a 1.8Tb totales, dependiendo de los detalles de secuenciación utilizados), a la forma como se recuperan las secuencias (300-3000 millones de fragmentos, cada uno del orden de 75-250 pares de base de longitud) y a la amplitud de áreas que se entrelazan en su análisis. Varias disciplinas dentro de la biología se han visto beneficiadas en el uso de datos genómicos (por ejemplo: la genética, la evolución, la fisiología, la ecología, entre otras). En particular, dentro del estudio de la evolución de las poblaciones naturales, el uso de genomas y transcriptomas a permitido enfocarse a entender en qué medida la selección natural es importante para explicar los patrones observados de variabilidad dentro y entre especies (e.g. Marra et al. 2014; Thavamanikumar et al. 2014, usando transcriptomas; Worley et al. 2014, usando genomas). De igual forma, el estudio de la diversidad y diversificación de las especies se han visto potenciadas con la generación y análisis de los datos genómicos con un especial desarrollo de las aplicaciones que involucran la resolución de las relaciones filogenéticas de los organismos (e.g. Springer 2013; Song y Edwards 2012).

2. Ecologia de la especiación, selección natural y abordajes de estudio

La importancia de la selección natural como fuerza principal en la especiación, si bien ha sido corroborada en general, permanece poco entendida (Schluter 2000; Coyne and Orr 2004; Nosil et al. 2012). Una de las razones en que se fundamenta este desafío es que todas las especies

y situaciones son de cierta forma únicas con respecto a su ambiente y su historia evolutiva. Por esto, el ruido de los efectos locales y la estocasticidad en los procesos pueden sobrepasar la señal de selección. En esta situación, la mejor manera de incrementar la solidez estadística en los estudios de biología evolutiva es por medio del análisis de varios eventos en un contexto filogenético bien definido que permita comparar las bases genéticas de los rasgos adaptativos (Grant y Grant 2008). La principal interrogante en la actualidad es, ¿de qué forma la selección participa en el proceso de especiación?, más particularmente, ¿cuáles son los mecanismos de la selección natural, qué genes son los involucrados, y cómo estos genes llevan a incompatibilidades en el hábitat, en el comportamiento, de tipo químicas, mecánicas y fisiológicas que son las barreras reproductivas entre las nuevas especies?

La especiación ecológica se refiere a la generación de aislamiento reproductivo entre poblaciones o un subconjunto de individuos de una única población como consecuencia de la adaptación a diferentes ambientes o nichos ecológicos (Schluter 2000; Schluter 2001; Rundle and Nosil 2005). En este caso la selección natural tiene una función divergente, actuando en direcciones contrarias en los ambientes, que lleva a la fijación de diferentes alelos, siendo estos ventajosos en uno de los ambientes pero no en el otro (Schluter 2000; Schluter 2001; Rundle and Nosil 2005). Los agentes en los que se sustenta la selección divergente son extrínsecos e incluyen factores bióticos y abióticos, tales como recursos alimenticios, clima, hábitat, y relaciones inter-especificas, como ser enfermedades, competencia e interferencia. La especiación ecológica puede tener como resultado varios tipos de aislamiento reproductivo, incluyendo aislamiento pre-copulatorio, esterilidad del híbrido, inviabilidad intrínseca y extrínseca de híbridos y aislamiento ecológico tanto pre- como postcigótico. La especiación sexual es considerada como especiación ecológica si la selección divergente, con base ecológica, guía la divergencia de las preferencias de apareamiento, como es el caso de la hipótesis del impulso sensorial (Endler 1992). Una vez que las diferencias genéticas iniciales entre poblaciones se han acumulado por el proceso planteado, nuevas mutaciones podrían verse favorecidas en una población y no en la otra por medio de interacciones epistáticas (Mani and Clarke 1990). Entonces, la epistasis, incluyendo la que produce las incompatibilidades en híbridos entre especies del tipo Dobzhansky-Muller (Coyne and Orr 2004), pueden ser el resultado de la especiación ecológica. El proceso de especiación puede ser rápido bajo este modelo de especiación, ya que los alelos son dirigidos a su fijación por medio de la selección natural y esto puede ocurrir con o sin flujo génico entre las poblaciones en divergencia, aunque es

de mayor efectividad cuando el mismo está ausente (Feder et al. 2012, para detalles sobre el flujo génico).

Para investigar los mecanismos de especiación por selección natural en la naturaleza existen dos aproximaciones. El enfoque *de abajo-arriba* (*bottom-up*), que involucra (i) mapeo genético del aislamiento reproductivo entre especies cercanas, (ii) probar si los genes descubiertos exhiben evidencia de selección positiva, y (iii) identificar el fenotipo y la fuente de los efectos alternativos en el éxito reproductivo de los loci seleccionados. Ésta aproximación ha sido exitosa en la identificación de varios genes implicados en la inviabilidad y esterilidad de híbridos, y en el aislamiento sexual entre especies (por ejemplo en el género *Drosophila*, e.g. Nosil and Schluter 2011). Sin embargo ha tenido un efecto aún mayor en evidenciar las señales de selección positiva, probando la función de la selección natural en el proceso de especiación y revelando que la fijación de las diferentes variantes ocurre previo al completo aislamiento reproductivo de las poblaciones. El enfoque *de arriba-abajo* (*top-down*), que implica identificar (i) los rasgos fenotípicos bajo selección divergente, (ii) aquellos rasgos asociados con el aislamiento reproductivo, y (iii) los genes que sustentan esos rasgos y el aislamiento. Sin dudas el paso de identificación de genes ha sido el mayor desafío para ambas aproximaciones pero es necesario para entender como la selección lleva al aislamiento reproductivo. En este sentido, la falencia más obvia en el conocimiento actual del proceso de especiación es el conocimiento de las conexiones entre genes y la selección a nivel de organismos. Varios casos de selección han sido reportados como generadores de aislamiento reproductivo, pero poco se sabe de los cambios genéticos que lo permiten. De igual forma, señales de selección positiva en el patrón de variación de muchos genes han sido reportadas, pero poco se conoce de los mecanismos de selección detrás de esto.

Al igual que para otras áreas de investigación, las tecnologías detrás del estudio de los genomas, con especial énfasis en la genómica y su vínculo con el origen de las especies (Seehausen et al. 2014) permiten profundizar en el estudio de estos proceso, incluso en especies no-modelo, y especialmente en poblaciones naturales (Ellegren and Sheldon 2008; Rice et al. 2011).

3. Taxa seleccionados para este estudio

3.1 *Complejo Uroderma bilobatum*

La familia Phyllostomidae es un grupo diverso y amplio de murciélagos, compuesto por aproximadamente 56 géneros y 160 especies (Koopman, 1993, Simmons, 1998, Wetterer et al. 2000, Botero-Castro et al. 2013). La misma exhibe la mayor variación en características morfológicas y de dieta cuando se la compara con cualquier otra familia dentro de los mamíferos (Baker et al. 2003). Con particular interés, dentro de la subfamilia Stenodermatini se encuentra el complejo *Uroderma bilobatum* ("Peter's tent-making bat", "murciélago toldero") que se distribuye en la región de los trópicos del nuevo mundo. Para este complejo se han descripto tres subespecies en base a diferencias cromosómicas (Baker et al. 1972, 1975; Baker 1979.) *U. bilobatum bilobatum* (2n=42) se encuentra en América del Sur, al este de los Andes, mientras que *U. bilobatum davisi* (2n=44) está distribuida a lo largo de la vertiente del Océano Pacifico de El Salvador, Guatemala, Honduras y México. La tercera subespecie, *U. bilobatum convexum* (2n=38), se distribuye en el resto de América Central y en la vertiente del océano Pacifico de Colombia y norte de Ecuador. Estas subespecies son esencialmente alopátridas, y cada una se caracteriza por re-arreglos cromosómicos fijos (Baker et al. 1972, 1975; Baker 1979). Las subespecies *U. b. davisi* y *U. b. convexum* coexisten en simpatría únicamente en una localidad (Honduras, Departamento Valle, 17 Km. SSW de Nacaome) e hibridan localmente. Se han detectado ejemplares con cariotipos mostrando evidencia de retrocruzas en localidades ubicadas hasta a 400 Km. de distancia de la zona de simpátría (Baker 1981). Ambas subespecies están estrechamente relacionadas; por ejemplo, la divergencia genética del gen del citocromo b mitocondrial (cytb) es de 2.5% (Hoffmann et al. 2003). El estudio de este y otros marcadores moleculares ha mostrado una hibridación limitada y sugieren que la selección diversificadora estaría actuando para limitar la introgresión (propuesto por Greenbaum 1981, Baker 1981, Barton 1982; discutido en Lessa 1990 y Hoffmann et al. 2003). Considerando el área de distribución en simpatría, de ambas subespecies, *U. b. davisi*, se encuentra en ambientes más áridos; sin embrago, si se considera todo el rango de distribución no hay factores ecológicos claros para distinguir el hábitat de ambas (Baker et al. 1975). Tampoco se han reportado diferencias a nivel de dieta, reproducción u otros aspectos de la historia de vida. De todas formas, estas similitudes generales no descartan diferencias ecológicas aún por descubrir que den sustento al escenario propuesto de divergencia adaptativa incipiente.

3.2 *Conflictos filogenéticos dentro de Eutheria*

Como se ha señalado, la secuenciación de varios genes nucleares y mitocondriales ha permitido importantes avances en el conocimiento de las relaciones filogenéticas de los mamíferos, incluyendo el agrupamiento en superórdenes de placentados. Sin embargo, en la actualidad se mantienen las interrogantes en cuanto a las relaciones filogenéticas internas de Chiroptera así como sus vínculos con otros órdenes dentro del superorden Laurasiatheria. Trabajos previos proponen dos resoluciones alternativas para esto último (Fig. 1, A). La primera agrupa los murciélagos como grupo hermano del clado formado por los carnívoros y los perisodáctilos, denominándolo como Pegasoferae (Lindbland-Toh et al 2011; McCormack et al. 2012; Meredith et al. 2011). La hipótesis alternativa, los agruparía en Scrotifera, siendo los murciélagos el grupo hermano del clado constituido por ungulados, cetáceos y carnívoros (Fereuungulata) (Murphy et al 2001; Tsagkogeorga et al. 2013; Song et al. 2013; Zhou et al. 2012). A su vez, las relaciones subordinales de los murciélagos varían a lo largo de diversos trabajos, proponiéndose también dos hipótesis filogenéticas (Fig. 1, B). La hipótesis tradicional divide los murciélagos en Microchiroptera, capaces de ecolocalización laríngea, y Megachiroptera (murciélagos frugívoros del viejo mundo, agrupados en la familia Pteropodidae), con sistema visual desarrollado y sin capacidad de ecolocación (Koopman 1993; Simmons 1998). La segunda hipótesis postula la parafília de los murciélagos que ecolocalizan. Bajo esta hipótesis, los murciélagos estarían divididos en los clados Yinpterochiroptera (que comprendería algunas familias de los microquirópteros, agrupadas en la superfamilia Rhinolophoidea y todos los megaquirópteros) y Yangochiroptera (que estaría conformado por el resto de los microquirópteros (Springer et al. 2001; Teeling et al. 2005).

