UNIVERSITAT DE BARCELONA

Desarrollo de técnicas bioinformáticas para el análisis de datos de secuenciación masiva en sistemática y genómica evolutiva: Aplicación en el análisis del sistema quimiosensorial en artrópodos

Cristina Frías Lopez

Tesis doctoral 2019

Desarrollo de técnicas bioinformáticas para el análisis de datos de secuenciación masiva en sistemática y genómica evolutiva:

Aplicación en el análisis del sistema quimiosensorial en artrópodos

Cristina Frías López

Cristina Frías López
Tesis doctoral 2019

UNIVERSITAT DE BARCELONA

Desarrollo de técnicas bioinformáticas para el análisis de datos de secuenciación masiva en sistemática y genómica evolutiva:

Aplicación en el análisis del sistema quimiosensorial en artrópodos

Memoria presentada por **Cristina Frías López**
para optar al Grado de Doctor por la Universidad de Barcelona.

Departamento de Genética, Microbiología y Estadística

La autora de la tesis
Cristina Frías López

El director y tutor de la tesis
Dr. Julio Rozas Liras
Catedrático
Departamento de Genética,
Microbiología y Estadística
Facultad de Biología
Universidad de Barcelona

El codirector y tutor de la tesis
Dr. Miquel A. Arnedo Lombarte
Catedrático
Departamento de Biologia Evolutiva,
Ecologia y Ciencias Ambientales
Facultad de Biología
Universidad de Barcelona

Barcelona, Septiembre de 2019

A mi Yaya.

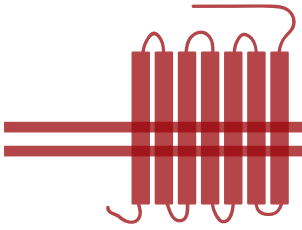
Abstract

The Next Generation Sequencing (NGS) technologies are providing powerful data to investigate fundamental biological and evolutionary questions including phylogenetic and adaptive genomic topics. Currently, it is possible to carry out complex genomic projects analyzing the complete genomes and/or transcriptomes even in non-model organisms.

In this thesis, we have performed two complementary studies using NGS data. Firstly, we have analyzed the transcriptome (RNAseq) of the main chemosensory organs of the chelicerate *Macrothele calpeiana*, Walckenaer, 1805, the only spider protected in Europe, to investigate the origin and evolution of the Chemosensory System (CS) in arthropods. The CS is an essential physiological process for the survival of organisms, and it is involved in vital biological processes, such as the detection of food, partners or predators and oviposition sites. This system, which has it relatively well characterized in hexapods, is completely unknown in other arthropod lineages. Our transcriptome analysis allowed to detect some genes expressed in the putative chemosensory organs of chelicerates, such as five NPC2s and two IRs. Furthermore, we detected 29 additional transcripts after including new CS members from recently available genomes in the HMM profiles, such as the SNMPs, ENaCs, TRPs, GRs and one OBP-like. Unfortunately, many of them were partial fragments.

Secondly, we have also developed some bioinformatics tools to analyze RNAseq data, and to develop molecular markers. Researchers interested in the biological application of NGS data may lack the bioinformatic expertise required for the treatment of the large amount of data generated. In this context, the development of user-friendly tools for common data processing and the integration of utilities to perform downstream analysis is mostly needed. In this thesis, we have developed two bioinformatics tools with an easy to use graphical interface to perform all the basics processes of the NGS data processing: i) **TRUFA** (TRanscriptome User-Friendly Analysis), that allows analyzing RNAseq data from non-model organisms, including the functional annotation and differential gene expression analysis; and ii) **DOMINO** (Development of Molecular markers in Non-model Organisms), which allows identifying and selecting molecular markers appropriated for evolutionary biology analysis. These tools have been validated using computer simulations and experimental data, mainly from spiders.

