Desarrollo de nuevos marcadores genómicos y su aplicación a la filogenia y variabilidad genética de mamíferos

Javier Igea de Castro

DESARROLLO DE NUEVOS MARCADORES GENÓMICOS Y SU APLICACIÓN A LA FILOGENIA Y VARIABILIDAD GENÉTICA DE MAMÍFEROS

Javier Igea de Castro

FACULTAT DE BIOLOGIA
DEPARTAMENT DE GENÈTICA
Programa de Doctorado de Genètica

Desarrollo de nuevos marcadores genómicos y su aplicación a la filogenia y variabilidad genética de mamíferos

Memoria presentada por **Javier Igea de Castro** para optar al título de Doctor por la Universidad de Barcelona

Trabajo realizado en el Instituto de Biología Evolutiva (CSIC-UPF)

Javier Igea de Castro
Barcelona, Noviembre de 2012

Director
José Castresana Villamor
Investigador Científico
Instituto de Biología Evolutiva (CSIC-UPF)

Tutor
Julio Rozas Liras
Catedrático de Genética
Universitat de Barcelona

AGRADECIMIENTOS

En primer lugar, quiero agradecerle a Jose Castresana el darme la oportunidad de hacer la tesis en el grupo y apoyarme durante todo el proceso, en sus épocas buenas y en las no tan buenas. Me llevo un buen carro de conocimientos y experiencias aprendidas que espero saber aprovechar. También me gustaría agradecer a Julio Rozas por haber aceptado ser mi tutor y ayudarme en el complicado mundo del papeleo universitario.

Por supuesto, no puede faltar una mención al P59 y asociados: Gerard, Víctor, Álex, Joan y Ana. Gracias a todos por los momentos compartidos, las risas, las birras, el *marmolet* y todos los consejos y ayudas. También me gustaría agradecer al resto de compañeros del instituto, especialmente a Dolors y la gente de su laboratorio, que acogieron mis inicios en el cacharreo en el laboratorio en los (lejanos) tiempos del CID. Y sin duda, también ha habido muchas otras personas sin cuyo aporte de ideas, muestras o ayuda en el laboratorio esta tesis no hubiera podido llegar a buen puerto. A todos ellos, muchas gracias.

Por último me gustaría agradecer a mis amigos, por estar siempre disponibles para una charla-cerveza y por tener la paciencia casi infinita ("¿infinita o...?") necesaria para soportar mi a veces peculiar sentido del humor. Y como no, a mi familia, especialmente a Luis y a Bea, por acogerme cuando llegué a Barcelona y cuidar tan bien de mí estos años y, por supuesto, a mi madre, gracias por TODO.

CONTENIDO

I.-INTRODUCCIÓN 1

1. MARCADORES MOLECULARES EN RECONSTRUCCIÓN FILOGENÉTICA 3

1.1 Alozimas 4

1.2 Polimorfismos de longitud de los fragmentos de restricción (RFLPs) 5

1.3 Microsatélites 6

1.4 Polimorfismos de longitud de los fragmentos amplificados (AFLPs) 7

1.5 ADN mitocondrial 8

1.6 Genes nucleares 9

1.7. Polimorfismos de nucleótido simple (SNPs) 11

2. ÁRBOLES DE GENES Y ÁRBOLES DE ESPECIES 13

2.1 Discordancia de los árboles de genes y el árbol de especies 14

2.2 Separación incompleta de linajes 15

3. ORGANISMOS DE ESTUDIO 18

3.1 Galemys pyrenaicus *como modelo de especie con requerimientos ecológicos estrictos* 18

3.2 Galemys pyrenaicus *(E. Geoffroy St. Hilaire, 1811)* 19

3.3 El género Neomys *Kaup 1829 como modelo para estudiar especiación y divergencia poblacional* 22

3.4 Los musgaños del género Neomys 23

II.-OBJETIVOS 27

III.-MATERIALES Y MÉTODOS 31

1. DESARROLLO DE NUEVOS MARCADORES GENÓMICOS 33

1.1 Protocolo de extracción y filtrado de intrones de mamíferos 33

1.1.1 Extracción de intrones de los genomas 33

1.1.2 Filtrado de los intrones 34

1.2 Validación experimental de los nuevos marcadores moleculares 37

1.2.1 Diseño de los cebadores y especies empleadas 37

1.2.2 Extracciones de ADN genómico a partir de tejido fresco 38

1.2.3 Amplificación por PCR de intrones 38

1.2.4 Clonación de productos de PCR 39

1.2.5 Análisis filogenéticos de los nuevos marcadores 39
2. FILOGEOGRAFÍA DEL DESMÁN IBÉRICO EMPLEANDO SECUENCIAS MITOCONDRIALES E INTRÓNICAS 40
2.1 Obtención de muestras 40
2.1.1 Prospección de excrementos de desmán 40
2.1.2 Muestras de tejido fresco 40
2.1.3 Muestras de colecciones de museo 40
2.2 Extracciones de ADN genómico y PCR 41
2.2.1 Extracciones de ADN genómico a partir de tejido fresco 41
2.2.2 Extracciones de ADN genómico a partir de excrementos 41
2.2.3 Extracción de ADN genómico a partir de muestras de museos 41
2.2.4 Amplificación por PCR de secuencias mitocondriales 42
2.2.5 Amplificación por PCR a partir de ADN de museo 42
2.2.6 Amplificación por PCR de intrones nucleares 43
2.3 Análisis filogenéticos del desmán ibérico 44
2.3.1 Filogenia mitocondrial 44
2.3.2 Análisis de secuencias nucleares 44
2.4 Análisis de diversidad genética, demografía y estructura genética 45
2.5 Estima del tiempo al ancestro común más reciente de las secuencias mitocondriales 46
2.6 Modelado de la distribución de la especies 48
3. ESTIMA DEL ÁRBOL DE ESPECIES DEL GÉNERO *NEOMYS* 50
3.1 Obtención de muestras 50
3.2 Extracciones de ADN genómico y PCR 50
3.2.1 Extracción de ADN genómico a partir de tejido fresco 50
3.2.2 Extracción de ADN genómico a partir de excrementos 50
3.2.2 Extracción de ADN genómico de un cráneo hallado en una egagrópila 50
3.2.3 Amplificación por PCR del citocromo *b* 50
3.2.4 Amplificación por PCR de intrones nucleares 51
3.3 Filogenia mitocondrial del género Neomys 51
3.4 Estima del árbol de especies del género Neomys 53
3.4.1 Obtención de las tasas evolutivas correspondientes al género *Neomys* 53
3.4.2.Estima de árboles de especies con *BEAST 55
3.4.3 Efecto de distintos priors en las tasas evolutivas sobre la estima de los tiempos de divergencia en *BEAST 56
3.5 Aplicación de un modelo de aislamiento con migracón a la divergencia de los linajes de Neomys anomalus 57

IV. RESULTADOS Y DISCUSIÓN 59

1. DESARROLLO DE UN NUEVO CONJUNTO DE MARCADORES GENÓMICOS PARA LA FILOGENIA DE ESPECIES CERCANAS DE MAMÍFEROS 61

1.1 Resultados 61

1.1.1 Obtención del conjunto de intrones 61

1.1.2. Análisis de las características genómicas del conjunto de intrones 63

1.1.3. Análisis de las distancias genéticas y los polimorfismos de nucleótido simples (SNPs) 65

1.1.4 Validación experimental de los nuevos marcadores moleculares: diseño de cebadores y PCR 66

1.1.5 Uso de los intrones seleccionados como marcadores para la filogenia de especies cercanas 68

1.2 Discusión 71

1.2.1 Características del conjunto final de intrones 71

1.2.1 Utilidad experimental de los intrones seleccionados para la filogenia de especies cercanas 72

2. FILOGEOGRAFÍA DEL DESMÁN IBÉRICO EMPLEANDO DATOS MITOCONDRIALES Y NUCLEARES 74

2.1 Resultados *74*

2.1.1. Análisis filogeográfico mitocondrial 74

2.1.2. Diversidad genética mitocondrial 76

2.1.4. Estima del tiempo al ancestro común más reciente (tmrca) de las secuencias mitocondriales 81

2.1.5. Modelado de la distribución de la especie en el LGM 84

2.2 Discusión 84

2.2.1. Evolución pleistocénica de las poblaciones de desmán ibérico 84

2.2.2 Influencia de los requerimientos acuáticos en la estructura genética del desmán ibérico 87

2.2.3. Fuertes señales de aislamiento en las zonas de contacto 88

2.2.4. Subespecies 89

2.2.5. Implicaciones en la conservación del desmán ibérico 90

3. CÁLCULO DE TIEMPOS DE DIVERGENCIA EN EL GÉNERO *NEOMYS* MEDIANTE ESTIMA DE ÁRBOL DE ESPECIES 92

3.1 Resultados 92

3.1.1 Filogenia mitocondrial del género *Neomys* 92

3.1.2. Análisis de secuencias nucleares en el género *Neomys* 94

3.1.3. Cálculo de tasas evolutivas de intrones y citocromo *b* en el género *Neomys* 96
3.1.4. Árbol de especies del género *Neomys* 99
3.1.5. Efecto de distintos priors, tipos de genes, y cálculo de las tasas evolutivas sobre la estima de los tiempos de divergencia en *BEAST 99
3.1.6. Aislamiento con migración 101
3.2.Discusión 102
3.2.1.Relaciones filogenéticas y divergencia dentro del género *Neomys* 102
3.2.2. Divergencia de los dos linajes de *Neomys anomalus* 103
3.2.2. Consideraciones sobre los priors de las tasas evolutivas en *BEAST 105
3.2.3. Implicaciones taxonómicas de los resultados 107

V. CONCLUSIONES 109

VI. APÉNDICES 115
PUBLICACIÓN 1 117
PUBLICACIÓN 2 133

VII. BIBLIOGRAFÍA 149

LISTA DE FIGURAS

Figura 1. Publicaciones indexadas en la base de datos de Pubmed.....14
Figura 2. Distribuciones de tiempo medios de coalescencia. 16
Figura 3. Ejemplo de discordancia entre árbol de genes y de especies.....17
Figura 4. *Galemys pyrenaicus*.....19
Figura 5. Mapa de distribución de *G. pyrenaicus* 20
Figura 6. Mapas de distribución del género *Neomys* 23
Figura 7. *Neomys anomalus* y *N. fodiens* 25
Figura 8. Esquema de los procesos de extracción y filtrado de los intrones..... 62
Figura 9. Árbol filogenético de referencia..... 63
Figura 10. Localización genómica de los 224 intrones del conjunto final. 64
Figura 11. Distancias genéticas intrónicas entre humano y chimpancé..... 66
Figura 12. Árboles filogenéticos de siete intrones seleccionados..... 70
Figura 13. Distancias por pares de intrones en las parejas de especies cercanas analizadas.....71
Figura 14. Filogeografía de *G. pyrenaicus*..... 75
Figura 15. Reconstrucción por parsimonia de los cambios aminoacídicos a lo largo de la filogenia del citocromo *b* de los linajes de los tálpidos. 76
Figura 16. Mapa de contornos de la diversidad genética de *G. pyrenaicus*..... 78
Figura 17. Muestras de *G. pyrenaicus* agrupadas por cuencas para el análisis AMOVA 79
Figura 18. Genealogías haplotípicas de cinco intrones en *G. pyrenaicus* 80
Figura 19. Distribución geográfica de las variantes de 3 SNPs en intrones de *G. pyrenaicus* 81
Figura 20. Datación del tmrca de los haplotipos mitocondriales de *G. pyrenaicus*.... 83
Figura 21. Distribución potencial de *G. pyrenaicus* durante el Último Máximo Glacial 84
Figura 22. Representación esquemática de la historia evolutiva de *G. pyrenaicus* ... 87
Figura 23. Filogenia mitocondrial del género *Neomys*. 93
Figura 24. Genealogías haplotípicas de los 13 intrones amplificados en el género *Neomys* 95
Figura 25. Obtención de tasas evolutivas en el género *Neomys* 98
Figura 26. Árbol de especies del género *Neomys*..... 99
Figura 27. Tiempos de divergencia en el árbol de especies con distintas estrategias de obtención de tasas..... 100

LISTA DE TABLAS

Tabla 1. Priors de calibración usados en los análisis de BEAST de laurasiaterios 48
Tabla 2. Intrones seleccionados para su amplificación y secuenciado en seis especies 68
Tabla 3. Diversidad genética mitocondrial de *G. pyrenaicus* 77
Tabla 4. Diversidad genética nuclear de *G. pyrenaicus* 80
Tabla A1. Cebadores usados para amplificar genes mitocondriales 139
Tabla A2. Cebadores usados para amplificar intrones 141
Tabla A3. Muestras biológicas empleadas en el estudio de *Galemys pyrenaicus* 142
Tabla A4. Muestras biológicas empleadas en el estudio del género *Neomys* 146

I.-INTRODUCCIÓN

1. MARCADORES MOLECULARES EN RECONSTRUCCIÓN FILOGENÉTICA

La clasificación de los seres vivos requiere de criterios de medida para poder ser realizada de forma objetiva. Se hace necesario, por tanto, comparar características comunes entre los organismos para determinar cuáles son más parecidos. La observación de los caracteres morfológicos externos fue históricamente la base conceptual de la taxonomía, por ser éstos de observación inmediata. El ejemplo clásico de esta clasificación sería la de Linneo en el siglo XVIII, que sentó las bases de la Taxonomía Moderna (Linné 1758).

En el siglo XIX se fue extendiendo el concepto de "Árbol de la Vida" y con él la idea de establecer una clasificación que agrupara a los seres vivos según su parentesco evolutivo (Lamarck 1830; Darwin 1859). Sin embargo, esta idea no llegaría a cristalizar hasta mediados del siglo XX, con la formulación de la sistemática filogenética, propuesta por Willi Hennig (Hennig 1966), que explicitaba los principios a seguir a la hora de elaborar las clasificaciones de los organismos. En concreto, se establecía una distinción clara entre semejanza fenotípica (que podía estar dada por fenómenos de homoplasia) y parentesco, que debía determinarse usando caracteres homólogos derivados (sinapomorfías).

Por otro lado, al mismo tiempo estaban teniendo lugar una serie de importantes descubrimientos que sentaban las bases de la biología molecular moderna, como el experimento Hershey-Chase (Hershey & Chase 1952), que confirmó la naturaleza del ADN como material hereditario o la caracterización de la estructura de doble hélice del ADN de Watson y Crick (Watson & Crick 1953), cristalizando todos ellos finalmente en el llamado "dogma central de la biología molecular", que describe el flujo de información en los sistemas biológicos desde el ADN a las proteínas, con el ARN como intermediario. Todo ello impulsó la creación de una nueva corriente, encabezada por los bioquímicos Pauling y Zuckerkandl, que abogaba por el uso de los marcadores moleculares para realizar clasificaciones de los seres vivos. Según esta línea de pensamiento, los caracteres moleculares eran una evidencia del proceso evolutivo más clara o directa que los morfológicos y eran compartidos por todos los seres vivos, y resultaban más fácilmente medibles y comparables en todos ellos (Zuckerkandl 1964; Suarez-Diaz & Anaya-Munoz 2008) Los marcadores moleculares proporcionaban, en definitiva, un instrumento de comparación más estandarizado y cuantificable, además de suponer un número casi ilimitado de caracteres de estudio con una menor tendencia

a la homoplasia (aunque ni mucho menos inexistente, como se comprobaría más adelante). La ventaja del enfoque molecular de cara a realizar análisis de carácter cuantitativo se vio enseguida reforzada por los primeros trabajos que incorporaron los algoritmos computacionales a la reconstrucción filogenética (Fitch & Margoliash 1967; Fitch & Margoliash 1968).

Desde esos primeros momentos, se fueron desarrollando gran cantidad de tipos de marcadores moleculares, desde los trabajos iniciales basados en proteínas hasta la popularización de los estudios centrados en el ADN. Éstos últimos se basaron sobre todo en el ADN mitocondrial durante los años 90 y ya en el siglo XXI, se introdujeron las técnicas de secuenciación de nueva generación, que permiten la obtención de enormes cantidades de información a un coste cada vez menor. Cabe destacar que ante la gran variedad de marcadores moleculares actualmente disponibles, con características muy distintas a nivel de requisitos técnicos, facilidad de análisis y precio, la elección de uno u otro tipo de marcador ha de venir determinada por el tipo de información y las preguntas que se quieren responder con su uso (Schlotterer 2004). Los niveles de resolución requeridos por el estudio que se llevará a cabo determinarán en gran medida el tipo de marcador que hay que emplear. En este sentido, los marcadores moleculares adecuados para estudios filogenéticos deben cumplir una serie de características (Sunnucks 2000; Avise 2000): deben ser abundantes y estar distribuidos por todo el genoma; su modo de evolución debe estar bien analizado, de forma que se pueda modelar e incorporar a análisis; y los datos obtenidos en distintos laboratorios y estudios deben poder ser comparables, para poder realizar inferencias generales. A continuación se describen algunos de los tipos de marcadores moleculares más utilizados en los últimos años.

1.1 Alozimas

Los primeros marcadores moleculares empleados fueron las alozimas. Esta técnica se basa en estudiar las diferencias de carga eléctrica y tamaño entre las distintas variantes alélicas de un enzima (de ahí el nombre “alozimas”), que provocan diferencias en la movilidad dentro de un campo eléctrico generado en una electroforesis. Los primeros estudios con alozimas en poblaciones naturales fueron llevados a cabo en humanos y *Drosophila*, en la década de los 60 del siglo XX (Harris 1966; Hubby & Lewontin 1966). Los resultados revelaron una sorprendente, y hasta entonces no anticipada, cantidad de polimorfismo. De hecho, la incorporación de estos nuevos datos facilitó el desarrollo de nuevas hipótesis como la teoría neutral de la evolución molecular (Kimura 1968). A

partir de entonces, impulsados por las ventajas de coste y aplicación de esta nueva técnica, se llevaron a cabo estudios de este tipo en gran variedad de organismos. Sin embargo, el riesgo de emplear marcadores afectados por la selección natural (y por tanto de evolución no neutra) es mucho mayor en el caso de los marcadores basados en proteínas.

Además, las alozimas suponen una forma indirecta de detectar variación heredada. Los marcadores basados en el estudio de la variación del ADN, por el contrario, presentan la ventaja de suponer un estudio de la variación más directa y, además, cuantificable, ya que se pueden medir (con más o menos precisión según la metodología empleada) el número de mutaciones ocurridas. Debido a esto, a partir de la década de los 80 pasaron a emplearse mayoritariamente marcadores moleculares basados en ADN.

1.2 Polimorfismos de longitud de los fragmentos de restricción (RFLPs)

El descubrimiento en 1968 de las endonucleasas de restricción (Linn & Arber 1968) permitió analizar diferencias en los patrones de corte de estas enzimas en el ADN visualizadas mediante una electroforesis. Las diferencias observadas entre los perfiles de digestión de distintos individuos reflejan mutaciones en los lugares de corte de la enzima de restricción empleada, tanto a nivel de sustituciones nucleotídicas como a nivel de inserciones o deleciones. La aplicación de esta técnica, extendida durante los años 80, permitió, por primera vez, analizar zonas no codificantes del ADN, así como variación silenciosa (sin efecto fenotípico) en las zonas codificantes. Fueron empleados en la elaboración de los primeros mapas genéticos basados en ADN y en estudios de asociación, y también en muchos trabajos filogenéticos y de genética de poblaciones. Sin embargo, la necesidad de emplear una sonda de hibridación para detectar el polimorfismo (y la imposibilidad de disponer de una adecuada en muchos casos) impidió que la aplicación de esta técnica fuera más generalizada.

Sin duda, el mayor impulso a la llamada revolución molecular fue el desarrollo de la reacción en cadena de la polimerasa (PCR) en los años 80 (Saiki *et al.* 1985; Saiki *et al.* 1988), que permitía amplificar cualquier región del genoma en un gran número de individuos de forma rápida y empleando una mínima cantidad de ADN en el proceso. Esto, unido al desarrollo de la secuenciación del ADN en 1977 (Sanger *et al.* 1977) y su posterior popularización, permitió la generación de enormes cantidades de datos y el despegue de los campos de la filogenia y la filogeografía. Cabe destacar también que, a

estos dos avances en el campo de la obtención de datos, se les unió un gran desarrollo en la computación, lo que permitió el procesamiento rápido de todos estos datos y la aplicación de modelos estadísticos para su análisis cada vez más complejos.

1.3 Microsatélites

Los microsatélites fueron uno de los primeros marcadores moleculares que se popularizaron aprovechando la aplicación de la tecnología de la PCR (Tautz 1989). Consisten en secuencias de periodo corto (de entre 2 y 6 nucleótidos) repetidas en tándem, hasta un tamaño típicamente de alrededor de menos de 500 pares de bases. Se trata de unos marcadores altamente polimórficos y abundantes a lo largo del genoma. Además, su tasa evolutiva es elevada (alrededor de 1 mutación cada 1000 generaciones, un valor varias órdenes de magnitud superior al de otros tipos de marcadores), lo que los hace muy adecuados para estudios a nivel de poblaciones (Ellegren 2004). El proceso de obtención de los microsatélites puede ser relativamente costoso pero, una vez realizado, pueden aplicarse los marcadores obtenidos a taxones cercanos al empleado en el estudio inicial.

Sin embargo, los microsatélites poseen un modelo evolutivo complejo: los cambios de longitud en los alelos se producen por el deslizamiento durante la replicación durante la meiosis. Esto complica la modelación adecuada de su evolución en los procesos de inferencia filogenética, dificultando su análisis (Selkoe & Toonen 2006). Por otro lado, presentan tasas muy variables entre loci y alelos (Brinkmann *et al.* 1998), lo que complica la realización de estudios comparativos entre linajes, debido a esta falta de aplicabilidad general. Además, debido a su elevada tasa de cambio y a su complejo modelo evolutivo, tienden a acumular cambios homoplásicos (Primmer & Ellegren 1998). Además, dificultades técnicas como el riesgo de aparición de alelos nulos (fallos en la amplificación de un microsatélite o una variante del mismo por mutaciones en la zona de unión del cebador) o de las llamadas “stutter bands” (amplificaciones de productos de PCR algo menores que el alelo real y que se producen por errores de la polimerasa al trabajar con secuencias tan repetitivas) complican en ocasiones el uso de estos marcadores y su aplicación a especies no estrechamente relacionadas entre sí.

Otro problema añadido es que, debido al proceso de desarrollo de estos marcadores, existe un riesgo de incurrir en un sesgo de muestreo (*ascertainment bias*), ya que se seleccionan los marcadores más polimórficos en la población de estudio. En este sentido, se ha señalado que las estimas de variabilidad o heterocigosidad obtenidas a

partir de estudios con microsatélites pueden no reflejar de forma precisa la variabilidad del genoma, y por tanto, éstas deberían ser analizadas con precaución (Vali *et al.* 2008).

1.4 Polimorfismos de longitud de los fragmentos amplificados (AFLPs)

Los AFLPs se basan en una digestión genómica completa usando enzimas de restricción seguida por una amplificación selectiva por PCR de una serie de los fragmentos obtenidos (Vos *et al.* 1995). Se obtiene así un perfil único para cada individuo, compuesto de marcadores distribuidos por todo el genoma, y que normalmente pertenecen a regiones no codificantes. No requieren, por tanto, conocimiento *a priori* sobre los genomas de los organismos que se van a analizar (Bensch & Akesson, 2005) y, por consiguiente, se han aplicado con éxito en un amplísimo conjunto de organismos distintos (Bonin *et al.* 2007).

Sin embargo, el mayor inconveniente que presenta su uso es el riesgo de homoplasia, que puede resultar de la comigración de fragmentos de ADN distintos entre los perfiles de AFLP de individuos diferentes, o de la comigración de fragmentos dentro del propio perfil de un individuo (Caballero & Quesada 2010). Estas asignaciones incorrectas de la homología de algunos fragmentos pueden provocar que las inferencias filogenéticas realizadas sean incorrectas y derivar en sesgos de los parámetros calculados, sobre todo en contextos de estudios filogenéticos profundos (Garcia-Pereira *et al.* 2010). Por otro lado, ciertos trabajos han señalado la importancia de establecer protocolos para asegurar la reproducibilidad de los resultados (Meudt & Clarke 2007).

Además de todos estos marcadores ya descritos, probablemente la técnica más usada en los estudios filogenéticos y filogeográficos es la secuenciación de fragmentos conocidos (genes o regiones no codificantes) en una o varias especies. El primer estudio de este tipo fue anterior al desarrollo de la PCR, con la clonación de la región Adh de 11 especímenes de *Drosophila melanogaster* (Kreitman 1983), pero fue con la PCR cuando el número de estudios con secuencias aumentó exponencialmente.

1.5 ADN mitocondrial

A finales de los años 70, empezó a gestarse un cuerpo de evidencias que señalaban al ADN mitocondrial como una molécula muy adecuada para estudiar las relaciones filogenéticas entre los organismos, en particular los muy estrechamente relacionados. Concretamente, en 1979, Brown *et al.* describieron una sorprendentemente elevada tasa evolutiva para el genoma mitocondrial de animales, que en aquel momento se cuantificó como del 2% de divergencia entre un par de linajes por cada millón de años. La posterior aplicación de la PCR, unida a la relativa facilidad de amplificar el ADN de este orgánulo comparado con el ADN nuclear hizo que se popularizara su uso hasta el punto de constituir la herramienta de trabajo principal de una nueva disciplina, la filogeografía (Avise *et al.* 1987), que estudia los principios y procesos que gobiernan las distribuciones geográficas de los linajes genealógicos, especialmente a nivel intraspecífico y entre especies cercanas. (Avise 2000). Un estudio fundamental para entender esta importancia capital es el de Kocher *et al.*, que en 1989, y sirviéndose de cebadores de PCR universales diseñados en las zonas más conservadas amplificaron segmentos homólogos de más de 100 especies de vertebrados e invertebrados. Particularmente, descubrieron que en mamíferos el gen del citocromo *b* poseía una tasa evolutiva especialmente elevada y contenía información filogenética útil desde niveles intraspecíficos a intergenéricos. Este descubrimiento y la facilidad de amplificación usando los cebadores de PCR universales propuestos lo convirtieron en el gen secuenciado en mayor número de especies de mamíferos.

El uso de genes mitocondriales en estudios filogeográficos y filogenéticos conlleva una serie de ventajas claras. En primer lugar, como ya se ha comentado, tiene una mayor tasa de sustitución que el genoma nuclear, lo que le proporciona una mayor señal filogenética. Además, debido a su herencia maternal y a su carácter haploide, su tamaño efectivo es cuatro veces menor que el de un gen nuclear en los organismos diploides. Esto implica que la probabilidad de tener monofilia recíproca entre dos especies será mayor con el árbol de un gen mitocondrial que con uno nuclear y se fijarán nuevos alelos con mayor rapidez (Palumbi *et al.* 2001). Además, ya que está presente en un número mayor de copias, la amplificación resulta más sencilla a nivel técnico y se puede obtener a partir de muestras biológicas más degradadas. Por último, el genoma mitocondrial se hereda como un único bloque porque no tiene recombinación (Ballard & Whitlock 2004), lo que es muy útil ya que la historia evolutiva de las secuencias se rompe en un grupo de árboles cuando este fenómeno

ocurre en vez de ser representada por un único árbol (Wiuf *et al.* 2001), y la mayoría de los métodos de reconstrucción filogenética parten de la asunción de ausencia de recombinación.

Sin embargo, el ADN mitocondrial también posee varias desventajas. Primeramente, como ya se ha comentando, toda la molécula de ADN mitocondrial está ligada y se transmite como un solo bloque de herencia. Esto implica que las genealogías obtenidas con cualquiera de los genes mitocondriales serán idénticas ya que pertenecen a la misma molécula. Con la llegada de la teoría de la coalescencia (Kingman 1982) comenzó el reconocimiento explícito de que la genealogía de una molécula (de un bloque no recombinante) particular es una muestra aleatoria de la distribución de todas las genealogías posibles, entre las cuales muchas reflejarán la historia real de las especies, pero otras no lo harán (a esta discordancia y sus implicaciones se hará referencia con más detalle más adelante). Por tanto, realizar inferencias empleando un solo árbol de genes (*gene tree*) puede llevar a conclusiones erróneas. Además, precisamente la herencia materna del ADN mitocondrial implica que las inferencias realizadas basadas en él son aplicables solamente a la historia evolutiva de las hembras y no a la de la especie completa. Así, por ejemplo, los resultados obtenidos en caso de dispersión ligada al sexo pueden ser muy diferentes si sólo se considera el ADN mitocondrial en vez de añadir información del ADN nuclear (Tosi *et al.* 2003). Todas estas críticas han hecho que los estudios filogeográficos hayan ido pasando de estar basados en un solo locus a ir paulatinamente incorporando información de *loci* independientes distribuidos por todo el genoma nuclear (Brito & Edwards 2009).

1.6 Genes nucleares

Los genes nucleares presentan unas tasas de mutación, en general, sensiblemente inferiores a las del ADN mitocondrial. Esto lleva implícito que, a menudo, no sean suficientemente informativos por sí mismos para resolver filogenias de especies cercanas o para ser utilizados a nivel intraespecífico. Por ello, inicialmente la secuenciación de genes nucleares y, más concretamente, de exones se empleó para resolver, de forma exitosa, relaciones filogenéticas a nivel profundo (Stanhope *et al.* 1992; Porter *et al.* 1996), debido fundamentalmente a la ventaja que suponía su menor propensión a la saturación, relacionado esto con la mencionada menor tasa evolutiva al compararla con el genoma mitocondrial.

Sin embargo, esta tasa mutacional baja, aunque minimice la saturación, también disminuye la cantidad de sitios informativos para los estudios filogenéticos. Esto es especialmente notable en las filogenias de especies cercanas. En este sentido, enseguida se señalaron a los intrones como unos buenos candidatos a ser menos susceptibles a estos problemas, debido a que, por carecer, en general, de constricciones funcionales, la posible acumulación de variabilidad será mucho mayor que en el caso de las regiones codificantes, que sí están sujetas a la acción de la selección natural (Zhang & Hewitt 2003). Además, el hecho de que los intrones estén flanqueados por zonas altamente conservadas (exones) facilitó el diseño de cebadores conservados sobre estas zonas que permitieran la amplificación en varias especies distintas. A esta estrategia se la denominó cebadores EPIC (*exon primer intron crossing*) (Slade *et al.* 1993; Lessa 1992; Palumbi & Baker 1994).

Por otro lado, como ya se ha comentado con anterioridad, el tamaño efectivo de los genes nucleares es cuatro veces mayor que el de los mitocondriales, lo que implica que están más afectados por los procesos estocásticos asociados a la coalescencia y, por ello, la necesidad de emplear más de un marcador es aún más patente (Moore 1995).

El uso de las secuencias nucleares en filogenia animal se fue extendiendo, ejemplificado por el uso de marcadores como el BRCA1 (Delsuc *et al.* 2002; Meredith *et al.* 2008; Adkins *et al.* 2001) o RAG1 (Murphy *et al.* 2001; Steppan *et al.* 2004a; Steppan *et al.* 2004b). Por otro lado, la incorporación creciente a los análisis filogenéticos de particiones independientes de datos con la secuenciación de genes nucleares como alternativa o complemento al ADN mitocondrial, implicó disponer de mayor cantidad de sitios informativos para aumentar el soporte estadístico de los nodos estudiados. Además, supuso también tener que incorporar distintos modelos evolutivos para las distintas particiones empleadas, que pueden tener además tasas evolutivas bastante diferentes. La práctica más empleada a la hora de tratar con la cantidad creciente de conjuntos de datos multilocus de muchas especies es la concatenación, que consiste en inferir la filogenia a partir de una única supermatriz de datos obtenida de la combinación de todas las secuencias. De esta forma, se asume que la historia filogenética de todas las particiones es la misma (o muy similar) (William & Ballard 1996). Sin embargo, como se comentará más adelante, esta asunción, bajo ciertos supuestos, no se cumple, llevando a errores en la reconstrucción de la historia evolutiva de los organismos estudiados.

Sin embargo, a pesar de este uso creciente de genes nucleares en la filogenia de mamíferos, la mayoría de los marcadores más empleados no provenían de una búsqueda sistemática destinada a seleccionar las partes del genoma más adecuadas para resolver las hipótesis filogenéticas planteadas.

Con la secuenciación de genomas, se empezaron a realizar búsquedas sistemáticas de marcadores a nivel genómico, de cara a seleccionar los marcadores más adecuados para el nivel de resolución deseado en una serie de grupos taxonómicos. Como ejemplos, cabe citar estudios en distintos grupos de organismos, como peces (Li *et al.* 2007), aves (Backstroem *et al.* 2008), reptiles (Townsend *et al.* 2008; Thomson *et al.* 2008) o mamíferos (Lyons *et al.* 1997) en los que se han seleccionado regiones codificantes (Ranwez *et al.* 2007) y no codificantes (Peng *et al.* 2009).

En general, como se ha mencionado, los intrones poseen menos constricciones funcionales que las partes codificantes del genoma nuclear y, por tanto, pueden acumular mayor número de sustituciones. Sin embargo, se ha comprobado que existen intrones con un alto grado de conservación por estar implicados en algún tipo de función (Gazave *et al.* 2007; Rodova *et al.* 2003), y se ha detectado una gran variabilidad en sus tasas evolutivas. Por ejemplo, una comparación entre intrones ortólogos de humano y ratón mostró distancias genéticas que variaban en casi un orden de magnitud (entre 0.2 y 1.7 sustituciones por sitio) (Castresana 2002). Además, algunos intrones muestran tasas evolutivas muy elevadas en algunos linajes pero no en otros, indicando un reloj molecular imperfecto y una falta de ajuste a la evolución global del genoma que puede dificultar las medidas precisas de diversidad genética en algunas especies (Soria-Carrasco *et al.* 2007). Estas aceleraciones o deceleraciones del reloj molecular pueden venir mediados por cambios en la posición cromosómica de un gen o por la aparición de nuevas isoformas en algun linaje por procesos de splicing alternativo (Plass & Eyras 2006).

1.7 Polimorfismos de nucleótido simple (SNPs)

Un SNP es un cambio puntual en la composición nucleotídica de una secuencia de ADN. Típicamente están presentes en una alta frecuencia en todos los genomas estudiados (aproximadamente uno cada 300 – 1000 pares de bases) constituyendo el tipo de polimorfismo más frecuente en el genoma. Normalmente se trata de marcadores bialélicos, por lo que su información a nivel individual es escasa, y han de analizarse una gran cantidad de ellos de forma conjunta (Schlotterer 2004). Requieren

un conocimiento extensivo de la secuencia del organismo de estudio, y su aplicación interespecífica no es factible a no ser que se trate de especies muy cercanas. Sin embargo, la creciente disponibilidad de genomas de una gran variedad de organismos está haciendo que actualmente su uso esté cada vez más extendido, en parte debido al potencial de generar una enorme cantidad de este tipo de marcadores (Aitken *et al.* 2004; Morin *et al.* 2004).

La llegada de los métodos de nueva generación de secuenciación de ADN, que permiten paralelizar la reacción de secuenciación, ha supuesto incrementar el rendimiento en varias órdenes de magnitud e igualmente disminuir los costes. Esto ha tenido su reflejo en una gran variedad de técnicas que permiten generar nuevos tipos marcadores para emplear en estudios filogenéticos (McCormack *et al.* 2011; Ekblom & Galindo 2011). Un ejemplo de éxito reciente lo constituyen los marcadores S-RAD (*sequenced restricted-site-associated DNA*), que se obtienen generando librerías de fragmentos de restricción (obtenidos a partir de ADN genómico) marcados con adaptadores que permiten su posterior identificación tras el paso de secuenciación masiva (Baird *et al.* 2008). De esta forma se pueden generar grandes cantidades de SNPs con un coste y esfuerzo mucho menor que empleando otras metodologías. De igual manera, con estas técnicas de última generación se han desarrollado conjuntos de otros marcadores, como microsatélites (Allentoft *et al.* 2009).

Además de la generación de nuevos marcadores moleculares, otra aplicación fundamental de estos nuevos métodos es la secuenciación dirigida o secuenciación de amplicones, que permite amplificar un número elevado de *loci* conocidos empleando cebadores marcados específicamente. El marcado de los cebadores puede hacerse tanto a nivel de individuo, lo que permite el genotipado para el *loci* de interés de muchas muestras simultáneamente (Binladen *et al.* 2007), pero también pueden añadirse marcajes a cada *loci* amplificado, de forma que puede obtenerse, en una sola reacción de secuenciación masiva, información multilocus para múltiples individuos (Kloch *et al.* 2010).

2. ÁRBOLES DE GENES Y ÁRBOLES DE ESPECIES

Los "árboles de especies" (*species trees*) en los que la topología y las longitudes de las ramas reflejan la historia evolutiva de los organismos, definen los patrones de separación de los distintos linajes, poblaciones o especies y los tiempos en los que ocurren esas divergencias. El objetivo de la sistemática molecular es ayudar a recuperar el llamado "árbol de la vida", que refleja este patrón único de ramificaciones a lo largo del tiempo, empleando para ello la información que proporciona el ADN. Lógicamente, existe una relación estrecha entre un árbol filogenético calculado a partir de un gen (*gene tree*) y el árbol de las especies que portan esos genes, pero no deben considerarse como equivalentes, ya que se trata de entidades distintas influenciadas por procesos evolutivos diferentes (Knowles & Kubatko 2010). En la reconstrucción filogenética se puede observar y calcular de forma directa la divergencia entre dos copias de genes pero no así la de las especies que los llevan, que siempre será más reciente que cualquier divergencia calculada con un árbol de genes, ya que es muy improbable que las dos moléculas compartan un ancestro común justo en el momento en el que se separaron las especies. Las posibles discordancias entre los árboles de genes y el árbol de especies correspondiente no deben considerarse como errores asociados al proceso de reconstrucción filogenética, si no más bien se debe tratar de incorporar esta heterogeneidad para obtener información sobre procesos temporales y demográficos relacionados con la divergencia de las especies (Nichols 2001). Debería considerarse que a partir del genoma se obtiene una colección de árboles de genes, cada uno con su historia evolutiva independiente y cuya forma puede coincidir o no con la del árbol de especies "real" (Maddison 1997).

Tradicionalmente, se ha hecho mucho énfasis en la inferencia de los árboles de genes a partir de los datos moleculares obtenidos de las secuencias de ADN. Este enfoque ha sido el prioritario en los campos de la filogenia y filogeografía, de forma que se han centrado los esfuerzos en los desarrollos teóricos y prácticos necesarios para obtener las estimas más precisas de los árboles de genes. Esta visión ha determinado, además, los planteamientos de los estudios realizados, tanto a nivel de muestreo como de hipótesis planteadas. No ha sido hasta los últimos años en los que se ha comenzado a reconocer de forma más explícita el problema de la estima de árboles de especies a partir de la información contenida en los árboles de genes. Este gran interés, reflejado en el creciente número de publicaciones aparecidas desde el año 2007 (figura 1), ha traído consigo el desarrollo de un conjunto de metodologías destinadas a integrar la información de varios árboles de genes para obtener el árbol de especies que los

contiene. Inicialmente, se comprobó que los métodos de estima de árboles de especies producían resultados con menor soporte estadístico que los mismos análisis realizados con métodos que empleaban la concatenación (Edwards 2009; Liu *et al.* 2009). Sin embargo, también se ha comprobado que, cuando hay gran heterogeneidad de árboles de genes, los métodos de concatenación, que asumen que todos los *loci* tienen la misma historia evolutiva, pueden llevar a conclusiones erróneas (Kubatko & Degnan 2007).

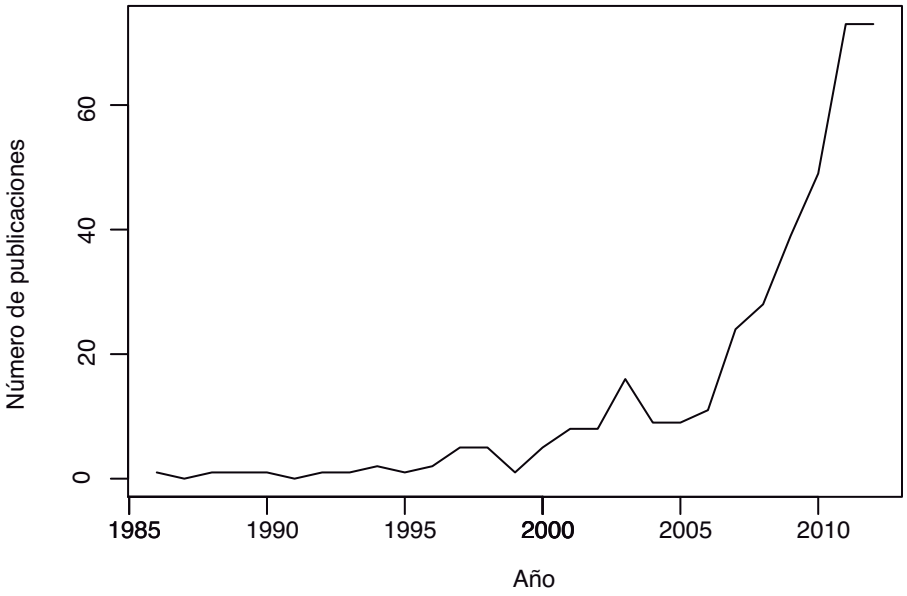

Figura 1. Publicaciones indexadas en la base de datos de Pubmed hasta octubre de 2012 con la expresión "species tree(s)" en su título o abstract

2.1 Discordancia de los árboles de genes y el árbol de especies

Existen varias causas que pueden explicar la discordancia a nivel topológico de los árboles de genes individuales con el árbol de especies correspondiente (Maddison 1997). Por ejemplo, cuando tiene lugar una duplicación génica, se genera una copia adicional del locus de estudio, que por tanto tendrá una historia evolutiva independiente de éste, y si se reconstruye un árbol de genes partiendo del muestreo de ambos duplicados, puede producirse la discordancia con el árbol de especies (Knowles & Kubatko 2010). Esto de nuevo indica la importancia de intentar seleccionar marcadores cuya condición de ortología sea lo más segura posible (por búsquedas genómicas en especies cercanas, por ejemplo). Por otro lado, la introgresión de material genético de una especie en otra cercana por procesos de hibridación también puede producir este tipo de patrones de discordancia. Puede afectar, además, de forma diferencial a distintas regiones del genoma. El último factor que puede provocar discordancias entre un árbol de genes y el árbol de especies correspondiente es la separación incompleta de linajes (*incomplete lineage sorting*) o coalescencia profunda

(*deep coalescence*) que, por su ubicuidad, está centrando actualmente los esfuerzos teóricos y prácticos en la modelización de las estimas de árboles de especies (Brito & Edwards 2009).

Por otro lado, relacionado además con todos los fenómenos mencionados que causan discordancia topológica entre el árbol de genes y árbol de especies, también hay que reseñar la heterogeneidad de longitudes de rama entre distintos árboles de genes (y el árbol de especies). Incluso cuando todos los árboles de genes fueran coincidentes entre sí y con el árbol de especies en lo referente a topología, puede haber diferencias significativas en la longitud de las ramas, también determinadas por los tamaños poblacionales ancestrales (Edwards 2009).

2.2 Separación incompleta de linajes

La modelización del fenómeno de la separación incompleta de linajes está intimamente ligada al desarrollo de la teoría de la coalescencia en los años 80 (Kingman 1982), que permite predecir el tiempo que ha de transcurrir para que dos alelos de una población encuentren su ancestro común más cercano en el pasado. Según esta teoría, el tiempo medio hasta el ancestro común varía por cada *loci* como se puede ver en la figura 2. El tiempo medio de coalescencia para un par de secuencias en un locus autosómico es $2N_e$ generaciones, pero como se puede observar, existe bastante varianza en estos tiempos. Cuando se realiza este cálculo para más de 2 secuencias, el tiempo transcurrido aumenta, siendo la media de $4N_e$ generaciones cuando el número de secuencias es mayor de 10 (representado por la línea de puntos en la figura 2). Además, el tiempo medio de coalescencia aumenta con el tamaño efectivo poblacional (y también se ve modificado por la acción de la selección) (Hein *et al.* 2005). Las relaciones entre los distintos alelos y sus patrones de coalescencia se representan mediante genealogías de genes. Por tanto, estudiando el polimorfismo y los tiempos de coalescencia de distintos *loci* independientes se pueden obtener estimas de los tiempos de divergencia de las especies o poblaciones, así como de los tamaños efectivos de las poblaciones ancestrales que las generaron.

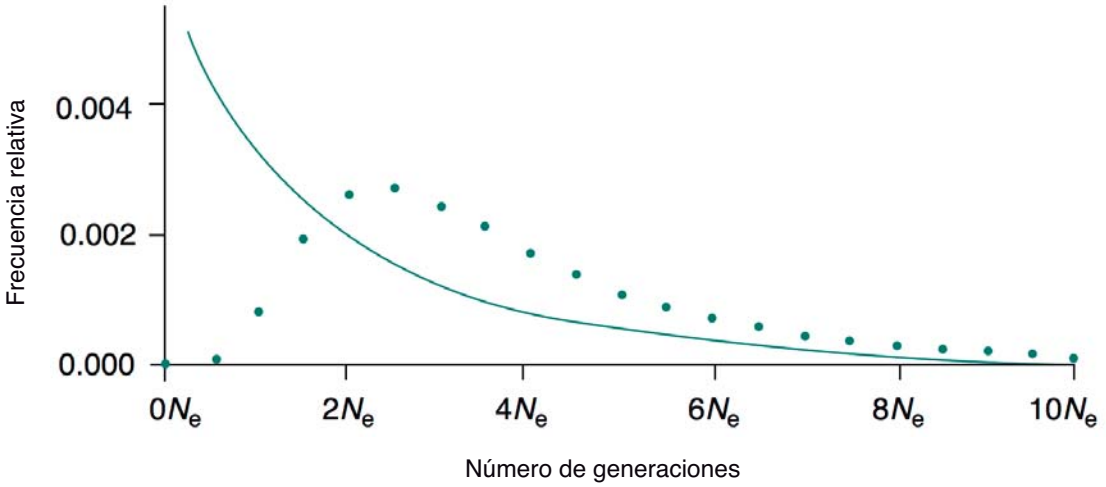

Figura 2. Distribuciones de tiempos medios de coalescencia (tiempo hasta el ancestro común) de un par de secuencias dentro de la misma especie (línea contínua) y de una muestra de varias secuencias (linea discontínua). (Modificado de Nichols 2001).

La separación incompleta de linajes ocurre cuando las copias de un gen dentro de una misma especie no encuentran su ancestro común (es decir, no coalescen) hasta un tiempo anterior al evento de especiación. De esta forma, los polimorfismos existentes en una población ancestral pueden mantenerse después de sucesivos eventos cladogenéticos. Así, al estimar las relaciones filogenéticas de una muestra de alelos de varias especies, puede ocurrir que algunos alelos de una especie estén más emparentados (es decir, compartan un ancestro común más cercano) con otros alelos de otra especie distinta que con el resto de alelos de su propia especie (figura 3).

La causa última de la separación incompleta de linajes es la acción de la deriva genética, que lleva a la fijación de alelos (y por tanto a la pérdida de polimorfismo) y cuya intensidad está determinada por el tamaño efectivo de la población de estudio. Así, cuanto menor es una población, más importante es el efecto de la deriva genética y menos posibilidad existe de que se den fenómenos de coalescencias profundas. Si dos especies divergen por un periodo de tiempo menor de $4N_e$ generaciones, la probabilidad de que un árbol de gen cualquiera coincida con el árbol de especies es baja (Pamilo & Nei 1988; Brower *et al.* 1996). En general, se ha asumido que la separación incompleta de linajes sólo es susceptible de causar una discordancia entre árboles de genes y de especies en radiaciones recientes. Sin embargo, la teoría indica que lo importante para las topologías es la longitud (en unidades de coalescencia) de las ramas entre los eventos de divergencia y no la profundidad en el árbol a la que ocurren esos eventos. En este sentido, recientemente se ha demostrado que la separación incompleta de linajes también puede ser un factor a tener en cuenta a la hora de

estimar las relaciones filogenéticas profundas entre los clados de mamíferos euterios (Song *et al.* 2012), y se ha sugerido que deberían incorporarse estas metodologías de forma más explícita en estudios filogenómicos similares.

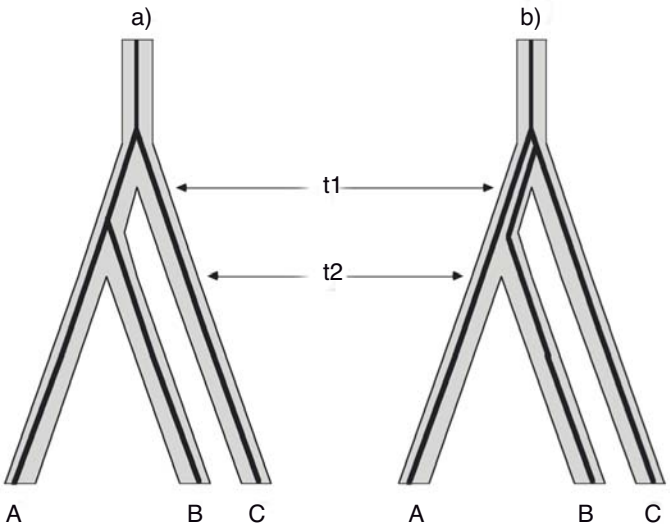

Figura 3. Ejemplo de discordancia entre árbol de genes (representados por líneas negras continuas) y el árbol de especies que las contiene (de color gris) para 3 especies A, B y C. El árbol del gen a) tiene la misma topología que el árbol de especies, al contrario que el árbol del gen b). La longitud de rama entre t1 y t2 relativa al tamaño poblacional determinará la incidencia de la separación incompleta de linajes (modificado de Avise & Robinson 2008)

3. ORGANISMOS DE ESTUDIO

En este trabajo, se han escogido dos taxones para realizar estudios filogenéticos y filogeográficos empleando los nuevos marcadores intrónicos desarrollados. El primero de ellos es el desmán ibérico, *Galemys pyrenaicus*, una especie relativamente desconocida de la que no existen apenas datos genéticos. El segundo de ellos es el género de musgaños europeos *Neomys*, que contiene varias especies con poblaciones muy divergentes.

3.1 *Galemys pyrenaicus* como modelo de especie con requerimientos ecológicos estrictos

Los patrones de diversidad genética de las especies son consecuencia de su historia evolutiva y de las barreras actuales que limitan la dispersión de los organismos. La historia evolutiva de muchos organismos del continente europeo está modelada en gran parte por las glaciaciones pleistocénicas (Taberlet *et al.* 1998; Gómez & Lunt 2007). El aislamiento de poblaciones en refugios aislados durante los episodios glaciales y las posteriores expansiones y recolonizaciones han dejado una señal típica en la estructura genética de las especies, caracterizada por la subdivisión en linajes diferenciados (que son producto del aislamiento en los refugios) y una mayor diversidad actual retenida en los linajes correspondientes a esos refugios, en contraposición a los linajes que ocupan zonas de colonización más reciente, con una mayor homogeneidad genética (Hewitt 2000). En este sentido, el desmán ibérico, un endemismo de la Península Ibérica, constituye un excelente modelo para testar la importancia de todos estos factores en la estructuración de la diversidad genética en especies de distribución reducida.

Sin embargo, las barreras actuales que limitan los movimientos de las especies, y por tanto el flujo genético entre las distintas poblaciones, también pueden ejercer una gran influencia en la estructura genética. Ésta podría verse muy influida por la naturaleza fragmentada de los habitats disponibles para el desmán, que tiene una fuerte dependencia del medio acuático y unos estrictos requerimientos de habitat. De este modo, la diversidad genética estaría particionada, como en el caso de otros organismos acuáticos, de acuerdo con las cuencas fluviales (Avise 2000), al igual que ocurre en otras especies que ocupan habitats fragmentados por naturaleza, como las regiones montañosas (Shafer *et al.* 2011).

Por tanto, la determinación de los patrones filogeográficos de *Galemys pyrenaicus* puede permitir establecer si es la historia evolutiva reciente o, por el contrario, las barreras actuales asociadas a los requerimientos ecológicos los factores más determinantes en la estructuración de la variabilidad génetica de esta especie endémica de la Península Ibérica.

3.2 *Galemys pyrenaicus* (E. Geoffroy St. Hilaire, 1811)

El desmán ibérico (*Galemys pyrenaicus*) es un mamífero perteneciente a la familia de los tálpidos, englobada dentro del orden Eulipotyphla (tradicionalmente denominado Insectivora). Dentro de la familia Talpinae, se ubica en la subfamilia Desmaninae, constituyendo, junto con el desmán ruso (*Desmana moschata*), los dos únicos representantes actuales de este linaje, que durante el Mioceno, Plioceno y Pleistoceno constaba de tres géneros (*Archaeodesmana* además de los ya mencionados *Galemys* y *Desmana*) con varias especies descritas en cada uno (McKenna *et al.* 1997). La monofilia de los desmaninos está bien soportada (Cabria *et al.* 2006) y el registro fósil más antiguo para este grupo de organismos está datado hace 8.2 millones de años (Fortelius 2012).