Al igual que para Chiroptera, diversos grupos dentro de los mamíferos placentados presentan interrogantes en cuanto a sus relaciones, lo cual resalta la importancia de los análisis filogenéticos de los mismos a escala genómica combinando los datos disponibles (ver Figura 1). En este sentido, tres grupos presentan gran interés, Dermoptera (lémures voladores o colugos), Pholidota (pangolines) y Arctoidea (osos, pinnípedos y mustélidos). En cuanto a Dermoptera, las hipótesis alternativas sobre su la relación dentro de los Euarchontoglires siguen siendo viables (Nie et al. 2008) (Fig. 1, C). Éstas son, (a) su ubicación como grupo hermano de los primates (Pozzi et al.

2014) o (b) como grupo hermano de suborden Anthropoidea (monos, simios y humanos) (Arnason et al. 2008). Para el caso de Pholidota, las dos hipótesis propuestas sobre su relación filogenética lo ubican, sea como grupo hermano de los xenarthros (osos hormigueros, armadillos y perezosos) (Novacek 1992), conformando junto con éstos el grupo hermano del resto de los mamíferos euterios, o alternativamente, sea como grupo hermano de los carnívoros (Arnason et al. 2008) (Fig. 1, E). Finalmente, existe también incertidumbre sobre las relaciones internas de los arctoideos (Fig. 1 D), estando en juego tres hipótesis. La primera de éstas apoya la unión de Ursidae (osos) y Pinnipedia (pinnípedos) con exclusión de los mustélidos (Musteloidae) (Luan et al. 2013). Alternativamente, se plantea el agrupamiento de Pinnipedia y Musteloidea (Luan et al. 2013). Por último, la tercera hipótesis de las vigentes propone la unión de Musteloidea con Ursidae y propone a Pinnipedia como grupo hermano de éste (Yu y Zhang 2006).

Esta tesis se enmarca en el contexto teórico y en las aplicaciones genómicas descriptos previamente, y se centra en dos líneas principales de investigación. La primera, busca generar datos transcriptómicos de *U. bilobatum* para estudiar las señales de selección positiva a nivel molecular. Para esto, se utilizarán muestras de glándulas salivares submandibulares, las cuales representan entre otras características, un nexo directo con las características de la dieta, y las cuales han sido claves en la diversificación de los mamíferos (Tandler y Phillips 2004). La segunda línea surge de los datos generados en la primera y plantea la posibilidad de definir un enfoque para resolver las relaciones filogenéticas dentro de los mamíferos. Si bien este enfoque se sustenta en la combinación de los datos generados junto con datos transcriptómicos y genómicos disponibles para otras especies de murciélagos, el mismo representó la posibilidad de escalar su aplicación a los grandes grupos de mamíferos

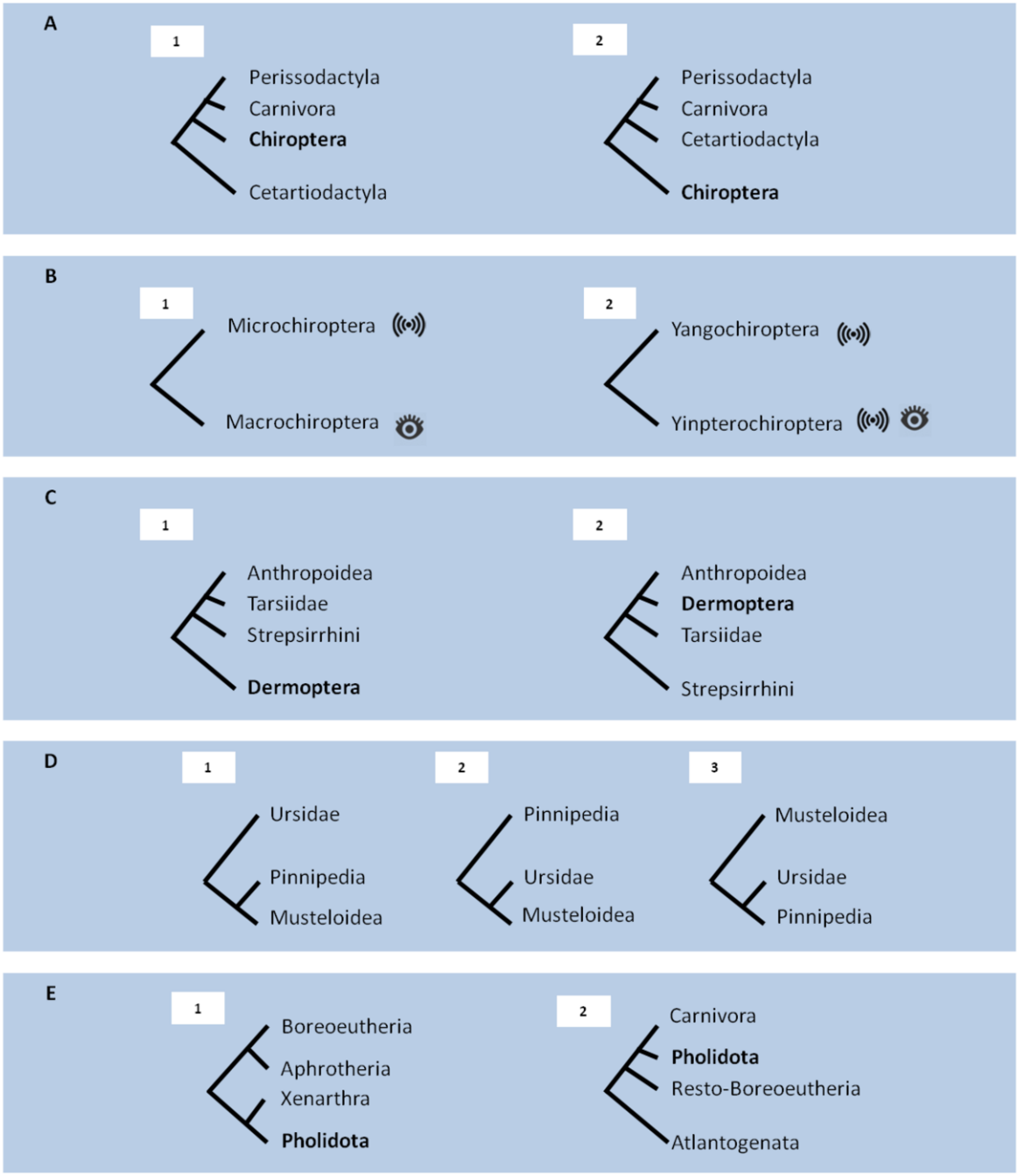


Figura 1. Hipotesis alternativas de relaciones filogeneticas analizadas con respecto a: A) el grupo de los quiropteros; B) las relaciones dentro de Chiroptera; C) Dermoptera; D) el grupo de los arctoideos; y E) Pholidota.

OBJETIVOS

4. Justificación

Como se ha señalado en la introducción, la reciente divergencia entre las dos subespecies de *Urodrema bilobatum* forma parte de la radiación de los murciélagos filostómidos, caracterizada por un gran número de especies muy diversificadas en su nicho trófico. Por otra parte, se sabe que las especializaciones anatómicas y fisiológicas de las glándulas salivares submandibulares están fuertemente involucradas en la diversificación trófica. Más en general, los estudios de los procesos de especiación han generado hipótesis que destacan el papel de la divergencia adaptativa en las fases tempranas de la formación de las especies. En este contexto, la posibilidad de secuenciar el transcriptoma completo de las glándulas submandibulares permite, potencialmente, caracterizar el conjunto de genes expresados, ligar su expresión con el nicho trófico (y más en general, con características de la historia natural) de los organismos, y examinar el papel de la selección positiva en la divergencia temprana de las especies. El caso de *Uroderma* bilobatum reúne las características deseables para abordar este conjunto de temas interconectados.

Al mismo tiempo, como se ha señalado más arriba, los análisis de la acción de la selección natural basados en árboles de genes y especies requieren colocar el caso de estudio en un contexto filogenético mayor. En esta tesis se realiza este ejercicio a dos niveles de divergencia diferentes. En primer lugar, en el marco de la evolución de los murciélagos, para poner a prueba la acción de la selección natural positiva en la divergencia temprana en *Uroderma* y vincularla con procesos adaptativos en diversas ramas del árbol de los Chiroptera. A estos efectos se combinan datos de genomas y de transcriptomas obtenidos, en este último caso, tanto a partir de nuestros propios datos de *HTS* como de datos disponibles en bases de libre acceso. Como extensión natural de este abordaje, se lo aplica para poner a prueba diversas hipótesis sobre las relaciones filogenéticas entre mamíferos placentados, con énfasis en algunos problemas no resueltos.

5. Objetivo general

Mediante el análisis del complejo *Uroderma bilobatum* con herramientas a escala genómica, contribuir al conocimiento de los procesos de especiación y divergencia en los murciélagos y a la resolución de problemas pendientes en la filogenia de los mamíferos placentados.

6. Objetivos específicos

- Ensamblar, anotar y comparar los transcriptomas de dos subespecies hermanas (*U. bilobatum davisi* y *U. bilobatum convexum*) a partir de muestras de glándulas salivares sub-mandibulares.
- Detectar loci sujetos a selección natural positiva durante la diversificación de los murciélagos, incluyendo, particularmente, la divergencia de las dos subespecies de *U. bilobatum*
- Contribuir a la resolución de relaciones filogenéticas aún conflictivas dentro de los euterios, utilizando los datos generados en combinación con aquellos disponibles en bases de datos públicas.
- Contribuir al desarrollo de protocolos de trabajo para el uso combinado de datos genómicos y transcriptómicos,con el objetivo de responder preguntas sobre la evolución biológica en un contexto filogenético.

MÉTODOS

7. Manejo de datos genómicos

7.1 *Edición, ensamblado y anotación del transcriptoma*

Al igual que para los clásicos trabajos con secuenciación del tipo Sanger (Sanger et al. 1977), las secuencias crudas generadas por las nuevas tecnologías necesitan de edición previo a su análisis. Para esto, se evalúan las calidades en las bases secuenciadas de cada fragmento de lectura (*read*) y se recortan aquellas por debajo del límite de calidad establecido. Al mismo tiempo, se revisa el contenido de bases no definidas y el contenido en GC a lo largo del *read*, entre otras cosas, para ajustar el recorte o la eliminación total del mismo para los análisis.

El principal desafío que presenta el análisis de transcriptomas en especies no-modelos (sin genoma de referencia) es su ensamblado a partir de los *reads* generados. Para esto, uno de los métodos utilizados es el implementado en el software TRINITY (Grabherr et al. 2011). El mismo no requiere del alineamiento de los reads a un genoma de referencia (produce un ensamblado *de novo*), se basa en los grafos de *de Bruijn* (de Bruijn 1946; Good 1946), en un proceso de tres etapas para generar fragmentos ensamblados (*contigs*) y ha mostrado ser igual o más efectivo que otros métodos, requiriendo incluso menos tiempo de análisis e insumos informáticos (Grabherr et al. 2011; Henschel et al. 2012; Zhao et al. 2011). Las tres etapas son *Inchworm, Chrysalis* y *Butterfly*. *Inchworm*, reconstruye contigs lineales a partir de los *reads* iniciales. Para esto, crea un diccionario de k-meros (con k=25), elimina los k-meros que posiblemente contienen errores de secuenciación y tomando el k-mero más frecuente en el diccionario se inicia el ensamblado de un *contig* el cual se extiende con los k-meros más frecuentes que se solapen en k-1 sitios en ambas direcciones del *contig* hasta que no se puede continuar. Una vez que un k-mero es utilizado para extender un *contig* se elimina del diccionario y el proceso de ensamblado se repite con el siguiente k-mero en frecuencia hasta que los mismos se agotan. *Chrysalis* utiliza los contigs generados en la etapa previa y agrupa de forma recursiva aquellos que se solapan perfectamente en al menos un k-mero-1 y cuya unión cuente con un mínimo de *reads* que lo soporten en ambos lados de la misma. Esto resulta en grupos que derivan de un mismo transcripto con variantes de *splicing* o con porciones únicas de genes parálogos. Luego, para cada agrupamiento de contigs definidos, *Chrysalis* construye el grafos de *de Bruijn*. Por último, *Butterfly* analiza las conexiones de los *reads* y sus *reads* pareados en el contexto del grafos de de Bruijn, reportando todas las posible secuencias y resolviendo las distintas isoformas de un transcripto así como los transcriptos de genes parálogos (ver detalles del método en Grabherr et al. 2011).