Índice



INTRODUCCIÓN

1. Impacto en la Biología Evolutiva y la Sistemática de las tecnologías NGS

La Biología Evolutiva y la Sistemática son dos ramas de la biología que investigan los procesos evolutivos responsables de generar la biodiversidad. Estas dos disciplinas estudian, entre otras cuestiones, la variabilidad del DNA para determinar los mecanismos evolutivos subyacentes y reconstruir relaciones filogenéticas entre individuos de poblaciones o especies diferentes. Sin embargo, la primera está más orientada al estudio de los mecanismos responsables de la evolución y la segunda en revelar los patrones característicos resultado de dicha evolución. Las principales fuerzas evolutivas que "moldean" los patrones de la biodiversidad son: la mutación (único proceso responsable de generar nuevas variantes), la recombinación (introduciendo combinaciones de variantes ya existentes), la deriva genética (cambiando las frecuencias de las variantes de forma estocástica), la migración o flujo génico (transfiriendo variantes de una población a otra) y, finalmente, la selección natural (único mecanismo que explica la adaptación de las especies a nuevos ambientes). Además, los patrones evolutivos también pueden ser moldeados por efectos demográficos, como expansiones, extinciones o cuellos de botella, etc. Para poder estudiar estos procesos es necesario obtener, analizar y comparar las secuencias genómicas de diversos organismos.

En 1977, a través de la tecnología de Sanger se obtiene la primera secuencia genómica completa del bacteriófago ΦX 174, que está compuesto por 5.386 pb (Sanger et al. 1977). Sin embargo, utilizando esta tecnología no era posible conseguir secuencias superiores a las 40 kb. Para alcanzar genomas más grandes, fue necesario el desarrollo de nuevas técnicas de secuenciación como el *shotgun sequencing* y la creación de secuenciadores automáticos basados en la técnica de Sanger (AB 373 DNA). Mediante la combinación de estas aproximaciones en 1995 se obtiene el genoma de la bacteria *Haemophilus influenzae Rd* que tiene una longitud de 1.830.137 pb (Fleischmann et al. 1995). Para ello, fue necesario el uso de 14 secuenciadores AB 373 DNA durante tres meses. Las lecturas se ensamblaron mediante el uso del software TIGR ASSEMBLER, programa que implementa una versión modificada del algoritmo de Smith-Waterman (alineamiento local) (T. F. Smith y Waterman 1981). Para este paso tardaron 30 horas utilizando un ordenador con un único procesador y 512 Mb de memoria RAM. Dado que el proceso de obtener un genoma era un proceso caro y laborioso, durante esa época, los recursos genómicos eran escasos.

No obstante, debido a los esfuerzos de numerosos investigadores para secuenciar el genoma humano, el cual tiene un tamaño de 3 Gb, se empiezan a desarrollar numerosas aproximaciones de secuenciación. A partir de 2005, con la llegada de las denominadas tecnologías de *Next Generation Sequencing* (NGS) la capacidad de secuenciar genomas y/o transcriptomas completos de cualquier organismo, se pone al alcance de muchos laboratorios, gracias a la reducción de costes y tiempo de producción. Estas tecnologías han permitido un avance importante en estudios evolutivos (Hudson, 2008), especialmente en aquellos que se centran en organismos no modelo. A partir de 2007, se publican los primeros estudios en los que se aplican metodologías de NGS para caracterizar los patrones evolutivos en poblaciones naturales, y estudios filogenéticos empleando un gran número de loci y de individuos (Gilad, Pritchard, y Thornton 2009). Desde entonces, la publicación de genomas anotados sufre un aumento vertiginoso (Figura 1), ampliando enormemente la disponibilidad de recursos para estudios de biología y genómica evolutiva. Sin embargo, dado que secuenciar genomas completos sigue siendo un proceso complejo, se han desarrollado técnicas para la secuenciación dirigida (o reducida) de una parte del genoma, que dependiendo de la técnica utilizada puede ser una región aleatoria, o específica (E. M. Lemmon y Lemmon 2013).