El desmán ibérico (al igual que el desmán ruso) es un animal de hábitos semiacuáticos, en contraste con las otras dos subfamilias de tálpidos, que ocupan otros nichos ecológicos (la mayoría son de tipo fosorial). La adaptación especializada al medio acuático comporta la presencia de estructuras muy particulares en la anatomía del desmán. Así, presenta una trompa alargada que emplea para capturar presas en el lecho del río y sus extremidades posteriores son de gran tamaño, así como la cola, que es de gran tamaño y recubierta por pelos duros, lo que facilita el desplazamiento dentro del agua (Palmeirim & Hoffmann 1983; Richard 1985).

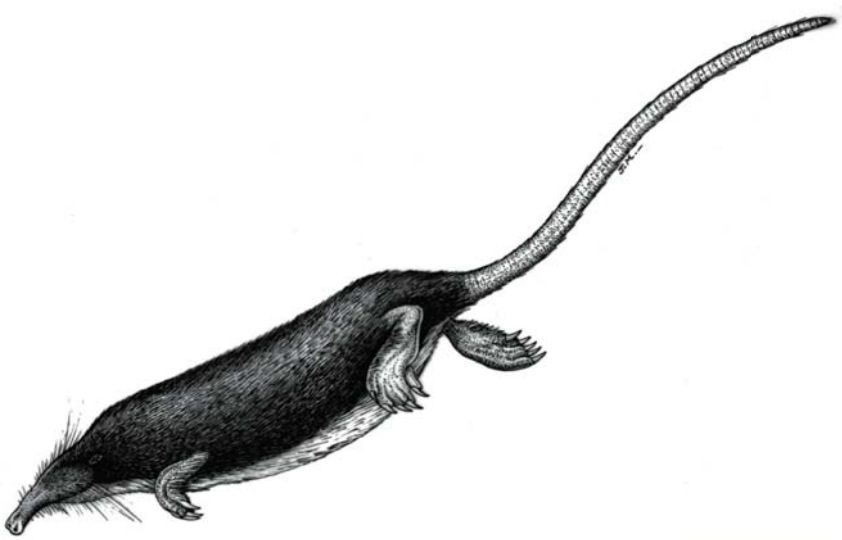

Figura 4. *Galemys pyrenaicus* (tomado del Atlas y Libro Rojo de los Mamíferos de España)

Normalmente ocupa ríos de montaña, si bien puede hallarse desde el nivel del mar hasta altitudes superiores a los 2000 metros. Más que la altitud, por tanto, parecen ser factores determinantes la pendiente y la profundidad del río (que no ha de ser excesiva) y, sobre todo, un régimen pluvial que permita el mantenimiento de un flujo regular de agua sin acusadas variaciones estacionales (Aymerich & Gosálbez 2002). En lo referente a su alimentación, se ha descrito que está casi exclusivamente basada en larvas de macroinvertebrados acuáticos de tres órdenes: Tricoptera, Diptera y Ephemeroptera (Castién 1994). Por tanto, la distribución del desmán está fuertemente ligada a la presencia de estos organismos y a los factores que les puedan afectar.

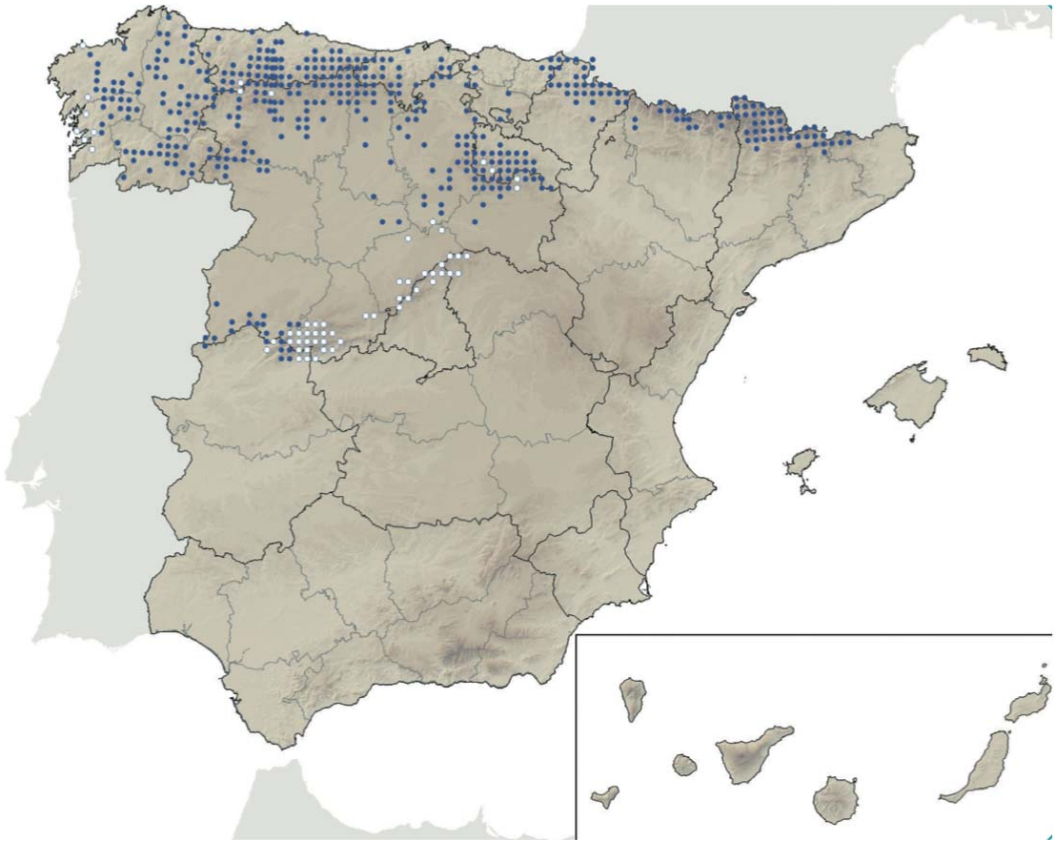

Figura 5. Mapa de distribución de *Galemys pyrenaicus* según el Atlas y Libro Rojo de los mamíferos terrestres de España (Ministerio de Medio Ambiente). Los puntos azules corresponden con los datos del Atlas actual y los blancos con puntos con presencia en el anterior Atlas (2002) pero no en el actual. No se muestran datos de Francia y Portugal

Galemys pyrenaicus es una especie endémica del norte de la Península Ibérica. Ocupa una franja que se extiende, de forma fragmentada, desde los Pirineos, pasando por todo el Arco Cantábrico y la Costa Atlántica de Galicia hasta el norte de Portugal. Además, también se encuentra en el Sistema Central y el Sistema Ibérico, completándose así una distribución de forma circular, ya que se encuentra ausente de la zona central de la Meseta (figura 5). Además, *Galemys pyrenaicus* es una especie polimórfica: actualmente se reconocen dos subespecies, *Galemys pyrenaicus pyrenaicus* y *G. pyrenaicus rufulus*, que presentan diferencias a nivel de tamaño y coloración del pelaje, si bien existe cierta controversia respecto a la magnitud de estas diferencias y la asignación de las distintas poblaciones a cada una de las subespecies. (Juckwer 1990; González-Esteban *et al.* 1999; López-Fuster *et al.* 2006).

El desmán ibérico está actualmente catalogado como "Vulnerable" en la Lista Roja de la IUCN (Fernandes *et al.* 2011), ya que parece estar sufriendo una regresión generalizada en todas sus poblaciones, creando una distribución fragmentada. La situación se ha agravado particularmente en las áreas de clima más mediterráneo, hasta el punto de que actualmente está catalogado en el Sistema Central como "en peligro de extinción" por el gobierno español. Probablemente, la causa fundamental de esta reducción hay que buscarla en el deterioro de su hábitat. Como ya se ha comentado, se trata de una especie con unos requerimientos ecológicos restrictivos y, por tanto, las perturbaciones que afectan a su hábitat pueden tener efectos muy importantes sobre su viabilidad a largo plazo. En este sentido, se han citado como factores amenazantes sobre la conservación del desmán ibérico la contaminación, que afecta directamente a los macroinvertebrados acuáticos (las principales presas del desmán), la modificación de los cursos y niveles de agua de los ríos por canalizaciones y presas, las alteraciones sobre los márgenes fluviales, como destrucción de la vegetación de ribera, que disminuyen el número de refugios para la especie e incluso la presencia de depredadores como la nutria (Nores *et al.* 2007).

3.3 El género *Neomys* Kaup 1829 como modelo para estudiar especiación y divergencia poblacional

Como ya se ha comentado, la estima de árboles de especies usando múltiples genealogías independientes e incorporando las predicciones de la teoría de la coalescencia permiten obtener medidas más precisas de los tiempos de especiación. La incorporación de estas metodologías en las estimas de tiempos de divergencia se hacen más necesarias cuanto más cercana al presente ha tenido lugar el fenómeno de especiación, puesto que es entonces cuando se intensifica el efecto de la estocasticidad asociada a la varianza en los tiempos de coalescencia de los linajes (Sanchez-Gracia & Castresana 2012). Por tanto, el riesgo de obtener estimas erróneas de tiempo basadas en un solo árbol de genes es mayor.

El género *Neomys* comprende tres especies de musarañas acuáticas de distribución paleártica y con divergencias estimadas de unos pocos millones de años. Además, en el caso de *Neomys anomalus*, existen dos subespecies, *N. anomalus anomalus* (restringida a la Península Ibérica) y *N. anomalus milleri* (en el resto del continente europeo), inicialmente descritas como dos especies distintas, *Neomys anomalus* y *N. milleri*, y posteriormente sinonimizadas.

Las especies y subespecies de este género, debido a su divergencia reciente (fechada en el Plioceno-Pleistoceno), constituyen un buen modelo para estudiar patrones de especiación y de divergencia poblacional empleando la información procedente de varios marcadores, tanto nucleares como mitocondriales e integrándola en el contexto de la teoría de la coalescencia para obtener dataciones precisas de las distintas fechas de divergencia dentro del género.

3.4 Los musgaños del género *Neomys*

El género *Neomys* se encuentra dentro de la familia Soricidae, en el orden Eulipotyphla. Esta familia comprende actualmente 385 especies de musarañas, lo que la convierte en una de las de mayor riqueza específica de los mamíferos (Wilson & Reeder 2011). Actualmente se aceptan tres grandes subfamilias dentro de los sorícidos: Myosoricinae (musarañas africanas), Crocidurinae (musarañas de dientes blancos) y Soricinae (musarañas de dientes rojos), a la que pertenecen los musgaños.

El género *Neomys* comprende tres especies de musarañas acuáticas (figura 6):

- *Neomys fodiens* (Pennant, 1771) (musgaño patiblanco), que tiene una amplia distribución que ocupa la mayor parte de la Europa continental y Gran Bretaña y se extiende hasta Asia Central.
- *Neomys anomalus* Cabrera 1907 (musgaño de Cabrera), con una distribución bastante más dispersa pero igualmente extensa y coincidente en muchas zonas con la de *N. fodiens* y que va desde la Península Ibérica hasta el Sudoeste Asiático, si bien está ausente del Norte de Europa.
- *Neomys teres* Miller 1908 (musgaño transcaucásico), de distribución mucho más limitada, que habita en el Caúcaso (Armenia, Georgia) y Asia Menor.

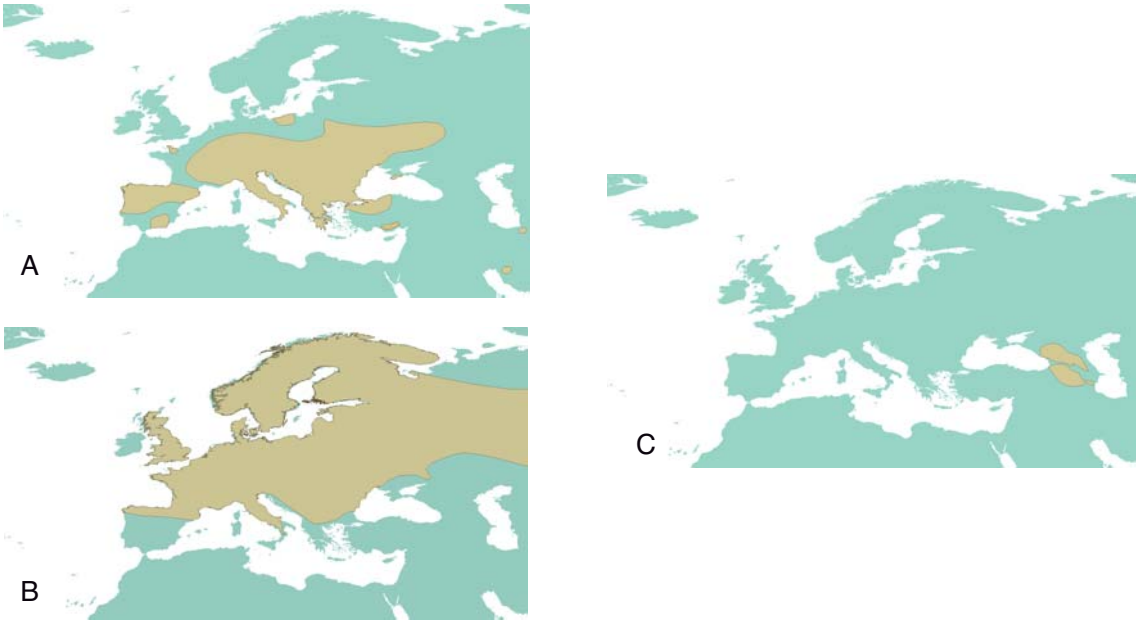

Figura 6. Mapas de distribución europea según la Unión Internacional para la Conservación de la Naturaleza (IUCN) de (A) *Neomys anomalus* (B) *N. fodiens* y (C) *N. teres*.

Las especies del género *Neomys* son los únicos sorícidos adaptados al medio acuático del continente europeo. Esta adaptación al medio acuático, aunque no es tan acusada como la ya descrita para el desman ibérico, se refleja en su morfología, con unas

extremidades posteriores de mayor tamaño que otras musarañas terrestres, que además están recubiertas de pelos rígidos que facilitan el desplazamiento dentro del agua. Asimismo, la cola también está recubierta de pelos que actuarían a modo de quilla, para proporcionar estabilidad durante el nado. Por otra parte, los lóbulos olfativos del cerebro de los musgaños han experimentado una reducción considerable respecto al resto de sorícidos terrestres debido a que la detección de presas no se realiza mediante el olfato sino mediante el tacto (Krystufek *et al.* 2000).

Este trabajo se centra en *Neomys anomalus* y *N. fodiens (*figura 7). Estas dos especies presentan leves diferencias a nivel morfológico: en general, el musgaño patiblanco es más grande y posee una serie de pelos blancos recubriendo la cola. La coincidencia en muchas áreas europeas de la distribución de ambas especies supone que, a menudo, ambas especies se dan en simpatría. Se ha postulado la ocurrencia de un desplazamiento de caracteres que evitaría la competencia (Kryštufek & Quadracci 2008). Así *Neomys anomalus*, más adaptada al comportamiento terrestre, pasa a emplear más este tipo de recursos, quedando el medio acuático más disponible para *Neomys fodiens*, que está más adaptado a él (Mendes-Soares & Rychlik 2009).

No ha sido hasta la realización de estudios con datos moleculares (Krystufek *et al.* 2000) (Castiglia 2007) cuando se han establecido las relaciones filogenéticas entre las tres especies del género *Neomys*. Hasta entonces, la ausencia de grandes diferencias morfológicas, siendo la más destacable de ellas las diferencias en la morfología peneana (Pucek 1964; Krystufek *et al.* 2000), dificultaban el establecimiento de estas relaciones, y se consideraba *Neomys anomalus*, por ser la especie menos adaptada al medio acuático, como externa al clado formado por *N. fodiens* y *N. teres* (Krystufek *et al.* 2000). Sin embargo, la filogenia molecular de Castiglia *et al.*, realizada con el citocromo *b*, postula que el linaje de *Neomys fodiens* se separó hace unos 7 millones de años del clado formado por *N. anomalus* y *N. teres*, que a su vez divergieron hace aproximadamente 1 millón de años. Cabe resaltar que estas estimas de divergencia fueron realizadas en base a un solo gen (el del citocromo *b*) y, además, sin tener en cuenta el efecto de la estocasticidad asociado a la coalescencia de los linajes que, como ya se ha comentado, puede tener una gran influencia en este tipo de cálculos, provocando notables sobreestimas de los tiempos de especiación.

Figura 7. A) *Neomys anomalus*. B) *Neomys fodiens*. Se puede apreciar el mayor recubrimiento de pelos en la cola de *N. fodiens*. (modificado de MacDonald & Barrett 2008).

II.-OBJETIVOS

Las filogenias que incorporan información procedente de distintos *loci* del genoma son fundamentales para resolver las relaciones evolutivas de los seres vivos, así como para determinar con precisión los tiempos de divergencia de las distintas especies y poblaciones y estudiar fenómenos como el flujo genético. Para el caso de los mamíferos, se dispone de una serie de marcadores moleculares que se emplean de forma rutinaria en estudios filogenéticos, pero ninguno de ellos son producto de un estudio exhaustivo que seleccione los más adecuados mediante un análisis de genomas completos.

El objetivo principal de esta tesis doctoral ha sido analizar varios genomas de mamíferos y seleccionar, de acuerdo a varios criterios, un conjunto de intrones para poder ser empleados como marcadores filogenéticos de especies cercanas y, después, emplear varios de estos intrones para estudiar y caracterizar la variabilidad genética y la historia evolutiva de varias especies de mamíferos.

De forma más concreta, se pueden describir tres objetivos:

1. Desarrollar, mediante el análisis de los intrones de genomas completos, un nuevo conjunto de marcadores moleculares útiles para la filogenia de especies cercanas de mamíferos mediante la aplicación de una serie de filtros que garanticen que tienen las características adecuadas (tamaño, copia única, evolución ajustada a la del resto del genoma, alta divergencia). Además, la amplificación y secuenciación de varios de estos intrones en un panel de varias especies de mamíferos que incluían varias parejas de especies congenéricas permitiría probar su utilidad como marcadores para estudios filogenéticos de especies cercanas. Este estudio supondría añadir una importante cantidad de nuevos marcadores moleculares para su uso en posteriores estudios filogenéticos de mamíferos.

2. Determinar la estructura y variabilidad genética del desmán ibérico, *Galemys pyrenaicus*, así como caracterizar su historia evolutiva reciente, empleando para ello secuencias mitocondriales y de varios intrones. El desmán ibérico es una especie endémica y amenazada de la Península Ibérica, de la que no existen apenas datos genéticos y, en este sentido, este estudio, realizado en gran parte con muestreo no invasivo, aportaría información esencial de cara a una gestión y conservación adecuadas. Además, y por tratarse de una especie con estrictos requerimientos acuáticos, puede resultar un buen modelo para estudiar la influencia de estos requerimientos sobre la estructuración de la variabilidad

genética de las especies, y compararlos con la influencia ejercida por la historia evolutiva reciente condicionada por otros factores como las glaciaciones pleistocénicas.

3. Realizar un estudio de los tiempos de divergencia y especiación dentro del género europeo de musgaños *Neomys* usando información procedente de intrones distribuidos por todo el genoma para estimar, incorporando las predicciones de la teoría de la coalescencia, el árbol de especies de este género a partir de los correspondientes árboles de genes, obteniendose cálculos precisos para los tiempos de separación de los distintos grupos de organismos. Asimismo, la incorporación de modelos de aislamiento con migración a este sistema permitiría evaluar el grado de flujo genético existente entre diversas poblaciones diferenciadas del género. Por último, un estudio comparativo de varias estrategias de cálculo de tasas evolutivas para su uso como *priors* en análisis bayesianos permitiría evaluar la influencia de éstas sobre los cálculos de tiempos de divergencia en los árboles de especies y sugerir las estrategias más adecuadas de cara a obtener las estimas más precisas y menos sesgadas.

III.-MATERIALES Y MÉTODOS

1. DESARROLLO DE NUEVOS MARCADORES GENÓMICOS

1.1 Protocolo de extracción y filtrado de intrones de mamíferos

1.1.1 Extracción de intrones de los genomas

En primer lugar se descargaron de la base de datos ENSEMBL (Hubbard *et al.* 2007) los genomas, en formato GenBank, de las siguientes especies de mamíferos: humano (*Homo sapiens*) en la versión del genoma NCBI 36 (Lander *et al.* 2001) (Venter *et al.* 2001); chimpancé (*Pan troglodytes*) versión Pan_troglodytes-2.1 (Mikkelsen *et al.* 2005); macaco Rhesus (*Macaca mulatta*) versión Mmul_1 (Gibbs *et al.* 2007); perro (*Canis familiaris*) versión CanFam 2.0 (Lindblad-Toh *et al.* 2005); vaca (*Bos* taurus) versión Btau_3.1 (Elsik *et al.* 2009); ratón (*Mus musculus)* versión NCBI m37 (Waterston *et al.* 2002) y rata (*Rattus norvegicus*) (Gibbs *et al.* 2004). Todos estos genomas estaban anotados con detalle y secuenciados con una gran cobertura. Además, había disponibles otros genomas correspondientes a otras especies de mamíferos, pero debido a que estaban secuenciados con una cobertura menor (2x) no fueron incluidos en este estudio. Estos genomas tienen, además, una menor densidad de anotaciones y en muchas ocasiones presentan los denominados intrones *frame shift,* que son artificialmente introducidos durante el proceso de anotación respecto al genoma de referencia, de mayor calidad de anotación, que se emplea como guía en el proceso (Hubbard *et al.* 2007; Green 2007). Por otro lado, también estaba disponible el genoma del marsupial *Monodelphis domestica* (Mikkelsen *et al.* 2007), que sí había sido secuenciado con una cobertura suficiente para este estudio. Sin embargo, no fue empleado en los análisis, ya que se trata de una especie muy divergente respecto al resto de mamíferos presentes en este estudio, y su inclusión podría causar problemas en los procesos de reconstrucción filogenética, particularmente a la hora de alinear las secuencias.

Se llevó a cabo un análisis preliminar con las siete especies y se confirmó, como ya ha sido previamente demostrado, que los roedores mostraban unas tasas de evolución mucho más altas que las del resto de las especies empleadas (Waterston *et al.* 2002; Gibbs *et al.* 2004). Por tanto, se decidió no emplear para este estudio los genomas de la rata y el ratón.

A continuación se desarrolló un conjunto de programas de Perl con el objetivo de extraer todos los intrones y exones de cada uno de los cinco genomas (humano, chimpancé, macaco Rhesus, vaca y perro) en formato GenBank. Asimismo, se almacenó

información relativa a la descripción, a la localización genómica y a la longitud de cada uno de los intrones y exones. En esta fase preliminar se descartaron todos los intrones con una longitud superior a 50000 nucleótidos.

1.1.2 Filtrado de los intrones

En primer lugar, se determinaron las relaciones de ortología existentes entre los intrones de las cinco especies de mamíferos empleadas. Para ello, se obtuvo la información relativa a la ortología almacenada en la base de datos ENSEMBL. Esta base de datos emplea análisis filogenéticos para determinar las relaciones de ortología, siendo dicha metodología más precisa que las más tradicionales basadas en las semejanzas de la secuencia (usando, por ejemplo, búsquedas BLAST) (Hubbard *et al.* 2007). Así pues, se descargaron de ENSEMBL una serie de tablas de ortología, que son listas de genes ortólogos entre un par de especies determinado, correspondientes a las parejas de las especies de interés. Cruzando las tablas de ortología de humano-chimpancé y humano-macaco Rhesus, se obtuvo una tabla de genes ortólogos uno-uno para los primates. A partir de la tabla de perro y vaca, se obtuvo la correspondiente a los laurasiatéridos, que a su vez fue cruzada con la ya obtenida de los primates, resultando en una tabla final de genes ortólogos uno-uno para las cinco especies consideradas. A partir de esta tabla de genes, se calculó la correspondiente a los intrones, incluyéndose sólo aquellos intrones pertenecientes a genes con igual número de intrones en todas las especies comparadas. También se comprobó que la posición relativa de cada uno de los intrones dentro del gen al que pertenecían estaba conservada a lo largo de las cinco especies, y se eliminaron los casos en los que esto no ocurría.

En segundo lugar, se limitó el tamaño de los intrones seleccionados para hacerlos fácilmente amplificables en el laboratorio. Se seleccionaron, por tanto, sólo aquellos intrones cuya longitud en *Homo sapiens* fuera de entre 200 y 1600 nucleótidos. La variación en el tamaño de cada intrón a lo largo de la filogenia también fue controlada: el máximo porcentaje de diferencia en longitud del intrón y su ortólogo fue fijado en un 10% para la pareja humano-chimpancé; 20% para humano-macaco Rhesus; 50% para la pareja vaca-chimpancé; y 100% para las parejas humano-vaca y humano-perro. Después de este paso, se creó un único archivo con las cinco secuencias de los intrones ortólogos en las cinco especies.

A continuación se construyeron los alineamientos múltiples de las secuencias intrónicas ortólogas usando el programa MAFFT versión 5.8 (Katoh *et al.* 2005), con

los parámetros por defecto. Después los alineamientos fueron tratados con Gblocks versión 0.91 (Castresana 2000), usando parámetros relajados (Talavera & Castresana 2007) para así descartar posibles posiciones mal alineadas que aumentaran artificialmente la divergencia entre las especies. Por último se reconstruyeron los árboles filogenéticos por máxima verosimilitud usando el programa PhyML versión 2.4.4 (Guindon & Gascuel 2003), con el modelo de sustitución GTR y cuatro categorías de tasa de sustitución.

Después se eliminaron aquellos intrones que estaban flanqueados por exones de menos de 40 nucleótidos de longitud, ya que se consideró éste como el mínimo tamaño requerido para poder diseñar un cebador en un exón.

El siguiente filtro que se aplicó fue el destinado a eliminar secuencias duplicadas a lo largo del genoma que puedan producir amplificaciones múltiples en la reacción de PCR. Para detectar estas duplicaciones, se realizó una búsqueda BLAST (Altschul *et al.* 1997) con las parejas de exones flanqueantes a cada intrón. Para cada especie, se lanzó la búsqueda de los exones contra su propio genoma y se seleccionaron sólo aquellos intrones flanqueados por exones que sólo producían un resultado (el propio exón) en una única región del genoma. Además, y debido a que en el formato de anotación de los genomas éstos están normalmente subdivididos en grandes regiones (usualmente cromosomas), se comprobó también que dentro de cada región de resultado único había ocurrido un único resultado positivo. Se emplearon los exones en lugar de los intrones para la búsqueda debido a que los cebadores se diseñaran en los exones, y además las posibles duplicaciones de intrones serían menos fácilmente detectables por búsquedas BLAST al estar menos sujetas a constreñimientos selectivos. El límite de *valor e* de la búsqueda BLAST se fijó en 10^{-4}.

Posteriormente, el objetivo era seleccionar los intrones con una evolución más similar a la global del genoma, que no mostraran aceleraciones o deceleraciones en la tasa de evolución de algún linaje y que mostraran, por tanto, un reloj molecular lo más neutro posible. Para ello, en primer lugar, se construyó un árbol de referencia que reflejara la evolución genómica global usando todos los intrones ortólogos de copia única, que resultaron ser un total de 1344 intrones. El alineamiento múltiple resultante del concatenado de estos intrones, de 768745 posiciones, fue usado para reconstruir, por máxima verosimilitud y utilizando el programa RAxML (Stamatakis 2006), el árbol filogenético de referencia. Éste se usó a continuación como base para establecer comparaciones con cada uno de los árboles filogenéticos de los intrones que queríamos

filtrar en esta última fase. Para ello se empleó la medida *K tree score* calculada por el software Ktreedist (Soria-Carrasco *et al.* 2007). En primer lugar, Ktreedist escala el árbol a comparar multiplicándolo por un factor (*scaling factor*) para que tenga una divergencia similar a la del árbol de referencia. Después calcula el *K tree score*, que refleja las diferencias topológicas y de longitud de ramas relativas del árbol a comparar frente al árbol de referencia (es decir, cuanto mayor es el valor de *K tree score*, mayores son las diferencias entre las longitudes de ramas relativas entre los dos árboles). Por consiguiente, se descartaron los intrones con árboles con valores altos de *K tree score*, ya que podría tratarse de casos de paralogía que los filtros anteriores no hubieran detectado o de intrones con tasas evolutivas muy distintas a la global del genoma en ciertas partes de la filogenia.

Por último, se aplicó un filtro para eliminar los intrones más conservados. Para ello se emplearon dos medidas de divergencia: el *scaling factor* calculado por el Ktreedist (que es una medida de cómo de grande es un árbol de un intrón respecto al árbol de referencia del genoma, y está por tanto muy correlacionada con la longitud total de ramas del árbol y la divergencia global) y la longitud total de ramas del árbol de primates (definida como la suma de todas las ramas del árbol del intrón correspondiente reconstruido sólo con las secuencias de humano, chimpancé y macaco Rhesus). Esta última medida está menos afectada por posibles defectos en el proceso de alineamiento de todas las secuencias, que podrían causar una sobreestimación de la distancias genéticas.

Seguidamente, se llevó a cabo una inspección visual de todos los intrones resultantes, para así poder detectar cualquier posible secuencia con alineamientos defectuosos que hubiera podido permanecer a pesar de todos los filtros aplicados.

El resultado final de estos procesos de filtrado fueron 224 intrones que pueden ser usados en la filogenia de especies cercanas. Para caracterizar estos intrones, se procedió a analizar la presencia de secuencias repetitivas usando el software RepeatMasker (Smit AFA. 2004), que incluye librerías específicas de elementos repetitivos de mamíferos. Por otro lado, se anotaron los polimorfismos de nucleótido simple (SNP) descritos para el genoma humano en cada uno de los 224 intrones, con el fin de caracterizar su variabilidad intraespecífica. Para ello, se descargó de la base de datos de ENSEMBL, usando la plataforma Biomart (Smedley *et al.* 2009), el listado completo de los SNPs descritos en intrones del genoma humano en su versión NCBI 36, y se mapearon sobre los intrones del conjunto final de este estudio. También se analizó

la presencia de microsatélites en los intrones de las cinco especies estudiadas mediante el programa Tandem Repeats Finder versión 4.04 (Benson 1999), con los siguientes parámetros: 2 7 7 80 10 50 6.

Por último, se generó un fichero PDF para los 224 intrones en el que se incluía toda la información relevante obtenida como los alineamientos del intrón y los exones flanqueantes, el árbol filogenético del intrón, la localización genómica, la descripción de la función que realiza el gen al que pertenece el intrón y la divergencia calculada, entre otras medidas.

1.2 Validación experimental de los nuevos marcadores moleculares

1.2.1 Diseño de los cebadores y especies empleadas

Con el objetivo de comprobar en el laboratorio si el conjunto de marcadores desarrollados *in silico* eran realmente útiles para su aplicación en filogenias de especies cercanas, se seleccionaron por inspección visual 12 de los 224 intrones para posteriormente secuenciarlos en varias parejas de especies de mamíferos.

A partir de los exones flanqueantes, se diseñaron cebadores de PCR degenerados que permitieran la amplificación en el mayor número de especies posible. Para ello se construyó un alineamiento con las secuencias correspondientes a los genomas ya analizados en este estudio (humano, chimpancé, macaco, vaca y perro), a los que se añadieron (cuando fue posible) otras secuencias, también obtenidas de la base de datos ENSEMBL, y correspondientes a otras especies de mamíferos cuyos genomas estaban secuenciados con una cobertura inferior (pero en muchos casos contenían información útil relativa a los exones). Estas especies fueron: el caballo (*Equus caballus*), el orangután de Sumatra (*Pongo abelii*), el murciélago *Myotis lucifugus*, el erizo europeo (*Erinaceus europaeus*), el gato doméstico (*Felis catus*), la tupaya de Belanger (*Tupaia belangeri*), el lémur naranja (*Microcebus murinus*), el elefante africano (*Loxodonta africana*), el tenrec (*Echinops telfairi*), el armadilllo de nueve bandas (*Dasypus novemcinctus*) y la musaraña común (*Sorex araneus*). Por tanto, se construyeron alineamientos múltiples de los exones que contenían secuencias, según el caso, de entre 11 y 15 especies. Se diseñaron los cebadores correspondientes, colocando siempre la parte 3' del cebador en la región más conservada del exón, para favorecer la amplificación interespecífica. Además se incluyeron degeneraciones, hasta un total de 48, para hacer que los cebadores fueran válidos para el mayor número de especies

posible. Para testar estos cebadores, se obtuvieron muestras biológicas de seis especies de mamíferos, pertenecientes a diferentes órdenes y que, conjuntamente con las ya descargadas de las bases de datos, permitieron disponer de varias parejas de especies cercanas en las que estudiar la variabilidad de los intrones. Las seis especies elegidas fueron: el murciélago ratonero gris (*Myotis escalerai*), el orangután de Borneo (*Pongo pygmaeus*), el leopardo de las nieves (*Uncia uncia*), el tigre (*Panthera tigris*), la comadreja común (*Mustela nivalis*) y el turón europeo (*Mustela putorius*).

1.2.2 Extracciones de ADN genómico a partir de tejido fresco

El ADN genómico total fue extraído a partir de fragmentos de tejido (cola, músculo, hígado, pelos) de un peso siempre inferior a 25 mg. Se empleó el kit DNEasy Blood and Tissue Kit (QIAGEN), siguiendo las instrucciones del fabricante. Las eluciones finales del ADN se hicieron en un volumen total de 75 μl de agua, y en ocasiones se obtuvo una segunda elución, más diluida, en un volumen de 100 μl de agua. La concentración de ADN de cada muestra se analizó mediante el ND-1000 (Nanodrop Technologies) midiendo la absorbancia a 260 nm.

1.2.3 Amplificación por PCR de intrones

Las reacciones de PCR fueron preparadas en una sala especialmente dedicada y controlada, aislada de los lugares en los que se manipulan productos de PCR e irradiada regularmente con luz ultravioleta. Asimismo, el instrumental de laboratorio empleado en la preparación fue también irradiado con luz ultravioleta para eliminar cualquier resto posible de ácidos nucleicos. El volumen final de las reacciones fue de 25 μl, y se añadieron entre 50 y 100 ng de ADN. La concentración final de los cebadores varió, según cada marcador amplificado, entre 0.5 y 1 μM y se emplearon 0.75 unidades de ADN polimerasa Promega GoTaq. La reacción de PCR consistió, típicamente, en una fase inicial de desnaturalización a 95 ºC durante 3', seguida de entre 27 y 35 ciclos compuestos de una fase inicial a 95 ºC durante 30 segundos, una fase de alineamiento a temperatura variable (según la pareja de cebadores usada, ver tabla A2), y una fase de extensión a 72 ºC durante 30-60 segundos (según la longitud del marcador a amplificar). Los productos de PCR resultantes se cargaron para su comprobación en geles de agarosa al 1% preteñidos con SYBR Safe DNA Gel Stain (Invitrogen). A continuación se purificaron las reacciones exitosas incubándolas con Exo-SapIT (Affymetrix). La secuenciación de ADN de los productos de PCR se llevó a cabo en distintos servicios de secuenciación, usando los cebadores de PCR y el kit BigDye v 3.1.

Las secuencias fueron ensambladas e inspeccionadas posteriormente empleando el software Geneious v.4.0 (Drummond AJ 2009).

1.2.4 Clonación de productos de PCR

Se observaron, en varias secuencias, tanto mutaciones puntuales (reflejadas como dobles picos en los cromatogramas de las secuencias de ADN) como inserciones o deleciones ("indels") que dificultaban asignar una secuencia clara a cada uno de los alelos de los individuos secuenciados. Por tanto, se subclonaron en vectores plasmídicos los productos de PCR correspondientes. La ligación se llevó a cabo con el vector pSTBlue-1 (Invitrogen), siguiendo las instrucciones del fabricante. La transformación se realizó en células competentes *Nova Blue* de *Escherichia coli* mediante un choque térmico. Las células transformadas se seleccionaron posteriormente en placas de medio LB con carbenicilina y, mediante selección por color usando IPTG y X-gal, se distinguieron las colonias con vectores portadores del inserto. De cara a su posterior secuenciación, el ADN plasmídico correspondiente se purificó usando el kit GenElute Plasmid Miniprep Kit (Sigma Aldrich), siguiendo las instrucciones del fabricante, a partir de cultivos líquidos de LB- carbenicilina incubados durante 18 horas a 37ºC en agitación. Los posibles errores inducidos por la polimerasa, y que se ponen de manifiesto durante el proceso de clonación, fueron detectados mediante la secuenciación de varias colonias de transformantes por cada producto de PCR subclonado y la posterior comparación de las secuencias resultantes.

1.2.5 Análisis filogenéticos de los nuevos marcadores

Para cada intrón amplificado y secuenciado con éxito en el panel de especies de mamíferos, se creó un alineamiento múltiple incluyendo las seis secuencias obtenidas en laboratorio y las correspondientes descargadas de las bases de datos, hasta un total de 14 especies. Se usó MAFFT para obtener estos alineamientos, y como ya se ha descrito, se eliminaron las posibles posiciones mal alineadas aplicando Gblocks con parámetros relajados. Mediante PhyML, se obtuvieron los árboles filogenéticos de máxima verosimilitud, seleccionando como modelo de evolución GTR y cuatro categorías de tasas de sustitución. Además, para analizar con mayor precisión las diferencias acumuladas entre cinco parejas de especies cercanas, se construyeron nuevos alineamientos de pares que únicamente incluían las dos secuencias correspondientes a cada miembro de la pareja de especies.

2. FILOGEOGRAFÍA DEL DESMÁN IBÉRICO EMPLEANDO SECUENCIAS MITOCONDRIALES E INTRÓNICAS

2.1 Obtención de muestras

2.1.1 Prospección de excrementos de desmán

Se realizaron prospecciones de excrementos de *Galemys pyrenaicus* en tramos fluviales de la Península Ibérica. La particular morfología de las excrementos de esta especie, así como su localización (en piedras situadas en el centro del curso del río y no en los márgenes) permite, con unos ciertos conocimientos previos, maximizar la proporción de excrementos pertenecientes a esta especie sobre el total recogido. Los excrementos fueron en cada caso recogidos con pinzas de laboratorio y depositados individualmente en tubos con etanol de 96º para favorecer su preservación y minimizar la degradación del ADN. Además se anotaron las coordenadas GPS correspondientes a cada muestra. Para evitar obtener muestras múltiples de un mismo individuo, solo se consideraron para los estudios posteriores las secuencias procedentes de excrementos separados por más de un kilómetro de distancia, lo que supone más del doble del rango territorial del desmán ibérico (Stone 1987; Melero *et al.* 2012). Las muestras separadas por una distancia menor a ese valor, pero con secuencias de ADN diferentes (y por tanto pertenecientes a distintos individuos), sí fueron tenidas en cuenta posteriormente. Así, se emplearon un total de 69 muestras de excrementos para este estudio.

2.1.2 Muestras de tejido fresco

Se obtuvieron muestras de tejido fresco de 63 ejemplares procedentes de campañas de trampeo correspondientes a estudios previos (en las que se tomaron muestras de pelo o una pequeña sección de la piel de la parte distal de la cola) o de ejemplares bien conservados de diferentes colecciones biológicas.

2.1.3 Muestras de colecciones de museo

Finalmente, se completó el conjunto de muestras con dos fragmentos de hueso (una uña y una costilla) donados por la colección de la Estación Biológica de Doñana.

2.2 Extracciones de ADN genómico y PCR

2.2.1 Extracciones de ADN genómico a partir de tejido fresco

Las extracciones de ADN de tejido fresco se realizaron como ya se ha descrito anteriormente (apartado 1.2.2), con la excepción de las extracciones realizadas a partir de pelos, en las que la lisis del tejido empleando proteinasa K se llevó a cabo durante 16 horas y empleando el doble de la cantidad indicada en el protocolo (40 µl de una solución 20 mg/ml).

2.2.2 Extracciones de ADN genómico a partir de excrementos

Las extracciones se realizaron usando el kit QIAamp DNA Stool Kit (QIAGEN), diseñado especialmente para tratar con este tipo de muestras, maximizando la cantidad de ADN endógeno obtenido y al mismo tiempo, minimizando la concentración de compuestos presentes en las heces que pueden inhibir la reacción de PCR. El protocolo se llevó a cabo siguiendo las instrucciones del fabricante, y el volumen final de las eluciones fue de 50 µl. Los extractos de ADN resultantes de este tipo de muestras contienen ADN de menor concentración y más degradado que los obtenidos a partir de tejido fresco, debido a la degradación química a la que es sometido el ADN durante la digestión y la excreción por parte del animal y hasta que es recogido el excremento y conservado adecuadamente (Nsubuga *et al.* 2004). Debido a esto, el riesgo de contaminación y la necesidad de un mayor control de la esterilidad es especialmente importante (Taberlet *et al.* 1999) y, por ello, todas las extracciones de muestras fecales se llevaron a cabo en un espacio irradiado con luz ultravioleta y separado físicamente de posibles focos de ADN contaminante, y empleando material de laboratorio de uso exclusivo para este tipo de extracciones.

2.2.3 Extracción de ADN genómico a partir de muestras de museos

La extracción de material genético de las muestras procedentes de la Colección de Tejidos de la Estación Biológica de Doñana se realizó en un laboratorio de ADN antiguo, donde nunca antes se había extraído ADN de desmán, usando los protocolos y precauciones propias (Gilbert *et al.* 2005). Las muestras de hueso (un fragmento de uña y un fragmento de costilla de dos ejemplares obtenidos en los años 1980 y 1973, respectivamente) fueron pulverizadas y a continuación decalcificadas incubándolas durante una noche a 37 ºC con agitación en una solución de EDTA 10 M. Después se procedió a la incubación, de nuevo durante toda la noche, con una solución de lisis con proteinasa K y SDS. Por último, el ADN fue extraído mediante un procedimiento estándar de fenol – cloroformo y concentrado finalmente usando columnas centricon.

2.2.4 Amplificación por PCR de secuencias mitocondriales

Las reacciones de PCR se realizaron como se ha descrito en el apartado 1.2.3. Debido a la degradación del ADN extraído a partir de las muestras de excrementos, la amplificación del citocromo *b* se llevó a cabo usando tres fragmentos solapantes de 483, 278 y 516 pares de bases en el caso de *Galemys pyrenaicus*. Para ello, se diseñaron cebadores internos a partir de las correspondientes secuencias que se obtuvieron de GenBank (tabla A1). Además, se obtuvo la secuencia de un fragmento de 342 pares de bases de la región 5' del D-loop, de nuevo usando cebadores específicos de desmán. Se obtuvo también la secuencia del citocromo *b* de un ejemplar del desmán ruso, *Desmana moschata*. Las reacciones de amplificación a partir de excrementos incluyeron 16 μg de suero de albúmina bovina (BSA), recomendada para disminuir la acción de inhibidores presentes en los extractos de ADN de heces (Morin *et al*. 2001).

2.2.5 Amplificación por PCR a partir de ADN de museo

El ADN extraído de muestras de museo suele ser de menor concentración y suele presentar una mayor degradación que el obtenido a partir de muestras de tejido fresco (Pääbo *et al*. 2004; Wandeler *et al*. 2007). Los métodos empleados en la preservación de las muestras en las colecciones biológicas pueden, además, dificultar la reacción de PCR, lo que provoca que, en general, los productos amplificados suelan ser de pequeño tamaño (menos de 200 pares de bases).

Por estos motivos la amplificación de las secuencias del citocromo *b* y el fragmento del D-loop se llevó a cabo empleando la técnica de PCR multiplex en dos fases ("two-step multiplex PCR") (Römpler *et al*. 2006). Este protocolo está diseñado para solventar estos inconvenientes asociados al ADN de muestras de ADN antiguo y de museo, permitiendo la obtención de secuencias largas de ADN a partir de una mínima cantidad de extracto. Por consiguiente, se diseñaron dos conjuntos de cebadores independientes y no solapantes para cada una de las dos moléculas a amplificar (citocromo *b* y D-loop). En el caso del citocromo *b*, se usaron dos conjuntos (A y B), cada uno de los cuales constaba de 8 productos de PCR no solapantes. Para el fragmento del D-loop, fueron necesarios sólo 2 conjuntos A y B de 2 y 3 fragmentos, respectivamente. En la primera fase de la amplificación, todos los cebadores de cada conjunto independiente fueron utilizados en una única reacción de PCR. La concentración final de cada uno de los cebadores fue de 0.15 μM, y se usaron 5 μl de extracto de ADN y 2 unidades de promega GoTaq. Las condiciones de la reacción de PCR fueron las siguientes: una desnaturalización inicial de 9 minutos a 94 ºC, seguida de 30 ciclos consistentes en tres

fases de desnaturalización (94 ºC), hibridación (54 ºC) y extensión (72 ºC) de 30 segundos cada una. Se añadió una extensión final de 4 minutos a 72 ºC.

A continuación, se llevó a cabo la segunda fase de la amplificación, que consistió en amplificaciones "simplex" para cada uno de los fragmentos de PCR, usando como molde en la reacción una dilución 1:20 del producto de la amplificación multiplex de la primera fase correspondiente. Se siguieron las mismas condiciones de PCR que en la fase multiplex, pero en esta ocasión la concentración de cebadores fue de 1.5 μM. Las 21 reacciones de PCR resultantes (16 correspondientes al citocromo *b* y 5 al D-loop) fueron secuenciadas directamente, y las secuencias únicas de cada fragmento obtenidas fueron posteriormente ensambladas para obtener la secuencia completa correspondiente.

Además de contener ADN muy degradado y en baja concentración, los extractos de ADN procedentes de muestras de museo también han estado en muchas ocasiones sometidos a daño químico durante el proceso de preservación realizado durante la preparación de la muestra para ser incluida en la colección biológica (Wandeler *et al.* 2007). Esto puede provocar que se incorporen mutaciones inducidas a algunas de las moléculas de ADN presentes en la muestra. Para evitar que estas mutaciones artificiales fueran incluidas en la secuencia obtenida a partir de esa muestra, se obtuvieron dos réplicas independientes de estas secuencias. En el caso de la muestra IBE-C3159, la segunda réplica de las secuencias se realizó mediante subclonación en un vector plasmídico (como se ha detallado en el apartado 1.2.4), debido a algunas dificultades con la secuenciación directa de productos de PCR en el primer experimento. Por cada uno de los fragmentos se secuenciaron 3 clones con el inserto y se obtuvo la secuencia consenso correspondiente. Finalmente, para cada muestra se compararon las secuencias obtenidas en las dos réplicas independientes, y se comprobó que eran idénticas. Por tanto, no parece que se produjera daño químico de consideración durante el proceso de preservación de estas dos muestras de museo.

2.2.6 Amplificación por PCR de intrones nucleares

Se seleccionó un subconjunto de 29 muestras de tejidos que representaban todos los linajes mitocondriales y cubrían toda la distribución geográfica de la especie, y se secuenciaron 8 intrones nucleares. Los intrones amplificados fueron ACOX2-3, COPS7A-4, DHRS3-3, LANCL1-4, PRPF31-3, ROGDI-7 y SMYD4-5 del conjunto desarrollado en el capítulo X, y ACPT-4, que fue obtenido durante los procesos de filtrado preliminares de ese trabajo. Las condiciones de la reacción de PCR son

similares a las descritas en el apartado 1.2.3, siendo las condiciones particulares de temperatura las detalladas en el anexo (tabla A2). Asimismo, se amplificaron estos intrones en el desmán ruso *Desmana moschata* y en el topo ibérico *Talpa occidentalis*.

2.3 Análisis filogenéticos del desmán ibérico

2.3.1 Filogenia mitocondrial

Las secuencias de citocromo *b* y D-loop de 134 muestras de *Galemys pyrenaicus* se concatenaron para los análisis posteriores. Se evaluaron los distintos modelos de sustitución mediante el criterio de información de Akaike (AIC), con la metodología implementada en el software jModeltest (Posada 2008). El modelo que se ajustaba mejor a los datos fue el Hasegawa-Kishino-Yang (HKY), con una distribución gamma (G) y una proporción de sitios invariantes (I).

Se estimó el árbol filogenético de máxima verosimilitud correspondiente usando PhyML versión 3.0 (Guindon *et al.* 2010). A partir del árbol resultante, se derivó una genealogía de haplotipos usando Haploviewer (Salzburger *et al.* 2011). Por otro lado, también se estudiaron las relaciones filogenéticas entre las secuencias mitocondriales de *G. pyrenaicus* usando la inferencia Bayesiana, implementada en el software BEAST v.1.6.2 (Drummond & Rambaut 2007). En este caso se empleó un reloj molecular estricto (ya que no se esperan desviaciones a estas distancias evolutivas cortas), y de nuevo el modelo evolutivo HKY+G+I. Como *prior* del árbol se utilizó un modelo de coalescencia de tamaño poblacional constante, y se corrió una cadena de Markov de 50 millones de generaciones, muestreando parámetros cada 1000 generaciones. La convergencia se examinó usando el programa Tracer, y se comprobó que todos los parámetros de interés tenían un valor del tamaño efectivo de muestreo (Effective Sampling Size, ESS) mayor de 200. Se eliminó el 10% inicial de las muestras como *burn-in* y se obtuvo el árbol sumario de mayor credibilidad de clados (máximum clade credibility tree) usando la mediana de las alturas de los nodos.

2.3.2 Análisis de secuencias nucleares

Para las secuencias nucleares, en los casos en los que se obtuvieron 2 alelos distintos por individuo, las secuencias correspondientes a los dos haplotipos fueron deducidas manualmente, ya que sólo contenían una única posición variable. A continuación, se obtuvieron las genealogías haplotípicas a partir del árbol de máxima verosimilitud de cada marcador nuclear calculado usando RAxML.

2.4 Análisis de diversidad genética, demografía y estructura genética

Los estadísticos de diversidad nucleotídica y haplotípica fueron calculados con DnaSP versión 5 (Librado & Rozas 2009). Se midieron las desviaciones de la neutralidad usando la D de Tajima (Tajima 1989) y R2 (Ramos-Onsins & Rozas 2002). Se llevó a cabo un análisis de la varianza molecular (AMOVA) de las secuencias mitocondriales usando el software ARLEQUIN (Excoffier & Lischer 2010). La significación estadística del índice de fijación (Fst) se analizó con 1000 permutaciones.

Las barreras genéticas existentes a lo largo de la distribución de *G. pyrenaicus* se determinaron usando el algoritmo de máxima diferenciación de Monmonier (Monmonier 1973), (que identifica la mayor distancia genética existente entre dos localidades) implementado en el software Alleles in Space (Miller 2005). Se emplearon distancias genéticas puras calculadas a partir de las secuencias mitocondriales concatenadas y las coordenadas geográficas correspondientes, y se seleccionó la opción de detectar una sola barrera.

La diversidad genética mitocondrial fue estimada en cada punto de muestreo usando todas las secuencias recolectadas en menos de 1 grado (aproximadamente 100 km) alrededor de esa localidad. Este área permitió estimar la diversidad genética a partir de un buen número de muestras para cada punto, y al mismo tiempo la resolución permitía distinguir diferencias regionales en diversidad genética. Además, el hecho de centrar las medidas en torno a cada localidad de muestreo (en lugar de usar una cuadrícula fija procuró un agrupamiento más eficiente en áreas no densamente muestreadas. Para evitar aumentar artificialmente la diversidad genética debido a la presencia de linajes en contacto secundario, sólo se emplearon para este cálculo en primera instancia las secuencias pertenecientes al mismo linaje mitocondrial. Para cada subconjunto alrededor de una localidad con más de 2 muestras, se estimó la diversidad nucleotídica (π). A continuación, se interpoló una rejilla de valores de diversidad nucleótidica y se construyó un mapa de relieve usando Surfer (Golden Software Inc.).

2.5 Estima del tiempo al ancestro común más reciente de las secuencias mitocondriales

La estima del tiempo al ancestro común más reciente (o "mrca" *most recent common ancestor*) de las secuencias mitocondriales de *Galemys pyrenaicus* se llevó a cabo en dos fases.