Una vez finalizado el ensamblado del transcriptoma, el paso siguiente involucra la descripción y anotación funcional del mismo. Esto permite la asociación particular de los contigs ensamblados a secuencias ortólogas conocidas y la descripción general de las cascadas metabólicas y funciones representadas en la muestra en estudio. Para esto existen diversas bases de datos (ENSEMBL, UCSC, OMA-browser, Orthomam, entre otras), que pueden ser interrogadas con diversos algoritmos (e.g. BLAST) para asignar identidades por homología. Estas bases de datos nos permiten descargar secuencias codificantes (CDS), exones, intrones o completas de genes previamente anotados para, en su mayoría, especies modelos de estudio. Al mismo tiempo, la ontología genética (GO) desarrollada por el GO Consortium (Ashburner et al. 2000) permite acceder a un listado de términos que asocian genes con funciones biológicas conocidas. La asociación entre términos GO y las secuencias estudiadas puede realizarse de forma manual, mediante scripts o también de forma automatizada mediante software existente, tanto de forma local (BLAST2GO, Conesa et al. 2005) como *online* (DAVID, Huang et al. 2009). Esta etapa de anotación funcional del transcriptoma permite reconstruir desde una perspectiva sistémica las distintas funciones representadas en la muestra analizada y permite también comparar diferentes condiciones experimentales o ambientales que se pretendan estudiar.

Los métodos aquí detallados para las fases de edición, ensamblado y anotación del transcriptoma han sido definidos dentro de una amplia lista de *software* y métodos alternativos disponibles (ver, Grabherr et al. 2011 para métodos de ensamblado; Gordon y Hannon 2010 y Bolger et al. 2014 para edición). De todas formas y como antecedente del grupo en el cual se enmarca este trabajo, hemos publicado recientemente un artículo donde se presentan las etapas aquí descriptas para el análisis de datos de secuenciación de RNA (**Anexo I**, Giorello et al. 2014).

8. Análisis de datos

Una vez ensamblado y anotado el transcriptoma, es posible utilizar este tipo de información para responder preguntas de interés en diversas áreas de estudio, como ser la fisiología, la biología celular o la genética, entre otras. De éstas, nos centraremos en el uso de

datos a escala genómica para el estudio de las adaptaciones a nivel molecular y la resolución de las relaciones filogenéticas entre taxa.

8.1 *Selección positiva episódica*

La evolución adaptativa a nivel de genes y genomas es en última instancia responsable de la adaptación a nivel morfológico, etológico y fisiológico, y de la divergencia entre especies y las innovaciones evolutivas. Si bien ha sido un campo de controversia (Schluter 2009), no hay dudas que la identificación de regiones codificantes para proteínas que hayan estado sujetas a evolución adaptativa es un tema importante por resolver para los biólogos evolutivos. Los casos de evolución adaptativa a nivel molecular han sido identificados principalmente comparando las tasas de remplazo de tipo sinónimo (cambios silenciosos, *ds*) y no-sinónimo (cambio de aminoácidos, *dn*) en secuencias nucleotídicas codificantes de proteínas. Tradicionalmente, las tasas de remplazo sinónimo y no-sinónimo se definen en el contexto de comparación entre dos secuencias, con d_s y d_n como el número de sustituciones sinónimas y no sinónimas por sitio, respectivamente. De esta forma, la relación $d_n/d_s = \omega$ mide la diferencia entre las dos tasas de sustitución a nivel de codones, donde si un cambio de aminoácido es neutral el mismo se fijará con la misma tasa que las mutaciones sinónimas dando como resultado un $\omega=1$. Si el cambio de aminoácido es deletéreo, la selección purificadora reducirá su tasa de sustitución de forma tal que $\omega < 1$. Únicamente cuando el cambio aminoacídico represente una ventaja adaptativa se fijará con una tasa mayor que las mutaciones sinónimas, dando un $\omega>1$ y será evidencia que permita sugerir un evento donde esté actuando la selección diversificadora. En este escenario de detección a nivel molecular de regiones codificantes bajo selección, el codón es considerado como la unidad de evolución. De esta forma, los análisis basados en codones no pueden inferir cuando un cambio sinónimo es consecuencia de un evento de mutación o de selección pero no asume que los mismos sean neutrales. Por ejemplo, un sesgo importante en el uso de codones puede ser la consecuencia tanto de un sesgo mutacional como de las presiones de selección, y pueden estar afectando seriamente la tasa de sustituciones sinónimas, por lo que discriminar entre estas dos situaciones representa un desafío importante de los modelos utilizados para su abordaje. En este sentido, diversos métodos han sido diseñados para identificar procesos selectivos a nivel de secuencia (Nielsen 2005; Yang y Bielawski 2000). Esencialmente, dos clases de métodos han sido definidos para estimar d_n y d_s entre dos secuencias codificantes para proteínas. La primera clase incluye varios métodos intuitivos

desarrollados desde 1980 (e.g. Li 1993; Nei y Gojobori 1986; Miyata y Yasunaga 1980). Como generalidad estos métodos involucran las etapas de, conteo de sitios sinónimos (S) y no-sinónimos (N) en las dos secuencias, conteo de las diferencias sinónimas y no-sinónimas entre las dos secuencias, y la corrección por sustituciones múltiples en un mismo sitio. S y N se calculan al multiplicar el largo de la secuencia por la proporción de cambios sinónimos y no sinónimos previo a la selección en la proteína (Ina 1995; Goldman y Yang 1994). La mayoría de estos métodos hace asunciones simplistas del proceso de sustitución nucleotídica e implican un tratamiento de los datos que los define como métodos de aproximación (Yang y Bielawski 2000; Yang y Nielsen 2000; Ina 1995). La segunda clase de métodos es la de métodos de máxima verosimilitud que se basan en el modelaje las sustituciones de codones (Muse 1996; Goldman y Yang 1994). En estos métodos, los parámetros (por ejemplo: divergencia de secuencias, relación entre transiciones y transversiones y ω) en el modelo son estimados a partir de los datos por la función de máxima verosimilitud y se utilizan para calcular d_n y d_s. Una característica importante de estos métodos es que permiten no solo estimar los parámetros del modelo sino que también realizar correcciones segun el tipo de cambio. En esta clase, los primeros modelos desarrollados asumían tasas sinónimas y no-sinónimas constantes a lo largo del tiempo y de los sitios. Mientras que la mayoría de las proteínas evolucionan bajo selección purificadora durante la mayor parte del tiempo, la selección positiva puede guiar la evolución en algunos linajes. Durante los episodios de evolución adaptativa, solo una pequeña fracción de los sitios son los involucrados en incrementar el éxito reproductivo mediante el remplazo de aminoácidos. Por estos argumentos es que las aproximaciones que asumen presiones constantes de selección en el tiempo y sobre sitios, carecen de poder para detectar los genes sujetos a selección positiva. Como consecuencia, varios de estos escenarios han sido incluidos en estos modelos de codones, permitiéndo aumentar la sensibilidad para detectar, particularmente, selección positiva y episodios cortos de evolución adaptativa. De todas formas, si por la mayoría del tiempo un gen evoluciona bajo selección purificadora pero es ocasionalmente sujeto a episodios de cambios adaptativos, una comparación entre dos secuencias distantemente emparentadas, raramente alcanzaría una relación d_n/d_s significativamente mayor a uno. Por esto, se han desarrollado métodos de verosimilitud para detectar selección positiva en linajes específicos de una filogenia, así como modelos que permiten diferentes valores de ω para diferentes ramas (Yang 1998; Yang y Nielsen 1998). Usando este tipo de modelos, una prueba de la razón de verosimilitud (*likelihood-ratio test*) puede construirse para probar hipótesis. Por ejemplo, el valor de ω para un linaje predefinido puede ser fijado en uno o

estimado como un parámetro libre. Por lo tanto los valores de verosimilitud bajo los dos modelos pueden ser comparados para probar en cuál linaje ω es mayor a uno. De igual forma, un modelo asumiendo un único ω para todos los linajes (*one-ratio*) puede ser comparado con otro modelo que asuma ω independientes para cada linaje (*free-ratio*) para probar la predicción neutralista de un valor idéntico entre linajes.

En particular, y basado en el refinamiento del estudio de la relación d_n/d_s, existe el método filogenético descripto por Murrell et al. (2012) conocido como MEME (*mixed effects model of evolution*). Este método surge como respuesta a la falla de métodos previos, los cuales asumen que las presiones de selección diversificadora sobre un sitio particular permanecen constantes en el tiempo, lo que impide la detección de los eventos biológicamente más esperables de selección episódica. El mismo se basa en la clase de métodos de *efectos al azar por rama-sitio* (*branch-site random effects*, Kosakovsky Pond et al. 2011), modificándolo de forma tal que la distribución de *ω* puede variar de sitio a sitio (*fixed effect*) y también de rama a rama en un sitio. De esta forma, MEME captura de forma confiable las huellas moleculares de la selección positiva persistente así como de la episódica, tarea en la que los métodos previos fallaban. Particularmente, episodios de evolución diversificadora afectando un subconjunto de ramas en sitios particulares serían reportados como selección purificadora por los métodos *por sitios*, mientras que el MEME lograría identificarlo de forma correcta.