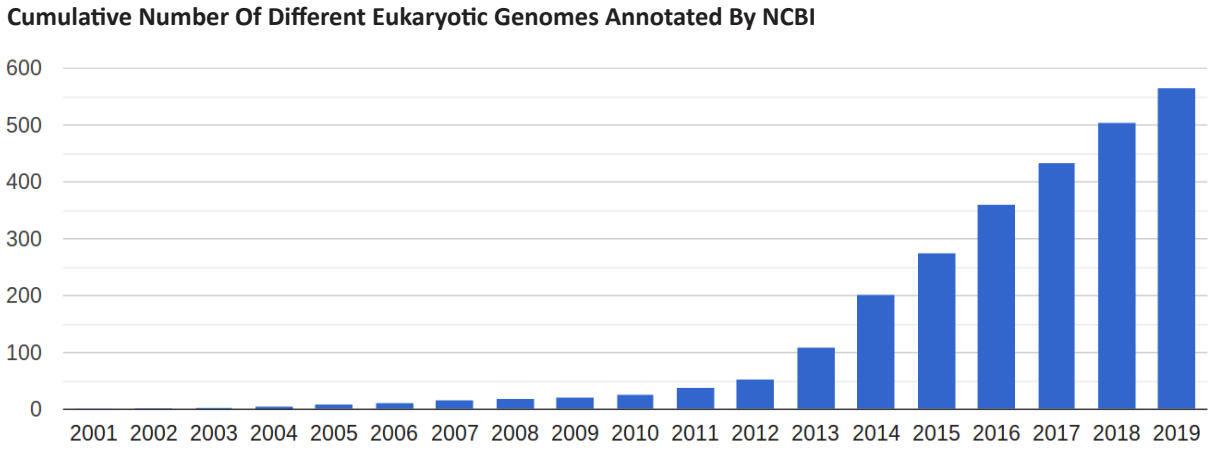


Figura 1. Número de genomas eucariotas anotados por NCBI desde 2001 hasta agosto de 2019. Fuente: (https://www.ncbi.nlm.nih.gov/genome/annotation_euk/).

Las técnicas basadas en NGS con más impacto en Biología Evolutiva son las siguientes:

a) **Secuenciación del transcriptoma (RNAseq)**: las tecnologías de RNAseq se desarrollaron para solventar algunas limitaciones asociadas a los *microarrays*, como la detección de isoformas poco abundantes, y la necesidad de diseñar sondas de hibridación para capturar los transcritos de interés (Wang, Gerstein, y Snyder 2009). Con las tecnologías de RNAseq podemos obtener el transcriptoma completo sin tener información genómica previa y se pueden utilizar para obtener los perfiles de expresión génica entre tejidos o individuos. Con esta tecnología es posible identificar genes involucrados en procesos biológicos concretos o genes con evidencias de selección positiva (Stapley et al. 2010), contribuyendo al estudio de los mecanismos que subyacen en los eventos de adaptación y de especiación (Ellegren 2008; Galindo, Grahame, y Butlin 2010).

b) **Secuenciación del genoma completo (*Whole genome sequencing* - WGS)**: las tecnologías de secuenciación genómica (DNA-seq) han permitido obtener de una forma sencilla genomas completos de organismos no modelo sin la necesidad de tener información previa de genomas cercanos. En los últimos años se han incrementado los estudios filogenéticos y de genética de poblaciones, utilizando individuos de poblaciones naturales, ya que es posible secuenciar el genoma de varios individuos en la misma carrera (*run*) de secuenciación (McCormack et al. 2013). Además, la resecuenciación a baja cobertura ha permitido caracterizar múltiples individuos en diferentes poblaciones y abordar aspectos de la evolución adaptativa como: detectar heterogeneidad en la tasa de recombinación, analizar el tamaño efectivo de la población, identificar expansiones de familias génicas involucradas en la adaptación específica de linaje y desvelar el papel de la divergencia genómica durante la especiación (Ellegren 2014).

c) **Identificación y desarrollo de marcadores moleculares a gran escala:** las tecnologías de NGS permiten detectar un gran número de SNP's (*Single nucleotide polymorphisms*), (Santosh Kumar, Banks y Cloutier 2012; R. Li et al. 2009), que puede ser de gran utilidad en estudios de genómica de poblaciones y/o de filogeografía (Townsend et al. 2008). Así mismo, se pueden desarrollar marcadores moleculares para estudios filogenéticos a través de la comparación de genomas completos, o mediante métodos de partición o reducción genómica. Estos métodos se basan en la preparación de librerías enriquecidas en fragmentos de DNA o RNA de interés. Para obtener las regiones de interés se utilizan cebadores, sondas o enzimas de restricción (Mamanova et al. 2010). Esta aproximación permite reducir la dimensión analítica, ya que enriquece la secuenciación en las regiones de interés y además aumenta la potencia del estudio, ya que se pueden combinar más individuos en el mismo análisis reduciendo así los costos económicos y de tiempo.