En un primer paso, el objetivo fue obtener una estima del tiempo de divergencia de *G. pyrenaicus* y *D. moschata* (el taxón más cercano existente actualmente). La separación de estos dos linajes de desmaninos está fijada, de acuerdo con el registro fósil y estudios moleculares previos, en 9-11.2 millones de años. Para obtener una estima más precisa de esta fecha basada en varias calibraciones fósiles, se construyó un árbol con secuencias nucleares (que son más fiables que las mitocondriales al no estar tan afectadas por la saturación a estas distancias evolutivas (Meredith *et al.* 2011)) de varios mamíferos del superorden Laurasiatheria usando inferencia bayesiana e implementando un reloj molecular relajado no correlacionado. Se incluyeron una serie de calibraciones basadas en fósiles propuestas anteriormente para nodos clave de la filogenia de mamíferos (Benton *et al.* 2009) que incluían mínimos estrictos (*hard minimum*) y máximos difusos (*soft bounds*). Se usaron los 8 intrones secuenciados para este estudio en *G. pyrenaicus* y *D. moschata*, a los que se añadieron las secuencias ortólogas correspondientes a mamíferos laurasiaterios con su genoma disponible en la base de datos ENSEMBL (Flicek *et al.* 2011): *Felis catus*, *Canis familiaris*, *Pteropus vampyrus*, *Equus caballus*, *Bos taurus*, *Tursiops truncatus* y *Sus scrofa*. No fue posible recuperar las secuencias de los 8 intrones para todas las especies, de modo que el porcentaje de datos no disponibles fue del 7.5%. Se construyeron los alineamientos múltiples de cada uno de los ocho intrones usando el programa MAFFT (con el método L-INS-i, que está indicado para una mayor precisión en la estima del alineamiento) (Katoh & Toh 2008). A continuación se aplicó Gblocks con parámetros relajados, para descartar posibles regiones que estuvieran mal alineadas. Los ocho alineamientos obtenidos se emplearon como particiones independientes en un análisis de inferencia bayesiana usando BEAST, versión 1.6.2. En cada caso, se aplicó el modelo de sustitución que mejor se ajustaba a la partición (sugerido por jModeltest) y un reloj relajado no autocorrelacionado con las tasas evolutivas modeladas por una distribución log-normal. Se calibraron los nodos correspondientes usando las dataciones descritas (Benton *et al.* 2009), usando como *priors* distribuciones lognormales definidas por el límite mínimo de divergencia, que determinó el *offset*; y el límite máximo difuso, cuyo valor se hizo coincidir con el percentil 95% de la densidad de probabilidad, determinando así la media. La desviación estándar de la distribución lognormal fue

fijada en 1 (tabla 1). Se forzó la monofilia en todos los clados usados en las calibraciones. El análisis se corrió durante 100 millones de generaciones, y se verificó la convergencia de todos los parámetros mediante Tracer. Se descartaron como *burn-in* el 10% inicial de los árboles antes de obtener el árbol de mayor credibilidad de clados usando TreeAnnotator. De este análisis se obtuvo la distribución posterior de edades estimadas del nodo *G. pyrenaicus – D. moschata*. Para comprobar posibles interacciones entre los *priors* de calibración empleados para los distintos nodos del árbol, se estimaron las distribuciones efectivas conjuntas de estos *priors* (*effective joint prior distributions*) (Drummond *et al.* 2006) realizando un análisis similar al descrito pero *vacío*, es decir sin incluir los datos de las secuencias.

En la segunda fase, se empleó la estima de tiempo de la divergencia *G. pyrenaicus – D. moschata* (obtenida en el paso anterior) en un análisis con secuencias mitocondriales de tálpidos, con el objetivo de estimar el tiempo hasta el ancestro común más reciente de los haplotipos de desmán obtenidos en este estudio. Para ello, se añadieron, a los 35 haplotipos de desmán ibérico y a la secuencia de desmán ruso obtenidas previamente, secuencias completas de citocromo *b* obtenidas de GenBank correspondientes a 18 especies de la subfamilia Talpinae (hasta un total de 77 haplotipos). Se realizó otro análisis usando el programa BEAST y en esta ocasión el modelo de sustitución que mejor se ajustaba a los datos fue el TrN + I. Los datos se dividieron en tres particiones, de acuerdo con las posiciones de los codones, y se asumió de nuevo un reloj molecular relajado no correlacionado. El nodo de *G. pyrenaicus – D. moschata* fue calibrado usando una distribución normal con la media y la distribución estándar calculadas a partir de los datos de la distribución posterior de edades obtenidas en el análisis anterior. Las condiciones del análisis fueron las mismas que las ya descritas anteriormente.

Tabla 1. Priors de calibración (en millones de años) usados en los análisis de BEAST de laurasiaterios para los trabajos de *Galemys pyrenaicus* y el género *Neomys*.

Clado (Código)	Límite mínimo estricto (*hard bound*)	Límite máximo difuso (*soft bound*)	Parámetros Lognormal	
			Media	Offset
Boreoeutheria	61.5	131.5	22.28	62.5
Laurasiatheria	62.5	131.5	21.95	62.5
Eulipotyphla	61.5	131.5	22.28	61.5
Ferungulata	62.5	131.5	21.95	62.5
Cetartiodactyla	52.4	65.8	4.27	52.4
Cetruminantia	52.4	65.8	4.27	52.4
Zooamata	62.5	131.5	21.95	62.5
Carnivora	39.68	65.8	8.28	39.68
Catarrhini	23.5	34	3.35	23.5

2.6 Modelado de la distribución de la especies

Para construir un modelo de distribución de *G. pyrenaicus*, se obtuvieron datos de presencia de la especie de los atlas de España (Nores *et al.* 2007), Portugal (Queiroz *et al.* 1996) y Francia (S.F.E.P.M. 1984). Las coordenadas correspondientes fueron obtenidas de los autores de los atlas o de la *Global Biodiversity Information Facility* (GBIF, **http://www.gbif.org**). Para cada registro, se tomó el centro del cuadrado UTM de 10x10 km correspondiente, resultando en un total de 680 puntos de datos. El área de estudio fue definida entre 39 y 44º de latitud y -10 y 4º de longitud. De esta forma, se incluía todo el área de distribución del desmán ibérico, además de áreas que pueden suponer lugares potenciales de dispersión, y que son adecuadas para la selección de datos de fondo (Elith *et al.* 2011).

Las 19 variables climáticas BioClim (Hijmans *et al.* 2005), que representan sumarios de medias y variación de temperaturas y precipitación, además de la altitud, fueron descargadas de la base de datos de clima global World Clim (versión 1.4) a una resolución espacial de 2.5 minutos de arco (**http://www.worldclim.org**). Las variables climáticas fueron descargadas tanto para las condiciones actuales como para el Último Máximo Glacial (*Last Glacial Maximum*, LGM). En este último caso se obtuvieron los datos para el modelo CCSM (*Community Climate System Model*) y para el modelo MIROC (*Model for Interdisciplinary Research On Climate*). La colinearidad entre las variables climáticas fue analizada mediante correlaciones por pares usando

1000 valores tomados al azar del área de estudio. Después de este análisis, se seleccionaron 11 variables con un coeficiente de correlación entre ellas menor de 0.9: BIO1 (media anual de temperatura), BIO2 (media de rango diurno de temperatura), BIO3 (isotermalidad), BIO4 (estacionalidad de la temperatura), BIO5 (temperatura máxima del mes más cálido), BIO6 (temperatura mínima del mes más frío), BIO8 (temperatura media del trimestre más húmedo), BIO9 (temperatura media del trimestre más seco), BIO12 (precipitación anual), BIO14 (precipitación del mes más seco) y BIO15 (estacionalidad de la precipitación).

Para predecir la distribución potencial de la especie, se usó Maxent versión 3.3.3 (Phillips *et al.* 2006), que permite obtener un modelo de probabilidad relativa de ocurrencia de la especie dentro de la cuadrícula del área de estudio seleccionada. Se emplearon las condiciones por defecto, con la excepción de que se corrió el modelo con 100 réplicas crosvalidadas, tomando el valor medio de las probabilidades de presencia. La precisión del modelo fue evaluada usando el 75 % de los datos de presencia para entrenar el modelo y el 25 % restante para testarlo. El área resultante debajo de la curva ROC (característica operativa del receptor, o *Receiving Operative Characteristic*) fue de 0.824, lo que se corresponde con un modelo predictivo útil (Phillips & Dudik 2008). El modelo MIROC predice condiciones muy suaves en el LGM en esta parte del planeta y, por tanto, se obtuvo con él una distribución predicha para el desmán muy similar a la actual. La presencia de muchas especies adaptadas al frío en la Península Ibérica durante el LGM (Alvarez-Lao & Garcia 2010) no encaja muy bien con este modelo y, por tanto, no se usó individualmente. La combinación de ambos modelos (CCSM y MIROC) para estimar el área mínima común bajo ambos (Hernandez-Roldan *et al.* 2011) produjo un resultado muy similar al obtenido usando sólo el modelo CCSM, ya que se trata de un modelo más restrictivo. Por tanto, sólo se empleó el modelo CCSM en los análisis finales para predecir la distribución potencial de la especie durante el Último Máximo Glacial.

3. ESTIMA DEL ÁRBOL DE ESPECIES DEL GÉNERO *NEOMYS*

3.1 Obtención de muestras

Se obtuvieron muestras de tejidos de las dos especies del genéro *Neomys* presentes en la Península Ibérica: *N. fodiens* (2 individuos) y *N. anomalus* (6 individuos, incluyendo una muestra de *Neomys anomalus milleri* de Rusia). El muestreo se completó mediante muestras no invasivas (excrementos en su gran mayoría), de forma que se obtuvo una representación de buena parte de la distribución del género en la Península Ibérica (tabla A4). También se obtuvo un cráneo hallado en una egagrópila de *Tyto alba* y asignado a *Neomys anomalus* de acuerdo a los caracteres diagnósticos correspondientes de la mandíbula inferior.

3.2 Extracciones de ADN genómico y PCR

3.2.1 Extracción de ADN genómico a partir de tejido fresco

Se llevaron a cabo como ya se ha descrito anteriormente, en el apartado 1.2.2.

3.2.2 Extracción de ADN genómico a partir de excrementos

El protocolo empleado para obtener ADN genómico a partir de los excrementos de musgaños fue el mismo que el ya descrito para el trabajo del desmán ibérico (ver apartado 2.2.2)

3.2.2 Extracción de ADN genómico de un cráneo hallado en una egagrópila

Se añadió nitrógeno líquido a la mandíbula inferior y se trituró con ayuda de un mortero hasta obtener un polvo fino. A continuación, se realizó una extracción de ADN empleando el kit DNEasy Blood and Tissue de QIAGEN, de forma similar a las de tejido fresco, aunque en este caso la incubación con proteinasa K y tampón de lisis se prolongó durante 6 horas para maximizar la eficiencia.

3.2.3 Amplificación por PCR del citocromo b

Se diseñaron cebadores específicos para *Neomys* para amplificar el citocromo *b* empleando secuencias obtenidas de Genbank de las tres especies del género (*N. anomalus*, *N. fodiens* y *N. teres*) (tabla A1). En el caso de las extracciones de ADN obtenidas a partir de tejidos se llevó a cabo una única reacción de PCR para amplificar los 1140 nucleótidos del citocromo *b*. Las muestras no invasivas (excrementos y cráneo de egagrópila) no permitieron, sin embargo, obtener un ADN de suficiente calidad y

concentración. Por ello, se diseñaron cebadores internos que permitieron la obtención de la secuencia completa mediante tres fragmentos de PCR solapantes de 403, 312 y 544 pares de bases. De nuevo se siguió un protocolo similar al detallado anteriormente (apartado 1.2.3).

3.2.4 Amplificación por PCR de intrones nucleares

Se seleccionaron 13 intrones del conjunto de marcadores desarrollados para filogenia de especies cercanas de mamíferos: ALAD-10, ASB6-2, CSF2-2, CST6-1, GALNT5-4, GDAP-1, HIF1AN-5, KDM8-2, MCM3-2, MYCBPAP-11, PRPF31-3, SLA-2 y TRAIP-8. Para cada uno de ellos se obtuvieron las secuencias correspondientes a los exones flanqueantes de tantas especies de mamíferos laurasiaterios como fue posible en la base de datos ENSEMBL. Se construyó entonces un alineamiento múltiple (usando MAFFT) para cada exón flanqueante, sobre el cual se diseñaron los cebadores correspondientes en las zonas más conservadas de cada exón (tabla A2). Se incluyeron bases degeneradas en los cebadores, hasta un máximo de 72, para intentar maximizar la complementariedad a los exones de *Neomys*. Los 13 intrones únicamente pudieron ser amplificados (siguiendo el protocolo descrito en el apartado 1.2.3) en las extracciones de ADN provenientes de tejidos, ya que, como ya se ha comentado, las procedentes de muestras no invasivas contenían ADN demasiado degradado y en insuficiente concentración.

Las fases haplotípicas de los intrones fueron determinadas usando PHASE versión 2.1.1 (Stephens & Donnelly 2003), y el umbral de probabilidad fue fijado en 0.9. A continuación, los haplotipos que quedaron sin resolver fueron subclonados en el vector plasmídico pstblue-1, siguiendo el protocolo mencionado en el apartado 1.2.4. Se secuenciaron varios clones por cada fragmento de PCR subclonado, para asegurarse la obtención de la secuencia de los dos alelos y minimizar el riesgo de presencia de errores de PCR asociados al proceso de clonación. Se obtuvieron a continuación tanto los árboles filogenéticos de máxima verosimilitud, empleando RAxML (Stamatakis 2006) con un modelo GTR+G de evolución, como las genealogías haplotípicas correspondientes usando Haploviewer (Salzburger *et al.* 2011).

3.3 Filogenia mitocondrial del género *Neomys*

Con el objetivo de evaluar de forma preliminar las relaciones filogenéticas existentes entre las distintas especies del género *Neomys*, y en particular entre las subespecies de *Neomys anomalus*, se empleó el gen del citocromo *b*. Para complementar los datos

obtenidas mediante el muestreo ya descrito, se descargaron de la base de datos del Genbank secuencias de citocromo *b* completas correspondientes a *Neomys teres*, *Neomys fodiens* y *Neomys anomalus* del resto de Europa y Asia. De esta forma, se obtuvo una representación de los linajes mitocondriales del género *Neomys*, con un mayor detalle en la Península Ibérica, de la siguiente manera: 10 individuos de *N. fodiens*, 20 individuos de *Neomys anomalus* (incluyendo ambas subespecies) y 4 individuos de *N. teres*. Además, se descargaron de GenBank 4 secuencias pertenecientes a *Chimarrogale platycephala* y *Chimarrogale himalayica* para usar como grupo externo. Ambas especies pertenecen a la tribu Nectogalini, en la que también se encuentra el género *Neomys*, del que divergieron hace 6.7 millones de años (He 2010). Se obtuvo el correspondiente alineamiento múltiple de las 38 secuencias resultantes mediante el programa MAFFT, empleando los parámetros por defecto.

A continuación, empleando el software RAxML, se calculó el árbol filogenético de máxima verosimilitud. El modelo de sustitución empleado fue GTR + G + I (sugerido por jModeltest como el que mejor se ajustaba a los datos), y el soporte estadístico de los nodos fue evaluado mediante 1000 réplicas de *bootstrap*. Para evaluar en detalle el nodo que agrupa a las dos subespecies de *Neomys anomalus* con su especie hermana *Neomys teres*, se investigaron las topologías alternativas existentes (monofilia de *Neomys anomalus*, y monofilia de *N. teres* con cada una de las subespecies) mediante un test AU (*approximate unbiased test*) (Shimodaira 2002). Para ello, además del árbol de máxima verosimilitud ya obtenido (y que reflejaba la monofilia de *Neomys anomalus*) se obtuvieron, suministrando las constricciones topológicas iniciales correspondientes a RAxML, los árboles con las topologías alternativas. Para el conjunto de las tres topologías resultantes, se computaron los log-likelihood por sitio, que fueron empleados a continuación por CONSEL (Shimodaira & Hasegawa 2001) para realizar el AU-test.

Además, se obtuvo también un árbol filogenético empleando inferencia bayesiana. En este caso se empleó el programa BEAST, usando un reloj molecular estricto e, igualmente, el modelo GTR+G+I. El análisis consistió en 25 millones de generaciones, de las cuales se eliminaron los tres millones iniciales, tras la comprobación de que se había alcanzado la distribución estacionaria, y se comprobó que los tamaños efectivos de muestreo (*ESS*, *effective sample size*) de los parámetros de interés fueran superiores a 200.

Por otro lado, se caracterizaron las distancias genéticas que separan a los distintos linajes de citocromo *b* hallados usando el software MEGA (Tamura *et al.* 2011). Se calculó la divergencia neta que separa los distintos grupos usando el modelo de Kimura-2-parámetros (K2P).

3.4 Estima del árbol de especies del género *Neomys*

Para obtener una estima precisa de la divergencia de las dos poblaciones diferenciadas de *Neomys anomalus* presentes en la Península Ibérica (correspondientes aproximadamente con las dos subespecies *anomalus* y *milleri*), se empleó la metodología de árboles de especies que, siguiendo la teoría de la coalescencia, coestima los árboles de genes y un árbol de especies común a todos ellos a partir de muestras de varios individuos por especie. Esta metodología está implementada en el software *BEAST (Heled & Drummond 2010), de tal forma que nos permite la aplicación de distribuciones sobre los valores *a priori* de gran variedad de parámetros relacionados con el modelo de sustitución, las tasas evolutivas o el reloj molecular, entre otros. De esta forma, incluyendo la información de los 13 intrones nucleares secuenciados en *Neomys anomalus anomalus* (3 individuos) y *Neomys anomalus milleri* (3 individuos) y, además, en *Neomys fodiens* (2 individuos) junto con la correspondiente al gen del citocromo *b*, se podría estimar la divergencia existente entre las especies o poblaciones en el árbol de especies.

3.4.1 Obtención de las tasas evolutivas correspondientes al género Neomys

Con el objetivo de obtener medidas rigurosas de las tasas evolutivas de los marcadores (citocromo *b* y los 13 intrones nucleares) que se emplearían en el análisis del árbol de especies del género *Neomys*, se realizaron dos análisis preliminares.

En primer lugar, para estimar las tasas correspondientes de cada uno de los intrones, se construyó un árbol filogenético de mamíferos laurasiaterios y euracontoglires, mediante inferencia bayesiana y empleando un reloj molecular relajado y varias calibraciones fósiles de referencia distribuidas a lo largo de toda la filogenia. Para ello, se emplearon las secuencias de los 13 intrones de *Neomys fodiens*, y, además, se obtuvieron, a partir de tejido fresco, las correspondientes a otros dos sorícidos: la musaraña tricolor (*Sorex coronatus*), miembro, al igual que los musgaños, de la subfamilia Soricinae; y la musaraña gris (*Crocidura russula*), miembro de la subfamilia Crocidurinae. A estas secuencias se les añadieron, empleando la base de datos ENSEMBL, las secuencias de los intrones ortólogos correspondientes a las siguientes

especies: *Erinaceus europaeus*, *Bos taurus*, *Tursiops truncatus*, *Pteropus vampyrus*, *Myotis lucifugus*, *Ailuropoda melanoleuca*, *Canis familiaris*, *Felis catus*, *Equus caballus*, *Homo sapiens* y *Macaca mulatta*. El porcentaje de datos ausentes en los alineamientos (ya que no todas las especies estaban representadas para todos los marcadores, bien por ausencias en las bases de datos o porque no se pudieron obtener en laboratorio para el caso de los sorícidos) fue de 13.2%. Se construyeron los alineamientos de todas las secuencias ortólogas disponibles para cada intrón usando el software MAFFT con el método L-INS-i (Katoh & Toh 2008). La aplicación posterior de Gblocks usando parámetros relajados permitió corregir posibles zonas con alineamiento defectuoso (Talavera & Castresana 2007). Los 13 alineamientos fueron incluidos como particiones independientes en un análisis de BEAST (Drummond & Rambaut 2007): a cada uno de ellos se le aplicó un reloj relajado no correlacionado, y se escogió en cada caso el modelo de sustitución que mejor se ajustaba a los datos, según lo indicado por jModeltest (Posada 2008). Asimismo, como ya se ha comentado con anterioridad, se incluyeron a lo largo del árbol calibraciones fósiles (Benton *et al.* 2009) correspondientes a varios nodos (tabla 1). Para ello, se crearon en cada caso los grupos monofiléticos definidos por cada nodo, y se definieron las calibraciones como ya se ha descrito en el apartado 2.5 e igualmente se determinó la distribución conjunta de los *priors* efectivos como se ha descrito. Se corrieron dos análisis independientes, cada uno de un total de 75 millones de generaciones. Tras una inspección con TRACER, se eliminaron 10 millones de generaciones iniciales de cada réplica como burn-in y se combinaron a continuación las dos muestras usando el software Logcombiner. A partir del conjunto de árboles muestreados de la distribución posterior obtenidos por BEAST se obtuvieron los datos de la tasa evolutiva de la rama correspondiente a *Neomys fodiens* para cada una de las 13 particiones del análisis, que correspondían a los 13 intrones. Estos datos se ajustaron en cada caso a una distribución normal, y se obtuvo, por tanto, para cada intrón, un valor medio de la tasa evolutiva de *Neomys fodiens* con su desviación estándar correspondiente. Además, se empleó también este análisis multilocus de mamíferos para datar la divergencia entre *Sorex coronatus* y *Neomys fodiens*, ya que esta fecha se emplearía en análisis posteriores. Para ello se extrajo la distribución posterior del nodo *S. coronatus – N. fodiens* y, de nuevo, tras eliminar el 10% inicial, se ajustaron los datos de la distribución posterior de edades a una distribución normal.

En segundo lugar, se procedió a la estima de la tasa evolutiva del citocromo *b* del linaje de *Neomys*. Para ello se descargó de la base de datos GenBank, una secuencia completa de citocromo *b* para cada una de las especies de la familia Soricidae (hasta un total de

76). Se obtuvo el alineamiento múltiple correspondiente usando MAFFT, y se llevó a cabo un análisis bayesiano empleando BEAST. El nodo de *Sorex coronatus* y *Neomys fodiens* se calibró empleando una distribución normal con la media y la desviación estándar obtenidas en el análisis anterior con los datos nucleares. Como modelo de sustitución se empleó GTR + G (tal como indicó jModeltest). Un análisis preliminar de las transiciones y transversiones observadas demostró una importante presencia de saturación en los datos. Por ello, y para contrarrestar el efecto que podría tener el hecho de estar calibrando con un nodo en una zona de la filogenia muy afectada por esta saturación, se escogió aplicar tres particiones a los datos, de acuerdo con las tres posiciones de los codones. El análisis se corrió durante 50 millones de generaciones y, tras comprobar que se había alcanzado la convergencia y todos los parámetros de interés tenían valores de ESS superior a 200, se obtuvo el árbol sumario de mayor credibilidad de clados. A partir de este análisis, se obtuvo la tasa evolutiva del citocromo *b* para la rama de *Neomys fodiens* con su correspondiente media y desviación estándar.

*3.4.2. Estima de árboles de especies con *BEAST*

En el análisis de *BEAST se definieron como especies o poblaciones independientes, los tres linajes diferenciados presentes en la Península Ibérica (*Neomys fodiens*, *Neomys anomalus anomalus* y *Neomys anomalus milleri*) encontrados en la filogenia mitocondrial del género *Neomys*. Un análisis de este tipo conlleva la asunción de que las especies o poblaciones independientes seleccionadas no han experimentado flujo genético significativo tras su divergencia (Heled & Drummond 2010). Además, también se asume que no existe recombinación intralocus en los marcadores empleados. Esto se confirmó en los 13 intrones secuenciados en el género *Neomys* empleando el test de 4 gametos implementado en DnaSP v.5.

Para cada una de las 13 particiones independientes de secuencias nucleares se empleó como modelo de sustitución el HKY, y el reloj molecular escogido fue el estricto. En el caso de la partición del citocromo *b* se empleó un muestreo aumentado respecto a los 7 individuos de los que se obtuvieron secuencias nucleares: se añadieron también las secuencias empleadas en la filogenia mitocondrial del género *Neomys* correspondientes a los tres linajes aquí analizados. En este caso, el modelo de sustitución escogido fue el HKY y, del mismo modo que anteriormente, se particionaron los datos de acuerdo a los codones. En cuanto a las tasas evolutivas, para cada una de las particiones se empleó como *prior* una distribución normal, con media y distribución estándar iguales a las estimas obtenidas para las tasas de *Neomys fodiens* a partir del análisis bayesiano

nuclear con mamíferos laurasiaterios (para el caso de los 13 intrones) y a partir del análisis del citocromo *b* de los sorícidos (para el caso de este gen mitocondrial). El análisis se corrió durante 50 millones de generaciones, tras las cuales se comprobó que se hubiera alcanzado la convergencia de todos los parámetros observando los tamaños efectivos de muestreo de los parámetros.

Se calculó el árbol sumario de mayor credibilidad de clados usando Tree Annotator y eliminando el 10% inicial de las muestras como burn-in, y se extrajeron las medias de los tiempos de divergencia entre *Neomys anomalus anomalus* y *Neomys anomalus milleri* por un lado; y *Neomys anomalus* y *Neomys fodiens* por el otro.

*3.4.3 Efecto de distintos priors en las tasas evolutivas sobre la estima de los tiempos de divergencia en *BEAST*

Con el objetivo de investigar el efecto del empleo de distintos *priors* en las tasas evolutivas en la estima de tiempos de divergencia y la conveniencia o no de la aplicación de tasas evolutivas estimadas de forma precisa y adaptadas a cada una de las particiones empleadas, se realizó una batería de experimentos para comparar los resultados empleados siguiendo distintas estrategias. Los *priors* de tasas evolutivas comparados fueron:

1) Tasa estimada de citocromo *b* para la rama de *Neomys fodiens* y tasas individuales de intrones estimadas para la rama de *Neomys fodiens*. Este es el experimento detallado con anterioridad.
2) Tasa estimada de citocromo *b* para *Neomys fodiens* y tasas individuales de intrones estimadas para el conjunto de mamíferos. En este caso, la tasa escogida para cada intrón fue obtenida de la partición correspondiente del análisis de laurasiaterios multilocus bayesiano con calibraciones fósiles. Para ello, se empleó la media calculada por BEAST para el reloj de cada intrón ("ucld.mean" individuales) y su desviación estándar como prior en forma de una distribución normal para el parámetro de la tasa del reloj de la partición correspondiente en el análisis de *BEAST.
3) Tasas estimada de citocromo *b* para *Neomys fodiens* y una única tasa de intrones estimada para el conjunto de mamíferos. Este caso es similar al anterior, salvo que a todas las particiones de intrones se les aplicó la misma tasa evolutiva, que fue la media de la tasa evolutiva de los 13 intrones en el análisis bayesiano multilocus ("ucld.mean" global).

4) Tasas de citocromo *b* y de intrones obtenidas de la bibliografía. En este caso, se empleó como prior de la tasa de evolución del citocromo *b* una distribución normal centrada en el 3%, con el intervalo de confianza del 95% abarcando entre 0.69% y 5.3% (quedando estos valores dentro de lo observado anteriormente para las tasas evolutivas de citocromo *b* de todos los mamíferos (Bininda-Emonds 2007)). Respecto a las tasas evolutivas de los intrones, se empleó la misma tasa para todos ellos. La distribución escogida fue una normal con media 0.3%, lo que supone una tasa evolutiva 10 veces menor de la fijada para la partición mitocondrial, siendo este valor la diferencia media aproximada entre las tasas evolutivas del citocromo *b* y las del conjunto de intrones nucleares a nivel general (Igea *et al.* 2010).
5) Tasas estimada de citocromo *b* para *Neomys fodiens* y una única tasa de intrones estimada también para *Neomys fodiens*. A todos los intrones se les aplicó un mismo *prior* de tasa evolutiva, obtenido a partir del valor medio de las 13 tasas evolutivas para los intrones en *Neomys fodiens*.
6) Tasas individuales de intrones estimadas para *Neomys fodiens;* y sin incluir secuencia alguna de citocromo *b*.

3.5 Aplicación de un modelo de aislamiento con migracón a la divergencia de los linajes de *Neomys anomalus*

Para comprobar hasta qué punto el cálculo de la divergencia entre los dos linajes diferenciados de *Neomys anomalus* puede verse afectado por la migración posterior a esa divergencia, y para cuantificar el posible flujo génico existente, se realizó una estima de los tiempos de divergencia y otros parámetros demográficos del género *Neomys* empleando el programa IMa2. Este software implementa un modelo de aislamiento con migración y, mediante cadenas de Markov Chain Montecarlo, permite estimar genealogías y parámetros como tiempos de divergencia, tamaños de población (tanto actuales como ancestrales) y migración (Hey 2010). De esta forma, se diferencia del *BEAST en que asume la presencia de flujo génico entre los taxones estudiados. Por otro lado, en el IMa2 las relaciones filogenéticas entre las distintas especies o poblaciones han de ser determinadas de antemano. Por ello, se fijó como árbol inicial el obtenido mediante el análisis de *BEAST, con 3 poblaciones: (*N. fodiens*, (*N. anomalus anomalus*, *N. anomalus milleri*)). Se incluyeron en el análisis las mismas particiones analizadas con el *BEAST, y se calcularon las tasas evolutivas correspondientes (en tasa de mutación por locus) multiplicando las tasas evolutivas calculadas con anterioridad para cada partición por la longitud de cada molécula en *Neomys fodiens*. Para el caso

de los intrones, se fijó en la mayoría de los casos como modelo mutacional el de sitios infinitos (*infinite sites model*), que resulta adecuado para datos nucleares de especies de divergencia reciente en los que las sustituciones múltiples no han tenido lugar. Sin embargo, en aquellos intrones que violaban este modelo (por detectarse más de dos posibles bases en un lugar nucleotídico) se escogió el modelo HKY. Éste fue también el modelo seleccionado para la partición del citocromo *b*. Se realizaron varios análisis preliminares con diferentes *priors* y esquemas para determinar los condiciones óptimas para las cadenas de Markov. Se valoró el estado de mezcla de las distintas cadenas y que se hubiera alcanzado la convergencia comprobando que las estimas de ESS de los parámetros fueran superiores a 200 en todos los casos; que los gráficos generados por el programa de los valores de varios parámetros (como la verosimilitud del modelo o el tiempo de divergencia, entre otros) no mostraran ninguna tendencia indicativa de que los análisis no hubieran alcanzado la convergencia o que no muestrearan toda la zona deseada del espacio paramétrico; y que las estimas posteriores generadas a partir de los dos sets de genealogías muestreadas que típicamente se generan en un análisis de IMa2 fueran similares, lo que es indicativo de un nivel de autocorrelación adecuado en el análisis. Estos análisis preliminares indicaron también que los datos no contenían suficiente información como para estimar parámetros referentes a migración ancestral, por lo que se procedió a estimar únicamente la migración existente entre las dos poblaciones actuales de interés (*Neomys anomalus milleri* y *N. a. anomalus*).

Los análisis finales consistieron en un total de 15 cadenas distintas que produjeron un total de 100000 genealogías muestreadas. Se aplicó el test de razón de la verosimilitud (LRT *likelihood ratio test*) descrito en (Nielsen & Wakeley 2001) para comprobar si las tasas de migración obtenidas eran significativamente distintas de cero. Asimismo, se empleó también un LRT para comparar un modelo sin limites de migración y uno con ausencia total de migración y analizar si alguno de los dos se ajustaba mejor a los datos obtenidos.

IV. RESULTADOS Y DISCUSIÓN

1.- DESARROLLO DE UN NUEVO CONJUNTO DE MARCADORES GENÓMICOS PARA LA FILOGENIA DE ESPECIES CERCANAS DE MAMÍFEROS

1.1 Resultados

1.1.1 Obtención del conjunto de intrones

Con el fin de seleccionar los intrones más útiles de cara a las filogenias de especies cercanas de mamíferos se desarrolló una serie de scripts en Perl para extraer los intrones y aplicar un conjunto de filtros destinados a seleccionar las características más deseables de los marcadores. La figura 8 detalla todo este proceso.

Inicialmente se descargaron los genomas de humano, chimpancé, macaco Rhesus, perro y vaca. Se extrajeron todos los intrones menores de 50.000 nucleótidos de estos genomas, que comprenden tres órdenes de mamíferos (Primates, Carnivora y Cetartiodactyla). El número total de intrones extraídos por genoma varió entre los 153659 de la vaca y los 173320 del perro (figura 8). Usando información de ortología extraída de la base de datos ENSEMBL, se llegó a un conjunto inicial de 11.835 intrones ortólogos 1-1 para las cinco especies.

A continuación se aplicaron varios filtros destinados a eliminar los intrones con un tamaño inadecuado para su posterior amplificación en el laboratorio. Para ello, se seleccionaron sólo aquellos intrones cuya longitud en *Homo sapiens* estaba entre 200 y 1600 nucleótidos. Además, se controló que la longitud del intrón estuviera relativamente conservada entre las distintas especies, tomando como referencia el tamaño del intrón en humano y relajando las restricciones de variación de tamaño a medida que aumentaba la distancia filogenética que separaba las especies a comparar. Tras la aplicación de estos filtros de tamaño, se obtuvo un total de 2750 intrones. Por otro lado, se eliminaron también los intrones que estaban flanqueados por algún exón de menos de 40 nucleótidos, ya que se consideró que el diseño de cebadores de PCR en estos exones cortos sería complicado. Con este filtro se eliminó el 5.2 % de los intrones presentes previamente a este filtro.

Se comprobó la presencia de intrones duplicados mediante búsquedas BLAST sobre los distintos genomas, usando los exones flanqueantes a cada intrón, ya que éstos son más fácilmente detectables por búsquedas BLAST incluso en genes divergentes (Alba & Castresana 2007). Sólo se consideraron intrones de copia única aquellos que tenían sus dos exones flanqueantes presentes una única vez en cada uno de los cinco

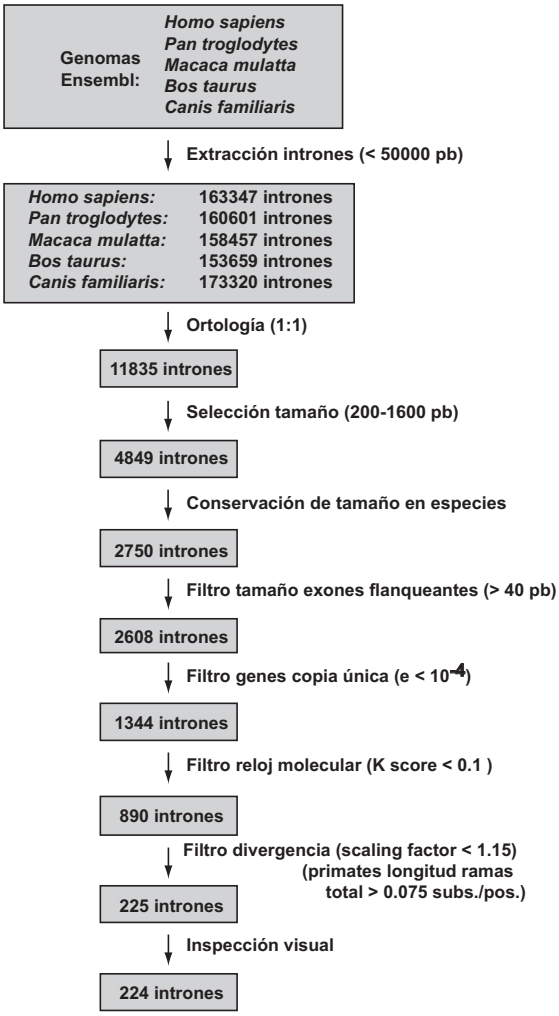

Figura 8. Esquema de los procesos de extracción y filtrado de los intrones (pb = pares de bases; subs/pos = sustituciones/posición).

genomas estudiados. Este paso descartó aproximadamente la mitad de los intrones con los que se contaba en ese momento, demostrando la amplia presencia de secuencias duplicadas en los genomas de los mamíferos. En este punto, se disponía de un total de 1344 intrones con un filtro preliminar de ortología, y que por tanto podían ser usados para estudiar diferentes procesos evolutivos. Sin embargo, para conseguir que estos intrones fueran útiles de cara a las filogenias de especies cercanas, se aplicaron una serie de filtros adicionales.

El siguiente paso fue eliminar intrones cuya tasa evolutiva no se ajustara a la global del genoma por presentar aceleraciones o deceleraciones en alguno de los linajes. Se construyó un árbol filogenético con los 1344 intrones ortólogos de copia única disponibles, que se usó para representar la referencia de la evolución global del genoma (figura 9).

Después se comparó cada uno de los intrones individuales con este árbol de referencia usando la medida del *K tree score*. Se estableció un valor máximo de este estadístico de 0.1 se descartaron todos los intrones que daban un valor superior, lo que indicaría que presentaban importantes diferencias a nivel de topología y longitud relativa de ramas al compararlos con el árbol de referencia. Esto supuso descartar el 34% de los intrones que había antes de este filtro.

Seguidamente, se excluyeron los intrones más conservados, que podrían realizar alguna función y por tanto no ser suficientemente variables para ser empleados como marcadores moleculares en filogenias de especies cercanas. Se emplearon dos medidas como criterios de divergencia: el *scaling factor*, que representa la divergencia global (y cuyo límite inferior se fijó en 1.15; ver Métodos) y la divergencia medida en las ramas de

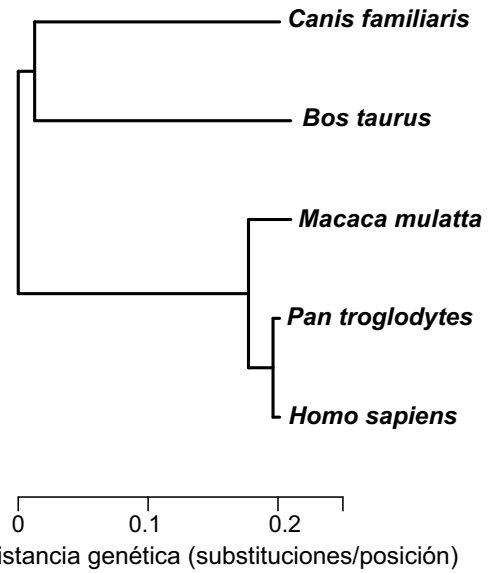

Figura 9. Árbol filogenético (máxima verosimilitud) de referencia que representa la evolución global del genoma. La raíz fue colocada en el punto medio del árbol.

primates (en este caso el mínimo fue de 0.075 sustituciones por posición). El resultado final de este filtro para eliminar los intrones más conservados fue un conjunto de 225 intrones (figura 8).

Por último, se procedió a la inspección visual de los 225 alineamientos y árboles filogenéticos para detectar cualquier secuencia mal alineada debido a defectos en el proceso de anotación o a asignaciones de ortología incorrectas. En este último filtro se eliminó un único intrón. El conjunto final resultante fue de 224 marcadores filogenéticos seleccionados para la filogenia de especies cercanas, fácilmente amplificables por PCR, variables y distribuidos por todo el genoma. Curiosamente, en este conjunto final está presente el intrón 1 de la transtiretina (TTR), que es uno de los intrones más usados tradicionalmente en los estudios filogenéticos de mamíferos.

1.1.2. Análisis de las características genómicas del conjunto de intrones

La localización genómica de los 224 intrones en *Homo sapiens* se muestra en la figura 10. Todos los cromosomas tienen al menos un intrón representado en el conjunto final, excepto el cromosoma 22 y el Y. Este último caso es esperable, ya que las secuencias correspondientes al cromosoma sexual masculino no estaban disponibles para los genomas de macaco Rhesus, vaca y perro. Por otro lado, sólo hay un único intrón localizado en el cromosoma X humano. La mayoría del resto de los intrones pertenecientes a este cromosoma fueron descartados durante el proceso de filtrado (en su mayor parte en el test de copia única y en el filtro final de divergencia).

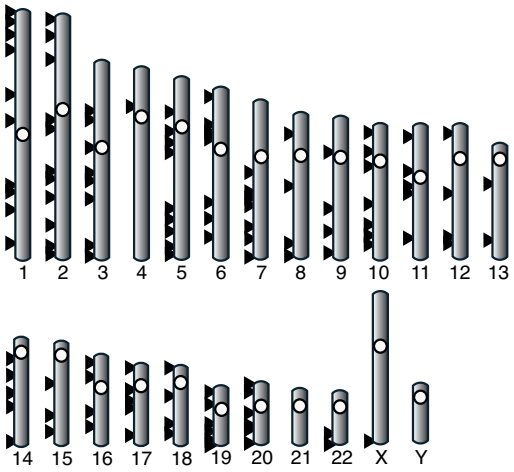

Figura 10. Cariotipo humano con la localización genómica de los genes a los que pertenecen los 224 intrones del conjunto final.

Se analizaron las secuencias repetitivas en los intrones con el programa RepeatMasker y se hallaron en 163 de los 224 intrones. En su mayor parte (un 70%) estas secuencias correspondieron a SINEs (elementos nucleares dispersos cortos, *Short Interspersed Nuclear Elements*), seguidos de LINEs (elementos nucleares dispersos largos o *Long Interspersed Nuclear Elements)*, que estaban presentes en el 26% de los intrones que portaban algún tipo de secuencia repetitiva. Las secuencias repetitivas constituían, en promedio, el 16% de la secuencia de los intrones que las portaban. Este tipo de secuencias evolucionan normalmente siguiendo un patrón neutro, mediante mutaciones puntuales, y por tanto su presencia no supone problema alguno para su inclusión en los análisis basados en secuencias que asumen modelos de sustitución nucleotídica tradicional. Si todos los organismos analizados presentan el elemento repetitivo (que sería la situación más probable en el caso de especies cercanas), este fragmento puede ser empleado con normalidad, como cualquier otra secuencia. Por el contrario, en los casos en los que no todas las secuencias presentan la secuencia repetitiva, se comprobó que los programas de alineamiento múltiple no mostraban ninguna dificultad al alinear este tipo de secuencias, pudiéndose por tanto emplear con normalidad en análisis filogenéticos posteriores.

Los microsatélites, sin embargo, sí pueden llegar a ser más problemáticos. Estos se definen como dos o más copias contiguas e idénticas (o aproximadamente idénticas) de un patrón de 1 a 6 nucleótidos. Los microsatélites, al contrario que lo expuesto para las secuencias repetitivas, evolucionan por un mecanismo de deslizamiento (*slippage*) en lugar de por mutaciones puntuales, y por ello no deben ser usados con los métodos filogenéticos o de coalescencia que asumen este tipo de modelos de sustitución. Por consiguiente, se comprobó su presencia en los 224 intrones mediante un análisis con el programa Tandem Repeats Finder. El resultado fue que 43 de ellos mostraban algún microsatélite en al menos una de las cinco especies analizadas. En general los microsatélites son exclusivos de un linaje, por lo que su presencia en una especie determinada no implica necesariamente que ocurra en otras especies. Por este motivo, no se eliminaron los intrones con microsatélites presentes en alguna de las cinco

especies. Sin embargo, si este tipo de repeticiones aparecieran en el intrón de la especie que se quiera estudiar, podría ser recomendable descartar este intrón para análisis posteriores. Otra opción sería eliminar las posiciones correspondientes a los microsatélites de los alineamientos correspondientes para poder emplear entonces programas que asumen mecanismos de evolución basados en mutaciones puntuales.

1.1.3. Análisis de las distancias genéticas y los polimorfismos de nucleótido simples (SNPs)

Con el objetivo de caracterizar el grado de variabilidad de los 224 intrones obtenidos con respecto a los conjuntos de los que se partieron antes de la aplicación de los sucesivos filtros y con el fin de determinar la eficacia de éstos, se estimó, como medida de la divergencia intrónica, la distancia genética entre humano y chimpancé para cada intrón. Esta distancia se estimó sobre árboles filogenéticos de máxima verosimilitud construidos a partir de alineamientos de humano, chimpancé y macaco. Se compararon así las distancias obtenidas para cada uno de los 224 intrones con las obtenidas en los conjuntos de los distintos pasos de filtrado de los intrones (figura 11). La divergencia media resultante en 3 de los conjuntos iniciales (los 2750 intrones ortólogos filtrados por tamaño; los 1344 intrones de copia única y los 890 intrones que pasaron el test de evolución neutra) fue de alrededor de 0.011 sustituciones/posición. Como era esperable, después de la aplicación del filtro de divergencia, en los 224 intrones la divergencia media aumentó, pero sólo ligeramente, hasta 0.014 sustituciones/posición.

En lo referente a los datos de polimorfismos, se calculó el número de polimorfismos de nucleótido simples (SNPs) presentes en cada intrón. La media de SNPs presentes en los 224 intrones del conjunto final fue de 4.35, frente a 3.19 en el conjunto inicial de 2750 intrones ortólogos filtrados por tamaño. Estos resultados absolutos se escalaron, dividiendo el número de SNPs en un intrón entre su longitud para obtener la densidad de SNPs. Con esta nueva medida, las diferencias entre los conjuntos comparados se mantuvieron: en el conjunto final era de 0.0055 SNPs por nucleótido frente a 0.0043 en el conjunto inicial. Por consiguiente, y a pesar de no haber empleado medidas directas de polimorfismo en los filtros, los intrones seleccionados también son ligeramente más variables intraespecíficamente que la media del genoma, lo que prueba de nuevo su utilidad en estudios filogenéticos de especies cercanas.

Además, se comparó también la distancia genética entre humano y chimpancé mostrada por cada intrón con la calculada usando el citocromo *b,* que es el marcador molecular más empleado en filogenia de mamíferos (Castresana 2001). Para ello, se

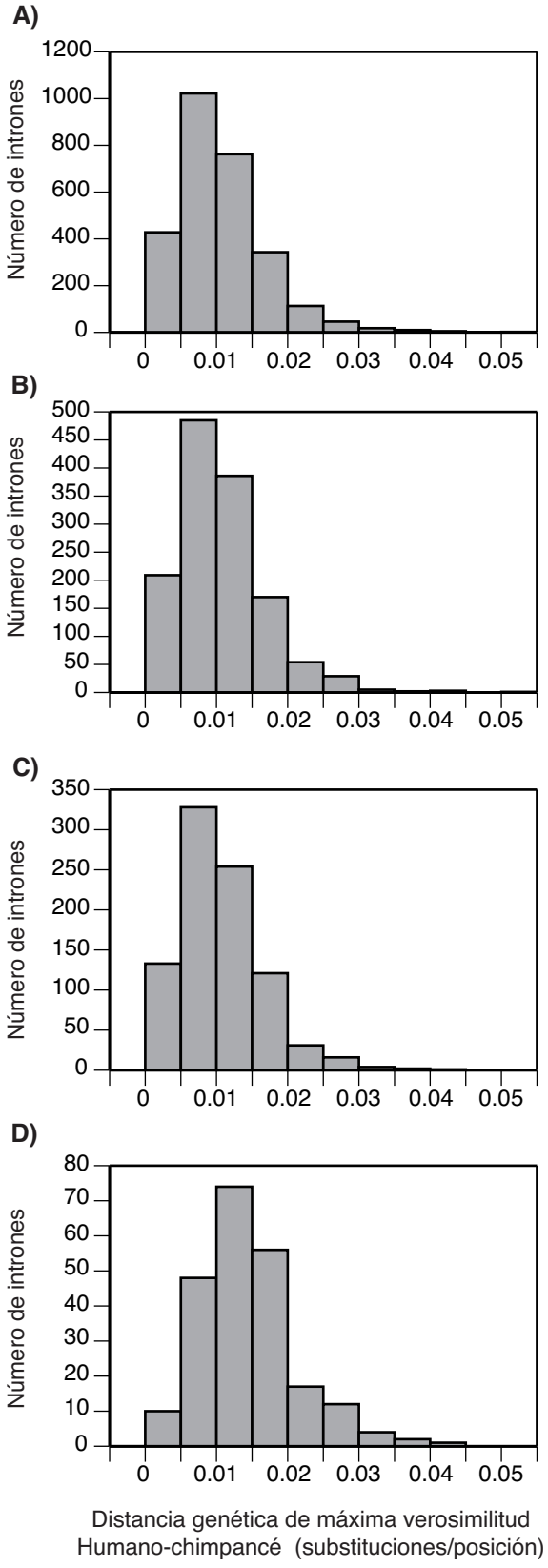

Figura 11. Distancias genéticas entre humano y chimpancé medidas sobre el árbol de las secuencias de primates correspondientes a distintos conjuntos: **(A)** 2750 intrones ortólogos con límites de tamaño impuestos, **(B)** 1344 intrones de copia única, **(C)** 890 intrones con evolución "neutra", y **(D)** el conjunto final de 224 intrones.

construyó, de forma similar a lo realizado con los intrones, un árbol filogenético con las secuencias de citocromo *b* de *Homo sapiens*, *Pan troglodytes* y *Macaca mulatta*. La distancia patrística medida para el citocromo *b* fue de 0.169 sustituciones por posición, lo que implica que el conjunto final de 224 intrones muestra, de media, una divergencia 12.1 veces menor que la del citocromo *b*.

Por último, de igual manera, se comparó la divergencia entre humano y chimpancé calculada para cada intrón con la correspondiente a los exones flanqueantes. La media de divergencia genética resultante fue de 0.006 sustituciones/posición. Esto se traduce en que los intrones muestran, de media, una divergencia 2.3 superior a la de sus exones flanqueantes.

1.1.4 Validación experimental de los nuevos marcadores moleculares: diseño de cebadores y PCR

Con el objetivo de comprobar en el laboratorio que el conjunto de marcadores desarrollados *in silico* eran realmente útiles para su aplicación en filogenias de especies cercanas, se seleccionaron por inspección visual 12 de los 224 intrones para posteriormente secuenciarlos en varias parejas de especies de mamíferos.

A partir de los exones flanqueantes, se diseñaron cebadores de PCR degenerados que permitieran la amplificación en el mayor número de especies posible. Para ello se construyó un alineamiento con las secuencias correspondientes a los genomas ya analizados en este estudio (humano, chimpancé, macaco, vaca y perro), a los que se añadieron, cuando fue posible, secuencias también obtenidas de la base de datos ENSEMBL, y correspondientes a otras especies de mamíferos cuyos genomas estaban secuenciados con una cobertura inferior.

Para probar estos cebadores, se usó ADN genómico de seis especies de mamíferos: el murciélago ratonero gris ibérico (*Myotis escalerai*), el orangután de Borneo (*Pongo pygmaeus*), el leopardo de las nieves (*Uncia uncia*), el tigre (*Panthera tigris*), la comadreja (*Mustela nivalis*) y el turón (*Mustela putorius*). Este grupo de especies, junto a las ya disponibles en las bases de datos, incluía varios pares de especies cercanas de distintos órdenes de mamíferos.

La tabla 2 muestra los cebadores empleados y los resultados de la amplificación en el panel de especies. De los 12 intrones para los que se diseñaron cebadores degenerados, 5 de ellos no produjeron una única banda del tamaño esperado en la reacción de PCR en alguna de las muestras tras sucesivas optimizaciones (como variaciones en la temperatura de alineamiento y otras modificaciones de las condiciones de la reacción de amplificación) y fueron, por tanto, descartados. En estos casos se engloban tanto amplificaciones de múltiples productos de PCR (visualizadas en forma de *smear* en el gel de agarosa), que pueden deberse a la existencia de duplicaciones génicas en un linaje determinado, como amplificaciones de productos de una longitud excesiva, superior a 2000 pares de bases. Los 7 intrones restantes produjeron una única banda de PCR secuenciable y del rango de tamaños esperados, si bien en muchas ocasiones la temperatura de hibridación de la PCR hubo de ser optimizada individualmente para cada especie testada. Además, en algunos casos, como el intrón SLC38A7-8 para *Panthera tigris* y CARHSP-1 y PNPO-3 para *Myotis escalerai*, se tuvieron que diseñar cebadores específicos para mejorar la reacción de PCR o la secuenciación posterior.

En todos los casos, las reacciones de PCR fueron secuenciadas y se comparó la secuencia obtenida con las ortólogas de otras especies, para corroborar así que se había amplificado realmente el intrón buscado. Además, para resolver las ambigüedades presentes en algunas bases nucleotídicas secuenciadas, se obtuvo mediante clonaje en vectores plasmídicos la secuencia exacta de uno de los alelos para cada uno de los intrones.

Tabla 2. Intrones seleccionados para su amplificación y secuenciado en seis especies, y resultado final. El nombre del intrón está tomado del nombre del gen según el HUGO Gene Nomenclature Committee (Comité de Nomenclatura de Genes del Genoma Humano) seguido de un guión y del número del intrón. En la columna de resultados, el símbolo (+) indica secuenciación correcta en todo el panel de especies; (1), amplificación inespecífica en una especie; (2), amplificación inespecífica en varias especies; (3), intrón de un tamaño excesivo en alguna especie; (4), banda de PCR múltiple en alguna especie.