8.2 *Métodos para reconstruir el árbol de las especies*

Uno de los principales objetivos de la sistemática como disciplina científica ha sido históricamente el de reconstruir y describir el *Árbol de la vida*, refiriéndose bajo este término a la historia única que describe la evolución de las relaciones entre especies. Ya que la evolución de las especies no se puede observar directamente, ni en las poblaciones naturales ni en el registro fósil, las relaciones entre estas se infieren generalmente por las características compartidas que estas presentan. Hasta 1970, para lograr este objetivo los trabajos de investigación en el área se basaban casi exclusivamente en el uso de caracteres morfológicos. Si bien esta aproximación tuvo inicialmente mucho éxito en los resultados obtenidos, la escasez en los caracteres disponibles así como la comparación entre especies sin aparentes homologías morfológicas fueron problemas que limitaron su funcionalidad (Hillis 1987). Con el desarrollo de las técnicas moleculares y de

análisis del ADN logrado a finales de 1960 se observó que los volúmenes de datos que las mismas permitían generar representaban un incremento significativo con respecto a las metodologías previas. De esta forma, cuando la secuenciación del ADN se tornó una técnica accesible y aplicable a un amplio rango de especies, las comparaciones moleculares se volvieron de rigor (e.g. Nei y Kumar 2000; Miyamoto et al. 1991). Así, la aparición de la sistemática filogenética ha permitido el desarrollo de un grupo de metodologías que hacen uso de la vasta información acumulada por las aproximaciones moleculares y la analizan para construir los árboles filogenéticos (Felsenstein 2004). Estas aproximaciones involucran diferentes métodos de reconstrucción de la filogenia de las especies. Uno de estos métodos es el que se basa en la *matriz de distancia* (Cavalli-Sforza y Edwards 1967). La idea general del método es calcular una medida de distancia entre cada par de especies estudias y luego encontrar una filogenia que prediga el conjunto de distancias observadas de la forma más aproximada posible. La reducción de la información molecular a una medida de distancia tiene como consecuencia la perdida de la información de los estados de los caracteres y cómo se combinan a distintos niveles. Otro de los métodos es el de *máxima parsimonia* (e.g. Fitch 1975), el cual minimiza el número de cambios en un árbol filogenético al asignar estados de caracteres a los nodos interiores del árbol. El largo de un carácter analizado se define como el número mínimo de cambios requeridos para explicar los estados presentes para ese carácter en el conjunto de los datos, y la sumatoria del número de cambios en todos los caracteres es el puntaje que recibe el árbol reconstruido. Por esto, el árbol de máxima parsimonia es aquel que minimiza el puntaje del árbol. Si bien los métodos de distancia y de máxima parsimonia se utilizan en la actualidad, su uso es restringido principalmente a trabajos morfológicos o que combinan secuencias y morfología. Esto no solo se debe a que presentan dificultades claramente conocidas (e.g. saturación de sustituciones y atracción de ramas largas), sino a que han sido desplazados por métodos que proponen la inclusión de modelos de evolución de los caracteres que se están analizando (Felsenstein 2004). Estos métodos son los más utilizados en la actualidad, particularmente para secuencias nucleotídicas y aminoacídicas, e incluyen los métodos de *máxima verosimilitud* y métodos *bayesianos*. El marco estadístico en el que se desarrollan los métodos de máxima verosimilitud fue inicialmente descripto por Fisher (1912; 1921; 1922). La función de verosimilitud se define como la probabilidad de los datos dado los parámetros, pero es vista como una función de los parámetros con los datos fijos y observados. Así, la función representa toda la información en los datos sobre los parámetros. De esta forma, las estimaciones por máxima verosimilitud son los valores de los parámetros que maximizan la verosimilitud. El uso de este

método para reconstruir filogenias fue originalmente presentado por Edwards y Cavalli-Sforza (1964) para datos de frecuencias génicas y por primera vez para secuencias nucleotídicas por Felsenstein (1981). Los métodos bayesianos están estrechamente relacionados con los de máxima verosimilitud. Difieren en que los parámetros en el modelo son considerados variables aleatorias con distribución estadística, donde en el contexto de máxima verosimilitud son contantes fijas desconocidas. De esta forma, antes del análisis de los datos, los parámetros se asignan a un distribución *a priori*, que luego se combina con los datos (o la verosimilitud) para generar la distribución *a posteriori*. Luego, todas las inferencias relacionadas con los parámetros están basadas en esta última distribución obtenida. La aplicación del método bayesiano fue originalmente presentada por Rannala y Yang (1996). En las últimas dos décadas, las inferencias bayesianas han ganado popularidad gracias a los avances de los métodos computacionales, especialmente los algoritmos de Cadenas de Markov Monte Carlo (MCMC). De igual forma, estos dos últimos métodos han incrementado su uso basados en el aumento del poder análisis computacional, en el desarrollo de software que permite su implementación y en el desarrollo de modelos de evolución de secuencia que se ajustan de mejor forma a la realidad del marcador utilizado.

A pesar de un desarrollo notable en los métodos de análisis, el uso de técnicas moleculares, con especial énfasis en el uso de secuencias de ADN, aún presenta el desafío de traducir la historia presente en la diversidad de secuencias (árboles de genes y alelos), lo que definíamos como el objetivo principal de la sistemática, la filogenia de especies y poblaciones (árbol de las especies). Este desafío surge ya que los genes y las especies pueden ser considerados entidades diferentes dentro de los niveles jerárquicos de la biología (Avise 2000). Los resultados obtenidos en los estudios de la evolución de las secuencias de ADN sobre las relaciones entre y dentro de las especies es, en última medida, una medida indirecta e incompleta de la historia evolutiva de las especies. Esto se da precisamente porque las especies son, por la mayoría de las definiciones, linajes evolutivos que comprenden múltiples genes, los cuales están representados en los organismos. El hecho de que las especies representen un nivel superior en la organización biológica con respecto a los genes significa que las tareas de la sistemática estarán incompletas hasta que los métodos de reconstrucción filogenética hagan una clara distinción entre árboles de genes y árboles de especies.

Los métodos que puedan surgir en este sentido deberán contemplar en otro orden de cosas, la homoplasia de los datos y las limitaciones intrinsecas de la reconstrucción filogenética, así como los factores principales de conflicto entre los arboles de genes y arboles de especies. Estos factores han sido extensamente estudiados en la bibliografía (Maddison 1997 realiza una revisión detallada). Las tres causas principales de este conflicto son, (i) la transferencia horizontal de genes; (ii) la duplicación génica; y (iii) la coalescencia profunda, y tienen diferentes niveles de importancia dependiendo de los taxa y genes que se estudien (ver Figura 2). En este sentido, la transferencia horizontal de genes es un fenómeno ampliamente conocido en los estudios sobre la filogenia de los microorganismos (e.g. Doolittle and Bapteste 2007; Bapteste et al. 2005). En cuanto a la duplicación de genes, la misma es común y ampliamente distribuida y puede invalidar los análisis si no es detectada. Sin embargo, al ser detectada puede ser una fuente importante de información para la resolución filogenética (e.g. Rasmussen and Kellis 2007; Sanderson and McMahon 2007; Page and Charleston 1997). El factor de la coalescencia profunda, el tercer causante del conflicto entre árboles de genes y especies, está mas ampliamente distribuido y no depende de eventos específicos que ocurren en algunos linajes sino que es una propiedad intrínseca de todas las poblaciones. La causa fundamental de este factor es la tasa de deriva génica, donde la coalescencia profunda prevalecerá si la tasa es baja (debido a poblaciones de mayor tamaño) si se la compara con el largo de los internodos en el árbol de las especies.

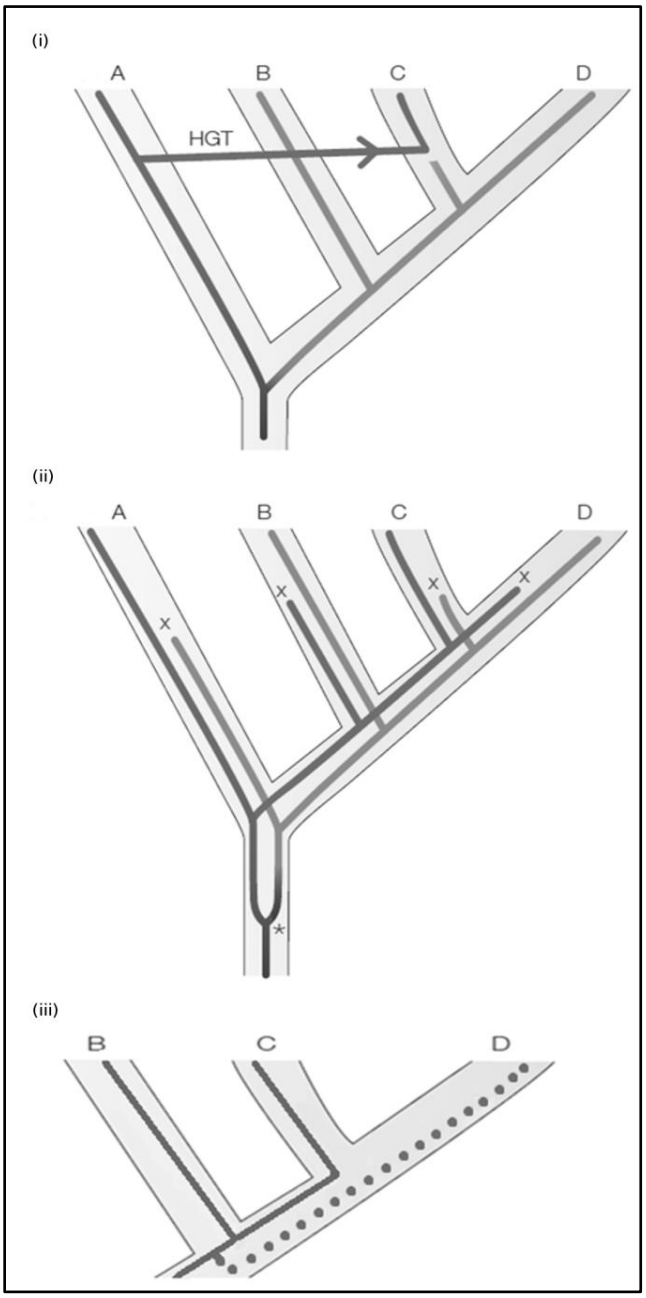


Figura 2. Posibles causas para la discordancia entre árboles de genes (lineas) y el árbol de las especies (sombreado gris). Esto puede ser por: (i) transferencia horizontal de genes (flecha con HGT), (ii) una duplicación genica ancestral (*) seguido de una perdida diferencial de genes (x), (iii) coalescencia profunda. Modificado de Gogarten y Townsend (2005).

La primera aproximación para responder a la necesidad de generar métodos que aborden la resolución de los arboles de genes y especies fue la de *evidencia total* o, como se la conoce a nivel de análisis de secuencias, el enfoque de *concatenación* de genes. Esta aproximación se caracteriza por la utilización de varios genes que son luego concatenados en una super-secuencia (o super-matriz) previo a su análisis. Esto surge en parte para incorporar grandes conjuntos de datos, pero también como un método que incorpore toda la información de secuencia disponible en los análisis filogenéticos (evidencia total, Kluge 2004). Sin embargo, en la práctica, diversas críticas surgieron a esta aproximación sustentándose en la creciente heterogeneidad observada en los árboles de genes en poblaciones naturales (e.g. Avise 1994). A su vez, por medio de simulaciones, se ha visto que las secuencias de ADN que evolucionan bajo tasas de sustitución diferentes pueden arrojar resultados erróneos cuando se los analiza con los algoritmos y modelos existentes de reconstrucción filogenética (Bull et al. 1993) y que los genes que podrían representar diferentes topologías no deberían ser combinados.

El dominio abrumador de los datos moleculares en la sistemática y el surgimiento de la filogenómica, hace que el desarrollo de métodos para la estimación del árbol de las especies sea una tarea clave para los próximos años. En este sentido, una nueva aproximación en la búsqueda de las relaciones filogenéticas de las especies ha sido formalmente descripta por Edwards (2009) y ha desplazado el análisis de *concatenación* apoyándose en los diversos trabajos que han abordado las dificultades que presenta esta aproximación (Degnan and Rosenberg 2006; Kubatko y Degnan 2007; Edwards et al. 2007). Esta nueva aproximación plantea el estudio de las relaciones entre los árboles de genes y árboles de especies en el contexto de la teoría del coalescente (Kingman 1982, 2000) y representa un desafío para el desarrollo de métodos que puedan manejar la magnitud de los datos que se están generando con las nuevas tecnologías. En este sentido, varios métodos han sido descriptos para el análisis de este tipo de datos, de los cuales el MP-EST (Liu et al. 2010) y el STAR (Liu et al. 2009) han sido los más utilizados en publicaciones recientes (e.g. Tsagkogeorga et al. 2013).