Nombre del intrón	Código ENSEMBL	Longitud (*H. sapiens*)	Secuencia cebadores (Forward y Reverse)	Resultado
SLC38A7-8	ENSG00000103042	877	RGGCCTRGCYGSCTGCTTCATCTT TCVGASAGYTTGGCTTGRATGAGGCA	+
COPS7A-4	ENSG00000111652	952	TACAGCATYGGRCGRGACATCCA TCACYTGCTCCTCRATGCCKGACA	+
CARHSP1-1	ENSG00000153048	698	ACYCGCCGSACSAGGACCTTCT GTRATGAAGCCRTGGCCCTTGGA	+
GAD2-1	ENSG00000136750	724	GGCTCHRGCTTYTGGTCYTTYGG YCCGAKGCCKCCSGTGAACTTCT	+
JMJD5-2	ENSG00000155666	1124	ACCABTGGCCVTGCATGMAGARGT TGATGAACTCRYTGACBGTCATGAG	+
OSTA-5	ENSG00000163959	535	TGMWGGYCATGGTGGAAGGCTTTG AGATGCCRTCRGGGAYGAGRAACA	+
PNPO-3	ENSG00000108439	843	GATGGCTTCCRHTTCTWCACTAACTT GGYTCCCARTAGAAGACMAKSGA	+
GAD2-3	ENSG00000136750	1051	TGCTCTAYGGRGAYKCMGAGAAG CAGAAACGCCARMGTGGSCCTTT	(1)
CLCN6-17	ENSG00000011021	1376	GTGGCCAAATGGACAGGGGACTTT TTGCCCTTCATGAACTCCTTCTCGT	(2)
TFPI2-2	ENSG00000105825	913	TACTACTAYGACAGGYACWYGCAGA CATGTCATRGAWSTTAGATTRAAGAA	(2)
CSE1L-12	ENSG00000124207	1496	CATGGRATYACAMAAGCWAATGA TAYTTRATRCCRTCAGCTTTAAG	(3)
TBC1D21-8	ENSG00000167139	1007	TCTTYCCCTGGTTCTGYYTCTGCTT CAKGCWGTAGGCCACCAGCACCT	(4)

1.1.5 Uso de los intrones seleccionados como marcadores para la filogenia de especies cercanas

Los siete intrones secuenciados se unieron a las secuencias ortólogas correspondientes a los genomas ya analizados de humano, chimpancé, macaco Rhesus, vaca y perro, así como a las correspondientes a los genomas del orangután de Sumatra (*Pongo abelii*), el caballo (*Equus caballus*) y el miotis norteamericano (*Myotis lucifugus*). Este conjunto de especies ampliado permitió disponer de cinco parejas de especies cercanas de mamíferos, pertenecientes a cuatro familias distintas: Mustelidae, con la pareja *Mustela nivalis* y *M. putorius*, con un tiempo de divergencia estimado de 2.8 millones de años (Koepfli *et al.* 2008); Felidae, con *Panthera tigris* y *Uncia uncia*, cuya separación se estima en 2.9 millones de años (Johnson *et al.* 2006); Hominidae, de la que se incluyeron la pareja *Pongo pygmaeus* y *P. abelii* (3.8 millones de años; (Raaum *et al.* 2005)) y la pareja *Homo sapiens* y *Pan troglodytes* (6 millones de años; (Kumar

et al. 2005)); y, por último, la familia Vespertilionidae, con *Myotis lucifugus* y *M. escalerai*, cuyos linajes se separaron hace 12.2 millones de años (Stadelmann *et al.* 2007).

Las secuencias obtenidas para cada uno de los intrones siempre presentaban diferencias entre todos los miembros de las parejas de especies cercanas, excepto en el caso del intrón PNPO-3, cuya secuencia en los dos orangutanes fue idéntica. Esto puede ser debido a la existencia de flujo génico reciente entre estas dos especies, a constricciones en la evolución de este marcador en este linaje o a factores estocásticos que impliquen que, por azar, no se han producido mutaciones en ninguna de las dos especies desde su separación. En el resto de los casos analizados se encontró al menos una diferencia nucleotídica entre las secuencias de las especies cercanas.

A continuación se procedió a la reconstrucción filogenética por máxima verosimilitud de cada uno de los 7 intrones (figura 12). Se obtuvieron así filogenias congruentes con la taxonomía aceptada para los mamíferos, particularmente a niveles intraordinales. A niveles supraordinales, en cambio, se observó que las relaciones entre los 4 órdenes de laurasiaterios estudiados (Carnivora, Perissodactyla, Cetartiodactyla y Chiroptera) no están bien resueltas, y hay varias alternativas topológicas en los distintos árboles filogenéticos de los intrones. Esto indica, como ya se ha demostrado, que los intrones individuales pueden no contener suficiente señal filogenética para reconstruir la historia de las relaciones interordinales de los mamíferos (Murphy *et al.* 2004) (Springer *et al.* 2004). Cabe destacar que se encontraron importantes diferencias en la divergencia global de los árboles filogenéticos obtenidos para los intrones. Como se puede percibir observando la escala de los árboles, los ejemplos más extremos de estas diferencias son, por un lado, GAD2-1 y OSTA-5, que son los intrones con una evolución más lenta y, por otro, JMJD5-2, que es el intrón más rápido.

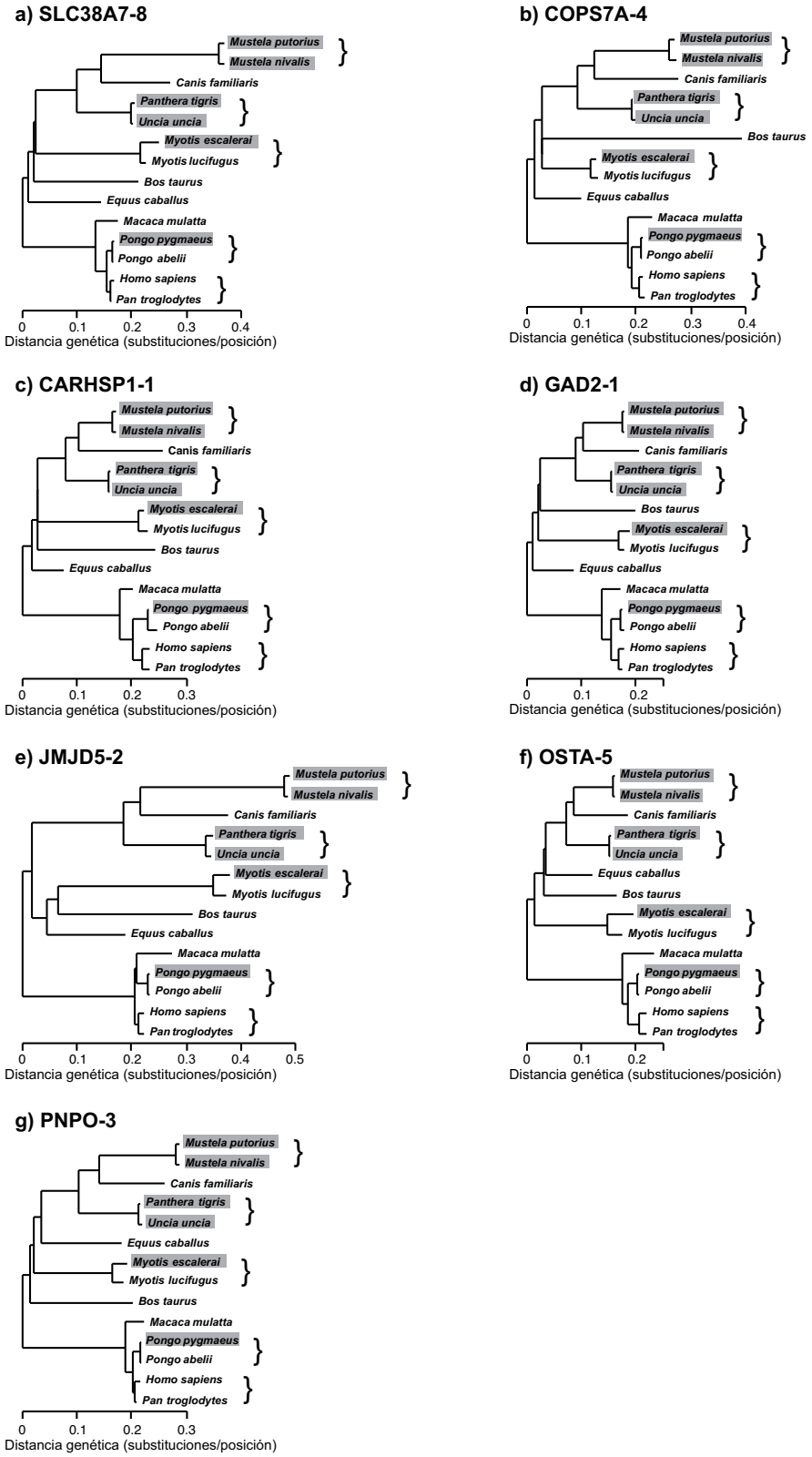

Figura 12. Árboles filogenéticos de máxima verosimilitud de siete intrones seleccionados. Las especies sombreadas indican nuevas secuencias obtenidas en este estudio y los corchetes señalan las parejas de especies cercanas analizadas. La raíz fue colocada en el 7% de la rama que separa los primates de los laurasiaterios (coincidiendo con el punto medio del árbol genómico global). Todos los árboles están dibujados a la misma escala.

Para valorar de forma más precisa las características de los siete intrones como marcadores para la filogenia de especies cercanas, se construyeron nuevos alineamientos para cada una de las parejas y se midieron las distancias genéticas que las separaban. De esta forma se evitaron problemas asociados a la inclusión de especies demasiado divergentes a la hora de medir las distancias genéticas. La figura 13 muestra las divergencias calculadas para cada intrón en las cinco parejas. Como era esperable, las parejas separadas por una divergencia mayor (como el caso de los murciélagos) acumulan mayor número de sustituciones que las especies más cercanas como la pareja de félidos. También cabe destacar la gran variabilidad existente en la divergencia intrónica acumulada por cada marcador al comparar las distintas parejas de especies.

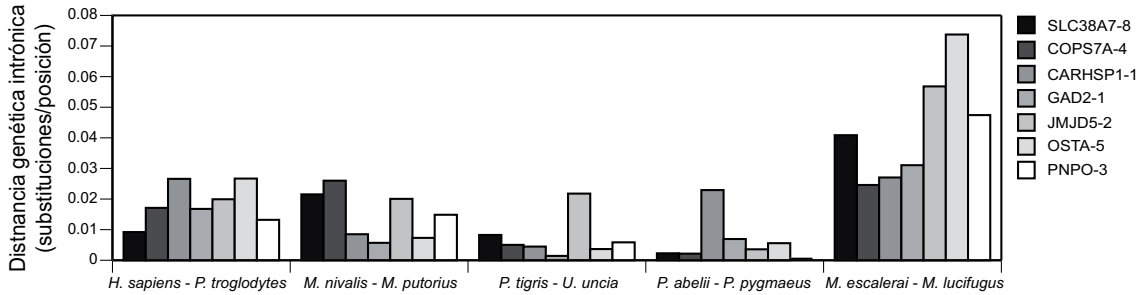

Figura 13. Distancias por pares de máxima verosimilitud para los diferentes intrones en las parejas de especies cercanas analizadas.

1.2 Discusión

1.2.1 Características del conjunto final de intrones

Mediante una serie de filtros aplicados al total de los intrones ortólogos de mamíferos, se seleccionó un conjunto de 224 intrones con características adecuadas para su uso en filogenia de especies cercanas de mamíferos. En este sentido, cabe destacar que los filtros seleccionados permitieron evitar seleccionar intrones presentes en más de una copia en el genoma. Los genomas de los mamíferos tienen un gran número de genes duplicados debido a diferentes procesos de duplicación a nivel génico o genómico ocurridos durante su historia evolutiva (Zhou & Mishra 2005). Estos genes podrían constituir un problema ya que los cebadores diseñados podrían hibridar en varios lugares del genoma, dando lugar a varios productos de PCR. Además, también se eliminaron los intrones cuyas tasas evolutivas no se ajustaran a las globales del genoma y presentaran aceleraciones o deceleraciones en algún grupo de organismos. Estos cambios en la tasa evolutiva en un linaje en particular podrían explicarse por fenómenos como cambios de función o de localización cromosómica, por ejemplo.

Igualmente se descartaron así parálogos ocultos que no hubieran sido detectados por los filtros anteriores. También fue fundamental descartar los intrones más conservados. Esto se puede comprobar en la figura 11, que representa la divergencia estimada para los intrones a lo largo de los distintos procesos de filtrado. En concreto, es apreciable la disminución relativa del tamaño de la primera barra del histograma de la figura correspondiente al conjunto de 224 intrones (figura 11D), que representaría a los intrones más conservados y que por tanto han acumulado menor divergencia. Probablemente, se trata de intrones implicados en alguna función importante, lo que impone constreñimientos que impide que tenga lugar variación. Por tanto, este tipo de intrones no son útiles en las filogenias de especies cercanas. Por otro lado, y a pesar de no haber empleado medidas de polimorfismo en los filtros, los intrones seleccionados también son ligeramente más variables intraespecíficamente que la media del genoma, ya que contienen más SNPs de lo esperado, lo que prueba de nuevo su utilidad en estudios filogenéticos de especies cercanas

1.2.1 Utilidad experimental de los intrones seleccionados para la filogenia de especies cercanas

Mediante la amplificación en laboratorio de varios intrones, se determinó que una buena parte de los marcadores testados fue efectivamente amplificable (tabla 2), con optimizaciones mínimas, en un grupo variado de especies correspondientes a varios órdenes de mamíferos. De cara a la aplicación de estos marcadores en estudios posteriores de filogenia y filogeografía, es importante destacar que la tasa de éxito de amplificación de los marcadores dependerá, lógicamente, de la especie escogida. Sin embargo, por norma general, cabría esperar que sean necesarios pequeños ajustes como variaciones en la temperatura de hibridación de la PCR o modificaciones en la especificidad de los cebadores variando el grado de degeneración de los aquí propuestos. Además, también se debe tener en cuenta que el objetivo de los cebadores diseñados era conseguir el mayor grado de amplificación interespecífica posible para poderlos emplear en un gran número de órdenes de mamíferos. Sin embargo, en estudios centrados en taxones más reducidos, como familias o géneros, puede resultar conveniente ajustar la especificidad y conseguir así reacciones de amplificación y secuenciación más eficaces.

Por otro lado, la variabilidad acumulada entre las distintas parejas especies para los marcadores amplificados reveló la ausencia de patrones claros en la acumulación de estas diferencias (figura 13). Esto sugiere que no es posible predecir que un intrón con muchas sustituciones en un linaje determinado las tendrá igualmente en otro linaje no

muy relacionado. Por ejemplo, los intrones más divergentes (y por tanto, más informativos) en la pareja humano – chimpancé son CARHSP-1 y OSTA-5, mientras que para la pareja comadreja – turón serían SLC38A7-8 y COPS7A-4. Además, el intrón JMJD5-2, que, como ya se ha comentado (figura 12), era el de mayor divergencia global calculada para los mamíferos no es siempre el más variable en las cinco parejas de especies. Por tanto, los intrones más variables son diferentes para cada linaje. Esto puede ser debido a la estocasticidad de las mutaciones, que afecta de forma especial a las ramas cortas, y también a diferencias en los tamaños poblacionales y a la presencia de polimorfismos ancestrales, que pueden constituir una importante parte de la divergencia existente entre genes de especies cercanas (Zhang & Hewitt 2003; Edwards & Beerli 2000). Por consiguiente, y dado que todos los intrones analizados resultaron ser suficientemente variables, puede no ser necesario primar el uso de intrones con tasas de evolución globales superiores en el conjunto de mamíferos o en un linaje concreto y no relacionado con los taxa de estudio. En cambio, sería más adecuado usar un número suficiente de intrones no ligados para atenuar los problemas estocásticos asociados a los procesos mutacionales y de coalescencia, especialmente importantes para las secuencias nucleares.

2. FILOGEOGRAFÍA DEL DESMÁN IBÉRICO EMPLEANDO DATOS MITOCONDRIALES Y NUCLEARES

2.1 Resultados

2.1.1. Análisis filogeográfico mitocondrial

Se obtuvieron 134 muestras de desmán ibérico recolectadas de 115 localidades que abarcaban gran parte de la distribución de la especie y todas las cuencas hidrográficas del área (tabla A3) (figura 14A). Para cada muestra, se obtuvo la secuencia completa del citocromo *b* y un fragmento del D-loop, llegando la concatenación de ambas a un total de 1482 bases nucleotídicas. Se obtuvo un total de 35 haplotipos distintos. La genealogía haplotípica correspondiente reconstruida a partir de un árbol filogenético de máxima verosimilitud (figura 14B) y un árbol calculado aplicando inferencia bayesiana y un reloj molecular (figura 14C) revelaron la presencia de cuatro clados o linajes mitocondriales distintos, denominados aquí A1, A2, B1 y B2. Además, se halló una importante correspondencia entre la pertenencia de las muestras a un clado mitocondrial determinado y su localización geográfica, lo que determinó una distribución parapátrica de estos linajes. Aún así, se encontró cierta mezcla de haplotipos en la zona de contacto de los linajes B1 y B2 en la parte oriental de la Cordillera Cantábrica. Sin embargo, el patrón más destacable se observó en la zona de contacto entre los linajes A2 y B1, situada en el centro del Sistema Ibérico, en la que los individuos correspondientes a ambos clados están separados por una estrecha franja de pocos kilómetros, sin que se haya encontrado hasta ahora intercambio alguno de haplotipos en las muestras secuenciadas.

La posición de la raíz en el árbol filogenético es importante para entender la evolución de las poblaciones de desmán y poder evaluar la importancia relativa de las zonas de contacto. Sin embargo, en este caso, la secuencia de la especie más cercana, *D. moschata*, presenta demasiada divergencia como para ser empleada como un grupo externo adecuado. El árbol enraizado que se obtuvo a partir del análisis bayesiano señalaba un alto soporte estadístico para la agrupación de los clados B1 y B2, pero no así para la agrupación de A1 y A2. Por ello, se analizaron de forma detallada las secuencias proteicas deducidas de las secuencias obtenidas del citocromo *b*, incluyendo además las correspondientes a *D. moschata* (que sí pueden resultar informativas a nivel proteico) y otras especies de tálpidos. Así, se comprobó que uno de los pocos cambios aminoacídicos encontrados en esta filogenia era compartido exclusivamente por todas las secuencias de los clados A1 y A2, localizado en la posición 329 de la proteína (figura 15). Por tanto, ya que un cambio no sinónimo (de alanina a treonina,

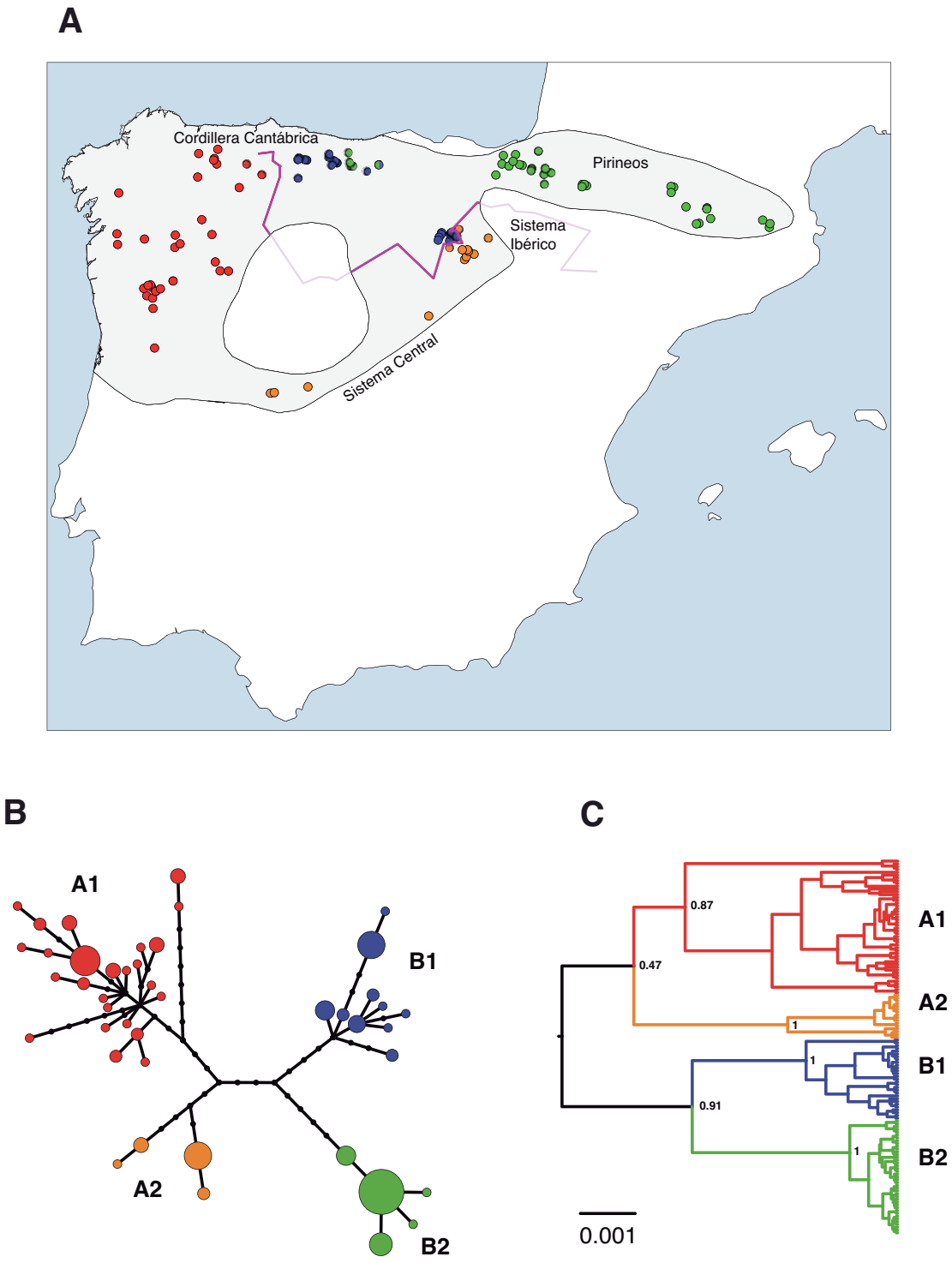

Figura 14. **(A)** Mapa de la Península Ibérica que muestra las 134 muestras de *G. pyrenaicus* usadas en este estudio. El área sombreada representa la distribución histórica de la especie de acuerdo a diversas fuentes de datos. Los colores de las muestras indican los cuatro clados mitocondriales obtenidos en los análisis filogenéticos (A1, A2, B1 y B2). La única localidad con dos muestras pertenecientes a clados distintos (B1 y B2) se muestra con los dos colores correspondientes. La línea de color púrpura indica la barrera genética identificada con el algoritmo de Máxima Diferenciación de Monmonier. **(B)** Genealogía haplotípica de las secuencias mitocondriales concatenadas basada en un árbol de máxima verosimilitud. Los círculos representan haplotipos, siendo el tamaño proporcional al número de individuos. Los puntos negros representan posibles haplotipos intermedios no muestreados. **(C)** Árbol de inferencia bayesiana de las mismas secuencias. Se muestran las probabilidades posteriores de los clados relevantes

concretamente) es muy raro, debe haber ocurrido una sola vez, en el linaje que lleva desde el ancestro común de A1 y A2 hasta el ancestro de todos los linajes mitocondriales de desmán. Por consiguiente, esta sustitución aminoacídica soporta la topología indicada en la Figura 14C como la más probable para representar las relaciones filogenéticas entre los cuatro linajes mitocondriales hallados en la especie. Además, y también de acuerdo a esta posición de la raíz, que indicaría que la mayor divergencia debería ocurrir entre los grupos de clados A (que contiene A1 y A2) y B (con B1 y B2), el algoritmo de máxima diferencia de Monmonier identificó la mayor distancia genética precisamente en las dos zonas de contacto de estas dos agrupaciones de clados (figura 14A).

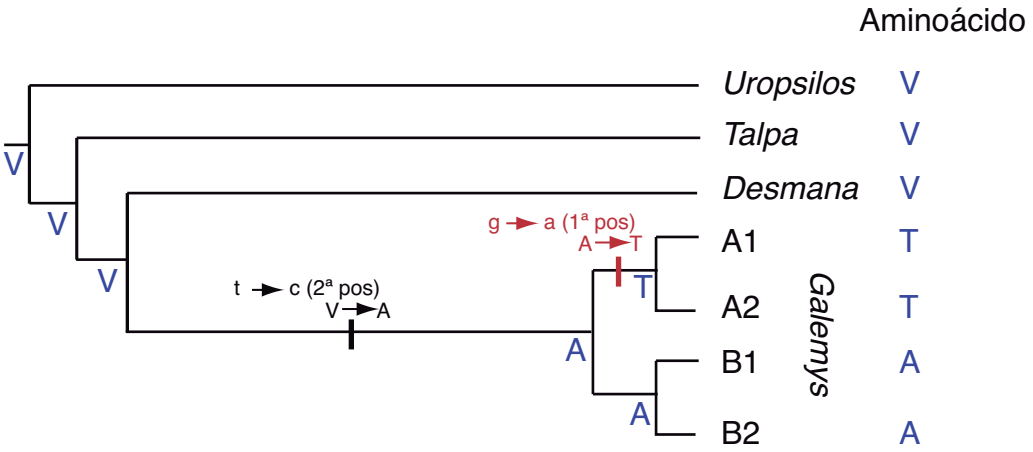

Figura 15. Reconstrucción por parsimonia de los cambios aminoacídicos a lo largo de la filogenia del citocromo *b* de los linajes de los tálpidos y *G. pyrenaicus*. Los estados actuales de los aminoácidos en la posición 329 de la proteína para cada especie se muestran en color azul, al igual que los estados ancestrales deducidos para cada nodo. Los dos cambios deducidos para esta posición se representan por una barra vertical en las ramas correspondientes. Para cada cambio, se muestra la posición en el codón y la sustición aminoacídica. El cambio compartido por los linajes A1 y A2 aparece en color rojo.

2.1.2. Diversidad genética mitocondrial

La diversidad genética total del desmán ibérico calculada a partir de las secuencias mitocondriales concatenadas es relativamente baja (tabla 3), con un valor de diversidad nucleotídica (π) de 0.0073 (para efectos comparativos, el resultado calculado únicamente a partir del citocromo *b* es de 0.006). Los valores de diversidad correspondientes a cada clado presentaron importantes diferencias entre sí, de hasta un orden de magnitud entre el clado más diverso (el A1) y el menos diverso (B2). Además, se detectó una expansión poblacional significativa en el clado A1 con los estadísticos R2 y Fs de Fu, pero no con la D de Tajima. En este sentido, se ha sugerido que la D de Tajima puede no ser el estadístico adecuado para detectar este tipo de

fenómenos en algunas condiciones, mientras que R2 y Fs de Fu han sido empleados para detectar desviaciones de un modelo de tamaño poblacional constante bajo una gran variedad de condiciones.

El uso de excrementos podría llevar a una subestima de los valores de diversidad genética si se emplearan muestras del mismo individuo. En este estudio, como se menciona en el apartado III.2.1.1, se seleccionaron sólo excrementos separados al menos por 1 kilómetro para evitar este problema, si bien también se incluyeron algunas muestras separadas por distancias inferiores si los haplotipos obtenidos eran distintos. Para comprobar si este esquema de selección de muestras podría estar generando una muestra sesgada, se calculó la diversidad genética de forma separada para excrementos y tejidos (tabla 3). Los resultados fueron muy similares para ambos conjuntos de datos, y también al compararlos con el total. Cuando los cuatros clados fueron analizados de igual manera, los valores de diversidad nucleotídica resultantes también eran muy parecidos, excepto para el clado A2 (probablemente debido al bajo tamaño muestral de este clado). Estos resultados indicaron que el esquema de muestreo de recolección de excrementos no introdujo sesgo alguno en los resultados de diversidad genética.

Tabla 3. Diversidad genética mitocondrial y estadísticos de expansión poblacional de las secuencias concatenada del citocromo *b* y el fragmento del D-loop. Los cálculos se realizaron para toda la especie y para los cuatro clados derivados de los análisis filogeográficos mitocondriales.
N = número de secuencias; S = número de sitios segregantes; h = número de haplotipos; Hd = diversidad haplotípica; R_2 = estadístico R_2 de Ramos & Rozas. Las desviaciones significativas ($p < 0.05$) del modelo se indican con un asterisco**.**

Parámetro	Total	A1	A2	B1	B2
N	134	48	16	29	41
S	72	40	7	15	4
h	44	25	4	10	5
Hd	0.935	0.927	0.592	0.842	0.534
π	0.0073	0.0036	0.0016	0.0024	0.0004
π (Tejidos)	0.0070	0.0038	0.0016	0.0028	0.0004
π (Heces)	0.0071	0.0029	0.0002	0.0024	0.0004
D Tajima	-0.545	-1.399	0.416	-0.202	-0.832
R2	0.074	0.062 (*)	0.163	0.112	0.083

Para apreciar mejor las diferencias en diversidad genética existentes a lo largo de todo el rango de distribución del desmán, se calcularon los valores de π para cada localidad, de forma que, para cada una de ellas, se incluyeron en el cálculo todas las muestras presentes en un círculo de radio de aproximadamente 100 km. El mapa de contornos derivado de estos cálculos demuestra claramente que los mayores valores de diversidad ocurren en el noroeste de la Península Ibérica (figura 16). A partir de esta área, los

valores de diversidad disminuyen gradualmente hacia las partes más orientales de la distribución del desmán, alcanzando valores mínimos en los Pirineos. En este gráfico, sólo se consideraron muestras del mismo linaje mitocondrial a la hora de calcular π en cada localidad. Si, por el contrario, se incluyeran también las muestras pertenecientes a linajes en contacto secundario, el gráfico cambiaría y las zonas de mayor diversidad serían, lógicamente, las zonas de contacto entre los linajes (no mostrado)

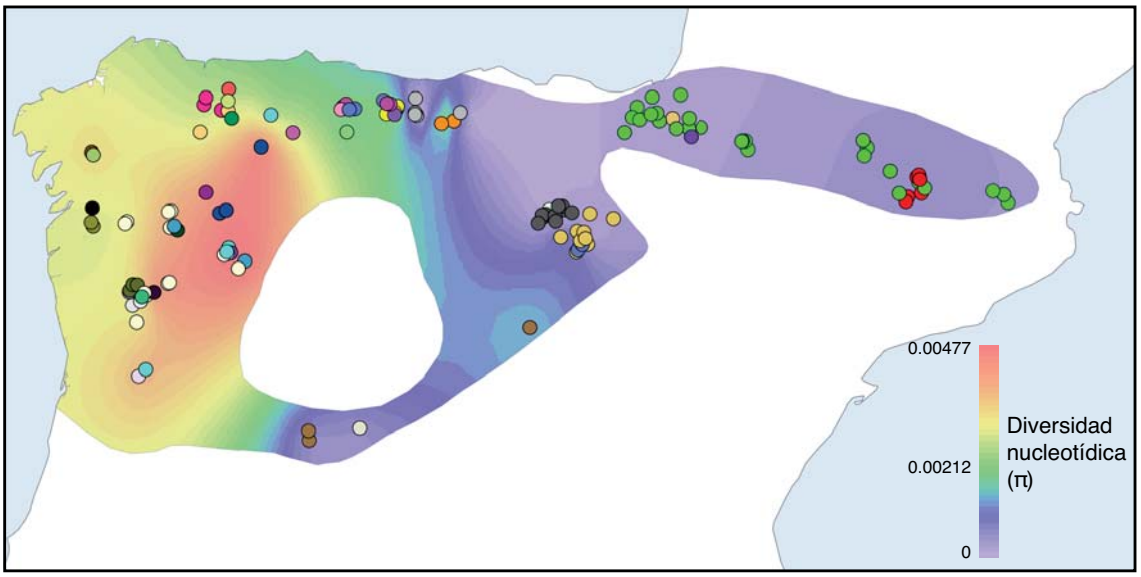

Figura 16. Mapa de contornos de la diversidad genética (π) de *G. pyrenaicus*. Sólo se consideraron muestras del mismo linaje mitocondrial para calcular π en cada localidad. Los colores indican la diversidad genética calculada. Para los puntos de muestreo, se usó un color diferente para cada haplotipo y, por tanto, una mayor variedad de colores en un área también indica mayor diversidad. Los puntos de muestreo se reposicionaron aleatoriamente en un radio de 5 km para poder mostrar las muestras de la misma localidad.

Por último, se analizó además la estructura de la diversidad genética mediante un análisis de la varianza molecular (AMOVA). Al agrupar las muestras de desmán de acuerdo a su cuenca hidrográfica (figura 17), se halló que estas agrupaciones explicaban un 32% de la variación genética. Sin embargo, este valor podría estar aumentado artificialmente por una elevada correlación entre las distancias genéticas y las geográficas ($r = 0.5$, $p = 0.001$), que indica que existe un pronunciado patrón de aislamiento por distancia (*isolation by distance*) en este sistema (Meirmans 2012). Restringiendo este análisis al clado A1, donde el efecto del aislamiento por distancia es mucho menor ($r = 0.15$, $p = 0.017$), sólo el 15.6% de la variación genética fue explicado por el agrupamiento de muestras siguiendo los sistemas fluviales principales.

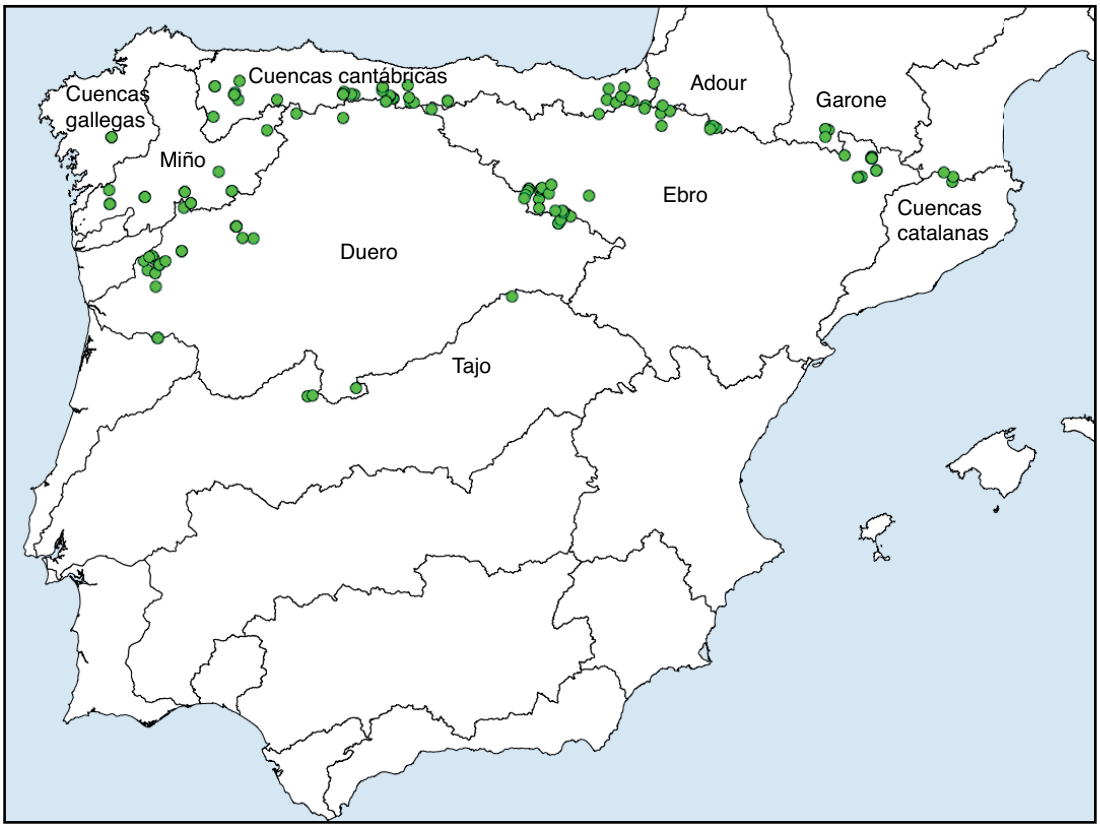

Figura 17. Muestras de *G. pyrenaicus* agrupadas por cuencas para el análisis AMOVA

2.1.3. Diversidad genética nuclear

Para un total de 29 desmanes procedentes de los cuatro clados mitocondriales, se secuenciaron ocho intrones nucleares (Igea *et al.* 2010), seleccionados por presentar evolución neutra y rápida. La idoneidad de estos intrones para este estudio se comprobó al detectar un importante número de diferencias entre las secuencias de *Galemys pyrenaicus* y *Desmana moschata*, descartándose así que estuvieran sujetos a conservación funcional en el linaje de los desmaninos. Sin embargo, el análisis intraespecífico de las secuencias obtenidas de estos intrones reveló que sólo cinco de ellos mostraban variabilidad en *G. pyrenaicus* (figura 18). Además, el número de alelos distintos para cada locus era muy bajo. Por consiguiente, la diversidad nucleotídica media de los intrones fue muy baja (π = 0.00034, tabla 4). Del total de 232 intrones secuenciados, únicamente 10 resultaron heterocigotos (lo que supone una heterocigosidad media de 0.043), y cada uno de ellos sólo contaba con una posición variable. Se examinaron también las diferencias a nivel de diversidad genética nuclear entre los distintos clados mitocondriales y, de nuevo, la mayor diversidad se observó en las muestras pertenecientes al clado A1 (tabla 4**)**.

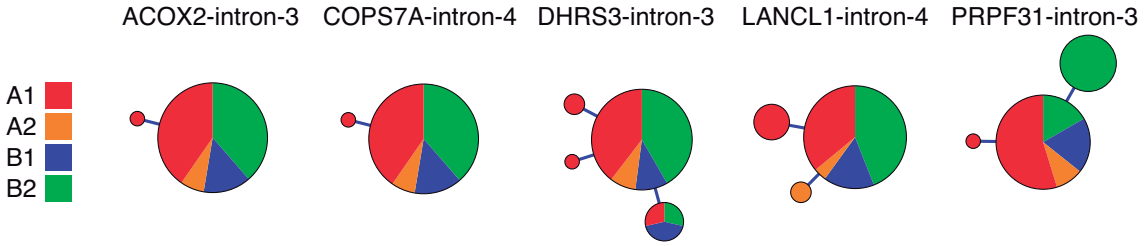

Figura 18. Genealogías haplotípicas de los cinco intrones que presentaron variabilidad en *G. pyrenaicus*. El tamaño de los círculos es proporcional al número de individuos. Los colores indican los cuatro linajes mitocondriales (A1, A2, B1 y B2) a los que pertenecen los individuos.

Tabla 4. Diversidad genética nuclear (π) de los 8 marcadores nucleares calculados para la especie y para los cuatro clados derivados del análisis filogeográfico mitocondrial

Intron	Total	A1	A2	B1	B2
ACOX2-3	0.00009	0.00022	0	0	0
ACPT-4	0	0	0	0	0
COPS-4	0.00005	0.00013	0	0	0
DHRS3-3	0.00122	0.00155	0	0.00206	0.00067
LANCL1-4	0.00049	0.00075	0.00127	0	0
PRPF31-3	0.00087	0.00017	0	0	0.00093
ROGDI-7	0	0	0	0	0
SMYD4-5	0	0	0	0	0
Media	0.00034	0.00035	0.00016	0.00026	0.00020

La comparación de los intrones del desmán ibérico con sus ortólogos en el desmán ruso permitió establecer las mutaciones derivadas en los SNPs de *G. pyrenaicus*. Si bien la presencia de la mayoría de las mutaciones estaba limitada a un solo individuo, tres de ellas estaban suficientemente dispersas por el rango geográfico de la especie como para ser informativas respecto al grado de conectividad de las poblaciones. Dos de estas mutaciones estaban confinadas en uno solo de los linajes mitocondriales, pero una tercera, en el intrón DHRS3-3, estaba presente en tres de los clados: A1, B1 y B2 (figura 19**)**.

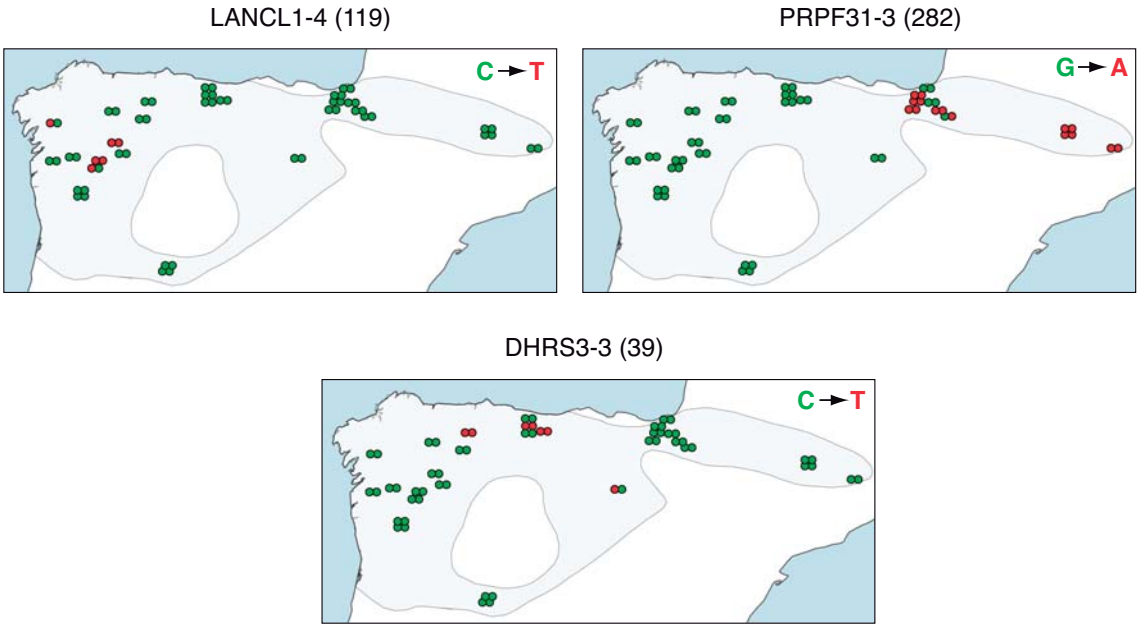

Figura 19. Distribución geográfica de las variantes de 3 SNPs. Para cada espécimen, los alelos de ambos cromosomas están representados con dos puntos adyacentes. La mutación derivada (deducida usando el grupo externo *Desmana*) se muestra en color rojo, Algunos puntos se han desplazado latitudinalmente para poder mostrar especímenes de la misma localidad.

2.1.4. Estima del tiempo al ancestro común más reciente (tmrca) de las secuencias mitocondriales

Para calcular el tiempo hasta el ancestro común más reciente de las secuencias mitocondriales, se siguió un proceso en dos fases. En una fase inicial, se estimó de forma precisa el tiempo de divergencia entre *Galemys pyrenaicus* y *Desmana moschata* usando las secuencias de ocho intrones nucleares junto con los ortólogos correspondientes a ocho especies de mamíferos con genomas secuenciados. Se empleó un análisis bayesiano con un reloj molecular relajado y una serie de fósiles de datación fiable como constreñimientos *a priori* para varios nodos del árbol filogenético. Previamente se evaluó la interacción de los diferentes *priors* de calibración, y se comprobó que las distribuciones de los *priors* efectivos estaban incluidas dentro de las correspondientes a los *priors*.

La media de la divergencia estimada entre el desmán ibérico y el ruso fue de 13.9 millones de años (figura 20A). Este valor es ligeramente superior a estimas previas basadas sólo en secuencias mitocondriales y con un menor número de calibraciones fósiles incluidas. Sin embargo, es congruente con la presencia más temprana en el registro fósil de Desmaninae, fechada en 8.2 millones de años (Fortelius 2012), si bien de esta forma se extiende el análisis del clado unos 5 millones de años hacia el pasado.

Esta estima de divergencia fue incluida en la segunda parte del análisis, que consistió en un análisis bayesiano de secuencias de citocromo *b* de *Galemys*, *Desmana* y secuencias de otros tálpidos empleadas como grupo externo. La tasa evolutiva obtenida para el citocromo *b* fue de 0.0145 sustituciones/posición/Ma, y la fecha resultante para el ancestro común más reciente de los haplotipos de *Galemys pyrenaicus* fue de 0.5 Ma, con un 95% de la máxima densidad posterior entre 0.2 y 0.8 Ma, lo que la sitúa en el Pleistoceno Medio (figura 20B). Cabe recordar que esta fecha representa la coalescencia de los linajes mitocondriales estudiados, pero las poblaciones pudieron haber divergido en fechas mucho más recientes. En definitiva, se trata de un límite superior a la divergencia máxima de las poblaciones de desmán ibérico actuales.

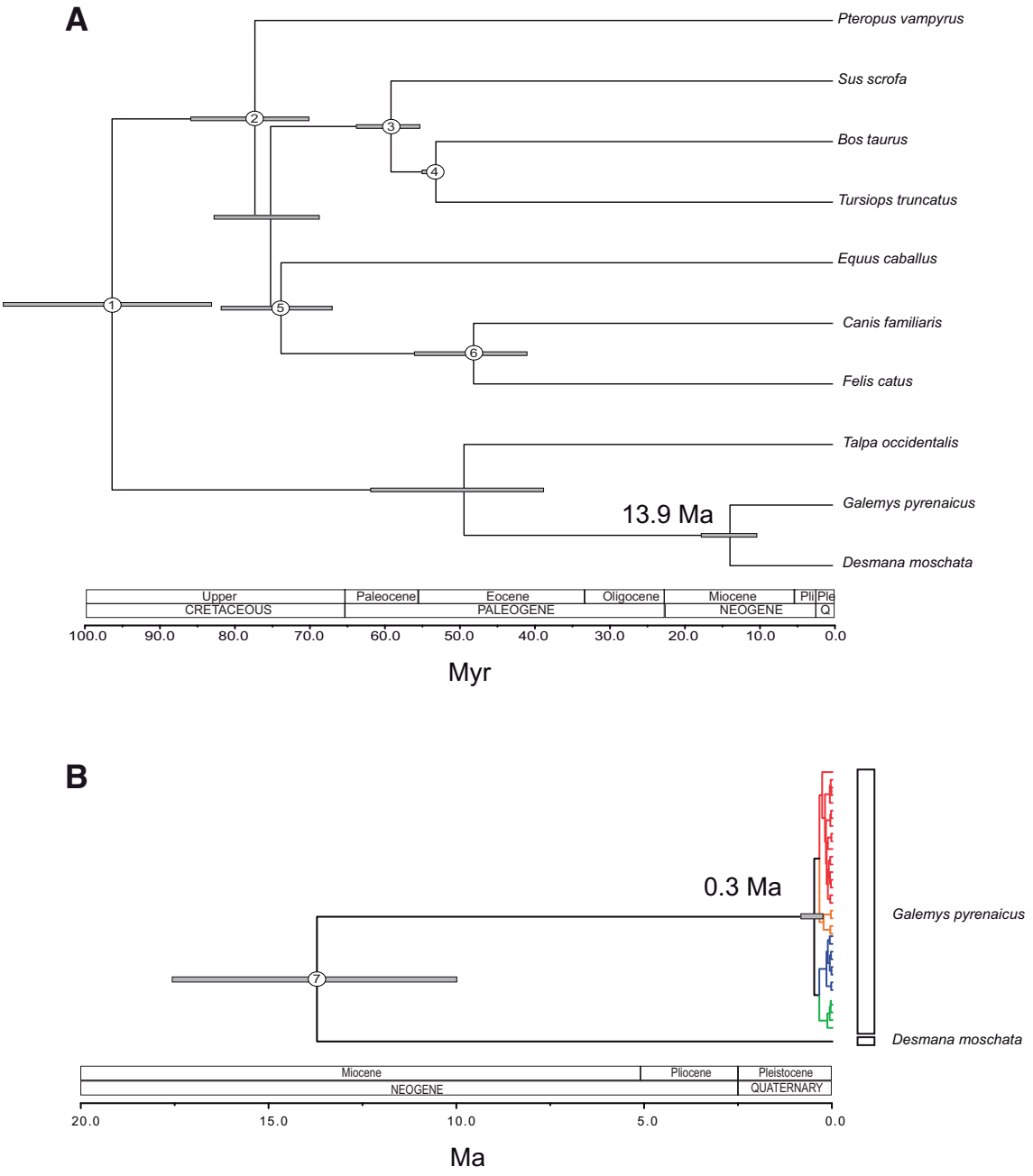

Figura 20. **(A)** Datación bayesiana con secuencias nucleares de mamíferos para estimar la divergencia de *Galemys* y *Desmana*. Se emplearon calibraciones fósiles correspondientes a Laurasiatheria (1), Ferungulata (2), Cetartiodactyla (3), Cetruminantia (4), Zooamata (5) y Carnivora (6). Las barras grises indican el intervalo del 95% HPD. **(B)** Datación bayesiana de las secuencias mitocondriales de *G. pyrenaicus* usando el nodo *Galemys – Desmana* como punto de calibración. Los colores indican los cuatro linajes mitocondriales de *G. pyrenaicus* obtenidos en el análisis filogeográfico. No se muestran las secuencias del grupo externo correspondientes a la subfamilia Talpinae.

2.1.5. Modelado de la distribución de la especie en el LGM

Con el objetivo de estudiar la relación entre el evidente gradiente de diversidad genética observado entre los linajes actuales de desmán (figura 16) y la posible localización de refugios glaciales, se construyó un modelo de distribución para la especie basado en los datos de ocurrencia actuales (figura 21). Cuando se proyectó este modelo en las condiciones del LGM, se observó que las máximas probabilidades de presencia potencial correspondían de nuevo con la zona noroeste de la Península, que es el área con una mayor diversidad genética en la actualidad (figura 16).

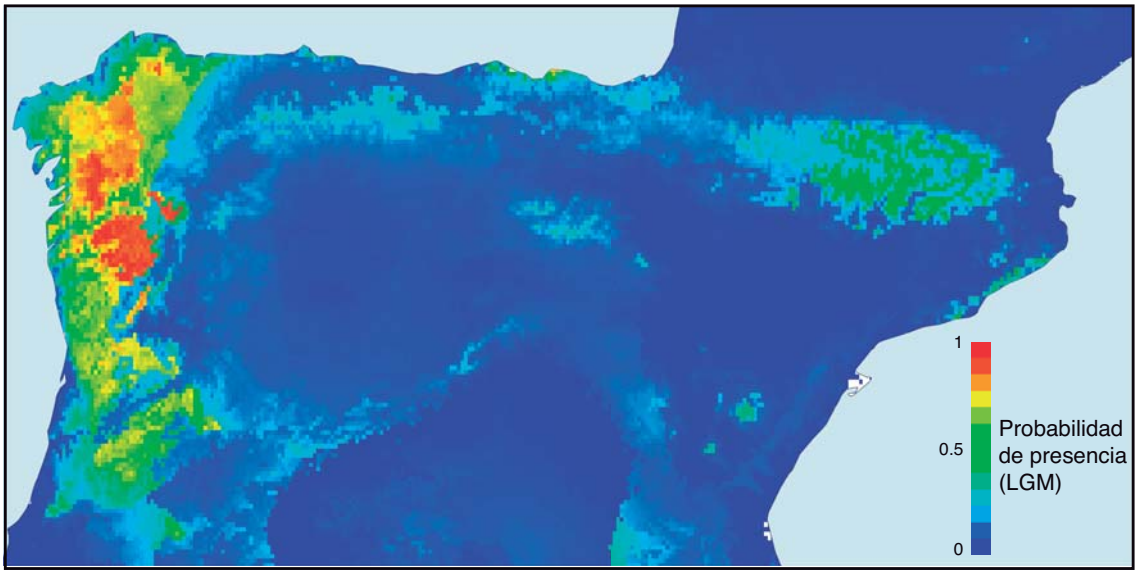

Figura 21. Distribución potencial de *G. pyrenaicus* durante el Último Máximo Glacial (LGM, *Last Glacial Maximum*) estimada por Maxent. Los colores indican la probabilidad de presencia.

2.2 Discusión

2.2.1. Evolución pleistocénica de las poblaciones de desmán ibérico

Se han presentado una serie de evidencias sólidas que indican que las glaciaciones pleistocénicas ejercieron una fuerte influencia en la historia evolutiva y la estructura genética de las poblaciones del desmán ibérico. En primer lugar, la filogenia de los genes mitocondriales exhibe un claro patrón geográfico, con cuatro linajes mitocondriales de marcada distribución parapátrica. Además, un análisis bayesiano usando múltiples calibraciones fósiles revela que la divergencia de los clados mitocondriales más separados se produjo en el Pleistoceno Medio. Es probable, por tanto, que estos linajes quedaran separados en cuatro refugios glaciales aislados (figura 22). A pesar de la existencia de varios ciclos glaciales recurrentes, se ha demostrado

que los datos mitocondriales reflejan probablemente la primera separación poblacional, debido a la mezcla incompleta de los linajes en los periodos interglaciales (Hofreiter *et al.* 2004).

Además, existen más evidencias que apoyan que el noroeste de la Península Ibérica sirvió como el refugio principal a las poblaciones de desmán. En primer lugar, los modelos de distribución de la especie proyectados a las condiciones del Último Máximo Glacial predicen las máximas probabilidades potenciales de ocurrencia precisamente en esta zona de la Península Ibérica (figura 21). Además, tal y como se esperaría en un escenario filogeográfico clásico en el que únicamente parte de la reserva genética perteneciente al refugio glacial coloniza nuevas áreas (probablemente a través de sucesivos cuellos de botella), la mayor diversidad genética actual se encontró en el Noroeste de la Península, disminuyendo gradualmente hacia la zona del Este. Por otra parte, se detectó una expansión poblacional pasada en este linaje mitocondrial. Esta zona de la Península Ibérica ha servido además como refugio glacial a otras especies dependientes del medio acuático como la salamandra rabilarga (*Chioglossa lusitanica*) (Alexandrino *et al.* 2000). Es probable que este área presentara, por tanto, unas condiciones óptimas de temperatura y pluviometría para estas especies durante los sucesivos ciclos glaciares. Aún así, y a pesar de la mayor importancia relativa de este refugio para las poblaciones de desmán, los datos recuperados y, en concreto, la presencia de cuatro linajes mitocondriales con divergencias datadas en el Pleistoceno Medio, sugieren también la presencia de otros tres refugios adicionales.

Las poblaciones pirenaicas del desmán ibérico, pertenecientes al clado B2, son muy homogéneas genéticamente, como se deduce de los mínimos valores de diversidad nuclear y mitocondrial. Por tanto, probablemente fueron originadas por la colonización desde un refugio distante, sufriendo algún proceso de cuello de botella. Este refugio pudo estar situado en el centro de la distribución actual de este clado, en los Montes Vascos, como ya se ha sugerido de acuerdo a la distribución actual de la especie (Aymerich & Gosálbez 2002). Desde este posible refugio, las poblaciones de este clado pudieron haberse dispersado hacia las áreas ocupadas actualmente por los individuos del clado B1, ya que se observa mezcla de ambos clados en algunos ríos en la actualidad.

Un refugio adicional, situado en la Cordilllera Cantábrica, pudo haber dado lugar a las poblaciones englobadas en el clado B1. Tanto el área de los Montes Vascos como el de la Cordillera Cantábrica han sido postulados con anterioridad como refugios glaciales potenciales para muchas especies, tanto de mamíferos (Deffontaine *et al.* 2009) como

de plantas (Magri 2008), asociados a climas húmedos. A partir de este refugio del clado B1, también debió ser colonizada la zona noroccidental del Sistema Ibérico (Sierra de la Demanda). Sin embargo, la dispersión, al menos a nivel mitocondrial, podría haberse visto frenada a continuar hacia las zonas más surorientales (sierras de Cameros, Urbión y Cebollera), como se indica por la ausencia total de linajes de este clado en estas áreas. En este sentido, la dispersión del clado B1 hacia la parte occidental de su posible refugio también se podría haber visto limitada, al menos en lo referente a las hembras, pues de nuevo en este caso no se observa presencia alguna del linaje B1 en las áreas ocupadas por el clado A1.

Por último, un tercer refugio situado en el Sistema Ibérico o en el Sistema Central pudo haber dado lugar a las poblaciones del clado A2, si bien su localización no estaría nada clara debido a la escasez de datos para este clado. Dada la poca distancia genética que separa este clado al noroccidental A1, el refugio pudo estar situado en el Sistema Central. En este caso, parte del Sistema Ibérico (las sierras de Cameros, Urbión y Cebollera) habrían sido colonizadas recientemente, pero de nuevo la dispersión, al menos de las hembras, no habría progresado hacia las zonas más noroccidentales como la Sierra de la Demanda.

La dispersión de los linajes mitocondriales a partir de sus refugios putativos periféricos en distintas direcciones y la ausencia de áreas apropiadas para la especie en la zona central de la distribución han creado una distribución de forma circular (figura 22). Además, y por efecto de las dispersiones interrumpidas de los distintos clados ya explicadas, se han creado dos importantes brechas genéticas. Esto implica que hay dos zonas de contacto entre los dos linajes principales (A y B), una en el medio del Sistema Ibérico y otra en el medio de la Cordillera Cantábrica (figura 14A). La distribución histórica de *G. pyrenaicus* (representada en gris en la figura 22) es mayor que las áreas muestreadas (mostradas en diferentes colores). En todo caso, en la mayoría de estas zonas fuera del área de muestreo, el desmán ibérico está presente en densidades muy bajas o incluso ha desaparecido. Por tanto, aunque parece improbable que existan otras zonas de contacto o mezcla en esta especie, no puede descartarse completamente, y es posible que muestreos adicionales en lugares no visitados puedan producir otros resultados interesantes.

Por tanto, la historia evolutiva del desmán coincide con un escenario de "refugios dentro de refugios", según el cual las penínsulas del Sur de Europa no pueden ser vistas como refugios glaciales homogéneos sino más bien como centros en los cuales se

formaron varios refugios relativamente independientes entre sí y que dieron lugar a linajes evolutivos diferenciados dentro de las especies (Gómez & Lunt 2007). Los resultados aquí presentados amplían esta hipótesis demostrando que las penínsulas no sólo fomentaron la creación de complejos mecanismos de aislamiento, sino también los procesos completos de contracción y dispersión asociados a las glaciaciones, dejando así profundas huellas en la estructura genética actual de especies endémicas como el desmán ibérico. Hasta ahora, estas claras señales genéticas sólo habían sido observadas en especies de distribución continental, pero en este caso quedan analizadas también a una escala peninsular.

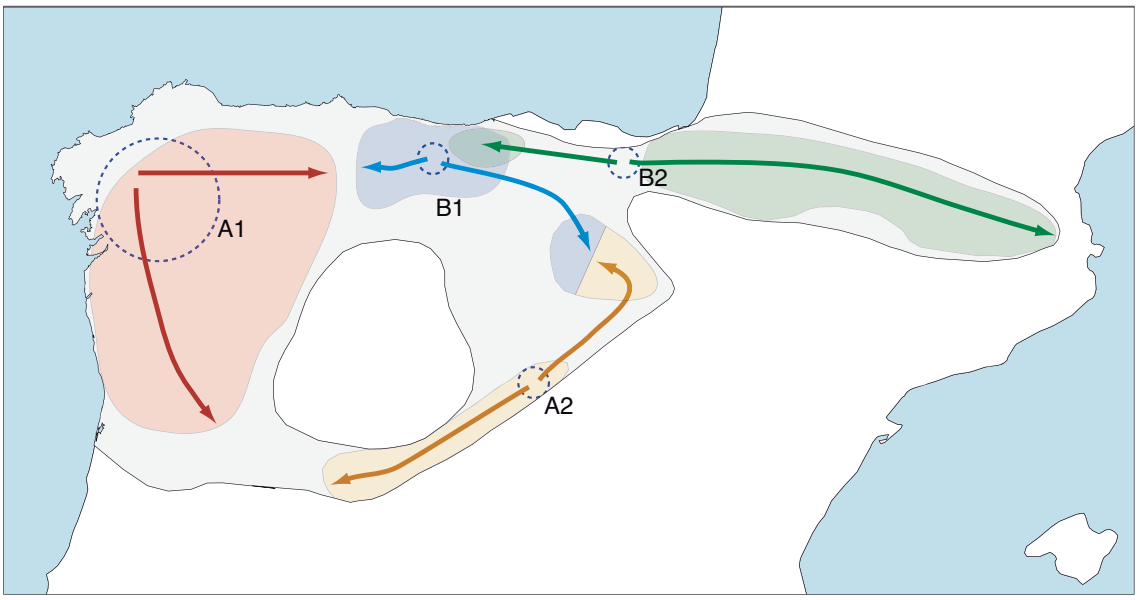

Figura 22. Representación esquemática de la historia evolutiva de *G. pyrenaicus*. El área sombreada representa la distribución histórica de la especie. Las áreas muestreadas de los cuatro linajes mitocondriales se muestran en distintos colores. Las localizaciones hipotéticas de los refugios glaciales se indican con círculos punteados dentro de la distribución actual de cada linaje. El tamaño de los círculos representa la importancia relativa de los refugios, y las rutas de colonización se muestran por medio de flechas.

2.2.2 Influencia de los requerimientos acuáticos en la estructura genética del desmán ibérico

El desmán ibérico es una especie de estrictos requerimientos acuáticos, con una gran dependencia de aguas limpias. Por tanto, cabría esperar que estos requerimientos tuvieran un efecto en su estructura genética, de forma que ésta estuviera muy influenciada por los ríos y las cuencas hidrográficas. Sin embargo, la proporción de la variación genética atribuible a los agrupamientos por cuencas fluviales fue menor de lo esperado según esta hipótesis. Aunque no se puede descartar la detección de mayores niveles de diferenciación empleando marcadores con un mayor grado de polimorfismo genético, los resultados sugieren que, en condiciones naturales, los desmanes pueden

cruzar fácilmente de una cuenca hidrográfica a otra a través de las montañas. Por ejemplo, se hallaron haplotipos mitocondriales idénticos a ambos lados de los Pirineos o a ambos lados de la Cordillera Cantábrica. Esta presencia de haplotipos idénticos compartidos entre distintas cuencas es lo que explica la baja diferenciación existente entre ellas. Por tanto, se puede concluir que la estructura genética del desmán ha estado más influenciada por la historia de las glaciaciones pleistocénicas que por su distribución actual de hábitat, independientemente de que este hábitat especializado esté muy fragmentado. La situación de esta especie es, por tanto, intermedia entre organismos estrictamente acuáticos, cuya diversidad genética está más determinada por las cuencas hidrográficas (Avise 2000), y animales semiacuáticos de gran movilidad, como la nutria europea, cuya diversidad genética no está nada influenciada por esas cuencas (Mucci *et al.* 2010).

2.2.3. Fuertes señales de aislamiento en las zonas de contacto

El descubrimiento más sorprendente en relación con la estructura genética del desmán fue la existencia de estrechas zonas de contacto entre linajes mitocondriales divergentes en contacto secundario después de la colonización postglacial. En estas zonas de contacto no se da, aparentemente, mezcla de linajes a nivel mitocondrial, lo que crea dos brechas genéticas. La brecha genética más pronunciada se halló en el medio del Sistema Ibérico (figura 14A). Se recolectaron un total de 23 muestras en los seis ríos de la zona, y se encontraron individuos pertenecientes a los dos clados mencionados. Éstos, sin embargo, estaban completamente segregados y no se encontraron juntos en el mismo río en ningún caso. Por tanto, se puede trazar una línea de separación, que discurre a lo largo del valle del río Najerilla, y que parece limitar la dispersión de las hembras de desmán. La segunda brecha genética afecta también a los linajes principales A y B, y está situada en el medio de la Cordillera Cantábrica. Sin embargo, a pesar de realizar varias prospecciones en esta zona, no se pudieron encontrar más muestras que estrecharan la distancia entre los dos clados, de forma que no se puede determinar si existe o no mezcla de linajes a nivel mitocondrial en este caso. En cualquier caso, la falta de penetración de hembras de un linaje en el área de distribución del otro linaje es una circunstancia infrecuente, ya que aparentemente no existen barreras a la dispersión en ninguna de las dos zonas de contacto. Se han observado situaciones similares en otras especies (Swenson & Howard 2005), incluyendo anfibios y reptiles de la Península Ibérica (Alexandrino *et al.* 2000; Godinho *et al.* 2008; Martínez-Solano *et al.* 2006; Recuero & García-París 2011), y también en mamíferos, como el oso pardo (Waits *et al.* 2000). Sin embargo, en la mayoría de estos casos se observa cierto grado de permeabilidad a través de las zonas

de contacto, en contraste con la situación observada en el desmán ibérico, mucho más estricta. Además, la zona de contacto del desmán, al contrario que en otras especies estudiadas, ocurre entre linajes mitocondriales con una divergencia relativamente reciente, lo que hace más sorprendente si cabe la falta de mezcla entre ellos. Se ha sugerido que la saturación del hábitat en las zonas de contacto podría inhibir la migración de las hembras (Pelletier *et al.* 2011; Recuero & García-París 2011). Esto podría explicar por qué algunas de estas especies se han dispersado a partir de refugios glaciales durante cientos de kilómetros pero actualmente parecen incapaces de cruzar una franja estrecha de unos pocos kilómetros.

El análisis de las zonas de contacto se ha basado en datos mitocondriales y por tanto se refiere exclusivamente a las hembras. Por tanto, no está representada la historia de la especie en su totalidad. De hecho, en muchas ocasiones, cuando se han complementado los estudios con datos nucleares, se ha observado que estas barreras no eran fuertes o ni siquiera existían para estos marcadores, lo que indica una dispersión diferencial ligada al sexo (Godinho *et al.* 2008; Nater *et al.* 2011; Waits *et al.* 2000). Los datos referentes a intrones aquí presentados no muestran suficiente variabilidad en *G. pyrenaicus* como para estudiar estos aspectos en su totalidad. Sin embargo, tres variantes de los SNPs obtenidos (figura 19) mostraron suficiente extensión geográfica como para ser útiles en el análisis de la dispersión (Novembre & Ramachandran 2011). Los tres mutantes derivados mostraban una distribución contigua, lo que sugiere que se trata de mutaciones de origen reciente. De hecho, una de las tres mutaciones (presente en el intrón DHRS3-3) aparece a ambos lados de la zona de contacto de la Cordillera Cantábrica, lo que podría implicar que ha ocurrido dispersión ligada a los machos a través de esta zona de contacto. Sin embargo, sería necesaria la obtención de mayor cantidad de datos genéticos nucleares para estudiar estos fenómenos de forma cuantitativa. Hasta el momento, el radio-seguimiento de las poblaciones de desmán ibérico no ha mostrado ningún patrón claro de dispersión diferencial ligada al sexo (Melero *et al.* 2012).

2.2.4. Subespecies

La existencia de dos grandes clados mitocondriales en el desmán ibérico podría en principio corresponderse con las dos subespecies descritas, *G. pyrenaicus pyrenaicus* y *G. pyrenaicus rufulus*, pero la distribución de los clados mitocondriales y la localización de las zonas de contacto no encajan bien con ninguna de las áreas de distribución propuestas, que han variado bastante en los distintos trabajos realizados (González-Esteban *et al.* 1999; Juckwer 1990; López-Fuster *et al.* 2006). Si bien los

estudios iniciales sobre el tema estaban basados en unos pocos ejemplares, los más recientes emplearon un mayor número de individuos, pero éstos fueron agrupados por grandes regiones geográficas que englobaban varios de los clados mitocondriales aquí descritos. Por consiguiente, estudios dirigidos específicamente a analizar posibles gradientes morfológicos en las zonas de contacto mencionadas podrían ayudar a determinar la validez del estatus de las subespecies descritas.

2.2.5. Implicaciones en la conservación del desmán ibérico

El desmán ibérico está protegido legalmente en los cuatro países en los que está presente (España, Portugal, Francia y Andorra) y está catalogado como "vulnerable" por la Unión Internacional para la Conservación de la Naturaleza (IUCN). Además, las poblaciones del Sistema Central, en la zona más meridional de su distribución, han sido recientemente clasificadas como "en peligro de extinción" (la categoría de protección más alta) por el Gobierno de España. Por tanto, se trata de uno de los mamíferos más amenazados en la Península Ibérica y, por extensión, en el continente europeo. De hecho, los datos más recientes parecen indicar una reducción sustancial de las poblaciones de la especie en el Sistema Central (Nores *et al.* 2007). Las propias prospecciones en la zona realizadas para este trabajo no obtuvieron ningún excremento de desmán en ninguna localidad, a pesar de tratarse de lugares donde se había capturado algún ejemplar de esta especie en las últimas décadas. Por tanto, se hubo de recurrir a las muestras procedentes de museos para completar el muestreo genético de la especie.

La diversidad genética del desmán ibérico es muy baja en todo su rango de distribución, tanto a nivel mitocondrial como nuclear. A nivel mitocondrial, para el cual hay gran cantidad de datos para su comparación, la diversidad nucleotídica de la especie es aproximadamente cuatro veces menor que la media de los mamíferos (Nabholz *et al.* 2008b) y es incluso menor en áreas como los Pirineos. A pesar de que estos marcadores no muestran suficiente resolución como para mostrar si este nivel de diversidad genética supone alguna amenaza directa, es importante tener estos datos en cuenta en caso de futuros cambios ambientales imprevistos, que podrían ser más perjudiciales para poblaciones de baja diversidad genética.

Las inferencias acerca de los movimientos y la dispersión del desmán ibérico también pueden tener implicaciones importantes para su conservación. En particular, los datos de diferenciación genética permiten inferir que los desmanes no están confinados a las cuencas de los ríos donde habitan, sino que se pueden mover a través de las cuencas

hidrográficas, probablemente incluyendo movimientos terrestres a larga distancia. Por tanto, estos datos indican que las unidades de gestión no deberían corresponder estrictamente a las cuencas fluviales ocupadas sino que deberían considerar grandes áreas que incluyan corredores entre cuencas para favorecer el intercambio natural entre las poblaciones.

Los programas de conservación de la especie también deberían tener en cuenta la delimitación de la especie en cuatro linajes mitocondriales distintos que se ha hallado en este trabajo. Estas poblaciones comenzaron a divergir en el Pleistoceno Medio y su integridad debería ser preservada para evitar problemas de depresión por *outbreeding*. (Frankham *et al.* 2011) A pesar de todo, se hacen necesarios estudios más detallados para determinar con claridad el grado de flujo genético existente entre los clados en los distintos sexos. Por tanto, y a falta de estos estudios detallados, deberían considerarse los distintos linajes mitocondriales como unidades de conservación, y tenerse esta información en cuenta, sobre todo, a la hora de llevar a cabo posibles translocaciones de individuos entre poblaciones, que podrían poner en peligro la integridad genética del desmán ibérico.

3. CÁLCULO DE TIEMPOS DE DIVERGENCIA EN EL GÉNERO *NEOMYS* MEDIANTE ESTIMA DE ÁRBOL DE ESPECIES

3.1 Resultados

3.1.1 Filogenia mitocondrial del género Neomys

Mediante una combinación de muestras procedentes de tejidos, muestreo no invasivo y secuencias obtenidas de las bases de datos (Tabla A4), se determinaron las relaciones filogenéticas para el gen del citocromo *b* dentro del género de musgaños europeos *Neomys*, empleando tanto metodología de inferencia bayesiana (BEAST) como de máxima verosimilitud (RAxML) (figura 23). En ambos casos la topología obtenida fue idéntica y, además, coincidente con la ya descrita con anterioridad para el grupo (Krystufek *et al.* 2000) (Castiglia 2007), siendo *Neomys fodiens* la especie externa a *Neomys teres* y *Neomys anomalus*. Se observaron además para *Neomys anomalus* dos linajes claramente diferenciados con una divergencia genética mitocondrial elevada, como ya había sido apuntado anteriormente con un muestreo más reducido (Castiglia 2007). Uno de ellos agrupaba a las muestras europeas y las de Catalunya y el otro contenía el resto de muestras de la Península Ibérica. Estos dos linajes se corresponden, a grandes rasgos, con dos subespecies descritas para *Neomys anomalus*: *N. anomalus anomalus* y *N. anomalus milleri*, con la diferencia de las muestras de Catalunya, que tradicionalmente se han agrupado con el resto de muestras de la Península Ibérica en *N. a. anomalus* y en este caso se hallaron dentro del clado *N. a. milleri,* con las muestras europeas.

El soporte de los agrupamientos entre estos taxones (calculado para los nodos correspondientes como los valores de bootstrap en el caso del análisis de máxima verosimilitud y como la probabilidad posterior en el caso del análisis bayesiano) es variable. En los casos de *Neomys fodiens* y *Neomys teres,* su monofilia para el citocromo *b* está muy bien soportada (bootstrap=100, probabilidad posterior=1). En el caso de *Neomys anomalus*, el soporte para el agrupamiento de *N. a. anomalus* y *N. a milleri*, es sensiblemente inferior (bootstrap=67, probabilidad posterior =0.9). En este sentido, la evaluación de las dos topologías alternativas para este nodo (que comportarían la parafilia de *Neomys anomalus* al formar un clado *Neomys teres* con cada una de las dos subespecies, respectivamente) mediante un AU test reveló que no

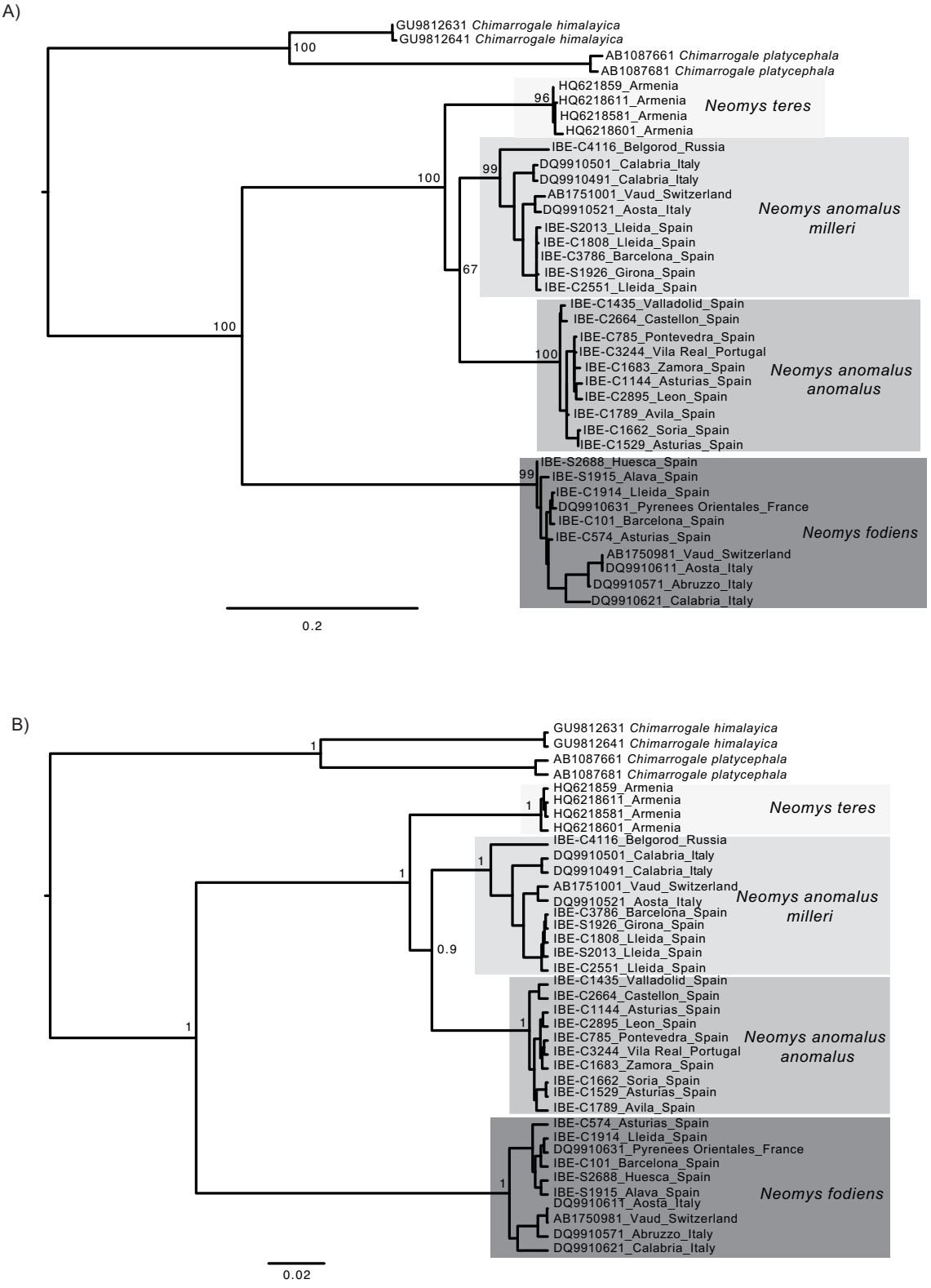

Figura 23. (A) Árbol de máxima verosimilitud de las secuencias de citocromo *b* del género *Neomys*. Los linajes se muestran en distintos tonos de gris, y en los nodos relevantes se incluye el soporte bootstrap correspondiente. Se emplearon secuencias de dos especies de *Chimarrogale* como grupo externo. **(B)** Árbol de inferencia bayesiana de las mismas secuencias, con las probabilidades posteriores de los nodos relevantes anotadas.

podían ser descartadas como explicación a los datos al compararlos con la topología obtenida en el árbol de máxima verosimilitud, que resultaba en la monofilia de *Neomys anomalus*.

La distancia genética media que separa el citocromo *b* de *Neomys fodiens* con el resto de los linajes mitocondriales estudiados es de un 16%. Por otro lado, los dos linajes encontrados para *Neomys anomalus* difieren en un 6.1%, una distancia bastante próxima a la que separa a ambos de *Neomys teres* (8%). En este sentido, cabe destacar que estudios anteriores señalaron que la divergencia existente entre las distintas especies del género *Sorex* (de la misma familia Soricidae a la que pertenece *Neomys*) varía entre el 1.32 y el 21.37%. (Fumagalli *et al.* 1999) .

3.1.2. Análisis de secuencias nucleares en el género Neomys

Para estudiar la divergencia entre las dos poblaciones diferenciadas de *Neomys anomalus* existentes en la Península Ibérica, correspondientes con las dos subespecies *N. a. anomalus* y *N. a. milleri*, se secuenciaron en 8 muestras de tejido de *Neomys* (3 *N. anomalus anomalus* y 3 *N. anomalus milleri* – 2 de ellos de la Península Ibérica y uno de Rusia - , además de 2 *N. fodiens como grupo externo*) 13 intrones seleccionados por ser variables a nivel interespecífico en mamíferos (Igea *et al.* 2010). Todos los intrones secuenciados mostraron efectivamente variabilidad de secuencia dentro del género y, en la mayoría de los casos, se detectó también variabilidad intraespecífica, demostrando así su utilidad como marcadores filogenéticos dentro de *Neomys* (figura 24). Un total de diez intrones no presentaron haplotipos compartidos entre *Neomys anomalus anomalus* y *Neomys anomalus milleri*, y el número mínimo de diferencias entre ambos osciló entre una y cinco mutaciones por intrón.

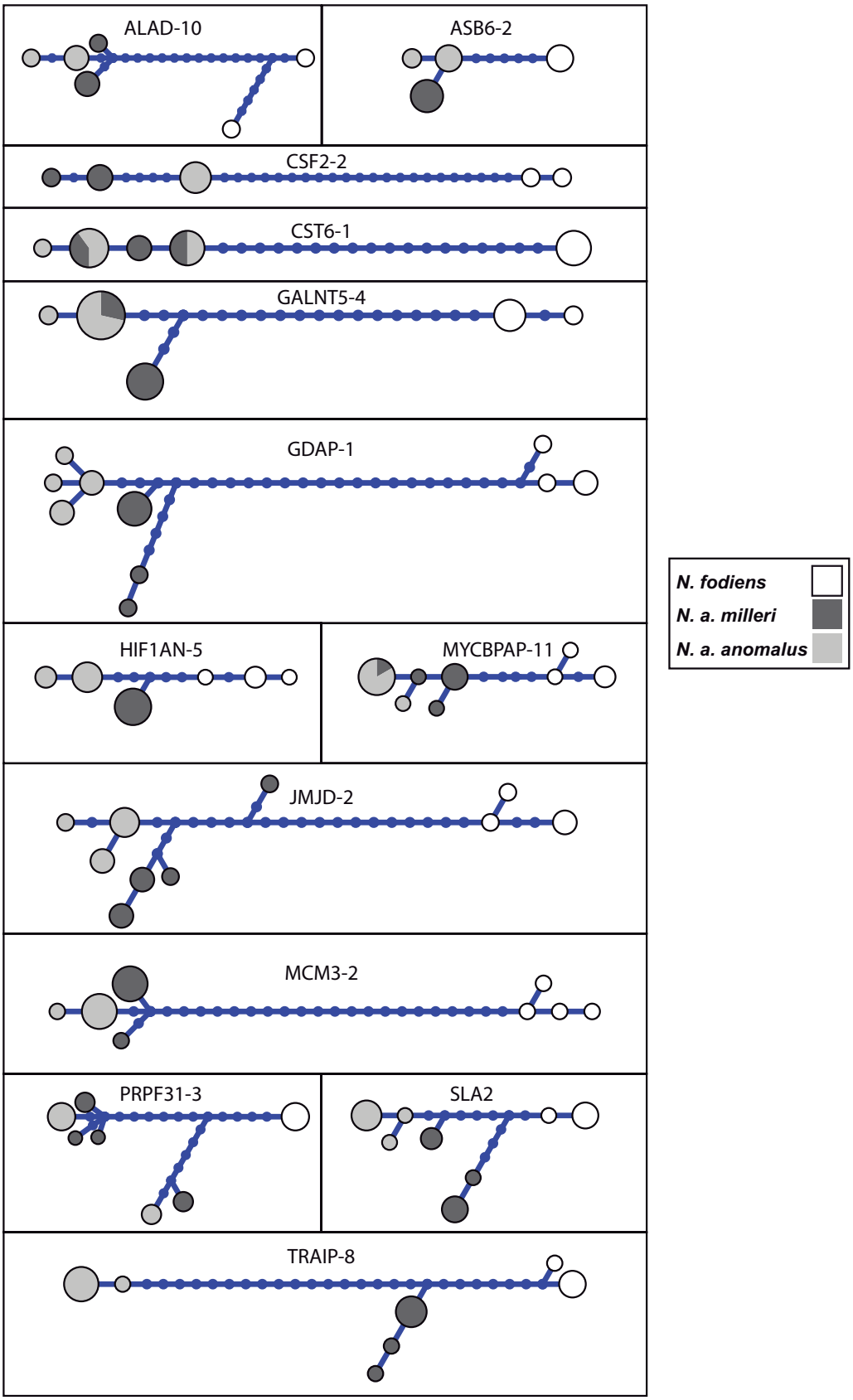

Figura 24. Genealogías haplotípicas de los 13 intrones amplificados en el género *Neomys* basadas en árboles filogenéticos de máxima verosimilitud. El tamaño de los círculos es proporcional al número de individuos con ese haplotipo, y los colores indican los linajes a los que pertenecen los individuos

3.1.3. Cálculo de tasas evolutivas de intrones y citocromo b *en el género* Neomys

Se procedió a estimar, mediante inferencia bayesiana usando el software *BEAST (Heled & Drummond 2010), el árbol de especies más probable calculando al mismo tiempo los árboles de genes de los 13 intrones nucleares y del citocromo *b*, así como los tiempos de divergencia y los tamaños poblaciones. Para poder estimar los tiempos de divergencia, es necesario introducir en el análisis o bien una calibración de algún nodo (proveniente de evidencias fósiles o de otros trabajos de datación molecular) o bien una estima de las tasas evolutivas de los marcadores empleados para el análisis. La fecha de divergencia entre *Neomys anomalus* y *Neomys fodiens* no está determinada de forma precisa, a tenor de los distintos resultados obtenidos con dataciones moleculares (Castiglia 2007; He *et al.* 2010) y lo escaso y poco fiable del registro fósil de los sorícidos (Dubey *et al.* 2007). Por lo tanto, se decidió emplear las tasas evolutivas calculadas de forma rigurosa.

Para estimar las tasas evolutivas correspondientes a cada uno de los intrones, se realizó un análisis bayesiano multilocus de laurasiaterios y euarcontoglires que incluía dos especies de sorícidos: *N. fodiens* y *Sorex coronatus*. Se usó un reloj molecular relajado y varios nodos de la filogenia calibrados con fósiles que incluían límites superiores e inferiores (Benton *et al.* 2009), y cada uno de los 13 intrones fue empleado como una partición independiente (figura 25a). Esto permitió extraer, para cada intrón, los valores de las tasas evolutivas globales de todo el árbol así como las tasas específicas del linaje de *Neomys fodiens*. Las tasas evolutivas estimadas para los distintos intrones presentaron una gran variabilidad (entre 0.002 y 0.016 sustituciones/sitio/ma). Al comparar las tasas individuales de cada intrón obtenidas para *Neomys* con las correspondientes al conjunto de los mamíferos, se comprobó existía una fuerte correlación positiva y significativa entre ambas (valor de la correlación = 0.9036). Sin embargo, la media de tasa evolutiva de los intrones en *Neomys fodiens* fue de 0.00773 sustituciones/sitio/ma, mientras en el árbol global de mamíferos fue de 0.00359. Los intervalos de confianza de ambos valores solapan de forma muy marginal, lo que sugiere que probablemente los sorícidos presentan una aceleración en sus tasas de sustitución nuclear al compararlos con la media de los mamíferos aquí estudiados.

Además, en este análisis multilocus nuclear, se obtuvo una datación precisa de la divergencia entre *Sorex coronatus* y *Neomys fodiens*. La media de la divergencia estimada fue de 40.95 millones de años. Este valor es bastante superior al sugerido por el registro fósil y al de otros trabajos de datación molecular, que lo fijan en

aproximadamente 17-20 millones de años (He *et al.* 2010; Fumagalli *et al.* 1999), si bien en estos dos últimos casos emplean como calibración la misma datación: la divergencia de Crocidurinae y Soricinae, fijada en 20 millones de años (Reumer 1989, 1994). Sin embargo, un análisis molecular bayesiano más reciente que empleó secuencias nucleares de gran cantidad de mamíferos y una serie de calibraciones fósiles (Meredith *et al.* 2011) determinó una fecha para esta divergencia de Crocidurinae y Soricinae mucho más antigua, de unos 45 millones de años, y más compatible con los resultados obtenidos en nuestro análisis, de 50.7 millones de años.

Debido a la saturación de secuencias del citocromo *b*, su tasa evolutiva se estimó en un análisis aparte con secuencias más próximas. Así, se obtuvo un árbol filogenético para el citocromo *b* de la familia Soricidae con el programa de inferencia bayesiana BEAST (figura 25b). Se calibró el análisis empleando la distribución posterior del nodo *Sorex coronatus – Neomys fodiens* obtenida en el análisis multilocus nuclear anteriormente descrito. El árbol filogenético obtenido coincide con la filogenia aceptada del grupo, con las subfamilias y la mayoría de los géneros monofiléticos. La tasa evolutiva del citocromo *b* así calculada para *Neomys fodiens* fue de 0.0152 sustituciones/sitio/ma, lo que corresponde con un 3.1% de divergencia por millón de años.

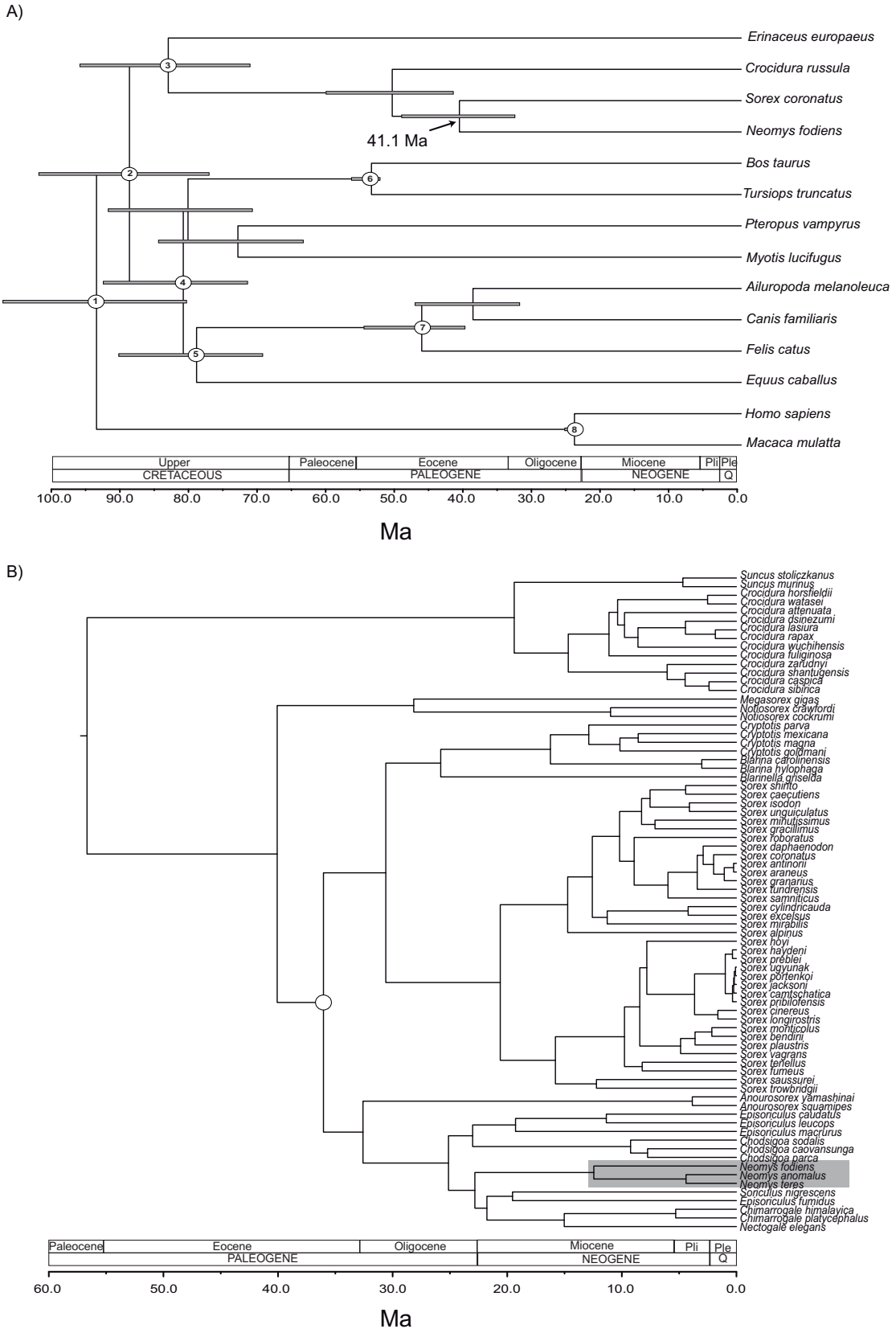

Figura 25. **(A)** Análisis bayesiano con secuencias nucleares de 13 intrones de mamíferos para extraer las tasas evolutivas correspondientes a *Neomys* fodiens. Se estimó también la fecha de divergencia de *Sorex* y *Neomys*. Se emplearon calibraciones fósiles correspondientes a Boreoeutheria (1), Laurasiatheria (2), Eulipotyphla (3), Ferungulata (4), Zooamata (5) y Cetruminantia (6), Carnivora (7) y Catarrhini (8). Las barras grises indican el intervalo del 95% HPD. **(B)** Análisis bayesiano para extraer la tasa del citocromo *b* correspondiente al género *Neomys* (indicado con color gris) empleando secuencias de sorícidos. Se calibró el nodo *Sorex* – *Neomys* con el resultado mostrado en (A).

3.1.4. Árbol de especies del género Neomys

Con las tasas nucleares y mitocondriales calculadas anteriormente, se procedió a estimar mediante *BEAST el árbol de especies para el género *Neomys*. El árbol sumario de mayor credibilidad de clados del árbol de especies de este análisis se muestra en la figura 26. Según este análisis, la separación de las dos subespecies de *N. anomalus* tuvo lugar hace unos 0.56 millones de años, situada (junto con la totalidad de los intervalos de la estima) en el Pleistoceno. Además, se dató la divergencia de *Neomys anomalus* y *N. fodiens* en 3.31 millones de años.

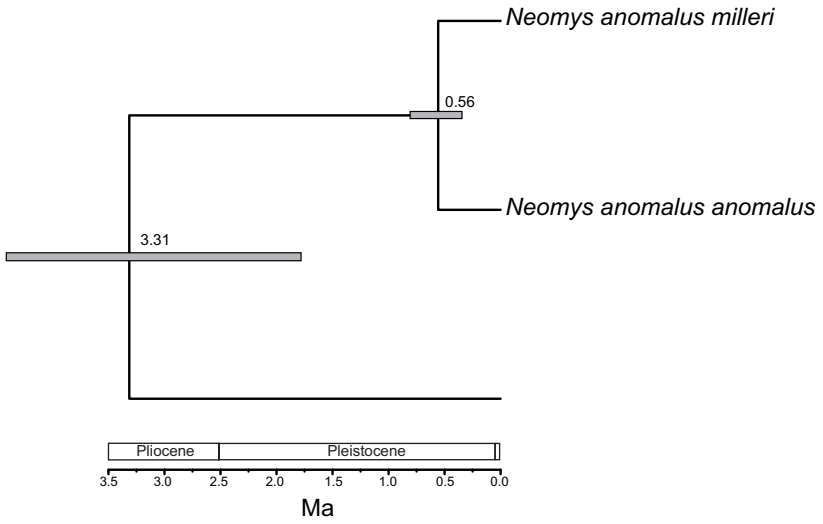

Figura 26. Árbol de especies de los linajes del género *Neomys* presentes en la Península Ibérica. Se muestran para cada nodo las medias correspondientes del tiempo de divergencia, y las barras grises indican el intervalo del 95% HPD de la estima

*3.1.5. Efecto de distintos priors, tipos de genes, y cálculo de las tasas evolutivas sobre la estima de los tiempos de divergencia en *BEAST*

La figura 27 muestra los resultados de distintas combinaciones de *priors* y conjuntos de datos en el análisis de *BEAST y los efectos que tienen, tanto sobre la estima del tiempo de divergencia *N. anomalus anomalus* y *N. anomalus milleri* (figura 27A) como sobre la estima de la divergencia *N. anomalus – N. fodiens* (la raíz del árbol de especies) (figura 27B). Se evaluaron las diferencias entre distintas estrategias, como el empleo o no de una tasa específica para cada uno de los marcadores nucleares empleados (o, por el contrario, el uso de una global para todos ellos); y, por otro lado, el empleo de una tasa específica para el organismo de estudio o una más general para el conjunto de mamíferos. Estas nuevas estimas se compararon con la que se ha detallado anteriormente, en el que se estimó una tasa evolutiva específica de *Neomys fodiens* para cada uno de los 13 intrones nucleares y también para el citocromo *b*.

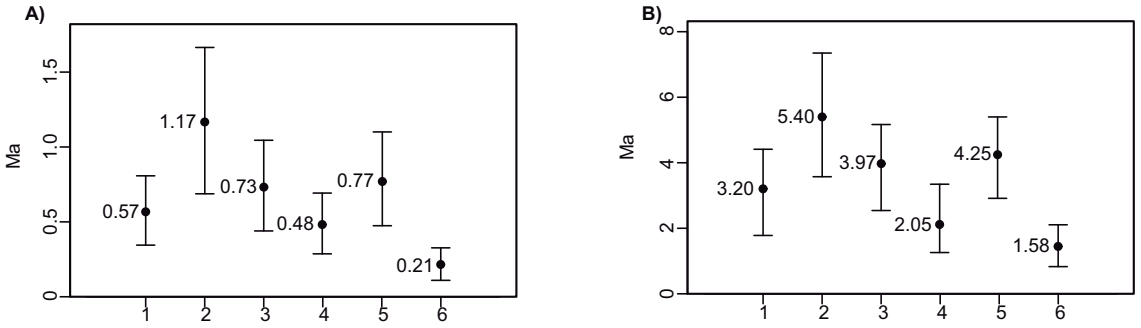

Figura 27. Media de las estimas de tiempos de divergencia con *BEAST de **(A)** *Neomys anomalus anomalus – Neomys anomalus millerii* y **(B)** *Neomys anomalus – Neomys fodiens* empleando distintos *priors* para las tasas evolutivas y conjuntos de datos: **1)** citocromo *b* (tasa medida para *Neomys*) + 13 intrones (tasas individuales medidas para *Neomys*), **2)** citocromo *b* + 13 intrones nucleares (tasas obtenidas de la bibliografía), **3)** citocromo *b* (tasa medida para *Neomys*) +13 intrones nucleares (un único rate, calculado como la media de mamíferos), **4)** citocromo *b* (tasa medida para *Neomys*) +13 intrones nucleares (un único rate, calculado como la media para *Neomys*), **5)** citocromo *b* (tasa medida para *Neomys*) + 13 intrones (tasas individuales medidas para mamíferos), y **6)** 13 intrones nucleares (tasas individuales medidas para *Neomys*).

En primer lugar, se emplearon como *priors* de las tasas mutacionales de citocromo *b* estimas obtenidas de la bibliografía y no estimadas directamente para los datos que se emplearon ni para el género *Neomys* (Bininda-Emonds 2007). Las tasas de nucleares se estimaron como 10 veces menores que esta, una valor aproximado también tomado de la bibliografía (Igea *et al.* 2010). En líneas generales, la principal diferencia de este enfoque con el de estimas más precisas residió en las tasas evolutivas de los genes nucleares. El valor resultante, de 0.00152 sustituciones/sitio/ma, era 4 veces menor que la media de las estimas más precisas realizadas individualmente para cada intrón en *Neomys fodiens* en el análisis bayesiano multilocus con varias calibraciones fósiles. El citocromo *b*, por otro lado, tenía una tasa mutacional muy similar a la estimada para *Neomys fodiens* con el análisis bayesiano de los sorícidos. El resultado (figura 27, caso 2) reveló que el tiempo de divergencia estimado entre los dos linajes de *N. anomalus* era aproximadamente el doble (1.17 frente a 0.57) y además el intervalo asociado a la estima aumentó sensiblemente. Asimismo, la estima de la edad de la raíz del árbol de especies mostró un comportamiento similar (5.40 millones de años frente a 3.26).

Por otro lado, se empleó como *prior* de la tasa evolutiva de todas las particiones nucleares un único valor calculado, o bien como la media de la tasa evolutiva para los intrones de todos los mamíferos (0.00359 sustituciones/sitio/ma), o bien para *Neomys fodiens* (0.00773 sustituciones/sitio/ma) (figura 27, casos 3 y 4, respectivamente). Los resultados para el tiempo de divergencia de *N. anomalus* fueron más similares en el caso de la tasa calculada como la media de los intrones de *Neomys fodiens*, que presentó unos intervalos de la estima que solapaban de forma notable con el

experimento control. Las estimas resultantes de emplear la tasa calculada para todos los mamíferos fueron algo superiores para los dos tiempos calculados, pero de nuevo los intervalos asociados eran bastante coincidentes. Por otro lado, tampoco se encontraron diferencias notables con respecto al control al aplicar a cada intrón su tasa individual correspondiente calculada también para mamíferos (figura 27, caso 5).

Por último, el uso de únicamente marcadores nucleares produjo una estima del tiempo de divergencia notablemente inferior a todos los experimentos que incluían una partición mitocondrial (figura 27, caso 6).

3.1.6. Aislamiento con migración

El modelo de aislamiento con migración (Hey 2010) permitió analizar los datos incluyendo la posibilidad de flujo genético entre las dos poblaciones de *Neomys anomalus* estudiadas (*N. anomalus anomalus* y *N. anomalus milleri*) y comprobar su efecto sobre la estima de tiempos de divergencia. Se restringió el modelo a la estima del flujo génico entre las 2 poblaciones mencionadas, puesto que la información contenida en los datos obtenidos (a nivel de polimorfismo en las secuencias) hacía difícil la estima precisa de parámetros relacionados con las poblaciones ancestrales y la población de *Neomys fodiens*. Las tasas de migración halladas (escaladas por la tasa mutacional) fueron $m_{milleri \rightarrow anomalus}$ = 0.001 (95% HPD = 0 – 0.33) y $m_{anomalus \rightarrow milleri}$ = 0.173 (95% HPD = 0.01 – 0.501). Únicamente ésta última resultó ser significativamente distinta de cero, según el *likelihoood ratio test* para los parámetros de migración de Nielsen y Wakeley (Nielsen & Wakeley 2001) implementado en el IMa2. Esta tasa de migración hallada se corresponde con un valor de tasa de migración poblacional (2NM, el producto de la tasa de migración por el número efectivo de copias de genes en la población receptora) de 0.1527. El tiempo medio de divergencia entre los dos linajes de *N. anomalus* fue de 332.834 años (95% HPD = 196.643 – 812.281 años. Por otra parte, el tiempo correspondiente a la divergencia *N. anomalus – N. fodiens* fue de 870649 años (95% HPD = 56.2135 – 1.198.619 años). En este último caso, sin embargo, el intervalo asociado a la estima no resulta fiable probablemente debido a la falta de información en las secuencias muestreadas para inferir parámetros relativos a la población ancestral, como se ha comentado anteriormente.

El análisis empleando el modelo menos parametrizado, es decir, de aislamiento sin migración entre las poblaciones (y por tanto, más similar a la metodología implementada en *BEAST) resultó, según un *likelihood ratio test*, ajustarse de igual manera a los datos que el modelo más complejo de aislamiento con migración. En este

caso, el tiempo de divergencia calculado fue, lógicamente, más reciente que los que tenían en cuenta la migración. La divergencia entre los dos linajes de *Neomys anomalus* se dató en 245.283 años (95% HPD = 161.900 – 363.407 años).

3.2.Discusión

3.2.1.Relaciones filogenéticas y divergencia dentro del género Neomys

La topología obtenida en este estudio para el género *Neomys* coincide con otros realizados con anterioridad. En concreto, Castiglia *et al.* (2007) realizaron un análisis de máxima verosimilitud empleando 252 pares de bases del citocromo *b* y un conjunto de datos más centrado en muestras italianas, y hallaron una topología similar y un soporte estadístico igualmente bajo para el nodo de *Neomys anomalus* y *N. teres*. Por otro lado, (Castiglia 2007) hallaron, en la única muestra de *Neomys anomalus* perteneciente a la Península Ibérica (concretamente al Sistema Central), un haplotipo muy divergente respecto al resto de muestras europeas de *N. anomalus*. En el presente estudio, con un muestreo muy ampliado, se ha confirmado la presencia de dos linajes mitocondriales de *N. anomalus* muy divergentes en la Península Ibérica. Uno de ellos correspondería con *N. a. milleri* y a él pertenecen las muestras de Cataluña (y las del resto de Europa), mientras el otro linaje se halla exclusivamente en el resto de la Península Ibérica, desde el Sistema Ibérico hasta la costa atlántica (Tabla A4). Además, la distancia genética que separa los ejemplares de *N. anomalus anomalus* de los de *N. anomalus milleri* es bastante similar a la que separa a cualquiera de estos dos linajes con *Neomys teres*, de nuevo en consonancia con la ya apuntada dificultad en resolver este nodo de la filogenia del género.

Por otro lado, la divergencia entre *Neomys anomalus* y *Neomys fodiens* se estimó en 3.31 millones de años. Esta fecha es sensiblemente inferior a la datación molecular realizada por (Castiglia 2007) de 6.9 millones de años. Cabe destacar que ésta fue calculada únicamente a partir de las distancias genéticas mitocondriales que separan a ambas especies, por lo que no se tuvieron en cuenta los efectos estocásticos asociados a la coalescencia de los distintos genes. Por otro lado, (He *et al.* 2010) obtuvieron, empleando en esta ocasión datos nucleares, una fecha de 1.69. Sin embargo, en este estudio tampoco calcularon esta divergencia empleando metodologías que estiman el árbol de especies, sino que usaron 2 genes nucleares concatenados.

3.2.2. Divergencia de los dos linajes de Neomys anomalus

Con el objetivo de establecer una datación precisa de la divergencia de las dos poblaciones ibéricas diferenciadas de *Neomys anomalus* empleando información de varios *loci* distribuidos por todo el genoma, se amplificaron 13 intrones nucleares en varios individuos pertenecientes a ambas subespecies así como a *Neomys fodiens*. Los intrones seleccionados resultaron ser efectivamente variables en el género, y los patrones de segregación de alelos resultaron compatibles con un historial de aislamiento genético sin un flujo genético importante, ya que 10 de los 13 nucleares no mostraron haplotipos compartidos entre los dos grupos.

Las tasas evolutivas estimadas para estos intrones mostraron que el género *Neomys* estaba más acelerado que la media de los mamíferos estudiados, como ya se ha sugerido anteriormente para los insectívoros (Bininda-Emonds 2007). En este sentido, es sabido que las tasas evolutivas varían en los distintos linajes por diversos motivos (Bromham 2009; Welch *et al.* 2008) y en concreto esta aceleración en los musgaños (que, como otros insectívoros, tienen altas tasas metabólicas y tiempos de generación y de vida cortos comparados con otros mamíferos) puede estar relacionada con hipótesis más clásicas como la de la tasa metabólica (Martin & Palumbi 1993), la del tiempo de generación (Bromham *et al.* 1996) y otras propuestas más recientemente como la de la longevidad (Nabholz *et al.* 2008a).