El método STAR (*species trees using average ranks of coalescences*) estima el árbol de las especies en dos etapas. En la primera, los árboles de genes son construidos para cada locus usando cualquier método tradicional, como ser, por ejemplo, el método de *Máxima verosimilitud* (*Maximum Likelihood*, ML). Los árboles de genes construidos son re-enraizados de igual manera y los valores de ranking de coalescencia entre todos los pares de especies son calculados para cada

uno de ellos. Para calcular el ranking se considera que el valor de coalescencia en el nodo raíz es igual al número total de taxa en el árbol y el mismo decrece de a 1 a medida que avanzamos a los nodos terminales del árbol de los genes. En la segunda etapa, se obtiene un árbol por el método de *Unión de vecinos* (*Neighbor Joining*, NJ), u otro método de distancias elegido, a partir de la matriz de distancia cuyas entradas son el doble del valor del ranking promedio de coalescencia calculado para todos los árboles de genes. Como se describe en detalle en la publicación del método (Liu et al. 2009), la topología del árbol resultante es un estimador consistente de la topología del árbol de las especies

Al igual que en el STAR, el MP-EST (*maximum Pseudo-likelihood for Estimating Species Trees*) toma como inicio una colección de árboles de genes con raíz para estimar el árboles de las especies. Luego, busca maximizar la función de pseudo-verosimilitud de los posibles tripletes de especies en el árbol de las especies. Este método surge de una reparametrización de la fórmula propuesta por Rannala and Yang (2003), donde el largo de las ramas pasa a medirse en unidades de coalescencia, T = 2τ/θ (Degnan y Salter 2005). De esta forma, la pseudo-verosimilitud del árbol de las especies S* dado los árboles de los genes G de define como el producto de la distribución múltinomial de todos los tripletes en el árbol de las especies, según:

$$L(S^* \mid \mathrm{G}) = w \times \prod_{j=1}^{\binom{N}{3}} \left\{ (1-(2/3)e^{-B_j})^{x_{j1}} ((1/3)e^{-B_j})^{x_{j2}} ((1/3)e^{-B_j})^{x_{j3}} \right\}$$

donde,

$w = \prod_{j=1}^{\binom{N}{3}} \left\{ \frac{M!}{x_{j1}!\, x_{j2}!\, x_{j3}!} \right\}$, N es el número de taxa, B_j es el largo de la rama interna del triplete j en el árbol de las especies y X_{j1}, X_{j2} y X_{j3} son las frecuencias de las tres categorías de tripletes en los árboles de genes y M es igual al número de genes para todos los j. Como w es una función de X, no tiene efecto en el proceso de maximizar L(S*|G), por lo que puede ser ignorado en la función de verosimilitud. Por último, por medio de la técnica del intercambio de vecinos más

próximos (nearest-neighbor interchanges, NNI) se realiza una búsqueda heurística para encontrar el valor máximo estimado de pseudo-verosimilitud del árbol de las especies.

RESULTADOS

9. Artículo I

9.1 Resumen artículo I

Título: Caracterización de genes seleccionados positivamente durante la evolución de los murciélagos con nuevos transcriptomas de *Uroderma bilobatum*

Nombre de los autores e instituciones:

Matías Feijoo - Departamento de Ecología y Evolución, Facultad de Ciencias, Universidad de la República, Montevideo, Uruguay
Caleb D. Phillips - Department of Biological Sciences, Texas Tech University, Lubbock, Texas, United States of America; Research and Testing Laboratory, Lubbock, Texas, United States of America
Facundo Giorello - Departamento de Ecología y Evolución, Facultad de Ciencias, Unversidad de la República, Montevideo, Uruguay
Robert J. Baker - Department of Biological Sciences, Texas Tech University, Lubbock, Texas, United States of America
Enrique P. Lessa - Departamento de Ecología y Evolución, Facultad de Ciencias, Universidad de la República, Montevideo, Uruguay

La evidencia de adaptación a nivel molecular es de interés para los evolucionistas que tratan de entender el papel de la divergencia adaptativa en el proceso de especiación. En los murciélagos (Orden Chiroptera), la evidencia de selección direccional ha sido asociada a cambios drásticos en el estilo de vida, especialmente relacionados con la evolución del vuelo y la ecolocalización. En este trabajo presentamos evidencia de selección direccional episódica en 100 genes involucrados en la divergencia adaptativa de dos subespecies cercanas de *Uroderma bilobatum*, las cuales hibridan en condiciones naturales. Estos hallazgos se enmarcan en un contexto más amplio que examina el papel de la selección direccional episódica a lo largo de las ramas que van desde el ancestro común de los murciélagos hasta el género *Uroderma*. Con este objetivo, se generaron datos del transcriptoma de las glándulas salivares submandibulares de *U. bilobatum*, que se combinaron con secuencias disponibles de Chiroptera y representantes de Laurasiatheria. Llamativamente, la ontología funcional de los genes seleccionados positivamente es compartida a dos niveles de divergencia (especiación en curso y linajes superiores de la filogenia). En particular, se encontró que los genes involucrados con los procesos de inmunidad y metabolismo lipídico han estado

continuamente sometidos a selección positiva a lo largo de la evolución de los murciélagos. La selección podría estar actuando en los murciélagos en relación a características biológicas tales como la hibernación, los mecanismos de sopor diario y la energética del vuelo, lo que implicaría la adaptación en genes vinculados al metabolismo lipídico. Los genes del sistema inmune son ampliamente conocidos como blanco de la selección positiva, sin embargo, las características de construcción de refugios, la carga patógena, la dieta y las estrategias de alimentación podrían tener un papel importante en la conformación del patrón de selección sobre el sistema inmunológico en el caso de los murciélagos.

9.2 Artículo I

Characterization of positively selected genes in bat evolution with new transcriptomes of tent-making bats (*Uroderma bilobatum)*

Author's name and institutions:

Matías Feijoo - Departamento de Ecología y Evolución, Facultad de Ciencias, Universidad de la República, Montevideo, Uruguay
Caleb D. Phillips - Department of Biological Sciences, Texas Tech University, Lubbock, Texas, United States of America; Research and Testing Laboratory, Lubbock, Texas, United States of America
Facundo Giorello - Departamento de Ecología y Evolución, Facultad de Ciencias, Unversidad de la República, Montevideo, Uruguay
Robert J. Baker - Department of Biological Sciences, Texas Tech University, Lubbock, Texas, United States of America
Enrique P. Lessa - Departamento de Ecología y Evolución, Facultad de Ciencias, Universidad de la República, Montevideo, Uruguay

Corresponding author's complete contact information:

matiasfeijoo@gmail.com
Igua 4225 piso 6, PC 11400
+(598)5258618-7-136
Facultad de Ciencias
Montevideo, Uruguay

Abstract

Molecular evidence of adaptation is of interest to evolutionists trying to understand the role of adaptive divergence in the speciation process. In bats (Order Chiroptera), evidence of directional selection has been associated to drastic changes in lifestyle, especially related to the evolution of flight and echolocation. Here we present evidence of episodic directional selection in 100 genes involved in the adaptive divergence of two closely related subspecies of tent-making bats

(*Uroderma bilobatum*) that are known to hybridize in nature. These findings are framed in a broader context by examining the role of episodic directional selection along successive branches from the common ancestor of bats to the genus *Uroderma*. For this propose, new *U. bilobatum* transcriptome data were generated and combined with available sequences of Chiroptera and representatives of other Laurasiatheria. Interestingly, functional ontologies of positively selected genes at both levels of divergence (ongoing speciation and at higher phylogenetic levels) are shared. In particular, genes involved in *immunity* and *lipid metabolism* pathways were found to be continuously under positive selection throughout the evolution of bats. Selection may be acting in bats in relation to biological characteristics such as hibernation, daily torpor mechanisms and energetic performance during active flapping flight, requiring adaptations in genes related to lipid metabolism. Genes of the immune system are well known targets of positive selection, but variation in roost characteristics, pathogen load, diet and feeding strategies may play important roles in the case of bats.

10. Artículo II

10.1 Resumen artículo II

Título: Macrosistemática de mamíferos euterios combinando datos de secuenciación masiva para expandir la cobertura taxonómica

Nombre de los autores e instituciones:

Matías Feijoo - Departamento de Ecología y Evolución, Facultad de Ciencias, Universidad de la República, Montevideo, Uruguay
Andrés Parada - Departamento de Ecología y Evolución, Facultad de Ciencias, Universidad de la República, Montevideo, Uruguay

Los avances en las tecnologías de secuenciación han permitido el abordaje de las relaciones filogenéticas de los mamíferos desde una perspectiva multigénica llegando a la escala genómica. Esto último se ha visto limitado a especies para las que se cuenta con genomas de referencia. A pesar de los avances logrados en años recientes en base a estudios multigénicos y genómicos, varios conflictos filogenéticos dentro del grupo de los mamíferos euterios (placentados)

permanecen aún sin resolver. Por esto y para aumentar el número de taxones utilizados para resolver las relaciones filogenéticas, se presenta un protocolo para el análisis de datos de secuenciación masiva (ARN o ADN) disponibles en bases de datos de libre acceso, en muchos casos generados con otros propósitos. Mostramos cómo este abordaje multilocus y con marcadores independientes contribuye a resolver las relaciones de Dermoptera, Pholidota, Chiroptera y Arctoidea. Si bien el número máximo de genes utilizados es moderado (95), se ha logrado incrementar sustancialmente los taxones incluidos en los análisis, llegando a duplicar en algunos casos los taxones trabajos previos en el tema. Los árboles obtenidos por medio de dos métodos basados en el coalescente (STAR y MP-EST) son consistentes y presentan buen apoyo estadístico para los nodos estudiados.

10.2 Artículo II

Macrosystematics of eutherian mammals combining HTS data to expand taxon coverage

Author's name and institutions:

Matías Feijoo - Departamento de Ecología y Evolución, Facultad de Ciencias, Universidad de la República, Montevideo, Uruguay

Andrés Parada - Departamento de Ecología y Evolución, Facultad de Ciencias, Universidad de la República, Montevideo, Uruguay

Corresponding author's complete contact information:

matiasfeijoo@gmail.com
Igua 4225 piso 6, PC 11400
+(598)5258618-7-136
Facultad de Ciencias
Montevideo, Uruguay

Abstract

In the last few years high-throughput sequencing technologies have permitted significant advances in mammalian phylogenetic studies from a genomic perspective. However, this have been limited to a limited number of species for which there are reference genomes. Thus, several issues inside eutherian mammals remain unresolved. This may be due in part to limited taxon sampling, as taxonomic density is known to affect phylogenetic resolution. In this context, we present a protocol to increase taxon coverage using high-throughput sequencing data (RNA or DNA) generated for other biological studies and available in public databases. Using multiple and independent loci, we addressed pending or controversial issues concerning the phylogenetic position of (and, in some cases, the relationships of major lineages within) Dermoptera, Pholidota,

Chiroptera and Arctoidea. Although the maximum number of genes used is moderate (95), in some cases taxon coverage doubles that of previous related studies. Both coalescent-based methods (STAR and MP-EST) used for species tree reconstruction were consistent to each other and interrogated nodes received high statistical support.