La datación para la divergencia de *Neomys anomalus anomalus* y *Neomys anomalus milleri* obtenida empleando la información de los 13 intrones nucleares y el citocromo *b* fue de 0.56 millones de años (95% HPD de 0.34 – 0.81), encuadrándose dentro del Pleistoceno. Esta datación, unida al hecho de que *Neomys anomalus anomalus* se localice exclusivamente en la Península Ibérica, sugiere que las glaciaciones pleistocénicas han podido ejercer un papel fundamental en la separación de las dos poblaciones de *Neomys anomalus*. Estos fenómenos han sido señalados, en particular en las penínsulas del Sur de Europa que sirvieron como refugio a muchas especies durante los periodos más fríos, como factores de gran influencia en la historia evolutiva y la estructura genética de muchas especies de mamíferos (Gómez & Lunt 2007; Melo-Ferreira *et al.* 2007). En el caso de *Neomys anomalus*, los resultados sugieren que la forma ibérica pudo haber sido el resultado de una población periférica que quedó aislada del resto en la Península Ibérica en alguno de los episodios glaciales que se sucedieron a lo largo del Pleistoceno.

El análisis empleando sólo los 13 intrones nucleares propone, por otro lado, una fecha de 0.22 millones de años (95% HPD: 0.11 – 0.33), sensiblemente inferior al análisis que incluye el citocromo *b*. El origen de esta discrepancia podría estar en que la partición de datos mitocondriales, al ofrecer mayor cantidad de información, tanto a nivel mutacional (por contar con una tasa evolutiva mayor y un mayor número de diferencias entre las secuencias estudiadas) como a nivel de tamaño muestral (por incluir un total de 10 secuencias por especie frente a las 6 secuencias obtenidas para los nucleares), proporciona mayor información y precisión a la estima de los tiempos de divergencia que la que se obtendría sólo empleando las particiones de datos nucleares. En este sentido, estudios con simulaciones han demostrado que la precisión de las estimas de tiempos de divergencia en metodologías de árboles de especies es mayor cuando se emplean marcadores más variables (Camargo *et al.* 2012) y, en concreto, cuando se incluyen marcadores mitocondriales (Sanchez-Gracia & Castresana 2012). Por tanto, cabe preguntarse si en el presente estudio no se ha alcanzado aún ese número suficiente de marcadores nucleares, lo que explicaría la discordancia entre ambas estimas. Por otro lado, podría ocurrir que el hecho de introducir el marcador mitocondrial puede estar influenciando negativamente en los cálculos del tiempo de divergencia. Así, podría ser que, a pesar de haber realizado una estima de la tasa lo más precisa posible y haber tenido en cuenta el efecto de la saturación, la tasa medida del citocromo *b* fuera menor de la real, lo que arrojaría una estima del tiempo de divergencia superior a la real al incluir el citocromo *b*. En cualquier caso, a la vista de los resultados, la divergencia de las dos poblaciones de *Neomys anomalus* se encuadra claramente dentro del Pleistoceno Medio, con los 215.000 años señalados por el experimento sólo con particiones nucleares como límite inferior.

En todo caso, la estima de los tiempos de divergencia empleando *BEAST, como ya se ha comentado, asume la ausencia total de intercambio génico entre las unidades evolutivas analizadas tras su separación. Los patrones de monofilia observados entre *N. a. anomalus* y *N. a.* milleri, recíproca en el caso del mitocondrial y la mayoría de los intrones nucleares, no sugerían un componente de flujo genético importante, ya que en ese caso sería esperable una mayor proporción de haplotipos compartidos. Para comprobar hasta qué punto se podría estar violando la asunción de ausencia de migración y evaluar el nivel de flujo génico existente entre las dos poblaciones, se analizaron los datos empleando un modelo de aislamiento con migración, implementado en el software IMa2. Cabe indicar que el hecho de no disponer de secuencias pertenecientes a *Neomys teres* en el análisis podría estar violando una de las asunciones del modelo de aislamiento con migración, que es que las poblaciones en

estudio no intercambien genes con ninguna otra población no muestreada e incluida en el análisis. Esto podría resultar en que las estimas se vieran afectadas, si bien estudios recientes con simulaciones han probado que el sofwtare IMa2 resulta ser bastante resistente a este tipo de violaciones de las asunciones del modelo (Strasburg & Rieseberg 2010).

Los resultados obtenidos con el IMa2 sólo detectaron una tasa de migración significativamente distinta de cero en la dirección *N. a. anomalus* -> *N. a. milleri*. El número de copias de genes migrantes por generación (2NM) asociado a esta tasa de migración es de 0.1527. En ausencia de selección, cuando este estadístico es mayor que 1, se considera que el flujo genético elevado contrarresta la acción de la deriva genética y limita por tanto la divergencia entre las poblaciones; mientras que si es menor que 1, la acción de la deriva genética prevalece y se produce diferenciación entre las poblaciones (Wright 1931). Además, el *likelihood ratio test* implementado en el programa IMa2 no pudo rechazar el modelo de aislamiento sin migración, con menos parámetros, como una explicación de igual ajuste a los datos. Por todo esto, se puede concluir que el flujo genético entre las dos poblaciones de *Neomys anomalus*, de haber existido, no ha tenido una importancia reseñable en las historias evolutivas de ambas que, por tanto, pueden considerarse como dos unidades evolutivas independientes.

*3.2.2. Consideraciones sobre los priors de las tasas evolutivas en *BEAST*

Las tasas evolutivas usadas como *priors* en los análisis bayesianos tienen una gran influencia en las estimas posteriores de los tiempos de divergencia. Para determinar las diferencias resultantes entre distintas estrategias para la obtención de las tasas *a priori* de las particiones empleadas en el programa *BEAST, se compararon los resultados alternativos obtenidos empleando tasas procedentes de la bibliografía (y no específicas por tanto de los marcadores moleculares ni de los organismos de estudio), y tasas medidas específicamente para cada marcador (o, por el contrario, como la media de los marcadores, sin tener en cuenta por tanto las diferencias de tasas existentes entre los distintos intrones) y para el grupo de estudio (o, por el contrario, empleando un valor medio obtenido del conjunto de los mamíferos).

Los resultados obtenidos (figura 27A y B) señalan, en primer lugar, los posibles errores asociados a emplear una tasa no medida específicamente para los marcadores empleados para estimar el árbol de especies. La tasa obtenida de la bibliografía para el citocromo *b* (y perteneciente a la familia Soricidae) era muy similar a la medida en este estudio para el género *Neomys*, pero la tasa nuclear era, por término medio, cuatro

veces inferior a la estimada de forma más precisa para cada intrón mediante un análisis multilocus con calibraciones fósiles. Esta diferencia tiene su reflejo en las estimas posteriores de los tiempos de divergencia de las dos subespecies de *Neomys anomalus*, que pasan a estar ubicadas claramente en el Pleistoceno Superior. Este efecto de aumento del tiempo de divergencia estimado se ve también en la estima de la raíz del árbol de especies, que en este caso se colocaría en el límite entre el Plioceno y el Mioceno, mientras que en el experimento control lo hacía entre el Pleistoceno y el Plioceno. Por tanto, el hecho de no estimar con precisión las tasas de los marcadores empleados lleva a unos errores en la estima de tiempos de divergencias que pueden conducir a la interpretación errónea de los factores que han generado esta divergencia.

En lo referente al resto de estrategias, que implican la medición de tasas específicas de marcador y/o grupo de estudio, el hecho de que el nodo que une las dos poblaciones de *Neomys anomalus* es un nodo muy reciente hace que las diferencias entre los distintos enfoques no se reflejen en unas diferencias acusadas en las estimas de los tiempos. Sin embargo, al analizar los resultados de la estima de tiempo de divergencia de *N. anomalus* y *N. fodiens*, las diferencias a nivel absoluto sí son más notables. El experimento con las tasas de los nucleares con una media obtenida para *Neomys* resulta en una fecha bastante más reciente (2.05 millones de años) que el experimento control (3.2), probablemente debido a que la tasa media aplicada es, para el caso de varios de los intrones empleados, hasta 3 veces superior a la tasa propia de ese intrón, lo que resulta en una disminución del tiempo de coalescencia calculado para esos árboles de genes, influyendo estos a su vez en el tiempo de especiación calculado. El ejemplo contrario es proporcionado por los experimentos con las tasas calculadas a partir de todo el conjunto de los mamíferos (o bien una única tasa o bien una tasa individual para cada intrón), que son significativamente distintas, y menores, de las tasas calculadas para *Neomys fodiens*. La aplicación de estas tasas menores explica, por consiguiente, el aumento de la divergencia calculada hasta los 4 millones de años, aproximadamente.

Por tanto, a la vista de estos resultados, se hace recomendable adoptar una estrategia que implique la estima lo más precisa posible de las tasas evolutivas para las particiones de datos empleadas, tanto a nivel de especificidad de marcadores moleculares empleados como de grupo taxonómico de estudio. La aplicación de tasas evolutivas no específicas, y aún más propias de otros marcadores moleculares o grupos taxonómicos, puede conllevar errores en la estima y estos, a su vez, pueden inducir a la

asociación errónea de los procesos de divergencia estudiados con ciertos fenómenos geológicos o biogeográficos.

3.2.3. Implicaciones taxonómicas de los resultados

Inicialmente, cuando en 1907 se describieron las dos subespecies de *Neomys anomalus* aquí estudiadas, fueron clasificadas como dos especies distintas, *Neomys milleri* (Mottaz 1907) y *Neomys anomalus* (Cabrera 1907), y ya se señaló, particularmente en la descripción de ésta última, la existencia de diferencias claras entre ambas en el color del pelaje y la forma del cráneo (Cabrera 1907). Sin embargo, la taxonomía más moderna (Spitzenberger 1990) considera ambas formas sinónimas y agrupadas bajo *Neomys anomalus*.

La divergencia entre las dos subespecies de *Neomys anomalus* que habitan actualmente en la Península Ibérica, según los resultados aquí presentados y derivados de la aplicación de metodologías de coalescencia para estimar y datar el árbol de especies usando información de múltiples *loci*, se produjo durante el Pleistoceno Medio, hace aproximadamente medio millón de años. Desde entonces el flujo genético entre ambas no ha sido reseñable. Tanto el tiempo transcurrido desde su separación como la ausencia de intercambios genéticos significativos indican que son dos unidades evolutivas independientes, afectadas por procesos de mutación, selección y deriva genética no comunes a ambas. En este sentido, se han propuesto recientemente, basándose en una revisión de estudios de divergencia que emplearon el modelo de aislamiento con migración para analizar especies cercanas y poblaciones, valores de 2NM y *t* que proporcionarían un medida objetiva para la diagnosis de especies (Hey & Pinho 2012). Los valores hallados para las dos subespecies de *Neomys anomalus* aquí estudiadas corresponderían más a dos especies distintas, lo que sugiere que la sinonimización actual de ambas bajo *Neomys anomalus* no reflejaría adecuadamente la realidad biológica de estos organismos.

Como ya se ha señalado, los procesos de divergencia de *Neomys teres*, *Neomys anomalus anomalus* y *Neomys anomalus milleri* fueron muy cercanos en el tiempo, como queda sugerido por la existencia de ramas cortas que separan ambos procesos de divergencia en la filogenia mitocondrial y el bajo soporte estadístico de los nodos (figura 23). La probabilidad de obtener árboles mitocondriales discordantes respecto a la topología real de separación de las especies o poblaciones es mayor precisamente en situaciones como esta, en la que la relación entre el tiempo entre dos nodos del árbol de especies y el tamaño efectivo poblacional es pequeña (Degnan & Rosenberg 2009). Por

tanto, la inclusión de secuencias pertenecientes a *N. teres*, la especie hermana de *N. anomalus*, y su análisis con técnicas de árboles de especies como las aquí presentadas, podría ayudar aún más a clarificar las relaciones taxonómicas mencionadas.

V. CONCLUSIONES

1. Se analizaron todos los intrones de los genomas de varios mamíferos disponibles en las bases de datos y se aplicaron una serie de filtros computacionales semiautomatizados (de tamaño, copia única, evolución molecular y divergencia) que produjeron finalmente un conjunto de 224 nuevos marcadores para estudios filogenéticos de especies cercanas de mamíferos.

2. Se caracterizó el conjunto final de intrones, comprobando que se trataba de intrones más divergentes y más variables intraespecíficamente que la media, y demostrando así su utilidad en estudios interespecíficos pero también en análisis intraespecíficos.

3. Se seleccionaron varios intrones del conjunto de 224 para su amplificación y secuenciación en un panel de mamíferos que contenía varias parejas de especies cercanas. La mayoría de los intrones seleccionados se amplificaron y secuenciaron, con ciertas optimizaciones previas, de forma correcta en todas las especies del panel e, igualmente, mostraron suficiente variabilidad para distinguir las especies cercanas.

4. No se observó un patrón claro en la acumulación de diferencias entre las distintas parejas de especies cercanas, no siendo posible predecir la variabilidad de un intrón en un grupo taxonómico determinado a partir de los datos de otros. Esto sugiere la necesidad de realizar estudios preliminares con los organismos de interés para seleccionar los intrones más adecuados, y remarca también la necesidad de emplear un número suficiente de marcadores independientes para disminuir los problemas estocásticos ligados a los procesos mutacionales.

5. Se realizó el primer estudio con datos genéticos del desmán ibérico (*Galemys pyrenaicus*) empleando secuencias mitocondriales y de 8 intrones obtenidas a partir de muestras de excrementos y tejidos frescos de gran parte de la distribución de la especie. Los niveles de diversidad genética hallados para esta especie fueron muy bajos comparados con la media de los mamíferos. Las secuencias mitocondriales revelaron la existencia de una fuerte estructuración filogeográfica, con zonas de contacto entre los linajes de distribución parapátrica a través de los cuales no se detectó prácticamente flujo genético. Los

datos nucleares, por otro lado, sí parecen sugerir más flujo genético en estas zonas de contacto.

6. Los diferentes linajes mitocondriales de *G. pyrenaicus* hallados mostraron grandes diferencias (de hasta un orden de magnitud) en sus valores de diversidad genética, con los máximos valores hallados en Galicia, Portugal y el Oeste de Asturias y los menores a lo largo de los Pirineos. Los datos de diversidad nuclear también confirmaron este patrón. Asimismo, la proyección para el Último Máximo Glacial de un modelo de distribución de la especie indicó también la presencia de un importante refugio glacial en el Noroeste de la Península Ibérica.

7. El tiempo de divergencia de los haplotipos mitocondriales del desmán obtenidos en este estudio se situó hace medio millón de años (Pleistoceno Medio). Para obtener esta estima se calculó previamente y de forma precisa la separación de los desmaninos mediante un análisis bayesiano empleando múltiples calibraciones fósiles e intrones nucleares.

8. Pese a tratarse de una especie estrictamente ligada al medio acuático, la diversidad genética del desmán ibérico no está tan afectada por estos requerimientos ecológicos como cabría pensar, existiendo flujo genético entre cuencas fluviales cercanas y siendo su historia evolutiva reciente (influenciada por las glaciaciones pleistocénicas) el factor más determinante. Esto tiene importantes implicaciones a nivel de conservación de la especie, ya que indica que las unidades de gestión no deberían corresponder estrictamente con los sistemas fluviales, y se deberían favorecer también corredores entre cuencas para posibilitar los intercambios entre poblaciones.

9. Las relaciones filogenéticas dentro del género de musgaños europeos *Neomys* fueron determinadas empleando muestras de tejidos, excrementos y secuencias obtenidas de las bases de datos. Se confirmó el carácter de especie externa de *Neomys fodiens* respecto al clado formado por *Neomys anomalus* y *Neomys teres*. Asimismo, se halló una notable distancia genética entre las dos subespecies de *N. anomalus*: *N. a. milleri* (localizada en Catalunya y Europa) y *N. a. anomalus* (en el resto de la Península Ibérica). Asimismo, se secuenciaron 13 intrones nucleares en varios individuos de *N. fodiens* y ambas subespecies de

N. anomalus, que resultaron ser efectivamente variables intra e interespecíficamente.

10. La estima del árbol de especies aplicando metodologías que tienen en cuenta las predicciones de la teoría de la coalescencia y empleando información de múltiples *loci* nucleares y mitocondriales permitió situar la divergencia de las dos subespecies de *N. anomalus* en el Pleistoceno Medio. Para ello, previamente, se calcularon las tasas evolutivas correspondientes específicas de *Neomys* para cada uno de los intrones escogidos mediante un análisis bayesiano multilocus con varias especies de mamíferos y calibraciones fósiles de referencia. Igualmente, la tasa evolutiva del citocromo *b* correspondiente fue obtenida de un análisis en el que se emplearon secuencias correspondientes a la familia de los sorícidos.

11. La aplicación de un modelo de aislamiento con migración al sistema de las dos subespecies de *N. anomalus* reveló la ausencia de flujo genético significativo entre ambas después de su divergencia. Esto, unido al hecho de que inicialmente ambas fueron descritas como subespecies distintas que fueron posteriormente sinonimizadas, sugiere que la actual clasificación taxonómica no refleja de forma correcta la realidad biológica de estos organismos.

12. La comparación de distintas estrategias a la hora de calcular las tasas evolutivas empleadas en las estimas de árboles de especies y su influencia en los tiempos de divergencia calculados reveló el riesgo asociado a no emplear tasas específicas de los organismos y de los marcadores empleados en este tipo de estudios.

VI. APÉNDICES

PUBLICACIÓN 1

Igea, J., Juste, J., and Castresana, J. (2010). **Novel intron markers to study the phylogeny of closely related mammalian species.** *BMC Evolutionary Biology 10, 369*

Igea et al. BMC Evolutionary Biology 2010, 10:369
http://www.biomedcentral.com/1471-2148/10/369

BMC Evolutionary Biology

METHODOLOGY ARTICLE Open Access

Novel intron markers to study the phylogeny of closely related mammalian species

Javier Igea[1], Javier Juste[2], Jose Castresana[1*]

Abstract

Background: Multilocus phylogenies can be used to infer the species tree of a group of closely related species. In species trees, the nodes represent the actual separation between species, thus providing essential information about their evolutionary history. In addition, multilocus phylogenies can help in analyses of species delimitation, gene flow and genetic differentiation within species. However, few adequate markers are available for such studies.

Results: In order to develop nuclear markers that can be useful in multilocus studies of mammals, we analyzed the mammalian genomes of human, chimpanzee, macaque, dog and cow. Rodents were excluded due to their unusual genomic features. Introns were extracted from the mammalian genomes because of their greater genetic variability and ease of amplification from the flanking exons. To an initial set of more than 10,000 one-to-one orthologous introns we applied several filters to select introns that belong to single-copy genes, show neutral evolutionary rates and have an adequate length for their amplification. This analysis led to a final list of 224 intron markers randomly distributed along the genome. To experimentally test their validity, we amplified twelve of these introns in a panel of six mammalian species. The result was that seven of these introns gave rise to a PCR band of the expected size in all species. In addition, we sequenced these bands and analyzed the accumulation of substitutions in these introns in five pairs of closely related species. The results showed that the estimated genetic distances in the five species pairs was quite variable among introns and that this divergence cannot be directly predicted from the overall intron divergence in mammals.

Conclusions: We have designed a new set of 224 nuclear introns with optimal features for the phylogeny of closely related mammalian species. A large proportion of the introns tested experimentally showed a perfect amplification and enough variability in most species, indicating that this marker set can be very helpful in multilocus phylogenetics of mammals. Due to the lower variability and stronger stochasticity of nuclear markers with respect to mitochondrial genes, studies should be designed to make use of several markers like the ones designed here.

Background

Phylogenetic analyses of closely related species are affected by specific problems that are different from those present in phylogenies of more distant species [1]. The first and most obvious difficulty is that nucleotide sequences must have enough variability within and among the studied species to obtain an adequate resolution of the phylogenetic tree. For this reason, mitochondrial genes, which show a high rate of nucleotide substitution, have been the main choice for the reconstruction of phylogenetic trees at the genus and family levels in all animals [2,3]. Another distinctive feature of the phylogenetic reconstruction of closely related species is that gene coalescence and the stochasticity associated with population genetic processes must be taken into account. For example, incomplete lineage sorting may cause the gene tree to have a different topology than the species tree [4-6]. This leads to the necessity of using multiple, unlinked genes, together with the integration of coalescent-based techniques, in the reconstruction of the species tree [7-10]. Phylogenies based on a broad representation of the genome can also help in species delimitation or analyses of genetic variability. Therefore, in these approaches it is essential to

* Correspondence: jose.castresana@csic.es
[1]Institute of Evolutionary Biology (CSIC-UPF), Passeig Maritim de la Barceloneta 37, 08003 Barcelona, Spain
Full list of author information is available at the end of the article

make use of the nuclear genome, where there are plenty of unlinked genes. However, nuclear genes in animals usually have much lower evolutionary rates than mitochondrial genes and, sometimes, they are not informative enough for assessing the variability within species or for phylogenetic reconstruction. In addition, due to the larger population size of nuclear genes with respect to mitochondrial genes, the former are more affected by coalescent stochasticity, so the necessity of using multiple genes is stronger [11,12]. Thus, the progress of multilocus phylogenetics requires, as a first step, an important effort of developing unlinked nuclear markers able to provide enough differences within and among species [13].

Systematic efforts to find novel markers for phylogenetic studies have been performed in different groups of organisms [14-21]. Markers selected include exons, introns and intergenic regions. Exons show very little variability for the phylogenies of closely related species, whereas intergenic regions present difficulties for the design of primers of wide-specificity. On the other hand, introns are a part of the genome with large nucleotide variability and, at the same time, they can be easily amplified with primers placed in the adjacent exons [20,22,23]. Thus, they are ideal candidates for multilocus phylogenies of closely related species. However, not all introns are equally valid for this task. Some introns are highly conserved due to their involvement in certain functions and they may not show enough differences between closely related species [24,25]. In fact, introns show a wide range of evolutionary rates. For example, a comparison of human and mouse introns showed genetic distances that ranged between 0.2 and 1.7 substitutions/site [26]. In addition, some introns show disparate rates of evolution in some lineages but not in others [27], indicating an imperfect molecular clock that may affect the measurement of genetic diversity in some species. Different processes, such as a change in chromosome position of the gene or the development of a new isoform by alternative splicing in certain lineages [28], may cause such variations in the evolutionary dynamics and thus in the molecular clock of the introns.

In mammals, introns have been mainly used to resolve deep groupings but also to study more recent phylogenies [29-33]. However, no attempt has been made so far to systematically select an optimal set of introns for mammals. In this work, we have devised a protocol to extract the best introns from the complete mammalian genomes of five species: human, chimpanzee, macaque, dog and cow [34-39]. We deliberately did not use the available genomes of rodents (mouse and rat) because they have genomic features that would have made the comparisons of all mammals problematic. For example, rodents have very attenuated isochors and show very fast evolutionary rates when compared to other mammals [40,41]. This is also true for introns, as previously shown [27]. The high evolutionary rates of rodent introns can complicate the alignments, phylogenetic reconstructions and measurement of genetic distances. To avoid the same problems we also decided not to include marsupials and monotremes. Thus, in this work we concentrated on the analysis of non-rodent eutherian genomes. From these genomes we obtained one-to-one orthologues, constructed alignments and trees from each individual intron and filtered out introns with inadequate features for shallow phylogenies. In addition, introns were selected to come from single-copy genes in order to avoid multiple bands in PCR reactions, to have an adequate length for PCR amplification, and to be surrounded by exons with enough space for primer design. From the resulting introns, we selected a small set that we used to experimentally verify that they work according to the expected features of ease of amplification and high evolutionary rate. Finally, we studied the variability of intron divergence in different species pairs.

Results and Discussion

Intron set acquisition

We extracted all introns smaller than 50,000 nucleotides from the genomes of human, chimpanzee, macaque, cow and dog, which comprise three mammalian orders (Primates, Carnivora and Cetartiodactyla). The total number of extracted introns per genome ranged between 153,659 in the cow and 173,320 in the dog. Using information from the ENSEMBL database, we arrived at an initial set of 11,835 one-to-one orthologous introns in the five mammalian species (Figure 1).

Most of these introns had an inadequate length for their amplification and were therefore discarded in our initial filtering processes. First, we restricted the intron length in *Homo sapiens* to a minimum of 200 and a maximum of 1600 nucleotides. Second, we controlled for intron length conservation among the five species used in this study, taking human as a referent and relaxing the constraints as the phylogenetic distance between the compared species increased. After the application of these size filters, 2750 introns remained. In addition, the introns flanked by small exons (<40 nucleotides) were discarded because the design of PCR primers could be difficult in them. This filter affected 5.2% of the introns available in the previous step.

Mammalian genomes have a large number of duplicated genes due to different gene or genome duplication processes [42]. These genes constitute a severe problem because primers could hybridize in a large number of genomic places, giving rise to multiple PCR products. Thus, a crucial step was to check for the duplication of the introns using BLAST over the different genomes. In

Igea et al. BMC Evolutionary Biology 2010, 10:369
http://www.biomedcentral.com/1471-2148/10/369

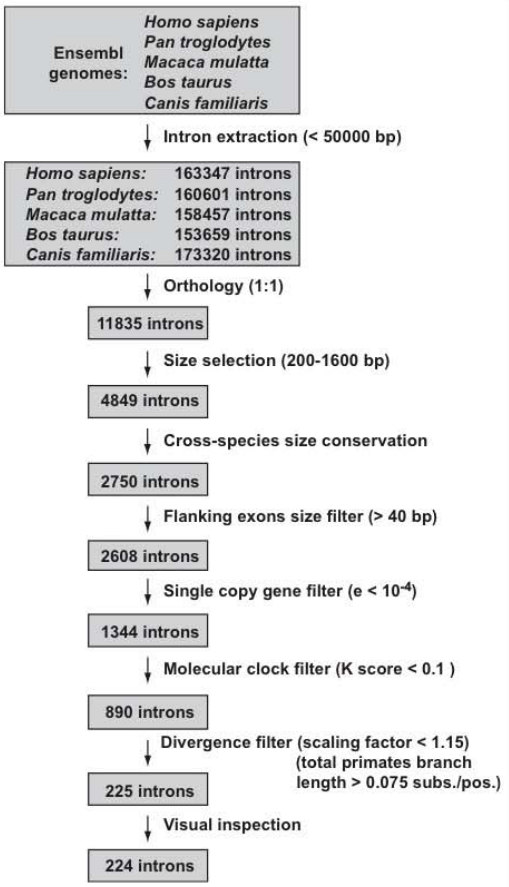

Figure 1 Scheme of the intron extraction and filtering processes. bp indicates base pairs. subs./pos. indicates substitutions/position.

fact, we performed the BLAST searches using the flanking exons instead of the introns because the exonic sequences are easier to detect by BLAST, even in divergent genes [43]. Only those introns that had both flanking exons present just once in the five genomes were considered to be single-copy. This step ruled out approximately half of the remaining introns. At this point, there were 1344 introns with a preliminary orthology filter that can be used to study different evolutionary processes. However, for these introns to be most useful in phylogenetic studies of closely related species we applied several additional filters.

Next, we eliminated introns that were too accelerated or decelerated in some lineages (and that could have been affected by changes in evolutionary processes in particular lineages such as a change of function, a change to a chromosome position with different evolutionary dynamics, alternative splicing, etc). That is, we selected introns whose rates of evolution were similar to the global genomic rate and therefore with a largely neutral molecular clock in mammals. Furthermore, trees with large branches in particular lineages may correspond to hidden paralogues that may have remained undetected up to this point. These paralogues are very problematic for estimations of evolutionary rates and other parameters necessary for their use in phylogenetic analyses. To detect this type of introns, we first constructed a global genomic tree from the concatenation of all introns available at this step (Figure 2). This tree was then used as a reference to assess if the phylogenetic tree of each individual intron had a rate of evolution similar to the global one in every lineage. This calculation was performed with the K tree score measured by the Ktreedist software, which reflects the topological and relative differences in branch length between a given tree and a reference tree [27]. That is, a high K tree score is indicative of a tree that has some highly accelerated or decelerated branches with respect to the reference tree regardless of the overall tree divergence. This score is also influenced by wrong topologies when the affected branches are large. By setting an arbitrary K tree score limit of 0.1 we removed approximately 34% of

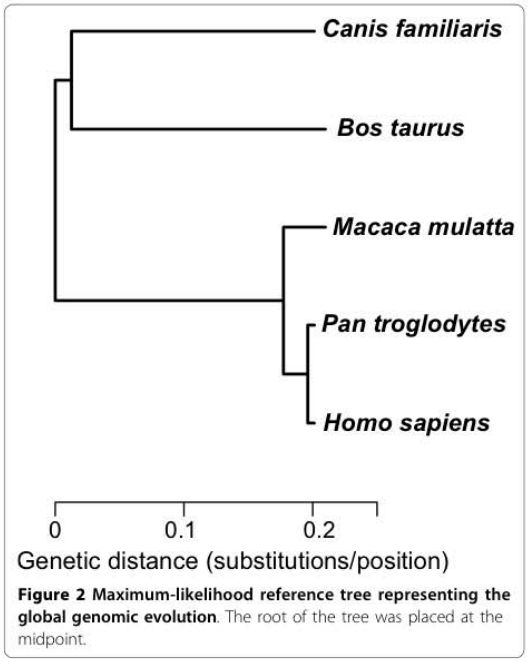

Figure 2 Maximum-likelihood reference tree representing the global genomic evolution. The root of the tree was placed at the midpoint.

Igea et al. BMC Evolutionary Biology 2010, 10:369
http://www.biomedcentral.com/1471-2148/10/369

the introns that were most different in shape from the global genomic tree.

The next step was designed to eliminate the most conserved introns, which could be involved in some function and are therefore not variable enough for the phylogeny of closely related species. To do this, we employed two different measures of divergence. First, we calculated with the Ktreedist software the scaling factor from each intronic phylogenetic tree to the global genomic tree. This measure allowed us to discard the introns that showed the slowest overall evolutionary rates. Second, we calculated another measure of divergence, the primates total branch length, by using the corresponding alignments and trees with only the three primate species (human, chimpanzee and macaque). This measure is more accurate and less affected by alignment imperfections possibly generated when comparing sequences that are too divergent. We then selected the introns that had a scaling factor lower than 1.15 and a primates total branch length higher than 0.075 substitutions per position. This rendered a dataset of 225 introns that excluded the most conserved ones.

Finally, a visual inspection step was performed to detect any poorly aligned sequences caused by wrong annotations or clearly incorrect orthology assignments. Only one intron had to be eliminated in this step. The resulting final set was thus composed of 224 new phylogenetic markers (Figure 1). In the additional file 1 we show all relevant information for each marker. This includes the alignments of both the intron and the flanking exons, the phylogenetic tree of the five mammals constructed with the intron sequences, and the genomic location and function of the gene to which the intron belongs. The examination of this information allows the selection of optimal markers for specific purposes and the design of exon primers with different degrees of specificity. Interestingly, one of the introns in our final set turned out to be intron 1 of transthyretin (TTR-1), which is one of the most widely used introns in mammalian phylogenetics [44-48]. To our knowledge, no other markers in our dataset have been used so far.

Analysis of genomic features of the new set of introns

The genomic location in *Homo sapiens* of each new marker is represented in Figure 3. All the human chromosomes carry at least one intron in our set, except chromosomes 21 and Y. The latter chromosome was expected to be missing from our set because no sequence for this chromosome was present in the available genome sequences of macaque, cow and dog. Regarding the X chromosome, only one intron was present in the final set. The rest of X-linked introns were discarded in the different filtering process, mainly in the single-copy test and divergence filters steps.

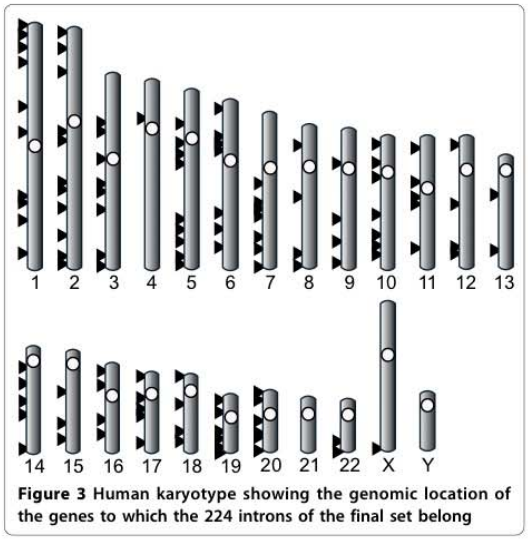

Figure 3 Human karyotype showing the genomic location of the genes to which the 224 introns of the final set belong

Repetitive sequences present in the intron set were analyzed with the RepeatMasker software. Of the 224 introns, 163 were found to contain repetitive elements. The most frequent elements were SINEs, which were present in 70% of the introns that bore any kind of element, followed by LINEs, which were found in 26% of them. Introns with repeats had, on average, 18% of their sequence corresponding to repetitive elements. These types of repetitive sequences normally evolve in a neutral way, mostly by point mutations, and therefore their presence is not a problem for sequence-based analysis methods that assume a normal nucleotide substitution model. If all species in the alignment have the repeat (which is the normal situation in closely related species) this fragment can be used normally as any other sequence. When only some species have the repetitive element (as it happened often in our set of five mammalian species), we observed that alignment programs do not have problems dealing with these repetitive elements, and therefore the alignments can also be used for further sequence or phylogenetic analyses.

More problematic are microsatellites (defined here as two or more contiguous, approximate copies of a pattern of 1 to 6 nucleotides [49]). Since microsatellites evolve by a slippage mechanism instead of by point mutations [50], they cannot be used with phylogenetic or coalescence methods that assume a normal point substitution model. We checked for the presence of microsatellites with the Tandem Repeats Finder program and found that 43 of the 224 introns had at least one microsatellite in some species. Microsatellites are, in general, specific for particular lineages, and their

Igea *et al. BMC Evolutionary Biology* 2010, **10**:369
http://www.biomedcentral.com/1471-2148/10/369

presence in one species does not mean that the intron cannot be used in other taxa. Therefore, introns with some microsatellites were not eliminated from the data set. However, when these sequences are found within the introns in a species to be studied it may be better not to use this intron and rather select a new one from the set designed here. Alternatively, the positions belonging to the microsatellites can be eliminated from the alignment for further processing with sequence analysis programs that assume a point mutation mechanism of evolution.

Analysis of genetic distances and single nucleotide polymorphisms

In order to test the degree of variability of the 224 final introns with respect to the initial data sets, we estimated the genetic distance between human and chimpanzee for each intron. To do this, we constructed the alignments of the primate species (human, chimpanzee and macaque), which could be used without further cleaning. Then, we estimated the corresponding maximum-likelihood trees and measured the patristic distances between human and chimpanzee. We then compared the distances obtained for the final set of 224 introns with the sets corresponding to the different filtering processes (Figure 4). The mean divergence between human and chimpanzee was around 0.011 substitutions/position in the three initial data sets: the size-constrained introns set, the single-copy gene introns set, and the set obtained after the application of the neutral evolution test. As expected, after the application of the divergence filter this distance increased, but only up to 0.014 substitutions/position. Thus, the overall gain in genetic distance in the final set was small. However, the main effect of the divergence filters was the elimination of highly conserved introns (as reflected in the relative decrease of the first bar of the histograms in Figure 4d), which could be involved in some important function and would be practically useless for studies of closely related species. Similarly, the analysis of human single nucleotide polymorphisms (SNPs) in the introns of our final set revealed that they had 4.35 SNPs on average. In comparison, the set of 2750 size-constrained orthologous introns (see Figure 1) contained 3.19 SNPs on average. When scaling these results by the length of each intron to obtain the SNP density, the difference between the means was still maintained: 0.0055 SNPs per nucleotide for the final set of introns *versus* 0.0043 for the initial set. Thus, human SNPs also showed a slight increase in the genetic variability of the introns in the final set.

To compare the mean intronic genetic distance with the divergence of cytochrome *b*, which is the most popular marker for mammalian phylogenetics [51], we

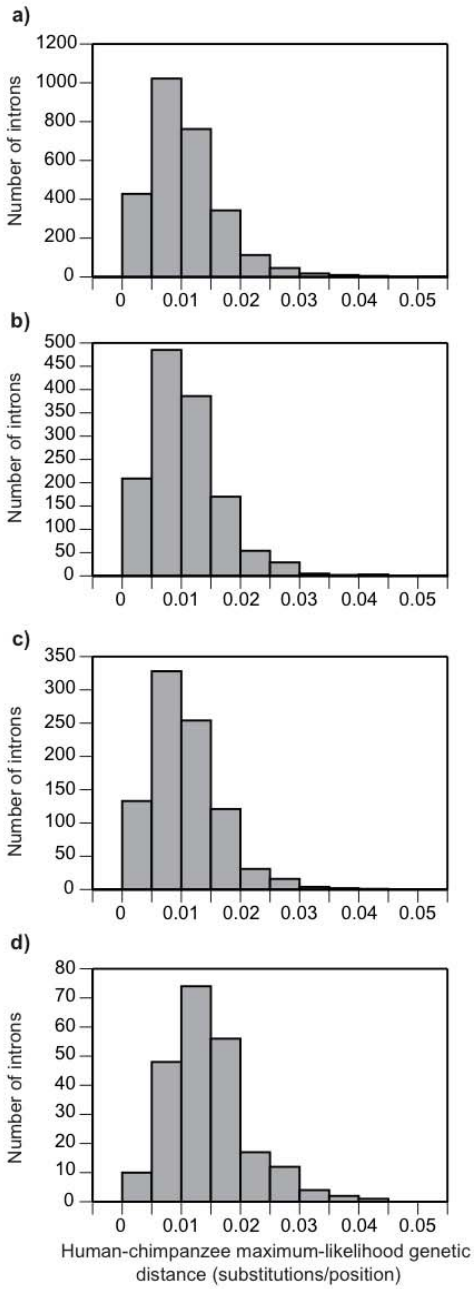

Figure 4 Human-chimpanzee genetic distances measured in the primates tree. The different sets correspond to: a) 2750 orthologous, size-constrained introns; b) 1344 single-copy introns; c) 890 orthologous introns of neutral evolution; and d) the final set of 224 introns.

Igea et al. BMC Evolutionary Biology 2010, 10:369
http://www.biomedcentral.com/1471-2148/10/369

obtained the cytochrome *b* distance between human and chimpanzee. To do this, we measured the patristic distance in a maximum-likelihood phylogenetic tree similarly reconstructed with human, chimpanzee and macaque cytochrome *b* sequences obtained from GenBank [52]. The resulting distance was 0.169 substitutions/position, which implies that the final set of 224 introns has, on average, 12.1 times less divergence than the mitochondrial cytochrome *b* gene. Furthermore, alignments containing the exons flanking the 224 introns were constructed for the three primates, maximum likelihood trees were built as above, and the resulting trees were used to calculate the distance between *Homo sapiens* and *Pan troglodytes* for both the upstream and downstream exons of each intron. The resulting mean genetic divergence was 0.006 substitutions/position, which implies that the 224 selected introns have, on average, 2.3 more divergence than their corresponding flanking exons.

Experimental validation of the newly developed phylogenetic markers: Primer design and PCR amplification

The intron filtering processes carried out above were designed to select a set of optimal introns for the phylogeny of closely related mammalian species. However, many unidentified factors may affect the amplification of these introns in different species. To experimentally test the real performance of these introns and the validity of the designed bioinformatic analysis we randomly selected twelve introns among the largest ones in the final data set for sequencing in different mammals.

In order to design primers of wide spectrum, the flanking exonic sequences already gathered (human, chimpanzee, macaque, cow and dog) were complemented by (whenever possible) a few others from the ENSEMBL database: horse (*Equus caballus*), Sumatran orangutan (*Pongo abelii*), little brown bat (*Myotis lucifugus*), European hedgehog (*Erinaceus europaeus*), domestic cat (*Felis catus*), Northern tree shrew (*Tupaia belangeri*), gray mouse lemur (*Microcebus murinus*), African bush elephant (*Loxodonta africana*), lesser hedgehog tenrec (*Echinops telfairi*), nine-banded armadillo (*Dasypus novemcinctus*) and common shrew (*Sorex araneus*). These genomes mostly corresponded to low-coverage genome projects but many of them were valid for certain exons. Exonic alignments, containing between 11 and 15 different species, were constructed. These alignments allowed us to design the 3' end of every primer in the most conserved part of the exonic alignment. Degenerate bases (up to a limit of 48 per primer) were used to make the primers suitable for as many different mammalian species as possible.

To test these primers, we used genomic DNA extracted from six mammalian species: Iberian mouse-bat (*Myotis escalerai*), Bornean orangutan (*Pongo pygmaeus*), snow leopard (*Uncia uncia*), tiger (*Panthera tigris*), least weasel (*Mustela nivalis*), and European polecat (*Mustela putorius*). This set included several closely related species in order to analyze their intronic divergence. Table 1 shows the primers used and the amplification results for this panel of species. Five out of the 12 introns failed to produce a single, clean PCR band of the expected size in one or more of the six analyzed species. This was due to several reasons, the main one being PCR misamplification in the form of gel smear. Some other problems were the excessive length of the amplified band in some species or the generation of double bands in the PCR reaction, which can reflect the existence of gene duplications in a particular lineage (Table 1). Some optimizations with different primers or hybridization temperatures did not solve the problems in these introns. The seven remaining introns did produce a clear single PCR band of the expected size in the six chosen species. It should be noted that, even for the successful primers, the annealing temperature had to be optimized for different species. Moreover, there were a few cases, namely SLC38A7-8 for *Panthera tigris* and CARHSP-1 and PNPO-3 for *Myotis escalerai*, where more specific primers had to be designed to improve the PCR band. The PCR bands of these introns were subsequently sequenced, and the sequences were compared to the known introns, which confirmed their correct amplification. In addition to the sequence of the PCR band, the exact sequence of one allele for each specimen was obtained by cloning the sequenced PCR product in a plasmid vector.

We can conclude that a large part of the markers tested were valid, with a minimum optimization, for a wide variety of mammals. Of course, the success rate will vary depending on the taxa of choice, but, to increase the chances that a primer works, it is important to take into account that this optimization may be necessary in all pilot studies and may include the adjustment of the hybridization temperature or the modification of the primer specificity with a different degree of degeneracy. In our experimental test of selected introns, we designed primers of broad specificity so that they could be used in a wide range of mammalian groups, but primers intended for specific taxa need not be based on such a variety of species and can have less degenerate bases. In addition to the species tested here, selected to include pairs of closely related species, we have successfully amplified other intron markers from our data set in other mammals, mainly belonging to Erinaceomorpha and Soricomorpha, as part of ongoing studies. They also produced the expected PCR band and the

Igea *et al. BMC Evolutionary Biology* 2010, **10**:369
http://www.biomedcentral.com/1471-2148/10/369

Table 1 Introns selected for amplification and sequencing in six species, and final outcome

Intron name	ENSEMBL Code	Length (*H. sapiens*)	Primer sequences (Forward and Reverse)	Result
SLC38A7-8	ENSG00000103042	877	RGGCCTRGCYGSCTGCTTCATCTT TCVGASAGYTTGGCTTGRATGAGGCA	+
COPS7A-4	ENSG00000111652	952	TACAGCATYGGRCGRGACATCCA TCACYTGCTCCTCRATGCCKGACA	+
CARHSP1-1	ENSG00000153048	698	ACYCGCCGSACSAGGACCTTCT GTRATGAAGCCRTGGCCCTTGGA	+
GAD2-1	ENSG00000136750	724	GGCTCHRGCTTYTGGTCYTTYGG YCCGAKGCCKCCSGTGAACTTCT	+
JMJD5-2	ENSG00000155666	1124	ACCABTGGCCVTGCATGMAGARGT TGATGAACTCRYTGACBGTCATGAG	+
OSTA-5	ENSG00000163959	535	TGMWGGYCATGGTGGAAGGCTTTG AGATGCCRTCRGGGAYGAGRAACA	+
PNPO-3	ENSG00000108439	843	GATGGCTTCCRHTTCTWCACTAACTT GGYTCCCARTAGAAGACMAKSGA	+
GAD2-3	ENSG00000136750	1051	TGCTCTAYGGRGAYKCMGAGAAG CAGAAACGCCARMGTGGSCCTTT	(1)
CLCN6-17	ENSG00000011021	1376	GTGGCCAAATGGACAGGGGACTTT TTGCCCTTCATGAACTCCTTCTCGT	(2)
TFPI2-2	ENSG00000105825	913	TACTACTAYGACAGGYACWYGCAGA CATGTCATRGAWSTTAGATTRAAGAA	(2)
CSE1L-12	ENSG00000124207	1496	CATGGRATYACAMAAGCWAATGA TAYTTRATRCCRTCAGCTTTAAG	(3)
TBC1D21-8	ENSG00000167139	1007	TCTTYCCCTGGTTCTGYYTCTGCTT CAKGCWGTAGGCCACCAGCACCT	(4)

The intron name is taken from the gene name according to the HUGO Gene Nomenclature Committee followed by a dash and the intron number. In the "Result" column, the (+) symbol indicates successful sequencing in the whole panel of species; (1), non-specific amplification in one species; (2), non-specific amplification in several species; (3), intron of very large size in some species; (4), double PCR band in some species.

corresponding sequence, indicating again the usefulness of this set of introns.

Use of the selected introns as markers for the phylogeny of closely related species

The sequenced introns were added to the introns already downloaded for human, chimpanzee, macaque, dog and cow, and this set was complemented by three additional genomes that had information for the seven successful introns: the Sumatran orangutan (*Pongo abelii*), the horse (*Equus caballus*) and the little brown bat (*Myotis lucifugus*). This extended species set allowed us to assess the variability of our introns in five pairs of closely related mammalian species, which comprised four different mammalian families: Mustelidae, which included *Mustela nivalis* and *M. putorius*, with an estimated divergence time of 2.8 million years [53]; Felidae, with *Panthera tigris* and *Uncia uncia*, which diverged 2.9 million years ago [54]; Hominidae, which included two pairs, namely, *Pongo pygmaeus* and *P. abelii*, with 3.8 million years of estimated divergence [55], and *Homo sapiens* and *Pan troglodytes*, with 6 million years of estimated divergence [56]; and Vespertilionidae, with *Myotis lucifugus* and *M. escalerai*, which diverged 12.2 million years ago [57].

All of the sequenced alleles for the seven introns were different in both members of every pair of closely related species, except, surprisingly, the PNPO-3 sequence obtained for *Pongo pygmaeus*, which was identical to the one downloaded from ENSEMBL for *P. abelii*. This can be due to several reasons such as a recent introgression, the existence of a constraint in the evolution of this particular marker in the orangutan lineage or the fact that, by mere chance, no mutations have accumulated in any of the two species since their recent divergence. All other intron pairs showed one or several substitutions between the species pairs selected, thus providing useful information for phylogenetic reconstruction.

Alignments and the corresponding maximum-likelihood phylogenetic trees were reconstructed for each of the seven intronic markers (Figure 5). The phylogenies obtained were largely congruent with the known taxonomy of the species, particularly within each order, showing that these introns have normal evolutionary dynamics, not only in the five species used for filtering the introns, but also in other mammals. In addition, it is interesting to note that we have four representative orders (Carnivora, Chiroptera, Cetartiodactyla and Perissodactyla) within laurasiatherians that could help resolve the phylogeny of this group. However, the relationships among them were different in each tree, and the bootstrap values for the different clades were very low (results not shown), indicating that individual introns may not have enough information for inter-ordinal mammalian phylogenies [58-62]. It is also important to observe that the trees estimated with the different markers show very apparent differences in their overall divergence. As can be seen from the scale of the trees, the most extreme examples were GAD2-1 and OSTA-5, which were the slowest introns, and JMJD5-2, which was the intron with the fastest rate of evolution.

To analyze the differences in divergence among introns in more detail, we constructed new alignments for each pair of closely related species to avoid possible problems associated with the alignments of more divergent taxa. We then estimated the intron genetic distance between species pairs from the corresponding intron pairwise alignment. Figure 6 compares the divergences between species pairs for the different introns. As expected, species pairs separated for longer times such as the two *Myotis* species accumulate an overall higher

Igea *et al. BMC Evolutionary Biology* 2010, **10**:369
http://www.biomedcentral.com/1471-2148/10/369

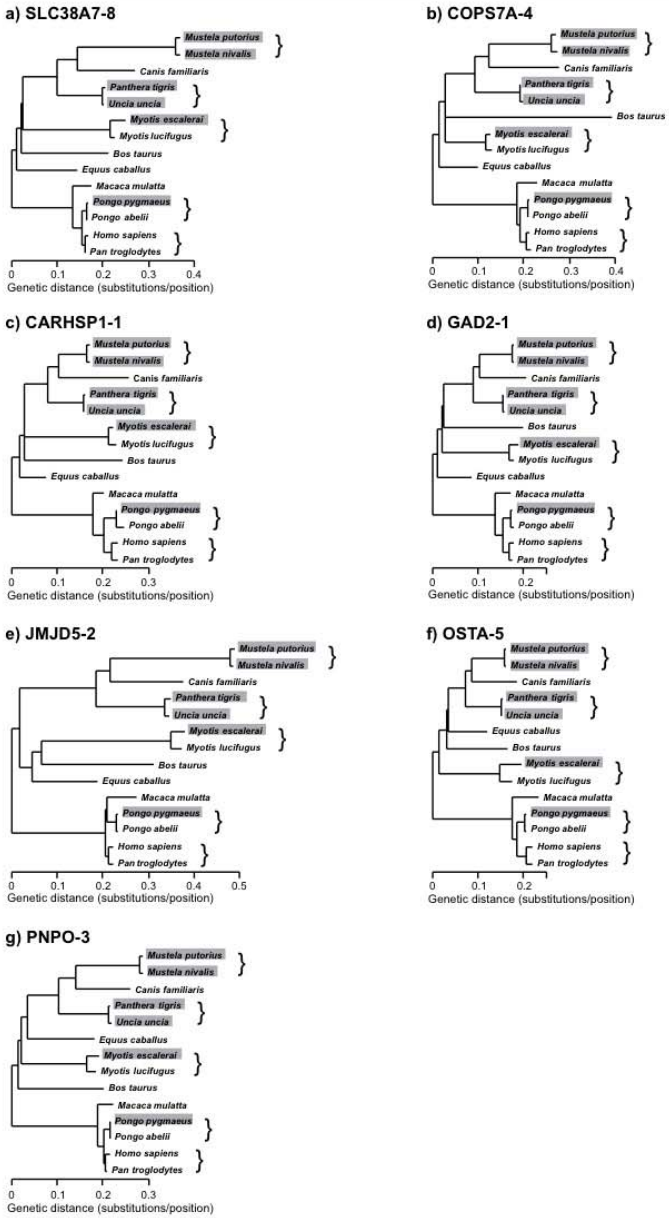

Figure 5 Maximum-likelihood phylogenetic trees of seven selected introns. Grey boxes indicate the new sequences obtained in this study and brackets show the pairs of closely related species specifically analyzed. The root was placed at 7% of the branch separating primates and laurasiatherians (coinciding with the midpoint of the global genomic tree). All trees are drawn at the same scale.

Igea et al. BMC Evolutionary Biology 2010, 10:369
http://www.biomedcentral.com/1471-2148/10/369

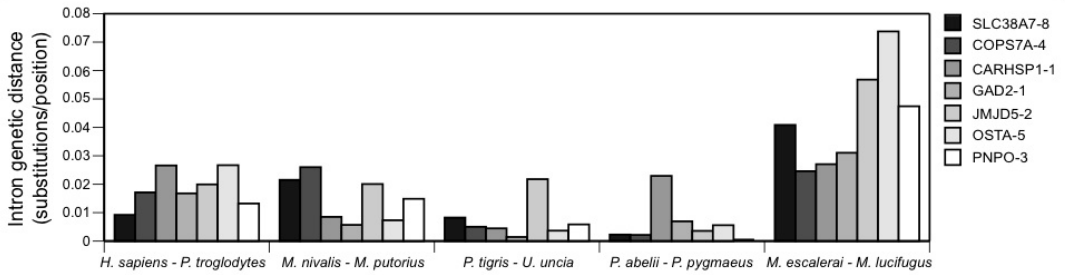

Figure 6 Pairwise maximum-likelihood distances of the different introns for the selected species pairs.

number of substitutions than the closest species such as *Panthera tigris* and *Uncia uncia*. It can also be deduced from this figure that there are important differences in intron divergence in different species pairs, indicating that it is not possible to predict that one intron with a high number of variable positions in one lineage will be equally variable in other lineages. For example, the most informative introns for *H. sapiens* and *P. troglodytes* would be CARHSP1-1 and OSTA-5 whereas for *Mustela* the best ones would be SLC38A7-8 and COPS7A-4. Furthermore, the intron JMJD5-2, despite being the fastest one according to the overall divergence in mammals (Figure 5), was not always the most variable between sister species. Therefore, the most variable introns are different for every lineage. This can be due to the stochasticity of mutations, which can specially affect to short branches, and to differences in population size and ancestral polymorphisms, which may constitute a large part of the divergence among genes in closely related species [13,63]. Thus, given the range of intron variability of our data set (with all introns being quite variable) it may not be worthwhile favoring introns with high overall rates in mammals. Rather, it may be better to use several unlinked introns to overcome different stochastic processes in any phylogenetic study of closely related species.

Conclusions

The development of multilocus phylogenetics requires the availability of a large number of adequate markers for different taxa. In this study, we have designed 224 intronic markers with optimal conditions for the phylogeny of closely related mammalian species (excluding rodents, marsupials and monotremes). Among the important criteria used to select these markers were that they belonged to single-copy genes, were not highly conserved, and did not show disparate rates of evolution in different lineages. The experimental validation of these introns showed that, after some optimizations, they could be amplified in different mammalian species, yielding a single PCR band. In addition, an analysis of the genetic distances estimated in several pairs of closely related mammalian species revealed that introns may show different degrees of divergence in different pairs, and that it is not possible to predict which introns will be more variable in each group. In any case, it may be necessary to carry out initial pilot studies with several introns to decide which ones perform best for each species or group of closely related species. In addition, the use of several introns will reduce the stochasticity associated to mutational and coalescence processes, which is particularly large for nuclear sequences. The availability of a large set of introns like the one provided here will greatly facilitate this work.

Methods

Databases and filters

The following mammalian genomes were downloaded from the ENSEMBL database [64] in GenBank format: human (*Homo sapiens*) version NCBI 36 [36,39], chimpanzee (*Pan troglodytes*) version Pan_troglodytes-2.1 [35], rhesus macaque (*Macaca mulatta*) version Mmul_1 [38], dog (*Canis familiaris*) version CanFam2.0 [37] and cow (*Bos taurus*) version Btau_3.1 [34]. These genomes are sequenced with high coverage and thoroughly annotated. There were also several low-coverage genomes available, but their lack of many annotated features made their use in studies like the present one very inconvenient [64,65]. The genome of the marsupial *Monodelphis domestica* [66] was also available and well annotated at the time this study was performed but we decided not to include it to avoid problems with alignments and phylogenetic reconstruction due to its large distance from the other mammalian species used. In addition, the mouse and rat genomes, which are also very divergent due to their peculiar genomic features [40,41], were also excluded from this study.

We developed a pipeline of Perl scripts to extract all the introns and exons of each mammalian genome (Figure 1). At the same time, features like the gene

Igea et al. BMC Evolutionary Biology 2010, 10:369
http://www.biomedcentral.com/1471-2148/10/369

description and the genomic location and length of every intron and exon were stored. In this initial step, introns of more than 50,000 nucleotides were discarded.

The ENSEMBL database includes orthology information obtained from a phylogenetic analysis, which makes the determination of the orthology relationships more accurate [64]. Therefore, we used this information to determine preliminary one-to-one orthologous genes between human, chimpanzee, macaque, dog and cow. To achieve this, we downloaded orthology tables containing the whole list of one-to-one orthologous genes between pairs of species. By crossing the information of the human-chimpanzee and human-macaque pairs we obtained a set of one-to-one orthologous genes for primates. From the table of dog and cow we obtained the corresponding set of orthologues for laurasiatherians. Finally, we crossed these two tables to obtain a set of putative one-to-one orthologous genes for the five considered species. The corresponding lists of introns were then assembled from this table. In addition, only those introns belonging to genes with equal numbers of introns and whose relative position inside the corresponding gene was conserved in every compared species were considered to be true one-to-one orthologues.

The size variation of each intron across the studied species was also controlled. To do this, the maximum difference in length allowed between human and chimpanzee, human and macaque, human and cow, human and dog, and cow and dog was established at 10%, 20%, 100%, 100% and 50%, respectively.

Multiple alignments of the orthologous intronic sequences were built using MAFFT version 5.8 [67]. Gblocks version 0.91 [68] was used with relaxed parameters [69] to discard poorly aligned positions. Maximum-likelihood phylogenetic trees were reconstructed using PhyML version 2.4.4, with the GTR model of evolution and four substitution rate categories [70].

To detect duplicated genes, BLAST searches [71] of the pair of flanking exons were performed for every intron. For each species, the exons were used as query against its own genome. The e-value limit was set at 10^{-4}. We selected only the introns with exons that produced a single hit (itself) against one region of the genome. Since each genome is subdivided into large fragments, normally chromosomes, we checked that every single hit was also composed of a single subalignment and therefore corresponded to a single exon.

A reference tree reflecting the global genomic evolution was constructed using all single-copy introns. To do so, the 1344 intron alignments that passed the BLAST filter were concatenated to generate an alignment of 768,745 positions. This alignment was used to build a maximum-likelihood phylogenetic tree using RaxML version 2.4.4, which can handle big alignments [72]. Phylogenetic trees of every individual intron were compared to this reference tree using the K tree score implemented in the Ktreedist program [27].

Repetitive sequences present in the final set of introns were analyzed with the program RepeatMasker version 3.2.8 using the corresponding library for each species [73]. The intron sequences were also scanned for microsatellites with Tandem Repeats Finder version 4.04 [49] using the following program parameters: 2 7 7 80 10 50 6. Moreover, the presence of human SNPs in these introns was analyzed using information obtained from the ENSEMBL database. To do this, the whole list of annotated intronic SNPs of the NCBI 36 version of the *Homo sapiens* genome was downloaded using the Biomart platform [74], and the corresponding SNPs were mapped in our set of introns.

Samples and laboratory procedures

Samples of six mammalian species spanning different orders were obtained. The chosen species were the Iberian mouse-bat (*Myotis escalerai*); the Bornean orangutan (*Pongo pygmaeus*), obtained from the INPRIMAT Consortium and the Biomedical Primate Research Centre in Rijswijk; the snow leopard (*Uncia uncia*) and the tiger (*Panthera tigris*), obtained from the Animal Tissue Bank of Catalunya; and the least weasel (*Mustela nivalis*) and the European polecat (*Mustela putorius*), obtained from Collection of Tissues and DNA of the Museo Nacional de Ciencias Naturales.

DNA extractions were performed using the DNEasy Blood and Tissue Kit (QIAGen), following manufacturer's instructions. PCR reactions were carried out in 25 μl of final reaction volume, containing 50 - 100 ng of template DNA, 0.5 - 1 μM of each primer, and 0.75 units of Promega GoTaq. Amplification conditions were as follows: initial denaturation at 95°C for 5 minutes, followed by 28 to 35 cycles comprising 30 seconds of denaturation at 95°C, 30 seconds of annealing at 52-65° C (this temperature turned out to be highly variable among different species and introns and had to be adjusted for each particular case) and extension at 72°C for 30-60 seconds. The resulting PCR products were loaded in 1% agarose gels stained with SYBR Safe DNA Gel Stain (Invitrogen). Bands were manually excised from the gel and subsequently purified using Illustra GFX PCR and DNA Gel Band Purification Kit (GE Healthcare). The resulting products were then sequenced using different sequencing services. Both heterozygous point mutations and indels that caused difficulties in assigning a clear sequence to each allele of the individuals were observed in several sequences. Therefore, we cloned each PCR product into the pstBlue-1 vector (Invitrogen) to be able to isolate one of the alleles. The plasmid containing one of the alleles was isolated from the bacterial cultures, purified using the

GenElute Plasmid Miniprep Kit (Sigma Aldrich), and subsequently sequenced. Resulting sequences were analyzed using Geneious version 4.5.5 [75]. To correct the errors induced by the polymerase and evidenced in the cloning process, both the PCR product and its corresponding cloned sequence were assembled. Intron sequences have been deposited in GenBank under accession numbers HM147892-HM147933.

Phylogenetic analyses of the new markers

For each successfully amplified and sequenced intron, a new multiple sequence alignment was created containing the six newly obtained sequences plus further sequences from ENSEMBL. As above, MAFFT was used to build these alignments, which were then cleaned using Gblocks with the same parameters as above. We constructed the phylogenetic trees using PhyML with the GTR model of evolution and four substitution rate categories. In addition, for each of these introns we analyzed the divergence for five pairs of closely related species. To do this, we constructed new pairwise intronic alignments for each species pair using MAFFT as above, but without applying Gblocks, and estimated maximum-likelihood distance using PAUP [76] with the GTR substitution model.

Additional material

Additional file 1: Final set of 224 introns selected for the phylogeny of closely related mammalian species. The first page of each intron lists the following information: description (function of the gene to which the intron belongs), gene name according to the HUGO Gene Nomenclature Committee (in parenthesis), intron number, chromosome where it is located in *Homo sapiens*, intron start (the location of the first base of the intron in the corresponding human chromosome), human intron length, intron alignment length (size of the alignment of the five species: human, chimpanzee, rhesus macaque, dog and cow), flanking exons length (the size of both the upstream and the downstream exons that flank this intron in *Homo sapiens*), SNP density (the number of single nucleotide polymorphisms described for human in this intron divided by its length), K tree score (a calculation that reflects topological and rate divergence of the intronic tree with respect to the genomic reference tree), scaling factor (the value that shows the global divergence of the intronic tree with respect to the genomic reference tree), human-chimpanzee distance in substitutions per position (the maximum-likelihood distance between the introns of *Homo sapiens* and *Pan troglodytes* measured in the primates phylogenetic tree), and total primate branch length in substitutions per position (the sum of all the branches of the corresponding phylogenetic tree built using only the three available primate species). For the representation of the exon and intron alignments, a few wrong definitions of intron ends and starts that we found were manually corrected. The five-species alignments of both upstream and downstream exons are represented, but only the 160 bases closest to the intron are displayed. To help assess sequence conservation, positions with more than 50% identities are highlighted. The maximum-likelihood phylogenetic tree of the intron for the five species is also shown. The root was placed at 7% of the branch separating primates and laurasiatherians (coinciding with the midpoint of the global genomic tree). The second page of each intron shows the five-species alignment of the intron. Introns are ordered in this document by total primate branch length.

Acknowledgements
We thank the collection of tissues and DNA of the Museo Nacional de Ciencias Naturales (CSIC), the EBD-CSIC's scientific collections, the Animal Tissue Bank of Catalunya (BTAC) of the Universidad Autónoma de Barcelona (UAB), and Dr. Ernst J Verschoor of the INPRIMAT Consortium and the Biomedical Primate Research Centre in Rijswijk for samples. We also thank M. Dolors Piulachs and people in her lab for technical advice. JI is recipient of a pre-doctoral research fellowship (I3P) from CSIC. JC is supported by grant number CGL2008-00434/BOS from the Plan Nacional I+D+I of the Ministerio de Ciencia e Innovación (Spain).

Author details
[1]Institute of Evolutionary Biology (CSIC-UPF), Passeig Maritim de la Barceloneta 37, 08003 Barcelona, Spain. [2]Estacion Biologica de Donana (CSIC), Avda. Americo Vespucio s/n, 41092 Seville, Spain.

Authors' contributions
JC and JI conceived and designed the experiments. JI did the experimental work and the main analyses. JJ contributed to the original design of the problem, participated in discussions and collaborated with samples. JC and JI wrote the paper. All authors reviewed and approved the final manuscript.

Received: 4 May 2010 Accepted: 30 November 2010
Published: 30 November 2010

References

1. Arbogast BS, Edwards SV, Wakeley J, Beerli P, Slowinski JB: **Estimating divergence times from molecular data on phylogenetic and population genetic timescales.** *Annu Rev Ecol Syst* 2002, **33**:707-740.
2. Brown WM, George M Jr, Wilson AC: **Rapid evolution of animal mitochondrial DNA.** *Proc Natl Acad Sci USA* 1979, **76(4)**:1967-1971.
3. Johns GC, Avise JC: **A comparative summary of genetic distances in the vertebrates from the mitochondrial cytochrome *b*.** *Mol Biol Evol* 1998, **15**:1481-1490.
4. Takahata N: **Gene genealogy in three related populations: consistency probability between gene and population trees.** *Genetics* 1989, **122(4)**:957-966.
5. Maddison WP: **Gene trees in species trees.** *Syst Biol* 1997, **46**:523-536.
6. Nichols R: **Gene trees and species trees are not the same.** *Trends Ecol Evol* 2001, **16(7)**:358-364.
7. Liu L, Pearl DK, Brumfield RT, Edwards SV: **Estimating species trees using multiple-allele DNA sequence data.** *Evolution* 2008, **62(8)**:2080-2091.
8. Carstens BC, Knowles LL: **Estimating Species Phylogeny from Gene-Tree Probabilities Despite Incomplete Lineage Sorting: An Example from Melanoplus Grasshoppers.** *Syst Biol* 2007, **56(3)**:400-411.
9. Degnan JH, Rosenberg NA: **Gene tree discordance, phylogenetic inference and the multispecies coalescent.** *Trends Ecol Evol* 2009, **24(6)**:332-340.
10. Brito PH, Edwards SV: **Multilocus phylogeography and phylogenetics using sequence-based markers.** *Genetica* 2009, **135(3)**:439-455.
11. Moore WS: **Inferring phylogenies from mtDNA variation: mitochondrial-gene trees versus nuclear-gene trees.** *Evolution* 1995, **49(4)**:718-726.
12. Zink RM, Barrowclough GF: **Mitochondrial DNA under siege in avian phylogeography.** *Mol Ecol* 2008, **17(9)**:2107-2121.
13. Zhang DX, Hewitt GM: **Nuclear DNA analyses in genetic studies of populations: practice, problems and prospects.** *Molecular ecology* 2003, **12(3)**:563-584.
14. Townsend TM, Alegre RE, Kelley ST, Wiens JJ, Reeder TW: **Rapid development of multiple nuclear loci for phylogenetic analysis using genomic resources: an example from squamate reptiles.** *Mol Phylogenet Evol* 2008, **47(1)**:129-142.
15. Thomson RC, Shedlock AM, Edwards SV, Shaffer HB: **Developing markers for multilocus phylogenetics in non-model organisms: A test case with turtles.** *Mol Phylogenet Evol* 2008, **49(2)**:514-525.
16. Ranwez V, Delsuc F, Ranwez S, Belkhir K, Tilak MK, Douzery EJ: **OrthoMaM: a database of orthologous genomic markers for placental mammal phylogenetics.** *BMC Evol Biol* 2007, **7**:241.
17. Peng Z, Elango N, Wildman DE, Yi SV: **Primate phylogenomics: developing numerous nuclear non-coding, non-repetitive markers for ecological and phylogenetic applications and analysis of evolutionary rate variation.** *BMC Genomics* 2009, **10**:247.

18. Li C, Ortí G, Zhang G, Lu G: **A practical approach to phylogenomics: the phylogeny of ray-finned fish (Actinopterygii) as a case study.** *BMC Evol Biol* 2007, **7**:44.
19. Kimball RT, Braun EL, Barker FK, Bowie RCK, Braun MJ, Chojnowski JL, Hackett SJ, Han KL, Harshman J, Heimer-Torres V, Holznagel W, Huddleston CJ, Marks BD, Miglia KJ, Moore WS, Reddy S, Sheldon FH, Smith JV, Witt CC, Yuri T: **A well-tested set of primers to amplify regions spread across the avian genome.** *Mol Phylogenet Evol* 2009, **50(3)**:654-660.
20. Aitken N, Smith S, Schwarz C, Morin PA: **Single nucleotide polymorphism (SNP) discovery in mammals: a targeted-gene approach.** *Molecular ecology* 2004, **13(6)**:1423-1431.
21. Backström N, Fagerberg S, Ellegren H: **Genomics of natural bird populations: a gene-based set of reference markers evenly spread across the avian genome.** *Molecular ecology* 2008, **17(4)**:964-980.
22. Creer S, Malhotra A, Thorpe R, Pook C: **Targeting optimal introns for phylogenetic analyses in non-model taxa: experimental results in Asian pitvipers.** *Cladistics* 2005, **21(4)**:390-395.
23. Hellborg L, Ellegren H: **Y chromosome conserved anchored tagged sequences (YCATS) for the analysis of mammalian male-specific DNA.** *Molecular ecology* 2003, **12(1)**:283-291.
24. Rodova M, Islam MR, Peterson KR, Calvet JP: **Remarkable sequence conservation of the last intron in the PKD1 gene.** *Mol Biol Evol* 2003, **20(10)**:1669-1674.
25. Gazave E, Marqués-Bonet T, Fernando O, Charlesworth B, Navarro A: **Patterns and rates of intron divergence between humans and chimpanzees.** *Genome Biol* 2007, **8(2)**:R21.
26. Castresana J: **Estimation of genetic distances from human and mouse introns.** *Genome Biol* 2002, **3**:research0028.0021-0028.0027.
27. Soria-Carrasco V, Talavera G, Igea J, Castresana J: **The K tree score: quantification of differences in the relative branch length and topology of phylogenetic trees.** *Bioinformatics* 2007, **23(21)**:2954-2956.
28. Plass M, Eyras E: **Differentiated evolutionary rates in alternative exons and the implications for splicing regulation.** *BMC Evol Biol* 2006, **6**:50.
29. Bardeleben C, Moore RL, Wayne RK: **A molecular phylogeny of the Canidae based on six nuclear loci.** *Molecular Phylogenetics and Evolution* 2005, **37(3)**:815-831.
30. Castillo A, Cortinas M, Lessa E: **Rapid diversification of South American tuco-tucos (*Ctenomys*; Rodentia, Ctenomyidae): contrasting mitochondrial and nuclear intron sequences.** *J Mammal* 2005, **86(1)**:170-179.
31. DeBry R, Seshadri S: **Nuclear intron sequences for phylogenetics of closely related mammals: an example using the phylogeny of *Mus*.** *J Mammal* 2001, **82(2)**:280-288.
32. Gilbert C, Ropiquet A, Hassanin A: **Mitochondrial and nuclear phylogenies of Cervidae (Mammalia, Ruminantia): Systematics, morphology, and biogeography.** *Mol Phylogenet Evol* 2006, **40(1)**:101-117.
33. Yu L, Zhang YP: **Phylogeny of the caniform carnivora: evidence from multiple genes.** *Genetica* 2006, **127(1-3)**:65-79.
34. Bovine Genome Sequencing and Analysis Consortium, Elsik CG, Tellam RL, Worley KC: **The genome sequence of taurine cattle: a window to ruminant biology and evolution.** *Science* 2009, **324(5926)**:522-528.
35. Chimpanzee Sequencing and Analysis Consortium: **Initial sequence of the chimpanzee genome and comparison with the human genome.** *Nature* 2005, **437(7055)**:69-87.
36. International Human Genome Sequencing Consortium: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409(6822)**:860-921.
37. Lindblad-Toh K, *et al*: **Genome sequence, comparative analysis and haplotype structure of the domestic dog.** *Nature* 2005, **438(7069)**:803-819.
38. Rhesus Macaque Genome Sequencing and Analysis Consortium: **Evolutionary and biomedical insights from the rhesus macaque genome.** *Science* 2007, **316(5822)**:222-234.
39. Venter JC, *et al*: **The sequence of the human genome.** *Science* 2001, **291(5507)**:1304-1351.
40. Mouse Genome Sequencing Consortium: **Initial sequencing and comparative analysis of the mouse genome.** *Nature* 2002, **420(6915)**:520-562.
41. Rat Genome Sequencing Project Consortium: **Genome sequence of the Brown Norway rat yields insights into mammalian evolution.** *Nature* 2004, **428(6982)**:493-521.
42. Zhou Y, Mishra B: **Quantifying the mechanisms for segmental duplications in mammalian genomes by statistical analysis and modeling.** *Proc Natl Acad Sci USA* 2005, **102(11)**:4051-4056.
43. Alba MM, Castresana J: **On homology searches by protein Blast and the characterization of the age of genes.** *BMC Evol Biol* 2007, **7**:53.
44. Gaubert P, Cordeiro-Estrela P: **Phylogenetic systematics and tempo of evolution of the Viverrinae (Mammalia, Carnivora, Viverridae) within feliformians: implications for faunal exchanges between Asia and Africa.** *Mol Phylogenet Evol* 2006, **41(2)**:266-278.
45. Ingram CM, Burda H, Honeycutt RL: **Molecular phylogenetics and taxonomy of the African mole-rats, genus *Cryptomys* and the new genus *Coetomys* Gray, 1864.** *Mol Phylogenet Evol* 2004, **31(3)**:997-1014.
46. Rowe DL, Honeycutt RL: **Phylogenetic relationships, ecological correlates, and molecular evolution within the cavioidea (mammalia, rodentia).** *Mol Biol Evol* 2002, **19(3)**:263-277.
47. Steiner C, Tilak MK, Douzery EJ, Catzeflis FM: **New DNA data from a transthyretin nuclear intron suggest an Oligocene to Miocene diversification of living South America opossums (Marsupialia: Didelphidae).** *Mol Phylogenet Evol* 2005, **35(2)**:363-379.
48. Yu L, Li QW, Ryder OA, Zhang YP: **Phylogeny of the bears (Ursidae) based on nuclear and mitochondrial genes.** *Mol Phylogenet Evol* 2004, **32(2)**:480-494.
49. Benson G: **Tandem repeats finder: a program to analyze DNA sequences.** *Nucleic Acids Res* 1999, **27(2)**:573-580.
50. Tautz D, Trick M, Dover GA: **Cryptic simplicity in DNA is a major source of genetic variation.** *Nature* 1986, **322(6080)**:652-656.
51. Castresana J: **Cytochrome *b* phylogeny and the taxonomy of great apes and mammals.** *Mol Biol Evol* 2001, **18(4)**:465-471.
52. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW: **GenBank.** *Nucleic Acids Res* 2009, , **37 Database:** D26-31.
53. Koepfli KP, Deere KA, Slater GJ, Begg C, Begg K, Grassman L, Lucherini M, Veron G, Wayne RK: **Multigene phylogeny of the Mustelidae: resolving relationships, tempo and biogeographic history of a mammalian adaptive radiation.** *BMC Biol* 2008, **6**:10.
54. Johnson WE, Eizirik E, Pecon-Slattery J, Murphy WJ, Antunes A, Teeling E, O'Brien SJ: **The late Miocene radiation of modern Felidae: a genetic assessment.** *Science* 2006, **311(5757)**:73-77.
55. Raaum RL, Sterner KN, Noviello CM, Stewart CB, Disotell TR: **Catarrhine primate divergence dates estimated from complete mitochondrial genomes: concordance with fossil and nuclear DNA evidence.** *J Hum Evol* 2005, **48(3)**:237-257.
56. Kumar S, Filipski A, Swarna V, Walker A, Hedges SB: **Placing confidence limits on the molecular age of the human-chimpanzee divergence.** *Proc Natl Acad Sci USA* 2005, **102(52)**:18842-18847.
57. Stadelmann B, Lin LK, Kunz TH, Ruedi M: **Molecular phylogeny of New World *Myotis* (Chiroptera, Vespertilionidae) inferred from mitochondrial and nuclear DNA genes.** *Mol Phylogenet Evol* 2007, **43(1)**:32-48.
58. Murphy WJ, Eizirik E, O'Brien SJ, Madsen O, Scally M, Douady CJ, Teeling E, Ryder OA, Stanhope MJ, de Jong WW, Springer MS: **Resolution of the early placental mammal radiation using Bayesian phylogenetics.** *Science* 2001, **294(5550)**:2348-2351.
59. Murphy WJ, Pevzner PA, O'Brien SJ: **Mammalian phylogenomics comes of age.** *Trends Genet* 2004, **20(12)**:631-639.
60. Springer MS, Stanhope MJ, Madsen O, de Jong WW: **Molecules consolidate the placental mammal tree.** *Trends Ecol Evol* 2004, **19(8)**:430-438.
61. Waddell PJ, Kishino H, Ota R: **A phylogenetic foundation for comparative mammalian genomics.** *Genome informatics International Conference on Genome Informatics* 2001, **12**:141-154.
62. Waddell PJ, Okada N, Hasegawa M: **Towards resolving the interordinal relationships of placental mammals.** *Systematic Biology* 1999, **48(1)**:1-5.
63. Edwards SV, Beerli P: **Perspective: gene divergence, population divergence, and the variance in coalescence time in phylogeographic studies.** *Evolution* 2000, **54(6)**:1839-1854.
64. Hubbard TJP, Aken BL, Beal K, Ballester B, Caccamo M, Chen Y, Clarke L, Coates G, Cunningham F, Cutts T, Down T, Dyer SC, Fitzgerald S, Fernandez-Banet J, Graf S, Haider S, Hammond M, Herrero J, Holland R, Howe K, Howe K, Johnson N, Kahari A, Keefe D, Kokocinski F, Kulesha E, Lawson D, Longden I, Melsopp C, Megy K, Meidl P, Overduin B, Parker A, Prlic A, Rice S, Rios D, Schuster M, Sealy I, Severin J, Slater G, Smedley D,

Spudich G, Trevanion S, Vilella A, Vogel J, White S, Wood M, Cox T, Curwen V, Durbin R, Fernandez-Suarez XM, Flicek P, Kasprzyk A, Proctor G, Searle S, Smith J, Ureta-Vidal A, Birney E: **Ensembl 2007.** *Nucleic Acids Res* 2007, **35**:D610-D617.

65. Green P: **2x genomes–does depth matter?** *Genome Res* 2007, **17(11)**:1547-1549.
66. Mikkelsen TS, Wakefield MJ, Aken B, Amemiya CT, Chang JL, Duke S, Garber M, Gentles AJ, Goodstadt L, Heger A, Jurka J, Kamal M, Mauceli E, Searle SM, Sharpe T, Baker ML, Batzer MA, Benos PV, Belov K, Clamp M, Cook A, Cuff J, Das R, Davidow L, Deakin JE, Fazzari MJ, Glass JL, Grabherr M, Greally JM, Gu W, Hore TA, Huttley GA, Kleber M, Jirtle RL, Koina E, Lee JT, Mahony S, Marra MA, Miller RD, Nicholls RD, Oda M, Papenfuss AT, Parra ZE, Pollock DD, Ray DA, Schein JE, Speed TP, Thompson K, VandeBerg JL, Wade CM, Walker JA, Waters PD, Webber C, Weidman JR, Xie X, Zody MC, Graves JA, Ponting CP, Breen M, Samollow PB, Lander ES, Lindblad-Toh K: **Genome of the marsupial *Monodelphis domestica* reveals innovation in non-coding sequences.** *Nature* 2007, **447(7141)**:167-177.
67. Katoh K, Kuma K, Toh H, Miyata T: **MAFFT version 5: improvement in accuracy of multiple sequence alignment.** *Nucleic Acids Res* 2005, **33(2)**:511-518.
68. Castresana J: **Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis.** *Mol Biol Evol* 2000, **17(4)**:540-552.
69. Talavera G, Castresana J: **Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments.** *Syst Biol* 2007, **56(4)**:564-577.
70. Guindon S, Gascuel O: **A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood.** *Syst Biol* 2003, **52(5)**:696-704.
71. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25(17)**:3389-3402.
72. Stamatakis A: **RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models.** *Bioinformatics* 2006, **22(21)**:2688-2690.
73. Smit AFA, Hubley R, Green P: **RepeatMasker Open-3.0. 1996-2004.** 2004 [http://repeatmasker.org].
74. Smedley D, Haider S, Ballester B, Holland R, London D, Thorisson G, Kasprzyk A: **BioMart–biological queries made easy.** *BMC Genomics* 2009, **10**:22.
75. Drummond AJ, Ashton B, Cheung M, Heled J, Kearse M, Moir R, Stones-Havas S, Thierer T, Wilson A: **Geneious v4.** 2009 [http://geneious.com].
76. Swofford DL: **PAUP*: Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4.** Sunderland, Massachusetts.: Sinauer Associates; 2003.

doi:10.1186/1471-2148-10-369
Cite this article as: Igea *et al.*: Novel intron markers to study the phylogeny of closely related mammalian species. *BMC Evolutionary Biology* 2010 **10**:369.

PUBLICACIÓN 2

Soria-Carrasco, V., Talavera, G., Igea, J. and Castresana, J. (2007). **The K tree score: quantification of differences in the relative branch length and topology of phylogenetic trees.** *Bioinformatics 23, 2954-2956*

BIOINFORMATICS APPLICATIONS NOTE Vol. 23 no. 21 2007, pages 2954–2956
doi:10.1093/bioinformatics/btm466

Phylogenetics

The K tree score: quantification of differences in the relative branch length and topology of phylogenetic trees

Víctor Soria-Carrasco, Gerard Talavera, Javier Igea and Jose Castresana*

Department of Physiology and Molecular Biodiversity, Institute of Molecular Biology of Barcelona, CSIC, Jordi Girona 18, 08034 Barcelona, Spain

Received on June 26, 2007; revised on August 13, 2007; accepted on September 6, 2007

Advance Access publication September 22, 2007

Associate Editor: Keith Crandall

ABSTRACT

Summary: We introduce a new phylogenetic comparison method that measures overall differences in the relative branch length and topology of two phylogenetic trees. To do this, the algorithm first scales one of the trees to have a global divergence as similar as possible to the other tree. Then, the branch length distance, which takes differences in topology and branch lengths into account, is applied to the two trees. We thus obtain the minimum branch length distance or K tree score. Two trees with very different relative branch lengths get a high K score whereas two trees that follow a similar among-lineage rate variation get a low score, regardless of the overall rates in both trees. There are several applications of the K tree score, two of which are explained here in more detail. First, this score allows the evaluation of the performance of phylogenetic algorithms, not only with respect to their topological accuracy, but also with respect to the reproduction of a given branch length variation. In a second example, we show how the K score allows the selection of orthologous genes by choosing those that better follow the overall shape of a given reference tree.

Availability: http://molevol.ibmb.csic.es/Ktreedist.html

Contact: jcvagr@ibmb.csic.es

1 INTRODUCTION

In phylogenetic reconstruction, the application of different methods or the use of different genes may lead to the estimation of different phylogenetic trees (Castresana, 2007; Hillis *et al.*, 2005; Huerta-Cepas *et al.*, 2007). In order to analyze if the resulting trees are congruent, it is fundamental to be able to quantify differences between such trees. Normally, only topology is taken into account for such task, for example, by means of the symmetric difference (Robinson and Foulds, 1981). Few methods have been developed that also take branch length information into account (Hall, 2005; Kuhner and Felsenstein, 1994). These methods have been successfully applied to quantify the performance of different phylogenetic methods in simulated alignments, but they have the drawback that they are not directly applicable to trees with different evolutionary rates. Here, we introduce a new phylogenetic comparison measure that takes branch length information into account after scaling the trees so that they have comparable global evolutionary rates.

*To whom correspondence should be addressed.

2 METHOD

The basis of our method to compare two phylogenetic trees, T and T', is the branch length distance (*BLD*) introduced by Kuhner and Felsenstein (Felsenstein, 2004; Kuhner and Felsenstein, 1994). This distance is sensitive to the similarity in branch lengths of both trees. Consider the set of partitions present in both trees, that is, the whole set of partitions present in T plus the set of partitions present in T' but not in T. Partitions for external branches are also included. For tree T, we can define an array B of branch lengths associated to each partition (b_1, b_2,..., b_N). Branches that do not appear in T (corresponding to partitions that are only present in T') are assigned to 0 in such array. We can similarly define the array B' associated to tree T'. The *BLD* between trees T and T' is the squared root of the sum of $(b_i - b'_i)^2$ for all partitions. However, *BLD* depends on the absolute size of the trees being compared, so that two trees with the same shape (topology and relative branch length) but different global rates will give rise to a very high *BLD* (Kuhner and Felsenstein, 1994), which may be unwanted.

In our method, we introduce a factor, K, to scale tree T' so that both trees, T and T', have a similar global divergence. Thus, we are interested in calculating *BLD* after scaling T' with a factor K:

$$BLD(K) = \sqrt{\sum_{i=1}^{N} (b_i - Kb'_i)^2} \quad (1)$$

To obtain the value of K that minimizes *BLD* we differentiate Equation (1). It can be shown that the value of K that makes this derivative zero is:

$$K = \frac{\sum_{i=1}^{N} (b_i b'_i)}{\sum_{i=1}^{N} b'^2_i} \quad (2)$$

We then substitute this value of K in Equation (1) and obtain the minimum branch length distance or K tree score. It should be taken into account that the K tree score is not symmetric, that is, the result from T to T' may not be the same than from T' to T, and, in consequence, the K score does not have the mathematical properties of a distance. Thus, this score is generally not useful to compare only two trees (although the K factor of Equation (2) can be very valuable for scaling purposes; see below). The K tree score is most useful when there is a tree that serves as reference (T) and several other trees (T') that will be scaled and compared to T. In such cases, trees T' that are similar in shape to T will receive a low K tree score whereas those that are very different will get a relatively higher K score, regardless of their overall rates.

The method that calculates the K tree score (as well as other tree comparison measures) is implemented in a Perl program called Ktreedist.

3 APPLICATIONS

There are several applications of the K tree score. First, it can be used to evaluate the quality of phylogenetic reconstructions in simulated alignments by comparing the true tree to the trees obtained with different phylogenetic methods. For example, the reference tree shown in Figure 1A was used to simulate with SeqGen (Rambaut and Grassly, 1997) 100 alignments of 1000 positions with a GTR model and gamma rate heterogeneity ($\alpha = 1.5$). We then constructed maximum-likelihood (ML) trees from such simulations using Phyml (Guindon and Gascuel, 2003) with two different conditions: without and with rate heterogeneity. To facilitate the comparison between both phylogenetic methods we imposed the topology of the reference tree during the ML reconstructions. After averaging the branch lengths of the 100 reconstructed trees, we obtained one tree for each phylogenetic method. Both trees differed in their overall rates (with the nonrate heterogeneity tree not capturing all substitutions, leading to a K scale factor $\gg 1$) but, importantly, they also differed in their shapes: see, for example, the relative lengths of sp3, sp4 and sp5. The differences in shape were reflected in the K scores: 0.197 for the average tree without rate heterogeneity and 0.030 for the average tree calculated with rate heterogeneity, indicating the better performance of the latter method. (Differences also appeared after averaging the K score from the 100 trees obtained with each method although, in this case, the magnitude of the difference was smaller.) Thus, the K tree score can be used to quantify the different quality in branch length reconstruction of different phylogenetic methods. The K score can also be used with trees that have different topologies. In such cases, nonshared branches that are relatively long will contribute to the K score much more than small conflicting branches. This is different from the symmetric difference (Robinson and Foulds, 1981), in which all topological differences count the same.

In a second example, we show how the K tree score can be used to make an accurate selection of orthologous genes. Orthologs should reflect the same topology of the species tree but they should also give rise, in principle, to a similar tree shape. We extracted from the ENSEMBL database (Hubbard *et al.*, 2007) the tables of pairs of orthologous genes of seven

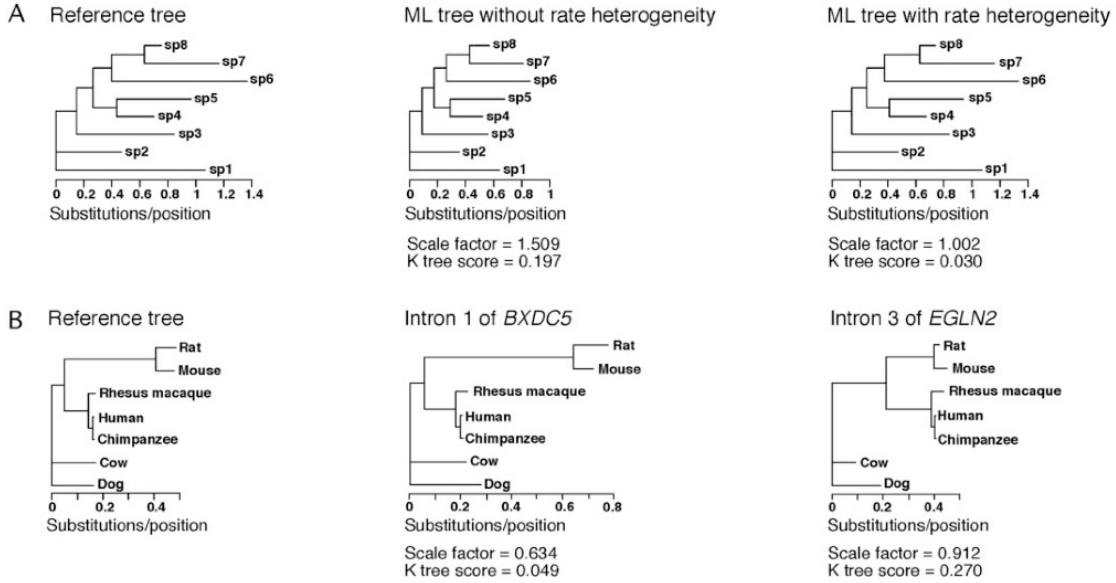

Fig. 1. (**A**) Reference tree used to simulate 100 alignments and the average reconstructions obtained by ML without and with rate heterogeneity. (**B**) Trees obtained with 472 concatenated introns (reference tree) and with two individual introns (intron 1 of *BXDC5* and intron 3 of *EGLN2*).

mammalian species. By matching the pairwise orthology tables, we constructed a set of one-to-one orthologs, and we downloaded the corresponding genes. We then extracted the introns and, after applying several filters (elimination of very long introns, those with problematic alignments, etc.), we obtained a set of 472 putative orthologous introns. Some of these introns produced ML phylogenetic trees that were of unusual shape, which could be due to different rates of evolution in different lineages (heterotachy) or could indicate that they do not come from orthologous genes (hidden paralogy). We then constructed a reference tree (Fig. 1B) with the concatenated alignment of the 472 introns using the RAxML program (Stamatakis, 2006), which can handle very long alignments, with a GTR model of evolution and four rate categories. This tree should reflect the average divergence of the seven genomes and, as expected, rodents showed a higher acceleration in their branches. We then calculated the K score of the trees of all individual introns with respect to the reference tree. We show in Figure 1B the trees of two putative orthologous introns. Intron 1 of *BXDC5*, despite having a high global rate, produced a phylogeny with the same topology and a very similar tree shape to the reference tree. This was reflected in a low K score: 0.049, smaller than the mean of the distribution of K scores of all individual introns (0.104), which is indicative of a very likely ortholog. (The K score would also be low in a similar tree but with a topological conflict affecting a small branch, which would not affect the high probability of orthology.) Intron 3 of *EGLN2* also reproduced the reference topology. However, this tree showed a relatively long basal branch in primates as well as a long branch connecting Euarchontoglires and Laurasiatherians. In consequence, the K score for this tree with respect to the reference is much higher: 0.270. In fact, this value is a clear outlier in the distribution of K scores. Although heterotachy cannot be discarded, the chances that the latter gene contains hidden paralogs in some species are higher than in the first gene. Thus, the K score can be used to establish a certain threshold and make a more accurate selection of orthologous genes.

If orthology is ensured for a set of genes, then a high K tree score with respect to a given reference will be indicative of trees with very fast-evolving species or with a significant amount of other types of heterotachy. These trees are of more difficult reconstruction, and thus the K tree score can be used to select (in a similar way as above) a set of the most reliable genes for estimating species phylogenies.

On a more practical side, the K scale factor [Equation (2)] can be used in instances where it is necessary to scale trees to have equivalent divergences. For example, the linearization of trees by means of a method like nonparametric rate smoothing produces trees with an arbitrary scale when no dates are known for the tree nodes (Sanderson, 1997). In such cases, one can make use of the K scale factor obtained from the comparison between the linearized tree and the original (reference) tree: the scaling of the linearized tree with this K factor will re-establish a genetic distance scale equivalent to that of the original tree.

ACKNOWLEDGEMENTS

J.C. is supported by grant number CGL2005-01341/BOS from the Plan Nacional I+D+I of the MEC (Spain), cofinanced with FEDER funds.

Conflict of Interest: none declared.

REFERENCES

Castresana,J. (2007) Topological variation in single-gene phylogenetic trees. *Genome Biol.*, **8**, 216.

Felsenstein,J. (2004) *Inferring Phylogenies*. Sinauer Associates, Sunderland, Massachusetts, pp. 531–533.

Guindon,S. and Gascuel,O. (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.*, **52**, 696–704.

Hall,B.G. (2005) Comparison of the accuracies of several phylogenetic methods using protein and DNA sequences. *Mol. Biol. Evol.*, **22**, 792–802.

Hillis,D.M. *et al.* (2005) Analysis and visualization of tree space. *Syst. Biol.*, **54**, 471–482.

Hubbard,T.J.P. *et al.* (2007) Ensembl 2007. *Nucleic Acids Res.*, **35**, D610–D617.

Huerta-Cepas,J. *et al.* (2007) The human phylome. *Genome Biol.*, **8**, R109.

Kuhner,M.K. and Felsenstein,J. (1994) A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Mol. Biol. Evol.*, **11**, 459–468.

Rambaut,A. and Grassly,N.C. (1997) Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Comput. Appl. Biosci.*, **13**, 235–238.

Robinson,D.F. and Foulds,L.R. (1981) Comparison of phylogenetic trees. *Math. Biosci.*, **53**, 131–147.

Sanderson,M.J. (1997) A nonparametric approach to estimating divergence times in the absence of rate constancy. *Mol. Biol. Evol.*, **14**, 1218–1231.

Stamatakis,A. (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, **22**, 2688–2690.

Tabla A1. Cebadores usados en esta tesis para amplificar genes mitocondriales. (1) cebadores empleados para amplificar el citocromo *b* en muestras de museo .(2) cebadores empleados para amplificar el D-loop en muestras de museo.

Primer	Secuencia	Gen	Taxa
Galpyr_cytb_Glu	ACTAATGACATGAAAAATCATCGTT	citocromo *b*	*G. pyrenaicus*
Galpyr_cytb_461r	TACAAGATCAGTTCCGATGTAAG	citocromo *b*	*G. pyrenaicus*
Galpyr_cytb_416f	TCTTACCATGGGGTCAAATATC	citocromo *b*	*G. pyrenaicus*
Galpyr_cytb_695r	AGATAATTAGAATGAGGATTAGTG	citocromo *b*	*G. pyrenaicus*
Galpyr_cytb_650f	GGATTATCATCCGACACTGATAA	citocromo *b*	*G. pyrenaicus*
Galpyr_cytb_Thr	TTTTCGTTTTTGGTTTACAAGAC	citocromo *b*	*G. pyrenaicus*
Galpyr_cytb_A1-Glu52F	TGACATGAAAAATCATCGTTGT	citocromo *b*(1)	*G. pyrenaicus*
Galpyr_cytb_A1-70R	TCATGATGAAATGTTTGATGG	citocromo *b*(1)	*G. pyrenaicus*
Galpyr_cytb_A2-136F	AATTTGCCTAGTGCTACAAATTA	citocromo *b*(1)	*G. pyrenaicus*
Galpyr_cytb_A2-250R	AAATATTGAAGCTCCGTTTGC	citocromo *b*(1)	*G. pyrenaicus*
Galpyr_cytb_A3-311F	CGTGGGACGAGGACTGTACTA	citocromo *b*(1)	*G. pyrenaicus*
Galpyr_cytb_A3-403R	TTGCACCTCAAAAGGATATTTG	citocromo *b*(1)	*G. pyrenaicus*
Galpyr_cytb_A4-464F	AATTTACTGTCAGCCATCCCTTA	citocromo *b*(1)	*G. pyrenaicus*
Galpyr_cytb_A4-539R	AATGGCAGGATAAAGTGGAAAG	citocromo *b*(1)	*G. pyrenaicus*
Galpyr_cytb_A5-589F	CAGCTCTAGCAGGCGTACACT	citocromo *b*(1)	*G. pyrenaicus*
Galpyr_cytb_A5-676R	ATTAGTGCTCCTAAAATGTCTTTAAT	citocromo *b*(1)	*G. pyrenaicus*
Galpyr_cytb_A6-738F	TCTCCCTACTCGTCCTATTTTCA	citocromo *b*(1)	*G. pyrenaicus*
Galpyr_cytb_A6-844R	CCTAGTTTGTTAGGAATTGATCG	citocromo *b*(1)	*G. pyrenaicus*
Galpyr_cytb_A7-895F	TTTTAGCCCTAGTCTTATCAATCC	citocromo *b*(1)	*G. pyrenaicus*
Galpyr_cytb_A7-969R	CGGCTGTTAAAAGTCAGAATAGG	citocromo *b*(1)	*G. pyrenaicus*
Galpyr_cytb_A8-1041F	ACAGCCTGTAGAACACCCATTC	citocromo *b*(1)	*G. pyrenaicus*
Galpyr_cytb_A8-1124-R	AAGACTCTTCATTTGAGCATGT	citocromo *b*(1)	*G. pyrenaicus*
Galpyr_cytb_B1-60F	CAACAACTCTTTTATTGATTTACCA	citocromo *b*(1)	*G. pyrenaicus*
Galpyr_cytb_B1-148R	TGATGTGTAATGTATTGCAAGGAA	citocromo *b*(1)	*G. pyrenaicus*
Galpyr_cytb_B2-226F	TTGCCGAGATGTAAACTACG	citocromo *b*(1)	*G. pyrenaicus*
Galpyr_cytb_B2-337R	AACAGGACGCCGATATTTCA	citocromo *b*(1)	*G. pyrenaicus*
Galpyr_cytb_B3-388F	TTATAGCCACCGCATTCATAG	citocromo *b*(1)	*G. pyrenaicus*
Galpyr_cytb_B3-465R	CATTCTACAAGATCAGTTCCGATG	citocromo *b*(1)	*G. pyrenaicus*
Galpyr_cytb_B4-522F	GCTTCTCAGTAGACAAAGCAACA	citocromo *b*(1)	*G. pyrenaicus*
Galpyr_cytb_B4-593R	GATCCTGTTTCGTGAAGGAATA	citocromo *b*(1)	*G. pyrenaicus*
Galpyr_cytb_B5-668F	ATAAAATCCCGTTCCACCCTTA	citocromo *b*(1)	*G. pyrenaicus*
Galpyr_cytb_B5-739R	TTATCTGGGTCTCCTAATAGATCTGG	citocromo *b*(1)	*G. pyrenaicus*
Galpyr_cytb_B6-832F	CCAGAATGATATTTCCTATTTGCAT	citocromo *b*(1)	*G. pyrenaicus*
Galpyr_cytb_B6-925R	CATGCTTCGTTGTTTTGAGGT	citocromo *b*(1)	*G. pyrenaicus*
Galpyr_cytb_B7-967F	ATATTCCGACCCCTAAGCCAAT	citocromo *b*(1)	*G. pyrenaicus*

Primer	Secuencia	Gen	Taxa
Galpyr_cytb_B7-1051R	AAGGACTGAGGCTACTTGTCC	citocromo *b*(1)	*G. pyrenaicus*
Galpyr_cytb_B8-1111F	TCCTAATTATTATACCACTAGCAAGCA	citocromo *b*(1)	*G. pyrenaicus*
Galpyr_cytb_B8-ThrR	GGTTTTCGTTTTTGGTTTACAA	citocromo *b*(1)	*G. pyrenaicus*
Galpyr_trnaPro	CCGCCACCAACACCCAAAGCTG	D-loop	*G. pyrenaicus*
Galpyr_Dloop_343R	GATGGTAGTCGAGAGATAACCT	D-loop	*G. pyrenaicus*
Galpyr_tRNAPro_A1_42F	CCGCCACCAACACCCAAAGCTG	D-loop(2)	*G. pyrenaicus*
Galpyr_Dloop_A1_49R	GTTTACTTCTCTTTGTTTGCGATT	D-loop(2)	*G. pyrenaicus*
Galpyr_Dloop_A2_ 116F	CCACCCCTATGTATTCGTGCAT	D-loop(2)	*G. pyrenaicus*
Galpyr_Dloop_A2_ 224R	TTGTAAGGTAAAGCATWTATTGTGTT	D-loop(2)	*G. pyrenaicus*
Galpyr_Dloop_A3_ 273F	GTACATACTCCTATCCCGTATGA	D-loop(2)	*G. pyrenaicus*
Galpyr_Dloop_B1_39F	TAAGAAYGAAAATTTTAAAGCCAAAT	D-loop(2)	*G. pyrenaicus*
Galpyr_Dloop_B1_122R	TTGGTATGCATGGGGACATGA	D-loop(2)	*G. pyrenaicus*
Galpyr_Dloop_B2_ 202F	RTAACCTAATGCATATCCAAGTGAT	D-loop(2)	*G. pyrenaicus*
Galpyr_Dloop_B2_ 278R	TATGGGTAATATGRTTTGTTTGTAT	D-loop(2)	*G. pyrenaicus*
Neomys_tRNAGlu	ATCGTTGTTATTCAACTATAAGAAC	citocromo *b*	*Neomys*
Neomys_cytb_403R	YCCYCARAATGATATTTGYCCTCA	citocromo *b*	*Neomys*
Neomys_cytb_389F	GTTATAGCCACTGCCTTTATAG	citocromo *b*	*Neomys*
Neomys_cytb_746R	TAATTGTCCGGGTCTCCGAGTA	citocromo *b*	*Neomys*
Neomys_cytb_614F	TWTTCCTYCATGAAACAGGATC	citocromo *b*	*Neomys*
Neomys_tRNAThr	TTTTGGTTTACAAGACCAGTGTAT	citocromo *b*	*Neomys*

Tabla A2. Cebadores usados para amplificar intrones nucleares en los distintos organismos de estudio de esta tesis. TD: PCR *touch-down*, en la que la temperatura de hibridación desciende 1ºC por ciclo, partiendo de 65ºC y durante 15 ciclos

Intrón	Secuencia	T (ºC)	Longitud	Taxa	Código ENSEMBL
ACOX2-3	CCTSGGCTCDGAGGAGCAGAT / GGGCTGTGHAYCACAAACTCCT	60	381	*G. pyrenaicus*	ENSG00000168306
ACPT-4	GAYTTTGACCGSACVCTGGAGAG / AGYAGYTCVYGGTATCGRGGACA	60	279	*G. pyrenaicus*	ENSG00000142513
COPS7A-4	TACAGCATYGGRCGRGACATCCA / TCACYTGCTCCTCRATGCCKGACA	62	634	*G. pyrenaicus*	ENSG00000111652
DHRS3-3	GACATCACCATCCTGGTGAACAA / AGGGTCAGGCTCTCCATGAAG	TD	259	*G. pyrenaicus*	ENSG00000162496
LANCL1-4	TTTGGARWRGARAARATTCCTCA / GRTACCAYTCRTACAWCAGTGG	TD	523	*G. pyrenaicus*	ENSG00000115365
PRPF31-3	GTCATYGTRGAYGCYAACAAC / BTTSACNGTGCGGATGTAATC	61	487	*G. pyrenaicus*	ENSG00000105618
ROGDI-7	CTGATGGAYGCYGTGATGCTGCA / CACGGTGAGGCASAGCTTGTTGA	60	291	*G. pyrenaicus*	ENSG00000067836
SMYD4-5	CAGCCTYCTGAAYCAYTCCTG / CAGBNRCAGTCAAAGAARTACTG	TD	402	*G. pyrenaicus*	ENSG00000186532
ALAD-10	AGAGTTYGCYATGYTGTGGCA / GGYGTGTAGTAGGTRATGATGA	TD	455	*Neomys*	ENSG00000148218
ASB6-2	TGYTGAAGATGGCYGAGCTG / TCCACCATGTCNGGCTGGTT	TD	319	*Neomys*	ENSG00000148331
CSF2-2	RAAACAGTARAWGTSRTCTCTG/ TNCAGACGNGTCTGCAGGCA	TD	673	*Neomys*	ENSG00000164400
CST6-1	RYTACAACATGGGCAGCAACA/ KGCMAGSGGGCARGTRGTGA	65	267	*Neomys*	ENSG00000175315
GALNT5-4	ATTYTTAGATTCTCAYGTGGAATG/ ACRTCYGGAGGAATKGTTCTC	60	727	*Neomys*	ENSG00000136542
GDAP-1	ACDCATTCYTTCASYTCBCAAAAG/ CAAWGCCTTTTCAGCAATTACCA	TD	688	*Neomys*	ENSG00000104381
HIF1AN-5	TACGAGAGGTTYCCYAATTTCCA/ CTTATACCAGAAGTTCACAGTGAT	TD	389	*Neomys*	ENSG00000166135
JMJD5-2	ACCABTGGCCVTGCATGMAGARGT/ TGATGAACTCRYTGACBGTCATGAG	TD	450	*Neomys*	ENSG00000155666
MYCBPAP-11	AAYAAYGGCACVGTGGYCATTT/ CAGCATYCRVAGAYTTRAAGAA	TD	341	*Neomys*	ENSG00000136449
MCM3-2	GGAATTTATCAGAGCAAAGTTC/ RTAGAAYTCYTCRTACTGCTTG	TD	335	*Neomys*	ENSG00000112118
PRPF31-3	GTCATYGTRGAYGCYAACAAC / BTTSACNGTGCGGATGTAATC	TD	481	*Neomys*	ENSG00000105618
SLA-2	AGGTGGCTGATGGCCTGTGCTGTG / TTCTTTCGATCAAAGGAGGTGTTGTC	TD	335	*Neomys*	ENSG00000155926
TRAIP-8	RGAGTAYGAGAAYCTDAAAGA / GCRCYCTGYAARTCCTTCTG	TD	685	*Neomys*	ENSG00000183763
SLC38A7-8	RGGCCTRGCYGSCTGCTTCATCTT / TCVGASAGYTTGGCTTGRATGAGGCA	62	883	Mamíferos	ENSG00000103042
COPS7A-4	TACAGCATYGGRCGRGACATCCA / TCACYTGCTCCTCRATGCCKGACA	66	1391	Mamíferos	ENSG00000111652
CARHSP1-1	ACYCGCCGSACSAGGACCTTCT / GTRATGAAGCCRTGGCCCTTGGA	64	709	Mamíferos	ENSG00000153048
GAD2-1	GGCTCHRGCTTYTGGTCYTTYGG / YCCGAKGCCKCCSGTGAACTTCT	62	728	Mamíferos	ENSG00000136750
JMJD5-2	GGCTCHRGCTTYTGGTCYTTYGG / YCCGAKGCCKCCSGTGAACTTCT	62	1126	Mamíferos	ENSG00000155666
OSTA-5	TGMWGGYCATGGTGGAAGGCTTTG / AGATGCCRTCRGGGAYGAGRAACA	60	544	Mamíferos	ENSG00000163959
PNPO-3	GATGGCTTCCRHTTCTWCACTAACTT / GGYTCCCARTAGAAGACMAKSGA	58	842	Mamíferos	ENSG00000108439