CONCLUSIONES Y PERSPECTIVAS

11. Conclusiones y perspectivas

Con el desarrollo de esta tesis se lograron definir protocolos precisos (**anexo II**) que nos permitieron, a partir del análisis de datos de secuenciación masiva, responder preguntas a nivel de la adaptación molecular en los quirópteros y las relaciones filogenéticas dentro de los mamíferos placentados. En este sentido, a partir de la secuenciación de ARN total con la tecnología de Illumina se logró ensamblar, describir y comparar el transcriptoma de dos subespecies del complejo *U. bilobatum*. Por otro lado, datos pertenecientes a 13 especies fueron descargados y analizados mediante métodos desarrollados recientemente para la resolución del árbol de las especies a partir de los árboles de genes. En detalle, se han podido identificar genes (>9000 para caracterizar y anotar el transcriptoma; 95 para filogenia; 547 para análisis de selección) que representan una base referencial para futuros trabajos. En cuánto a la búsqueda de evidencia de selección positiva a nivel de secuencias, encontramos que genes vinculados a algunas funciones biológicas (no necesariamente los mismos genes dentro de ellas) fueron seleccionadas a lo largo de la diversificación de los murciélagos, pero también en la etapa inicial de separación de dos especies incipientes. Sobre las relaciones filogenéticas dentro de los mamíferos euterios, los datos generados junto con los disponibles públicamente, permitieron abordar el análisis de cuatro clados con relaciones aún controversiales. De esta forma, se logró incrementar el número de taxa utilizados para resolver dichos conflictos con datos multilocus de regiones codificantes.

En esta tesis se abordaron dos áreas de estudio en la evolución de las especies, con especial énfasis en los mamíferos, y a partir de los resultados obtenidos se plantean varias nuevas interrogantes. Para el caso de *U. bilobatum,* que cuenta con subespecies que hibridan*,* el análisis detallado de los re-arreglos cromosómicos (Fig. 3) presenta un sistema adecuado para el estudio de la función de los cambios cromosómicos en el mantenimiento de la diversificación y la variación geográfica en mamíferos. En este sentido, hemos comenzado a trabajar a partir de muestras de embriones para la obtención de cromosomas re-arreglados mediante citometría de flujo. Sin embargo, hemos fallado en el intento de lograr cultivos estables de fibroblastos que permitan llegar al volumen celular requerido por la técnica de aislamiento, lo que nos imposibilita los pasos siguientes de secuenciación y análisis.

Por otro lado, el enfoque a nivel poblacional mediante el diseño de polimorfismos de base única (SNPs) a partir de datos generados en base a un amplio número de ejemplares parece necesario para futuros trabajos que analicen la variación de estos marcadores a lo largo de la distribución de

cada subespecie y en la zona de contacto. Tanto para los análisis propuestos como para posibles análisis filogenéticos del grupo, podría ser relevante la inclusión de datos de HTS de especies cercanas. Esto permitiría además profundizar en el estudio del grupo de los filostómidos, el cual presenta una amplia diversidad de especies, adaptaciones y cambios morfológicos asociados a distintos tipos de dieta (Fig. 4). De esta forma, no solo se han generado datos fundacionales para el estudio de un caso particular de hibridación en murciélagos, sino también para avanzar hacia el entendimiento de los procesos de diversificación a distintos niveles de los mamíferos.

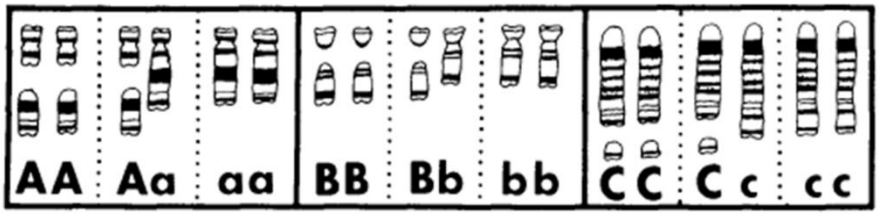


Figura 3. Cambios cromosómicos entre *U. b. davisi* (44, AA/BB/CC) y *U. b. convexum* (38, aa/bb/cc)

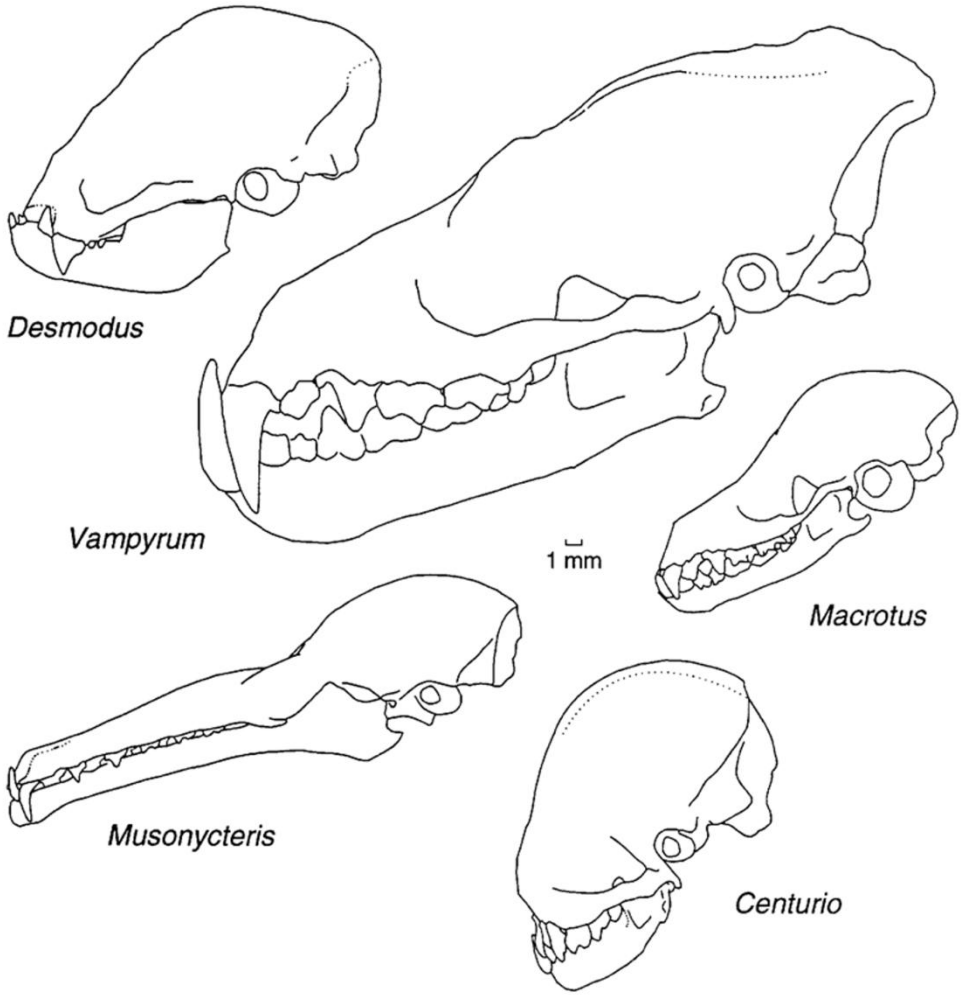


Figura 4. Cinco formas craneales extremas dentro de los filostómidos y con asociación a la dieta: Desmodus, un sanguívoro; Vampyrum, un carnívoro; Musonycteris, un nectarívoro; y Centurio, un frugívoro. También se muestra una estructura intermedia: Macrotus, un insectívoro y frugívoro (Freeman 2000).

BIBLIOGRAFÍA

12. Referencias

Arnason, U., Adegoke, J.A., Gullberg, A., Harley, E.H., Janke, A., Kullberg, M.(2008).Mitogenomic relationships of placental mammals and molecular estimates of their divergences. Gene 421, 37-51.

Avise, J. C. (1994). Molecular markers, natural history and evolution. Chapmanand Hall, New York.

Avise J. C.(2000). Phylogeography: the history and formation of Species.HarvardUniv. Press, Cambridge, MA.

Baker, RJ. (1979). Karyology.In Biology of the Bats of the New World Family Phyllostomatidae, Part III (ed. RJ Baker, JK Jones and DC Carter), pp. 107-155. Special Publications, Texas Tech University Museum, Texas.

Baker, RJ. (1981). Chromosome flow between chromosomally characterized taxa of a volant mammal, Uroderma bilobatum (Chiroptera: Phyllostomidae). Evolution35: 296-305.

Baker, R. J., Atchley, W. R.,Mcdaniel, V. R. (1972). Karyology and morphometrics of Peters' tentmaking bat, Uroderma bilobatum Peters (Chiroptera, Phyllostomatidae). Systematic Biology, 21(4), 414-429.

Baker, R. J., Bleier, W. J.,Atchley, W. R. (1975). A contact zone between characterized taxa of Uroderma bilobatum (Chiroptera: Phyllostomidae). Systematic Zoology, 24, 133-142.

Bapteste, E., E. Susko, J. Leigh, D. MacLeod, R. L. Charlebois,Doolittle, W. F. (2005). Do orthologous gene phylogenies really support treethinking?. BMC Evolutionary Biology. 5:33.

Barton, NH. (1982). The structure of the hybrid zone in Uroderma bilobatum (Chiroptera: Phyllostomidae). Evolution36: 863-866.

Bolger, A. M., Lohse, M.,Usadel, B. (2014).Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics, btu170.

Botero-Castro, F., Tilak, M. K., Justy, F., Catzeflis, F., Delsuc, F., Douzery, E. J. P. (2013).Next-generation sequencing and phylogenetic signal of complete mitochondrial genomes for resolving the evolutionary history of leaf-nosed bats (Phyllostomidae). Molecular Phylogenetic and Evolution 69 : 728-739.

Brown, J.R. (2001). Genomic and phylogenetic perspectives on theevolution of prokaryotes. Systematic Biology, 50:497-512.

Cavalli-Sforza, L. L., Edwards, A. W. (1967).Phylogenetic analysis.Models and estimation procedures. American journal of human genetics, 19(3 Pt 1), 233.

Collins, F. S., Green, E. D., Guttmacher, A. E., Guyer, M. S. (2003). A vision for the future of genomics research. Nature, 422(6934), 835-847.

Coyne, J. A.,Orr, H. A. (2004). Speciation.Sinauer. Sunderland, MA.

de Bruijn, N. G. (1946). A combinatorial problem. Proceedings of the Koninklijke Nederlandse Akademie van Wetenschappen 49: 758-764.

Da Wei Huang, B. T. S.,Lempicki, R. A. (2008). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nature protocols, 4(1), 44-57.

Degnan, J. H.,Salter, L. A. (2005). Gene tree distributions under the coalescent process. Evolution, 59(1), 24-37.

Degnan, J. H.,Rosenberg, N. A. (2006). Discordance of species trees with their most likely gene trees. PLoS genetics, 2(5), e68.

Dong, Q., Lawrence, C. J., Schlueter, S. D., Wilkerson, M. D., Kurtz, S., Lushbough, C.,Brendel, V. (2005). Comparative plant genomics resources at PlantGDB. Plant physiology, 139(2), 610-618.

Doolittle, W. F., Bapteste E. (2007). Pattern pluralism and the Tree of Lifehypothesis. Proceedings of the National Academy of Sciences USA, 104:2043–2049.

Duina, A. A., Miller, M. E., Keeney, J. B. (2014). Budding Yeast for Budding Geneticists: A Primer on the Saccharomyces cerevisiae Model System. Genetics, 197(1), 33-48.