Tabla A3. Muestras biológicas empleadas en el estudio de *Galemys pyrenaicus*

Código de especimen	Tipo de muestra	Localidad	Cuenca	Latitud	Longitud
IBE-C216	Excrementos	Ars	Ebro	42.4	1.4
IBE-C224	Excrementos	Ars	Ebro	42.4	1.4
IBE-C320	Tejido	Tor	Ebro	42.6	1.4
IBE-C321	Tejido	Tor	Ebro	42.6	1.4
IBE-C421	Tejido	Tor	Ebro	42.6	1.4
IBE-C422	Tejido	Tor	Ebro	42.6	1.4
IBE-C428	Tejido	Tor	Ebro	42.6	1.4
IBE-C437	Tejido	Barriomartin	Duero	42.0	-2.5
IBE-C438	Tejido	San Emiliano	Duero	43.0	-6.0
IBE-C448	Tejido	Rio Salentinos-Sil	Miño	42.8	-6.4
IBE-C449	Tejido	Rio Somiedo-Somiedo	Cuencas cantábricas	43.1	-6.3
IBE-C452	Tejido	Rio Ambroz-Hervas	Tajo	40.3	-5.9
IBE-C485	Tejido	Rio Ambroz-Hervas	Tajo	40.3	-5.8
IBE-C539	Excrementos	Confluencia Deva-Cicera	Cuencas cantábricas	43.2	-4.6
IBE-C600	Excrementos	Toran	Garona	42.8	0.8
IBE-C618	Excrementos	Rio Besaya-Ventorrillo	Cuencas cantábricas	43.1	-4.1
IBE-C660	Tejido	Ouro Zona Baja-Tamega	Duero	41.5	-7.8
IBE-C683	Excrementos	Rio Besaya-Ventorrillo	Cuencas cantábricas	43.1	-4.1
IBE-C720	Excrementos	Cantijan	Cuencas cantábricas	43.1	-4.8
IBE-C723	Excrementos	Deva	Cuencas cantábricas	43.1	-4.8
IBE-C737	Excrementos	Deva	Cuencas cantábricas	43.1	-4.8
IBE-C779	Excrementos	Cabecera Douro-Tamega	Duero	41.6	-8.0
IBE-C802	Excrementos	Lousas-Zona Baja Beca-Tamega	Duero	41.6	-7.8
IBE-C831	Excrementos	Gondiaes-Beca-Tamega	Duero	41.6	-7.9
IBE-C844	Excrementos	Oura	Duero	41.7	-7.5
IBE-C861	Excrementos	Oura	Duero	41.7	-7.5
IBE-C894	Excrementos	Ouro-Tamega	Duero	41.5	-7.8
IBE-C895	Excrementos	Avelares-Louredo-Tamega	Duero	41.5	-7.8
IBE-C919	Tejido	Ouro Zona Baja	Duero	41.5	-7.8
IBE-C922	Excrementos	Louredo-Tamega	Duero	41.5	-7.9
IBE-C964	Excrementos	Toran	Garona	42.8	0.8
IBE-C975	Excrementos	Rio Puerma-Yuso	Duero	43.1	-4.9
IBE-C979	Excrementos	Rio Puerma-Yuso	Duero	43.1	-4.9
IBE-C982	Excrementos	Escrita	Ebro	42.6	1.0
IBE-C990	Excrementos	Varrados	Garona	42.8	0.8
IBE-C1002	Excrementos	Puerto de Canto	Ebro	42.4	1.2
IBE-C1053	Excrementos	Ronas-Najerilla-Anguiano	Ebro	42.2	-2.8
IBE-C1068	Excrementos	Portilla-Najerilla-Villavelayo	Ebro	42.1	-2.9
IBE-C1069	Tejido	Rio Calamantio-Mansilla de la Sierra	Ebro	42.2	-2.9
IBE-C1070	Excrementos	Arroyo Usaya y Oja-Ezcaray	Ebro	42.3	-3.0
IBE-C1072	Excrementos	Arroyo Ortigal-Oja-Ezcaray	Ebro	42.2	-3.1
IBE-C1075	Excrementos	Oja y Arroyo Altuzarra-Ezcaray	Ebro	42.2	-3.0

Código de especimen	Tipo de muestra	Localidad	Cuenca	Latitud	Longitud
IBE-C1077	Excrementos	Arroyo Usaya y Oja-Ezcaray	Ebro	42.3	-3.0
IBE-C1132	Excrementos	Rio Razoncillo-Molinos de Razon	Duero	42.0	-2.6
IBE-C1154	Excrementos	Rio del Oro-Tremado	Cuencas cantábricas	43.2	-6.8
IBE-C1174	Excrementos	Rio Valledor-Villanueva	Cuencas cantábricas	43.1	-6.8
IBE-C1177	Excrementos	Rio Valledor-Villanueva	Cuencas cantábricas	43.1	-6.8
IBE-C1611	Excrementos	Rio Ter-Pastiura	Cuencas catalanas	42.4	2.3
IBE-C1612	Excrementos	Riu Ritort-Espinavell	Cuencas catalanas	42.4	2.4
IBE-C1661	Excrementos	Rio Razon-Sotillo del Rincon	Duero	41.9	-2.7
IBE-C1671	Excrementos	Rio Razon-Sotillo del Rincon	Duero	41.9	-2.7
IBE-C1687	Excrementos	Rio Razoncillo-Molinos de Razon	Duero	42.0	-2.6
IBE-C1729	Excrementos	Entrambasaguas-Rio Hijar	Ebro	43.0	-4.3
IBE-C1843	Excrementos	Rio del Coto-Arroyo de la Brana	Cuencas cantábricas	43.1	-6.7
IBE-C1851	Excrementos	Rio Deva-Convento del Naranco	Cuencas cantábricas	43.1	-4.8
IBE-C1869	Excrementos	Rio de Pumarin-Pontenova	Cuencas cantábricas	43.3	-6.7
IBE-C1880	Excrementos	Rio Cantijan-Puente Pontesque	Cuencas cantábricas	43.1	-4.8
IBE-C2517	Excrementos	Redes-Caleao	Cuencas cantábricas	43.2	-5.4
IBE-C2596	Excrementos	Rio Valvanera-Monasterio	Ebro	42.2	-2.9
IBE-C2597	Excrementos	Rio Cidacos-Peroblasco	Ebro	42.2	-2.3
IBE-C2607	Excrementos	Redes-Pendones	Cuencas cantábricas	43.1	-5.3
IBE-C2736	Tejido	Rio Caranyo-Covelo	Miño	42.3	-8.4
IBE-C2737	Tejido	Rio Riobo-A Estrada	Cuencas gallegas	42.7	-8.4
IBE-C2738	Tejido	Rio Riobo-A Estrada	Cuencas gallegas	42.7	-8.4
IBE-C2739	Tejido	Rio Cabras-Laza	Duero	42.1	-7.4
IBE-C2740	Tejido	Rio Meladas-Carballeda	Miño	42.2	-6.8
IBE-C2741	Tejido	Rio Meladas-Carballeda	Miño	42.2	-6.8
IBE-C2742	Tejido	Rio Ourille-Celanova	Miño	42.2	-7.9
IBE-C2743	Tejido	Rio Ourille-Celanova	Miño	42.2	-7.9
IBE-C2744	Tejido	Rio Pontigon-Fonsagrada	Cuencas cantábricas	43.2	-7.0
IBE-C2745	Tejido	Rio Pontigon-Fonsagrada	Cuencas cantábricas	43.2	-7.0
IBE-C2746	Tejido	Rio Queixa-Chandrexa de Queixa	Miño	42.2	-7.4
IBE-C2747	Tejido	Rio Queixa-Chandrexa de Queixa	Miño	42.2	-7.4
IBE-C2749	Tejido	Rio Termes-As Neves	Miño	42.1	-8.4
IBE-C2750	Tejido	Rio Termes-As Neves	Miño	42.1	-8.4
IBE-C2751	Tejido	Rio Ribeira Pequena-Laza	Miño	42.1	-7.4
IBE-C2752	Tejido	Rio Ribeira Pequena-Laza	Miño	42.1	-7.4
IBE-C2753	Tejido	Rio Bullan-Navia de Suarna	Cuencas cantábricas	42.9	-7.1
IBE-C2754	Tejido	Rio Marinan-Obarco	Miño	42.4	-7.0
IBE-C2755	Tejido	Rio Leitzaran-Berastegi	Cuencas cantábricas	43.2	-2.0
IBE-C2756	Tejido	Rio Aiaiturrieta-Ataun	Cuencas cantábricas	43.0	-2.1
IBE-C2757	Tejido	Rio Amundarain-Zaldibia	Cuencas cantábricas	43.1	-2.0
IBE-C2760	Tejido	Rio Erasote-Leitza	Cuencas cantábricas	43.1	-1.9
IBE-C2763	Tejido	Rio Olazar-Eugi	Ebro	43.0	-1.5
IBE-C2765	Tejido	Rio Ezpelura-Urrotz	Cuencas cantábricas	43.1	-1.7
IBE-C2781	Tejido	Rio Ameztia-Labaien	Cuencas cantábricas	43.1	-1.7

Código de especimen	Tipo de muestra	Localidad	Cuenca	Latitud	Longitud
IBE-C2791	Tejido	Rio Urrobi-Auritz	Ebro	43.0	-1.3
IBE-C2795	Tejido	Rio Irati-Artozki	Ebro	42.9	-1.3
IBE-C2796	Tejido	Rio Sasoaran-Eugi	Ebro	43.0	-1.6
IBE-C2798	Tejido	Rio Arrazola-Orbaitzeta	Ebro	43.0	-1.2
IBE-C2799	Tejido	Rio Elama-Artikutza	Cuencas cantábricas	43.2	-1.8
IBE-C2856	Tejido	Rio Lamedo	Cuencas cantábricas	43.1	-4.6
IBE-C2892	Tejido	Rio Cares-Cain	Cuencas cantábricas	43.2	-4.9
IBE-C2893	Tejido	Rio Cares-Corona	Cuencas cantábricas	43.2	-4.9
IBE-C2894	Tejido	Rio Cares-Casiellos	Cuencas cantábricas	43.2	-4.9
IBE-C2950	Excrementos	Redes-Caleao	Cuencas cantábricas	43.2	-5.4
IBE-C2975	Excrementos	Ribeira do Castelo	Duero	41.6	-7.7
IBE-C2990	Excrementos	Lousas-Cabecera Beca-Tamega	Duero	41.6	-7.8
IBE-C3159	Uña (museo)	Rio Lillas-Cantalojas	Tajo	41.2	-3.2
IBE-C3161	Costilla (museo)	Navacepeda de Tormes	Duero	40.4	-5.2
IBE-C3301	Tejido	Paiva-Fraguas	Duero	40.8	-7.8
IBE-C3303	Tejido	Sabor-Macas-Quintanilla	Duero	41.8	-6.6
IBE-C3305	Tejido	Paiva-Fraguas	Duero	40.8	-7.8
IBE-C3338	Excrementos	Rio Bullon	Cuencas cantábricas	43.1	-4.5
IBE-C3342	Excrementos	Rio Bullon-Valdeprado	Cuencas cantábricas	43.1	-4.5
IBE-C3455	Excrementos	Pedroso-Najerilla-Pedroso	Ebro	42.3	-2.7
IBE-C3473	Excrementos	Arroyo de Puente la Ra-Iregua	Ebro	42.1	-2.7
IBE-C3474	Excrementos	Tobia-Najerilla	Ebro	42.2	-2.9
IBE-C3475	Excrementos	Tobia-Najerilla	Ebro	42.3	-2.9
IBE-C3481	Excrementos	Tobia-Najerilla	Ebro	42.3	-2.9
IBE-C3482	Excrementos	Lumbreras-Iregua-Lumbreras	Ebro	42.1	-2.6
IBE-C3486	Excrementos	Arroyo la Vieja-Iregua-Lumbreras	Ebro	42.0	-2.6
IBE-C3489	Excrementos	Arroyo la Vieja-Iregua-Lumbreras	Ebro	42.0	-2.6
IBE-C3522	Tejido	Sabor-Franca-Soutelo	Duero	41.9	-6.8
IBE-C3541	Tejido	Sabor-Abaixo de Gimonde	Duero	41.8	-6.7
IBE-C3545	Tejido	Sabor-Franca-Soutelo	Duero	41.9	-6.8
IBE-C3602	Tejido	Corgo-Cabril-Arnal	Duero	41.3	-7.8
IBE-C3620	Excrementos	Curueno	Duero	42.9	-5.4
IBE-C3623	Excrementos	Redes-Caleao	Cuencas cantábricas	43.2	-5.4
IBE-C3637	Tejido	Sabor-Franca-Soutelo	Duero	41.9	-6.8
IBE-C3638	Tejido	Sabor-Franca-Soutelo	Duero	41.9	-6.8
IBE-C3641	Excrementos	Redes-Monasterio	Cuencas cantábricas	43.2	-5.3
IBE-C3647	Tejido	Sabor-Franca-Soutelo	Duero	41.9	-6.8
IBE-C3756	Tejido	La Soledad	Duero	42.2	-3.1
IBE-C3759	Tejido	Urumea	Cuencas cantábricas	43.1	-1.8
IBE-C3762	Tejido	Luzaide	Adour	43.0	-1.3
IBE-C3763	Tejido	Aritzakun	Adour	43.3	-1.4
IBE-S1925	Tejido	Riu Ritort-Camprodon	Cuencas catalanas	42.3	2.4
IBE-S2012	Excrementos	Riu del Canto-Soriguera	Ebro	42.4	1.2
IBE-S2026	Excrementos	Riu del Canto-Soriguera	Ebro	42.4	1.2

Código de especimen	Tipo de muestra	Localidad	Cuenca	Latitud	Longitud
IBE-S2762	Excrementos	Hecho-Aragon-Subordan	Ebro	42.8	-0.6
IBE-S2764	Excrementos	Hecho-Aragon-Subordan	Ebro	42.9	-0.7
IBE-S2768	Excrementos	Hecho-Aragon-Subordan	Ebro	42.9	-0.7
IBE-S2773	Excrementos	Hecho-Aragon-Subordan	Ebro	42.8	-0.7

Tabla A4. Muestras biológicas empleadas en el estudio del género *Neomys*

Código de especimen	Taxón	Tipo de muestra	Localidad
IBE-C3786	*Neomys anomalus milleri*	Tejido	Barcelona-España
IBE-S1926	*Neomys anomalus milleri*	Excrementos	Girona-España
IBE-C1808	*Neomys anomalus milleri*	Tejido	Lleida-España
IBE-S2013	*Neomys anomalus milleri*	Excrementos	Lleida-España
IBE-C2551	*Neomys anomalus milleri*	Excrementos	Lleida-España
IBE-C4116	*Neomys anomalus milleri*	Tejido	Belgorod-Rusia
IBE-C1435	*Neomys anomalus anomalus*	Craneo (egagrópila)	Valladolid-España
IBE-C2664	*Neomys anomalus anomalus*	Excrementos	Castellón-España
IBE-C1144	*Neomys anomalus anomalus*	Excrementos	Asturias-España
IBE-C2895	*Neomys anomalus anomalus*	Tejido	Leon-España
IBE-C785	*Neomys anomalus anomalus*	Excrementos	Pontevedra-España
IBE-C3244	*Neomys anomalus anomalus*	Excrementos	Vila Real-Portugal
IBE-C1683	*Neomys anomalus anomalus*	Excrementos	Zamora-España
IBE-C1662	*Neomys anomalus anomalus*	Excrementos	Soria-España
IBE-C1529	*Neomys anomalus anomalus*	Tejido	Asturias-España
IBE-C1789	*Neomys anomalus anomalus*	Tejido	Avila-España
IBE-C574	*Neomys fodiens*	Excrementos	Asturias-España
IBE-C1914	*Neomys fodiens*	Tejido	Lleida-España
IBE-C101	*Neomys fodiens*	Tejido	Barcelona-España
IBE-S2688	*Neomys fodiens*	Excrementos	Huesca-España
IBE-S1915	*Neomys fodiens*	Excrementos	Alava-España

VII. BIBLIOGRAFÍA

Adkins RM, Gelke EL, Rowe D, Honeycutt RL (2001) Molecular phylogeny and divergence time estimates for major rodent groups: evidence from multiple genes. *Molecular Biology and Evolution* **18**, 777-791.

Aitken N, Smith S, Schwarz C, Morin PA (2004) Single nucleotide polymorphism (SNP) discovery in mammals: a targeted-gene approach. *Mol Ecol* **13**, 1423-1431.

Alba MM, Castresana J (2007) On homology searches by protein Blast and the characterization of the age of genes. *BMC Evol Biol* **7**, 53.

Alexandrino J, Froufe E, Arntzen JW, Ferrand N (2000) Genetic subdivision, glacial refugia and postglacial recolonization in the golden-striped salamander, *Chioglossa lusitanica* (Amphibia: urodela). *Molecular Ecology* **9**, 771-781.

Allentoft M, Schuster SC, Holdaway R*, et al.* (2009) Identification of microsatellites from an extinct moa species using high-throughput (454) sequence data. *Biotechniques* **46**, 195-200.

Altschul SF, Madden TL, Schaffer AA*, et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389-3402.

Alvarez-Lao DJ, Garcia N (2010) Chronological distribution of Pleistocene cold-adapted large mammal faunas in the Iberian Peninsula. *Quaternary International* **212**, 120-128.

Avise JC, Robinson TJ (2008) Hemiplasy: a new term in the lexicon of phylogenetics. *Syst Biol* **57**, 503-507.

Avise JC (2000) *Phylogeography : the history and formation of species* Harvard University Press, Cambridge, Mass.

Avise JC, Arnold J, Ball RM*, et al.* (1987) Intraspecific Phylogeography: The Mitochondrial DNA Bridge Between Population Genetics and Systematics. *Annual Review of Ecology and Systematics* **18**, 489-522.

Aymerich P, Gosálbez J (2002) Factors de distribució de *Galemys pyrenaicus* (Insectivora, Talpidae) a Catalunya. *Orsis* **17**, 21-35.

Backstroem N, Fagerberg S, Ellegren H (2008) Genomics of natural bird populations: a gene-based set of reference markers evenly spread across the avian genome. *Mol. Ecol.* **17**, 964-980.

Baird NA, Etter PD, Atwood TS*, et al.* (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS One* **3**, e3376.

Ballard JW, Whitlock MC (2004) The incomplete natural history of mitochondria. *Mol Ecol* **13**, 729-744.

Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **27**, 573-580.

Benton MJ, Donoghue PCJ, Asher RJ (2009) Calibrating and constraining molecular clocks. In: *The Timetree of Life* (eds. Hedges SB, Kumar S), pp. 35-86. Oxford University Press, Oxford, New York.

Bininda-Emonds OR (2007) Fast genes and slow clades: comparative rates of molecular evolution in mammals. *Evol Bioinform Online* **3**, 59-85.

Binladen J, Gilbert MT, Bollback JP*, et al.* (2007) The use of coded PCR primers enables high-throughput sequencing of multiple homolog amplification products by 454 parallel sequencing. *PLoS One* **2**, e197.

Bonin A, Ehrich D, Manel S (2007) Statistical analysis of amplified fragment length polymorphism data: a toolbox for molecular ecologists and evolutionists. *Mol Ecol* **16**, 3737-3758.

Brinkmann B, Klintschar M, Neuhuber F, Huhne J, Rolf B (1998) Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat. *Am J Hum Genet* **62**, 1408-1415.

Brito PH, Edwards SV (2009) Multilocus phylogeography and phylogenetics using sequence-based markers. *Genetica* **135**, 439-455.

Bromham L (2009) Why do species vary in their rate of molecular evolution? *Biol Lett* **5**, 401-404.

Bromham L, Rambaut A, Harvey PH (1996) Determinants of rate variation in mammalian DNA sequence evolution. *J Mol Evol* **43**, 610-621.

Brower AVZ, DeSalle R, Vogler A (1996) Gene Trees, Species Trees, and Systematics: A Cladistic Perspective. *Annu Rev Ecol Syst* **27**, 423-450.

Brown WM, George M, Jr., Wilson AC (1979) Rapid evolution of animal mitochondrial DNA. *Proc Natl Acad Sci U S A* **76**, 1967-1971.

Caballero A, Quesada H (2010) Homoplasy and distribution of AFLP fragments: an analysis in silico of the genome of different species. *Molecular Biology and Evolution* **27**, 1139-1151.

Cabria MT, Rubines J, Gómez-Moliner B, Zardoya R (2006) On the phylogenetic position of a rare Iberian endemic mammal, the Pyrenean desman (*Galemys pyrenaicus*). *Gene* **375**, 1-13.

Camargo A, Avila LJ, Morando M, Sites JW, Jr. (2012) Accuracy and precision of species trees: effects of locus, individual, and base pair sampling on inference of species trees in lizards of the Liolaemus darwinii group (Squamata, Liolaemidae). *Syst Biol* **61**, 272-288.

Castién E, Gosálbez, J. (1994) Diet of *Galemys pyrenaicus* (Geoffroy, 1811) in the North of the Iberian Peninsula. *Netherlands Journal of Zoology* **45**, 422-430.

Castiglia RA, F.; Aloise, G.; Amori, G. (2007) Mitochondrial DNA reveals different phylogeographic structures in the water shrews *Neomys anomalus* and *N. fodiens* (Insectivora: Soricidae) in Europe. *journal of Zoological Systematics and Evolutionary Research* **45**, 255-262.

Castresana (2002) Estimation of genetic distances from human and mouse introns. *Genome Biol.* **3**, RESEARCH0028.

Castresana J (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.* **17**, 540-552.

Castresana J (2001) Cytochrome *b* phylogeny and the taxonomy of great apes and mammals. *Mol. Biol. Evol.* **18**, 465-471.

Darwin C (1859) *On the origin of species by means of natural selection* J. Murray, London,.

Deffontaine V, Ledevin R, Fontaine MC*, et al.* (2009) A relict bank vole lineage highlights the biogeographic history of the Pyrenean region in Europe. *Molecular Ecology* **18**, 2489-2502.

Degnan JH, Rosenberg NA (2009) Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends Ecol Evol* **24**, 332-340.

Delsuc F, Scally M, Madsen O*, et al.* (2002) Molecular phylogeny of living xenarthrans and the impact of character and taxon sampling on the placental tree rooting. *Molecular Biology and Evolution* **19**, 1656-1671.

Drummond AJ AB, Cheung M, Heled J, Kearse M, Moir R, Stones- Havas S, Thierer T, Wilson A (2009) Geneious v4. http://geneious.com.

Drummond AJ, Ho SY, Phillips MJ, Rambaut A (2006) Relaxed phylogenetics and dating with confidence. *PLoS Biol* **4**, e88.

Drummond AJ, Rambaut A (2007) BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology* **7**, 214.

Dubey S, Salamin N, Ohdachi SD, Barriere P, Vogel P (2007) Molecular phylogenetics of shrews (Mammalia: Soricidae) reveal timing of transcontinental colonizations. *Mol Phylogenet Evol* **44**, 126-137.

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

Edwards SV, Beerli P (2000) Perspective: gene divergence, population divergence, and the variance in coalescence time in phylogeographic studies. *Evolution* **54**, 1839-1854.

Ekblom R, Galindo J (2011) Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity (Edinb)* **107**, 1-15.

Elith J, Phillips SJ, Hastie T*, et al.* (2011) A statistical explanation of MaxEnt for ecologists. *Diversity and Distributions* **17**, 43–57.

Ellegren H (2004) Microsatellites: simple sequences with complex evolution. *Nat Rev Genet* **5**, 435-445.

Elsik CG, Tellam RL, Worley KC*, et al.* (2009) The genome sequence of taurine cattle: a window to ruminant biology and evolution. *Science* **324**, 522-528.

Excoffier L, Lischer HE (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources* **10**, 564-567.

Fernandes M, Herrero J, Aulagnier S, Amori G (2011) *Galemys pyrenaicus*. In: *IUCN Red List of Threatened Species. Version 2011.2*, pp. 1-3.

Fitch WM, Margoliash E (1967) Construction of phylogenetic trees. *Science* **155**, 279-284.

Fitch WM, Margoliash E (1968) The construction of phylogenetic trees. II. How well do they reflect past history? *Brookhaven Symp Biol* **21**, 217-242.

Flicek P, Amode MR, Barrell D*, et al.* (2011) Ensembl 2011. *Nucleic Acids Res* **39**, D800-D806.

Fortelius M (2012) *New and Old Worlds Database of Fossil Mammals (NOW)* University of Helsinki.

Frankham R, Ballou JD, Eldridge MDB*, et al.* (2011) Predicting the probability of outbreeding depression. *Conservation Biology* **25**, 465-475.

Fumagalli L, Taberlet P, Stewart DT*, et al.* (1999) Molecular phylogeny and evolution of Sorex shrews (Soricidae: insectivora) inferred from mitochondrial DNA sequence data. *Mol Phylogenet Evol* **11**, 222-235.

Garcia-Pereira MJ, Caballero A, Quesada H (2010) Evaluating the relationship between evolutionary divergence and phylogenetic accuracy in AFLP data sets. *Molecular Biology and Evolution* **27**, 988-1000.

Gazave E, Marques-Bonet T, Fernando O, Charlesworth B, Navarro A (2007) Patterns and rates of intron divergence between humans and chimpanzees. *Genome Biology* **8**, -.

Gibbs RA, Rogers J, Katze MG*, et al.* (2007) Evolutionary and biomedical insights from the rhesus macaque genome. *Science* **316**, 222-234.

Gibbs RA, Weinstock GM, Metzker ML*, et al.* (2004) Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428**, 493-521.

Gilbert MT, Bandelt HJ, Hofreiter M, Barnes I (2005) Assessing ancient DNA studies. *Trends Ecol Evol* **20**, 541-544.

Godinho R, Crespo EG, Ferrand N (2008) The limits of mtDNA phylogeography: complex patterns of population history in a highly structured Iberian lizard are only revealed by the use of nuclear markers. *Molecular Ecology* **17**, 4670-4683.

Gómez A, Lunt DH (2007) Refugia within refugia: patterns of phylogeographic concordance in the Iberian Peninsula. In: *Phylogeography of Southern European Refugia* (eds. Weiss S, Ferrand N), pp. 155-188. Springer, Amsterdam.

González-Esteban J, Castién E, Gosálbez J (1999) Morphological and colour variation in the Pyrenean desman *Galemys pyrenaicus* (Geoffroy, 1811). *Zeitschrift für Säugetierkunde* **64**, 1-11.

Green P (2007) 2x genomes--does depth matter? *Genome Res.* **17**, 1547-1549.

Guindon S, Dufayard JF, Lefort V*, et al.* (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* **59**, 307-321.

Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.* **52**, 696-704.

Harris H (1966) Enzyme polymorphisms in man. *Proc R Soc Lond B Biol Sci* **164**, 298-310.

He K, Li YJ, Brandley MC*, et al.* (2010) A multi-locus phylogeny of Nectogalini shrews and influences of the paleoclimate on speciation and evolution. *Mol Phylogenet Evol* **56**, 734-746.

Hein J, Schierup MH, Wiuf C (2005) *Gene Genealogies, Variation and Evolution: A Primer in Coalescent Theory*. Oxford University Press, USA.

Heled J, Drummond AJ (2010) Bayesian inference of species trees from multilocus data. *Molecular Biology and Evolution* **27**, 570-580.

Hennig W (1966) *Phylogenetic systematics* University of Illinois Press, Urbana,.

Hernandez-Roldan JL, Murria C, Romo H*, et al.* (2011) Tracing the origin of disjunct distributions: a case of biogeographical convergence in *Pyrgus* butterflies. *Journal of Biogeography*, 1-15.

Hershey AD, Chase M (1952) Independent functions of viral protein and nucleic acid in growth of bacteriophage. *J Gen Physiol* **36**, 39-56.

Hewitt G (2000) The genetic legacy of the Quaternary ice ages. *Nature* **405**, 907-913.

Hey J (2010) Isolation with migration models for more than two populations. *Molecular Biology and Evolution* **27**, 905-920.

Hey J, Pinho C (2012) Population genetics and objectivity in species diagnosis. *Evolution* **66**, 1413-1429.

Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A (2005) Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology* **25**, 1965–1978.

Hofreiter M, Serre D, Rohland N*, et al.* (2004) Lack of phylogeography in European mammals before the last glaciation. *Proceedings of the National Academy of Sciences, USA* **101**, 12963-12968.

Hubbard TJ, Aken BL, Beal K*, et al.* (2007) Ensembl 2007. *Nucleic Acids Res.* **35**, D610-617.

Hubby JL, Lewontin RC (1966) A molecular approach to the study of genic heterozygosity in natural populations. I. The number of alleles at different loci in Drosophila pseudoobscura. *Genetics* **54**, 577-594.

Igea J, Juste J, Castresana J (2010) Novel intron markers to study the phylogeny of closely related mammalian species. *BMC Evolutionary Biology* **10**, 369.

Johnson WE, Eizirik E, Pecon-Slattery J*, et al.* (2006) The late Miocene radiation of modern Felidae: a genetic assessment. *Science* **311**, 73-77.

Juckwer EA (1990) *Galemys pyrenaicus* (Geoffroy, 1811) – Pyrenäen-Desman. In: *Handbuch der Säugetiere Europas: Insektenfresser, Herrentiere* (eds. Niethammer J, Krapp F), pp. 79-92.

Katoh K, Kuma K, Toh H, Miyata T (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.* **33**, 511-518.

Katoh K, Toh H (2008) Recent developments in the MAFFT multiple sequence alignment program. *Briefings in Bioinformatics* **9**, 286-298.

Kimura M (1968) Evolutionary rate at the molecular level. *Nature* **217**, 624-626.

Kingman JFC (1982) The Coalescent. *Stochastic Processes and their Applications* **13**, 235-248.

Kloch A, Babik W, Bajer A, Sinski E, Radwan J (2010) Effects of an MHC-DRB genotype and allele number on the load of gut parasites in the bank vole Myodes glareolus. *Mol Ecol* **19 Suppl 1**, 255-265.

Knowles LL, Kubatko LS (2010) *Estimating species trees : practical and theoretical aspects* Wiley-Blackwell, Hoboken, N.J.

Kocher TD, Thomas WK, Meyer A*, et al.* (1989) Dynamics of mitochondrial DNA evolution in animals: amplification and sequencing with conserved primers. *Proc Natl Acad Sci U S A* **86**, 6196-6200.

Koepfli KP, Deere KA, Slater GJ*, et al.* (2008) Multigene phylogeny of the Mustelidae: resolving relationships, tempo and biogeographic history of a mammalian adaptive radiation. *BMC Biol.* **6**, 10.

Kreitman M (1983) Nucleotide polymorphism at the alcohol dehydrogenase locus of Drosophila melanogaster. *Nature* **304**, 412-417.

Kryštufek B, A. D, Griffiths HI (2000) Evolutionary biogeography of water shrews (*Neomys* spp.) in the western Palaearctic Region. *Canadian Journal of Zoology* **78**, 1616-1625.

Kryštufek B, Quadracci A (2008) Effects of latitude and allopatry on body size variation in European water shrews. *Acta Theriologica* **53**, 39-46.

Kubatko LS, Degnan JH (2007) Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Syst Biol* **56**, 17-24.

Kumar S, Filipski A, Swarna V, Walker A, Hedges SB (2005) Placing confidence limits on the molecular age of the human-chimpanzee divergence. *Proc. Natl. Acad. Sci. U S A* **102**, 18842-18847.

Lamarck JBPAdMd (1830) *Philosophie zoologique*, Nouv. éd. edn. G. Baillière; etc., Paris,.

Lander ES, Linton LM, Birren B, *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921.

Lessa EP (1992) Rapid surveying of DNA sequence variation in natural populations. *Molecular Biology and Evolution* **9**, 323-330.

Li C, Orti G, Zhang G, Lu G (2007) A practical approach to phylogenomics: the phylogeny of ray-finned fish (Actinopterygii) as a case study. *BMC Evol. Biol.* **7**, 44.

Librado P, Rozas J (2009) DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* **25**, 1451-1452.

Lindblad-Toh K, Wade CM, Mikkelsen TS, *et al.* (2005) Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* **438**, 803-819.

Linn S, Arber W (1968) Host specificity of DNA produced by Escherichia coli, X. In vitro restriction of phage fd replicative form. *Proc Natl Acad Sci U S A* **59**, 1300-1306.

Linné Cv (1758) *Caroli Linnæi ... Systema naturæ per regna tria naturæ, secundum classes, ordines, genera, species, cum characteribus, differentiis, synonymis, locis*, Ed. 10.,reformata. edn. impensis L. Salvii, Holmiæ,.

Liu L, Yu L, Pearl DK, Edwards SV (2009) Estimating species phylogenies using coalescence times among sequences. *Syst Biol* **58**, 468-477.

López-Fuster MJ, García-Perea R, Fernández-Salvador R, Gisbert J, Ventura J (2006) Craniometric variability of the Iberian desman, *Galemys pyrenaicus* (Mammalia: Erinaceomorpha: Talpidae). *Folia Zoologica* **55**, 29-42.

Lyons LA, Laughlin TF, Copeland NG*, et al.* (1997) Comparative anchor tagged sequences (CATS) for integrative mapping of mammalian genomes. *Nat Genet* **15**, 47-56.

MacDonald D, Barret P (2008) *Guía de Campo de los mamíferos de España y Europa*. Ed. Omega.

Maddison WP (1997) Gene Trees in Species Trees. *Syst Biol* **46**, 523-536.

Magri D (2008) Patterns of post-glacial spread and the extent of glacial refugia of European beech (*Fagus sylvatica*). *Journal of Biogeography* **35**, 450-463.

Martin AP, Palumbi SR (1993) Body size, metabolic rate, generation time, and the molecular clock. *Proc Natl Acad Sci U S A* **90**, 4087-4091.

Martínez-Solano I, Teixeira J, Buckley D, García-París M (2006) Mitochondrial DNA phylogeography of *Lissotriton boscai* (Caudata, Salamandridae): evidence for old, multiple refugia in an Iberian endemic. *Molecular Ecology* **15**, 3375-3388.

McCormack JE, Hird SM, Zellmer AJ, Carstens BC, Brumfield RT (2011) Applications of next-generation sequencing to phylogeography and phylogenetics. *Mol Phylogenet Evol*.

McKenna MC, Bell SK, Simpson GG (1997) *Classification of mammals above the species level* Columbia University Press, New York.

Meirmans PG (2012) The trouble with isolation by distance. *Mol Ecol* **21**, 2839-2846.

Melero Y, Aymerich P, Luque-Larena JJ, Gosálbez J (2012) New insights into social and space use behaviour of the endangered Pyrenean desman (*Galemys pyrenaicus*). *European Journal of Wildlife Research* **58**, 185-193.

Melo-Ferreira J, Boursot P, Randi E*, et al.* (2007) The rise and fall of the mountain hare (*Lepus timidus*) during Pleistocene glaciations: expansion and retreat with hybridization in the Iberian Peninsula. *Mol Ecol* **16**, 605-618.

Mendes-Soares H, Rychlik L (2009) Differences in swimming and diving abilities between two sympatric species of water shrews: *Neomys anomalus* and *Neomys fodiens* (Soricidae). *Journal of Ethology* **27**, 317-325.

Meredith RW, Janecka JE, Gatesy J*, et al.* (2011) Impacts of the Cretaceous Terrestrial Revolution and KPg extinction on mammal diversification. *Science* **334**, 521-524.

Meredith RW, Westerman M, Springer MS (2008) A timescale and phylogeny for "bandicoots" (Peramelemorphia: Marsupialia) based on sequences for five nuclear genes. *Mol Phylogenet Evol* **47**, 1-20.

Meudt HM, Clarke AC (2007) Almost forgotten or latest practice? AFLP applications, analyses and advances. *Trends Plant Sci* **12**, 106-117.

Mikkelsen TS, Hillier LW, Eichler EE*, et al.* (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**, 69-87.

Mikkelsen TS, Wakefield MJ, Aken B*, et al.* (2007) Genome of the marsupial Monodelphis domestica reveals innovation in non-coding sequences. *Nature* **447**, 167-177.

Miller MP (2005) Alleles in space (AIS): computer software for the joint analysis of interindividual spatial and genetic information. *Journal of Heredity* **96**, 722-724.

Monmonier MS (1973) Maximum-difference barriers: an alternative numerical regionalization method. *Geographical Analysis* **5**, 245-261.

Morin PA, Chambers KE, Boesch C, Vigilant L (2001) Quantitative polymerase chain reaction analysis of DNA from noninvasive samples for accurate microsatellite genotyping of wild chimpanzees (Pan troglodytes verus). *Mol Ecol* **10**, 1835-1844.

Morin PA, Luikart G, Wayne RK, group tSw (2004) SNPs in ecology, evolution and conservation. *Trends in Ecology & Evolution* **19**, 208-216.

Mucci N, Arrendal J, Ansorge H*, et al.* (2010) Genetic diversity and landscape genetic structure of otter (*Lutra lutra*) populations in Europe. *Conservation Genetics* **11**, 583-599.

Murphy WJ, Eizirik E, O'Brien SJ*, et al.* (2001) Resolution of the early placental mammal radiation using Bayesian phylogenetics. *Science* **294**, 2348-2351.

Murphy WJ, Pevzner PA, O'Brien SJ (2004) Mammalian phylogenomics comes of age. *Trends Genet* **20**, 631-639.

Nabholz B, Glemin S, Galtier N (2008a) Strong variations of mitochondrial mutation rate across mammals--the longevity hypothesis. *Molecular Biology and Evolution* **25**, 120-130.

Nabholz B, Mauffrey JF, Bazin E, Galtier N, Glemin S (2008b) Determination of mitochondrial genetic diversity in mammals. *Genetics* **178**, 351-361.

Nater A, Nietlisbach P, Arora N*, et al.* (2011) Sex-Biased Dispersal and Volcanic Activities Shaped Phylogeographic Patterns of Extant Orangutans (genus: *Pongo*). *Molecular Biology and Evolution* **28**, 2275-2288.

Nichols R (2001) Gene trees and species trees are not the same. *Trends Ecol Evol* **16**, 358-364.

Nielsen R, Wakeley J (2001) Distinguishing migration from isolation: a Markov chain Monte Carlo approach. *Genetics* **158**, 885-896.

Nores C, Queiroz AI, Gisbert J (2007) *Galemys pyrenaicus*. In: *Atlas y libro rojo de los mamíferos terrestres de España*, pp. 92-98.

Novembre J, Ramachandran S (2011) Perspectives on human population structure at the cusp of the sequencing era. *Annual Review of Genomics and Human Genetics* **12**, 245-274.

Nsubuga AM, Robbins MM, Roeder AD*, et al.* (2004) Factors affecting the amount of genomic DNA extracted from ape faeces and the identification of an improved sample storage method. *Mol Ecol* **13**, 2089-2094.

Pääbo S, Poinar H, Serre D*, et al.* (2004) Genetic analyses from ancient DNA. *Annual Review of Genetics* **38**, 645-679.

Palmeirim JM, Hoffmann RS (1983) *Galemys pyrenaicus*. *Mammalian Species* **207**, 1-5.

Palumbi SR, Baker CS (1994) Contrasting population structure from nuclear intron sequences and mtDNA of humpback whales. *Molecular Biology and Evolution* **11**, 426-435.

Palumbi SR, Cipriano F, Hare MP (2001) Predicting nuclear gene coalescence from mitochondrial data: the three-times rule. *Evolution* **55**, 859-868.

Pamilo P, Nei M (1988) Relationships between gene trees and species trees. *Molecular Biology and Evolution* **5**, 568-583.

Pelletier A, Obbard ME, White BN, Doyle C, Kyle CJ (2011) Small-scale genetic structure of American black bears illustrates potential postglacial recolonization routes. *Journal of Mammalogy* **92**, 629-644.

Peng ZG, Elango N, Wildman DE, Yi SV (2009) Primate phylogenomics: developing numerous nuclear non-coding, non-repetitive markers for ecological and phylogenetic applications and analysis of evolutionary rate variation. *BMC Genomics* **10**, -.

Phillips SJ, Anderson RP, Schapire RE (2006) Maximum entropy modeling of species geographic distributions. *Ecological Modelling* **190**, 231-259.

Phillips SJ, Dudik M (2008) Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. *Ecography* **31**, 161-175.

Plass M, Eyras E (2006) Differentiated evolutionary rates in alternative exons and the implications for splicing regulation. *BMC Evol Biol* **6**, 50.

Porter CA, Goodman M, Stanhope MJ (1996) Evidence on mammalian phylogeny from sequences of exon 28 of the von Willebrand factor gene. *Mol Phylogenet Evol* **5**, 89-101.

Posada D (2008) jModelTest: phylogenetic model averaging. *Molecular Biology and Evolution* **25**, 1253-1256.

Primmer CR, Ellegren H (1998) Patterns of molecular evolution in avian microsatellites. *Molecular Biology and Evolution* **15**, 997-1008.

Pucek Z (1964) The structure of the glans penis in *Neomys* Kaupp, 1929 as a taxonomic character. *Acta Theriologica* **9**, 374-377.

Queiroz AI, Quaresma CM, Santos CP, Barbosa A, Carvalho H (1996) Desman distribution in Portugal. Current knowledge. In: *Council of Europe Environmental Encounters Series, Nº 25*, pp. 19-27.

Raaum RL, Sterner KN, Noviello CM, Stewart CB, Disotell TR (2005) Catarrhine primate divergence dates estimated from complete mitochondrial genomes: concordance with fossil and nuclear DNA evidence. *J. Hum. Evol.* **48**, 237-257.

Ramos-Onsins SE, Rozas J (2002) Statistical properties of new neutrality tests against population growth. *Molecular Biology and Evolution* **19**, 2092-2100.

Ranwez V, Delsuc F, Ranwez S*, et al.* (2007) OrthoMaM: a database of orthologous genomic markers for placental mammal phylogenetics. *BMC Evol. Biol.* **7**, 241.

Recuero E, García-París M (2011) Evolutionary history of *Lissotriton helveticus*: multilocus assessment of ancestral vs. recent colonization of the Iberian Peninsula. *Molecular Phylogenetics and Evolution* **60**, 170-182.

Reumer JWF (1989) Speciation and evolution in the Soricidae (Mammalia: Insectivora) in relation with the paleoclimate. *Revue suisse de zoologie* **96**, 81-90.

Reumer JWF (1994) Phylogeny and distribution of the Crocidosoricinae (Mammalia: Soricidae). In: *Advances in the Biology of Shrews* (ed. Merritt JFK, G.L., Jr.; Rose, R.K.), pp. 345-356. Special Publication of Carnegie Museum of Natural History.

Richard B (1985) *Le desman des Pyrénées, un mammifère inconnu à découvrir* Le Rocher, Monaco.

Rodova M, Islam MR, Peterson KR, Calvet JP (2003) Remarkable sequence conservation of the last intron in the PKD1 gene. *Molecular Biology and Evolution* **20**, 1669-1674.

Römpler H, Dear PH, Krause J*, et al.* (2006) Multiplex amplification of ancient DNA. *Nature Protocols* **1**, 720-728.

S.F.E.P.M. (1984) *Atlas des mammifères sauvages de France*, Paris.

Saiki RK, Gelfand DH, Stoffel S*, et al.* (1988) Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science* **239**, 487-491.

Saiki RK, Scharf S, Faloona F*, et al.* (1985) Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science* **230**, 1350-1354.

Salzburger W, Ewing GB, Von Haeseler A (2011) The performance of phylogenetic algorithms in estimating haplotype genealogies with migration. *Molecular Ecology* **20**, 1952-1963.

Sanchez-Gracia A, Castresana J (2012) Impact of deep coalescence on the reliability of species tree inference from different types of DNA markers in mammals. *PLoS One* **7**, e30239.

Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* **74**, 5463-5467.

Schlotterer C (2004) The evolution of molecular markers--just a matter of fashion? *Nat Rev Genet* **5**, 63-69.

Selkoe KA, Toonen RJ (2006) Microsatellites for ecologists: a practical guide to using and evaluating microsatellite markers. *Ecol Lett* **9**, 615-629.

Shafer ABA, Côté SD, Coltman DW (2011) Hot spots of genetic diversity descended from multiple Pleistocene refugia in an alpine ungulate. *Evolution* **65**, 125-138.

Shimodaira H (2002) An approximately unbiased test of phylogenetic tree selection. *Syst Biol* **51**, 492-508.

Shimodaira H, Hasegawa M (2001) CONSEL: for assessing the confidence of phylogenetic tree selection. *Bioinformatics* **17**, 1246-1247.

Slade RW, Moritz C, Heideman A, Hale PT (1993) Rapid assessment of single-copy nuclear DNA variation in diverse species. *Mol Ecol* **2**, 359-373.

Smedley D, Haider S, Ballester B*, et al.* (2009) BioMart--biological queries made easy. *BMC Genomics* **10**, 22.

Smit AFA. HR, Green P. (2004) RepeatMaster Open-3.0. 1996-2004. http://repeatmasker.org.

Song S, Liu L, Edwards SV, Wu S (2012) Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. *Proc Natl Acad Sci U S A* **109**, 14942-14947.

Soria-Carrasco V, Talavera G, Igea J, Castresana J (2007) The K tree score: quantification of differences in the relative branch length and topology of phylogenetic trees. *Bioinformatics* **23**, 2954-2956.

Springer MS, Stanhope MJ, Madsen O, de Jong WW (2004) Molecules consolidate the placental mammal tree. *Trends Ecol Evol* **19**, 430-438.

Stadelmann B, Lin LK, Kunz TH, Ruedi M (2007) Molecular phylogeny of New World Myotis (Chiroptera, Vespertilionidae) inferred from mitochondrial and nuclear DNA genes. *Mol. Phylogenet. Evol.* **43**, 32-48.

Stamatakis A (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**, 2688-2690.

Stanhope MJ, Czelusniak J, Si JS, Nickerson J, Goodman M (1992) A molecular perspective on mammalian evolution from the gene encoding

interphotoreceptor retinoid binding protein, with convincing evidence for bat monophyly. *Mol Phylogenet Evol* **1**, 148-160.

Stephens M, Donnelly P (2003) A comparison of bayesian methods for haplotype reconstruction from population genotype data. *Am J Hum Genet* **73**, 1162-1169.

Steppan S, Adkins R, Anderson J (2004a) Phylogeny and divergence-date estimates of rapid radiations in muroid rodents based on multiple nuclear genes. *Syst Biol* **53**, 533-553.

Steppan SJ, Storz BL, Hoffmann RS (2004b) Nuclear DNA phylogeny of the squirrels (Mammalia: Rodentia) and the evolution of arboreality from c-myc and RAG1. *Mol Phylogenet Evol* **30**, 703-719.

Stone RD (1987) The social ecology of the Pyrenean desman (*Galemys pyrenaicus*) (Insectivora: Talpidae), as revealed by radiotelemetry. *Journal of Zoology* **212**, 117-129.

Strasburg JL, Rieseberg LH (2010) How robust are "isolation with migration" analyses to violations of the im model? A simulation study. *Molecular Biology and Evolution* **27**, 297-310.

Suarez-Diaz E, Anaya-Munoz VH (2008) History, objectivity, and the construction of molecular phylogenies. *Stud Hist Philos Biol Biomed Sci* **39**, 451-468.

Sunnucks P (2000) Efficient genetic markers for population biology. *Trends Ecol Evol* **15**, 199-203.

Swenson NG, Howard DJ (2005) Clustering of contact zones, hybrid zones, and phylogeographic breaks in North America. *American Naturalist* **166**, 581-591.

Taberlet P, Fumagalli L, Wust-Saucy AG, Cosson JF (1998) Comparative phylogeography and postglacial colonization routes in Europe. *Molecular Ecology* **7**, 453-464.

Taberlet P, Waits LP, Luikart G (1999) Noninvasive genetic sampling: look before you leap. *Trends Ecol Evol* **14**, 323-327.

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585-595.

Talavera G, Castresana J (2007) Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst Biol* **56**, 564-577.

Tamura K, Peterson D, Peterson N*, et al.* (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular Biology and Evolution* **28**, 2731-2739.

Tautz D (1989) Hypervariability of simple sequences as a general source for polymorphic DNA markers. *Nucleic Acids Res* **17**, 6463-6471.

Thomson RC, Shedlock AM, Edwards, Shaffer HB (2008) Developing markers for multilocus phylogenetics in non-model organisms: A test case with turtles. *Mol Phylogenet Evol* **49**, 514-525.

Tosi AJ, Morales JC, Melnick DJ (2003) Paternal, maternal, and biparental molecular markers provide unique windows onto the evolutionary history of macaque monkeys. *Evolution* **57**, 1419-1435.

Townsend TM, Alegre RE, Kelley ST, Wiens JJ, Reeder TW (2008) Rapid development of multiple nuclear loci for phylogenetic analysis using genomic resources: An example from squamate reptiles. *Mol. Phyl. Evol.* **47**, 129-142.

Vali U, Einarsson A, Waits L, Ellegren H (2008) To what extent do microsatellite markers reflect genome-wide genetic diversity in natural populations? *Mol Ecol* **17**, 3808-3817.

Venter JC, Adams MD, Myers EW*, et al.* (2001) The sequence of the human genome. *Science* **291**, 1304-1351.

Vos P, Hogers R, Bleeker M*, et al.* (1995) AFLP: a new technique for DNA fingerprinting. *Nucleic Acids Res* **23**, 4407-4414.

Waits L, Taberlet P, Swenson JE, Sandegren F, Franzén R (2000) Nuclear DNA microsatellite analysis of genetic diversity and gene flow in the Scandinavian brown bear (*Ursus arctos*). *Molecular Ecology* **9**, 421-431.

Wandeler P, Hoeck PEA, Keller LF (2007) Back to the future: museum specimens in population genetics. *Trends in Ecology & Evolution* **22**, 634-642.

Waterston RH, Lindblad-Toh K, Birney E*, et al.* (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520-562.

Watson JD, Crick FH (1953) Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* **171**, 737-738.

Welch JJ, Bininda-Emonds OR, Bromham L (2008) Correlates of substitution rate variation in mammalian protein-coding sequences. *BMC Evol Biol* **8**, 53.

William J, Ballard O (1996) Combining data in phylogenetic analysis. *Trends Ecol Evol* **11**, 334.

Wilson DE, Reeder DM (2011) "Class Mammalia Linnaeus, 1758. In: *Animal biodiversity: An outline of higher-level classification and survey of taxonomic richness* (ed. Zhang Z-Q), pp. 56-60. Zootaxa.

Wiuf C, Christensen T, Hein J (2001) A simulation study of the reliability of recombination detection methods. *Molecular Biology and Evolution* **18**, 1929-1939.

Wright S (1931) Evolution in Mendelian Populations. *Genetics* **16**, 97-159.

Zhang DX, Hewitt GM (2003) Nuclear DNA analyses in genetic studies of populations: practice, problems and prospects. *Mol. Ecol.* **12**, 563-584.

Zhou Y, Mishra B (2005) Quantifying the mechanisms for segmental duplications in mammalian genomes by statistical analysis and modeling. *Proc. Natl. Acad. Sci. U S A* **102**, 4051-4056.

Zuckerkandl E (1964) Perspectives in molecular anthropology. In: *Classification and human evolution* (ed. Washburn SL), pp. 243-272. Methuen & Co., London.