Edwards, S. V. (2009). Is a new and general theory of molecular systematics emerging?. Evolution, 63(1), 1-19.

Edward, A. W. F., Cavalli-Sforza, L. L. (1964). Reconstruction of evolutionary trees.Phonetic and Phylogenetic Classification. The Systematic Association: London.

Edwards, S. V., Liu, L.,Pearl, D. K. (2007). High-resolution species trees without concatenation. Proceedings of the National Academy of Sciences USA, 104(14), 5936-5941.

Ellegren, H.,Sheldon, B. C. (2008). Genetic basis of fitness differences in natural populations. Nature, 452(7184), 169-175.

Endler, J. A. (1992). Signals, signal conditions, and the direction of evolution. American Naturalist, S125-S153.

Feder, J. L., Egan, S. P.,Nosil, P. (2012).The genomics of speciation-with-gene-flow. Trends in Genetics, 28(7), 342-350.

Felsenstein, J. (1981). Evolutionary trees from DNAsequences: a maximum likelihood approach. J. Molecular Evolution, 17, 368–376.

Felsenstein, J., Felenstein, J. (2004). Inferring phylogenies (Vol. 2). Sunderland: Sinauer Associates.

Fisher, R. A. (1912). On an absolute criterion for fitting frequency curves. Messenger of Mathmatics, 41: 155-160.

Fisher, R. A. (1921). On the" Probable Error" of a Coefficient of Correlation Deduced from a Small Sample. Metron, 1, 3-32.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics.Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character, 309-368.

Fitch, W. M. (1975). Toward finding the tree of maximum parsimony. In Proc. Eighth International Conference on Numerical Taxonomy, GF Estabrook, ed. (pp. 189-230).

Freeman, P. W. (2000). Macroevolution in Microchiroptera: recoupling morphology and ecology with phylogeny. Mammalogy Papers: University of Nebraska State Museum, 8.

Giorello, F. M., Feijoo, M., D'Elía, G., Valdez, L., Opazo, J. C., Varas, V., ...Lessa, E. P. (2014). Characterization of the kidney transcriptome of the South American olive mouse Abrothrix olivacea.BMC genomics, 15(1), 446.

Goldman, N.,Yang, Z. (1994). A codon-based model of nucleotide substitution for protein-coding DNA sequences. Molecular biology and evolution, 11(5), 725-736.

Good, I. J. (1946). Normal recurring decimals. Journal of the London Mathematical Society, 21: 167-169.

Gordon, A., Hannon, G. J. (2010). Fastx-toolkit. http://hannonlab.cshl.edu/fastx_toolkit

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q. et al. 2011.Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature biotechnology 29: 644-652.

Grant, P. R.,Grant, B. R. (2008). How and why species multiply: the radiation of Darwin's finches. Princeton University Press.

Greenbaum, I. F. (1981). Genetic interactions between hybridizing cytotypes of the tent-making bat (*Uroderma bilobatum*). Evolution, 306-321.

Gogarten, J. P., Townsend, J. P. (2005). Horizontal gene transfer, genome innovation and evolution. Nature Reviews Microbiology, 3(9), 679-687.

Henschel, R., Lieber, M., Wu, L. S., Nista, P. M., Haas, B. J., LeDuc, R. D. (2012, July).Trinity RNA-Seq assembler performance optimization. In Proceedings of the 1st Conference of the Extreme Science and Engineering Discovery Environment: Bridging from the eXtreme to the campus and beyond (p. 45). ACM.

Hillis, D. M. (1987) Molecular Versus MorphologicalApproaches to Systematics. Annual Review of Ecology and Systematics, 18:23–42.

Hoffmann, F. G., Owen, J. G.,Baker, R. J. (2003). mtDNA perspective of chromosomal diversification and hybridization in Peters' tent-making bat (Uroderma bilobatum: Phyllostomidae). Molecular Ecology, 12(11), 2981-2993.

Hugenholtz, P. (2002). Exploring prokaryotic diversity in the genomic era. Genome Biology, 3(2), 1-0003.

Ina, Y. (1995). New methods for estimating the numbers of synonymous and nonsynonymous substitutions. Journal of molecular evolution, 40(2), 190-226.

Kingman, J. F. C. (1982). The coalescent. Stochastic processes and their applications, 13(3), 235-248.

Kingman, J. F.C. (2000). Origins of the coalescent: 1974-1982. Genetics, 156(4), 1461-1463.

Kluge, A. G. (2004). On total evidence: for the record. Cladistics 20:205–207.

Kosakovsky Pond, S. L., Murrell, B., Fourment, M., Frost, S. D., Delport, W.,Scheffler, K. (2011). A random effects branch-site model for detecting episodic diversifying selection. Molecular biology and evolution, 28: 3033–3043.

Koopman, K. F. (1993). Order Chiroptera. In *Mammals Species of the World, a Taxonomic and Geographic Reference* (ed. D Wilson and DM Reeder), pp. 137-241. Second Edition, Smithsonian Institution Press, Washington.

Kubatko, L. S.,Degnan, J. H. (2007). Inconsistency of phylogenetic estimates from concatenated data under coalescence. Systematic Biology, 56(1), 17-24.

Lessa, E. P. (1990). Multidimensional analysis of geographic genetic structure. Systematic Biology, 39(3), 242-252.

Li, W. H. (1993). Unbiased estimation of the rates of synonymous and nonsynonymous substitution. Journal of molecular evolution, 36(1), 96-99.

Lindblad-Toh, K., Garber, M., Zuk, O., Lin, M. F., Parker, B. J., Washietl, S., et al. (2011). A high-resolution map of human evolutionary constraint using 29 mammals. Nature, 478(7370), 476-482.

Luan, P. T., Ryder, O. A., Davis, H., Zhang, Y. P.,Yu, L. (2013). Incorporating indels as phylogenetic characters: Impact for interfamilial relationships within Arctoidea (Mammalia: Carnivora). Molecular Phylogenetic and Evolution, 66, 748-756.

Maddison, W. P. (1997). Gene trees in species trees. Systematic Biology, 46:523–536.

Mani, G. S.,Clarke, B. C. (1990). Mutational order: a major stochastic process in evolution. Proceedings of the Royal Society of London. B. Biological Sciences, 240(1297), 29-37.

McCormack, J. E., Faircloth, B. C., Crawford, N. G., Gowaty, P. A., Brumfield, R. T.,Glenn, T. C. (2012). Ultraconserved elements are novel phylogenomic markers that resolve placental mammal phylogeny when combined with species-tree analysis. Genome research, 22(4), 746-754.

Meredith, R. W., Janečka, J. E., Gatesy, J., Ryder, O. A., Fisher, C. A., Teeling, E. C., et al. (2011). Impacts of the Cretaceous Terrestrial Revolution and KPg extinction on mammal diversification. Science, 334(6055), 521-524.

Miyamoto, M. M., Cracraft, J. (1991) Phylogenyinference, DNA sequence analysis, and thefuture of molecular systematics. In: MiyamotoMM, Cracraft J (eds) Phylogenetic Analysis ofDNA Sequences. Oxford Univ. Press, NewYork.

Miyata, T., Yasunaga, T. (1980). Molecular evolution of mRNA: a method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its application. Journal of Molecular Evolution, 16(1), 23-36.

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L.,Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature methods, 5(7), 621-628.

Murphy, W. J., Eizirik, E., O'Brien, S. J., Madsen, O., Scally, M., Douady, C. J., et al. (2001).Resolution of the early placental mammal radiation using Bayesian phylogenetics. Science, 294(5550), 2348-2351.

Murrell, B., Wertheim, J. O., Moola, S., Weighill, T., Scheffler, K.,Pond, S. L. K. (2012). Detecting individual sites subject to episodic diversifying selection.PLoS genetics, 8(7), e1002764.

Muse, S. V. (1996). Estimating synonymous and nonsynonymous substitution rates. Molecular biology and evolution, 13(1), 105-114.

Nei, M., Gojobori, T. (1986).Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions.Molecular biology and evolution, 3(5), 418-426.Nei, M., Kumar, S. (2000). Molecular Evolutionand Phylogenetics, Oxford University Press,New York.

Nie, W., Fu, B., O'Brien, P. C., Wang, J., Su, W., Tanomtong, A. et al. (2008). Flying lemurs-The 'flying tree shrews'?. Molecular cytogenetic evidence for a Scandentia-Dermoptera sister clade. BMC Biology. 6, 18.

Nielsen, R. (2005). Molecular signatures of natural selection.Annu. Rev. Genet., 39, 197-218.

Nosil, P.,Schluter, D. (2011). The genes underlying the process of speciation. Trends in EcologyEvolution, 26(4), 160-167.

Novacek, M.J.(1992). Mammalian phylogeny: shaking the tree. Nature 356, 121-125.

Page, R., Charleston, M. A. (1997). From gene to organismal phylogeny:reconciled trees and the gene tree/species tree problem. Molecular Phylogenetic and Evolution, 7:231–240.

Pozzi, L., Hodgson, J. A., Burrell, A. S., Sterner, K. N., Raaum, R. L.,Disotell, T. R.(2014). Primate phylogenetic relationships and divergence dates inferred from complete mitochondrial genomes. Molecular Phylogenetic and Evolution, 75, 165-183.

Rannala, B., Yang, Z. (1996). Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. Journal of molecular evolution, 43(3), 304-311.

Rannala, B., Yang, Z. (2003). Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. Genetics, 164(4), 1645-1656.

Rasmussen, M. D., Kellis M. (2007). Accurate gene-tree reconstructionby learning gene- and species-specific substitution rates across multiplecomplete genomes. Genome Research, 17:1932-1942.

Rice, A. M., Rudh, A., Ellegren, H.,Qvarnström, A. (2011). A guide to the genomics of ecological speciation in natural animal populations. Ecology Letters, 14(1), 9-18.

Rudd, S. (2003). Expressed sequence tags: alternative or complement to whole genome sequences?. Trends in plant science, 8(7), 321-329.

Rundle, H. D.,Nosil, P. (2005). Ecological speciation. Ecology letters, 8(3), 336-352.

Russell, S. (2012). From sequence to function: the impact of the genome sequence on Drosophila biology. Briefings in functional genomics, 11(5), 333-335.

Sanderson, M. J., McMahon, M. M.(2007). Inferring angiosperm phylogenyfrom EST data with widespread gene duplication. BMC Evolutionary Biology 7(Suppl 1):S3.

Sanger, F., Nicklen, S.,Coulson, A. R. (1977).DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences USA, 74(12), 5463-5467.

Schluter, D. (2000). The ecology of adaptive radiation. Oxford University Press.

Schluter, D. (2001). Ecology and the origin of species. Trends in Ecology and Evolution, 16(7), 372-380.

Schluter, D. (2009). Evidence for ecological speciation and its alternative. Science, 323(5915), 737-741.

Seehausen, O., Terai, Y., Magalhaes, I. S., Carleton, K. L., Mrosso, H. D., Miyagi, R., et al. (2008). Speciation through sensory drive in cichlid fish. Nature, 455(7213), 620-626.

Simmons, N. B. (1998). A reappraisal of interfamilial relationships of bats. En, Bat Biology and Conservation (ed. TH Kunz and PA Racey), pp. 3-26. Smithsonian Institution Press, Washington.

Song, S., Liu, L., Edwards, S. V.,Wu, S. (2012). Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. Proceedings of the National Academy of Sciences USA, 109(37), 14942-14947.

Springer, M. S., Teeling, E. C., Madsen, O., Stanhope, M. J.,de Jong, W. W. (2001). Integrated fossil and molecular data reconstruct bat echolocation. Proceedings of the National Academy of Sciences USA, 98, 6241-6246.

Tandler, B.,Phillips, C. J. (2004). Microstructure of mammalian salivary glands and its relationship to diet.En: Glandular Mechanisms of Salivary Secretion. Garrett JR, Ekström J, Anderson LC (eds). Front Oral Biol. Basel, Karger, vol 10, pp 21-35.

Teeling, E. C., Springer, M. S., Madsen, O., Bates, P., O'brien, S. J.,Murphy, W. J.(2005). A molecular phylogeny for bats illuminates biogeography and the fossil record. Science 307, 580-584.

Thavamanikumar, S., Southerton, S., Thumma, B. (2014). RNA-Seq using two populations reveals genes and alleles controlling wood traits and growth in eucalyptus nitens. PloS one, 9(6), e101104.

Tsagkogeorga, G., Parker, J., Stupka, E., Cotton, J. A.,Rossiter, S. J. (2013). Phylogenomic analyses elucidate the evolutionary relationships of bats. Current Biology, 23(22), 2262-2267.

Vogel, J. P., Garvin, D. F., Mockler, T. C., Schmutz, J., Rokhsar, D., Bevan, M. W., et al. (2010). Genome sequencing and analysis of the model grass Brachypodium distachyon. Nature, 463(7282), 763-768

Wetterer, A. L., Rockman, M. V.,Simmons, N. B. (2000). Phylogeny of phyllostomid bats (Mammalia: Chiroptera): data from diverse morphological systems, sex chromosomes, and restriction sites. Bulletin of the american Museum of natural History, 1-200.

Yang, Z. (1998). Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Molecular biology and evolution, 15(5), 568-573.

Yang, Z., Nielsen, R. (1998).Synonymous and nonsynonymous rate variation in nuclear genes of mammals. Journal of Molecular Evolution, 46(4), 409-418.

Yang, Z., Bielawski, J. P. (2000).Statistical methods for detecting molecular adaptation. Trends in Ecology and Evolution, 15(12), 496-503.

Yang, Z., Nielsen, R. (2000).Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Molecular Biology and Evolution, 17(1), 32-43.

Yu, L., Zhang, Y.P., (2006). Phylogeny of the caniform carnivora: evidence from multiple genes. Genetica, 127, 65-79.

Zhao, Q. Y., Wang, Y., Kong, Y. M., Luo, D., Li, X.,Hao, P. (2011). Optimizing de novo transcriptome assembly from short-read RNA-Seq data: a comparative study.BMC bioinformatics, 12(Suppl 14), S2.

Zhou, X., Xu, S., Xu, J., Chen, B., Zhou, K., Yang, G. (2012). Phylogenomic analysis resolves the interordinal relationships and rapid diversification of the Laurasiatherian mammals. Systematic Biology, 61(1), 150-164.

ANEXOS

13. Anexos

13.1 Anexo I

Acceso libre: http://www.biomedcentral.com/content/pdf/1471-2164-15-446.pdf

Giorello *et al. BMC Genomics* 2014, **15**:446
http://www.biomedcentral.com/1471-2164/15/446

BMC Genomics

RESEARCH ARTICLE Open Access

Characterization of the kidney transcriptome of the South American olive mouse *Abrothrix olivacea*

Facundo M Giorello[1*], Matias Feijoo[1], Guillermo D'Elía[2], Lourdes Valdez[2], Juan C Opazo[2], Valeria Varas[2], Daniel E Naya[1] and Enrique P Lessa[1]

Abstract

Background: The olive mouse *Abrothrix olivacea* is a cricetid rodent of the subfamily Sigmodontinae that inhabits a wide range of contrasting environments in southern South America, from aridlands to temperate rainforests. Along its distribution, it presents different geographic forms that make the olive mouse a good focal case for the study of geographical variation in response to environmental variation. We chose to characterize the kidney transcriptome because this organ has been shown to be associated with multiple physiological processes, including water reabsorption.

Results: Transcriptomes of thirteen kidneys from individuals from Argentina and Chile were sequenced using Illumina technology in order to obtain a kidney reference transcriptome. After combining the reads produced for each sample, we explored three assembly strategies to obtain the best reconstruction of transcripts, TrinityNorm and DigiNorm, which include its own normalization algorithms for redundant reads removal, and Multireads, which simply consist on the assembly of the joined reads. We found that Multireads strategy produces a less fragmented assembly than normalization algorithms but recovers fewer number of genes. In general, about 15000 genes were annotated, of which almost half had at least one coding sequence reconstructed at 99% of its length. We also built a list of highly expressed genes, of which several are involved in water conservation under laboratory conditions using mouse models.

Conclusion: Based on our assembly results, Trinity's *in silico* normalization is the best algorithm in terms of cost-benefit returns; however, our results also indicate that normalization should be avoided if complete or nearly complete coding sequences of genes are desired. Given that this work is the first to characterize the transcriptome of any member of Sigmodontinae, a subfamily of cricetid rodents with about 400 living species, it will provide valuable resources for future ecological and evolutionary genomic analyses.

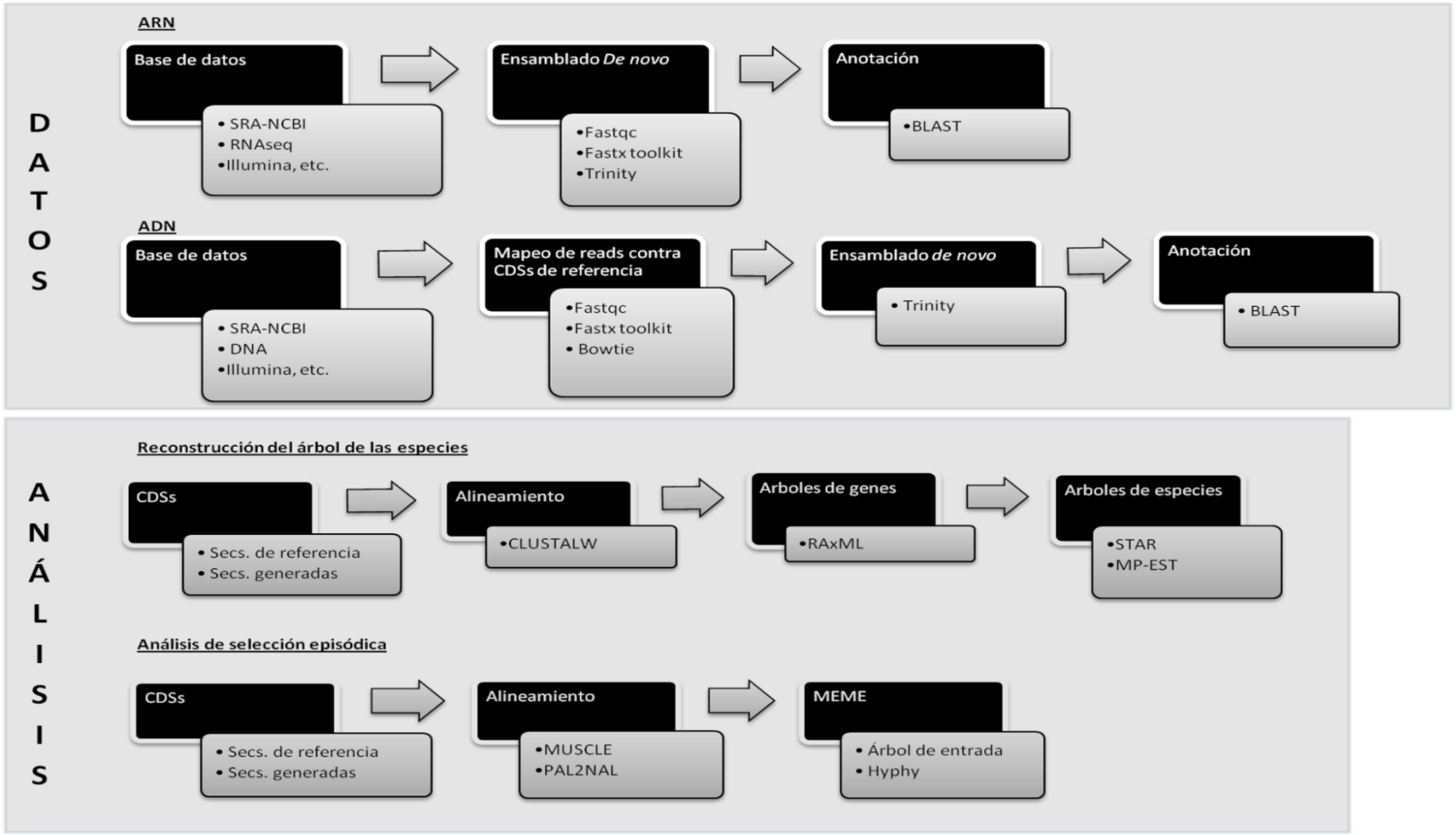
Keywords: *Abrothrix olivacea*, Abrotrichini, Cricetidae, Sigmodontinae, Muroidea, RNA-Seq, Gene expression, *De novo* assembly, Normalization methods

Background

The olive mouse *Abrothrix olivacea* [1] is a cricetid rodent of the subfamily Sigmodontinae, one of the largest mammalian subfamilies with about 400 species and 86 living genera [2,3]. The olive mouse is distributed along Chile and Argentinean Patagonia, from 18°S to 55°S latitude [4], extending for over 1000 km latitudinally, and encompassing a great variety of environments: coastal deserts in the north, Mediterranean scrubs in central Chile, Valdivian and Magallanic forests through the south of Chile and Argentina and Patagonian steppe towards the Atlantic coast. *A. olivacea* must withstand the arid Chilean north and the Patagonia steppe, as well as the Valdivian rain forest with 2700 mm or more of annual rainfall [5]. Given the striking biotic and abiotic differences among these environments, differences in thermoregulation and osmoregulation, among other physiological traits, are expected to occur. Higher tolerance to water shortage in populations from xeric habitat has already been demonstrated [6]. On the basis of variation in morphology, coloration patterns, and more recently DNA sequence data [4,7], many *A. olivacea* subspecies have been described and at

* Correspondence: fagire@gmail.com
[1]Departamento de Ecología y Evolución, Facultad de Ciencias, Universidad de la República, Montevideo, Uruguay
Full list of author information is available at the end of the article

13.2 Anexo II


DATOS
ARN
Base de datos
• SRA-NCBI
• RNAseq
•Illumina, etc.
Ensamblado De novo
•Fastqc
•Fastx toolkit
•Trinity
Anotación
•BLAST
ADN
Base de datos
• SRA-NCBI
• DNA
•Illumina, etc.
Mapeo de reads contra CDSs de referencia
•Fastqc
•Fastx toolkit
• Bowtie
Ensamblado de novo
• Trinity
Anotación
• BLAST
ANÁLISIS
Reconstrucción del árbol de las especies
CDSs
• Secs. de referencia
• Secs. generadas
Alineamiento
•CLUSTALW
Arboles de genes
•RAxML
Arboles de especies
•STAR
•MP-EST
Análisis de selección episódica
CDSs
• Secs. de referencia
• Secs. generadas
Alineamiento
•MUSCLE
•PAL2NAL
MEME
• Árbol de entrada
• Hyphy