En esta tesis nos centraremos en estudios realizadas con técnicas de secuenciación de partición genómica, en concreto con datos de RNAseq y librerías genómicas reducidas generadas con enzimas de restricción.

2. Tecnologías de secuenciación del DNA

2.1 Origen de las primeras metodologías para secuenciar DNA

Desde el descubrimiento de la estructura del DNA en 1953 (Watson y Crick 1953), como la molécula que almacena y transmite el material hereditario, se empiezan a desarrollar diversas técnicas para determinar su secuencia. Por secuenciación de ácidos nucleicos entendemos, la determinación del orden de los nucleótidos de la secuencia lineal de la molécula de DNA (o RNA). Los principales descubrimientos que permitieron el desarrollo de las primeras técnicas de secuenciación fueron: el uso de las enzimas ADN polimerasas (síntesis de la cadena molde) (Wu 1972) y de las enzimas de restricción (fragmentación del ADN) (H. O. Smith y Welcox 1970), desarrollo de la reacción de polimerización en cadena (PCR) (Mullis y Faloona 1987; Saiki et al. 1988) y la aplicación de los geles de electroforesis para separar los fragmentos sintetizados en base a su tamaño (Gilbert y Maxam 1973; Sanger F. 1975).

- Las primeras técnicas de secuenciación del DNA surgen en el 1977. Ese año se publican en paralelo dos métodos pioneros para secuenciar el DNA: Walter Maxam y Allan Gilbert publican una técnica basada en la degradación química (Maxam y Gilbert 1977)**,** y Fred Sanger y Alan R. Coulson publican una técnica de reacción de "terminación de cadena" o "técnica dideoxi" (Sanger y Coulson 1975). La estrategia de estas dos técnicas era bastante similar, primero se obtenían fragmentos de DNA de diferente longitud (a través de digestión química o polimerización enzimática), donde el último nucleótido de cada fragmento se marcaba con un isótopo de fósforo radiactivo (32P). A continuación, se separaban los fragmentos en base a su tamaño en un gel de acrilamida, y por último para inferir la secuencia se revelaba el gel a través de una autorradiografía. El método de Sanger se convirtió en el más utilizado dado que, en comparación con el de Maxam y Gilbert, era más rápido y utilizaban menos productos tóxicos. En ambos métodos se realizaban cuatro reacciones independientes para obtener los fragmentos de DNA acabados en único tipo de nucleótido. En la técnica de Maxam y Gilbert, realizaban varias digestiones químicas y en la de

Sanger aplicaban los principios biológicos de la replicación del DNA. El sistema de Sanger para amplificar la secuencia era más eficiente. Utilizaban dideoxinucleótidos trifosfatos (ddNTPs) modificados, nucleótidos que carecen del grupo hidroxilo OH en el carbono 3', para inhibir la elongación de la cadena. Esto ocurre porque la ADN polimerasa necesita la presencia del grupo 3' OH para insertar el siguiente nucleótido. La técnica se basaba en realizar cuatro reacciones de síntesis independientes, en cada una de ellas se añadía un sólo tipo de ddNTP marcado radioactivamente (ddNTPs: ddGTP, ddATP, ddCTP, ddTTP) y el resto de desoxinucleótidos (dNTP). Cuando uno de los ddNTPs modificados se insertaba en la secuencia la polimerización finalizaba y se iban generando fragmentos de diferente longitud con el último nucleótido insertado marcado radioactivamente. Actualmente, el método de Sanger se continúa utilizando para obtener secuencias con una longitud comprendida entre 500 - 800 pb mediante amplificación por PCR.

- En 1986, aparece la primera generación de tecnologías de secuenciación (*First-Generation Sequencing* - FGS). Leroy Hood en colaboración con Applied Biosystem Instrument (ABI) implementan la secuenciación de Sanger de forma automatizada y desarrollan las primeras plataformas de secuenciación, Applied Biosystems 370A. Adaptaron la metodología de Sanger con algunas modificaciones, como por ej. la substitución del marcaje radiactivo por fluorescencia y añadieron un sistema de electroforesis capilar (L. M. Smith et al. 1986). En lugar de utilizar un único tipo de marcaje radiactivo se utilizaba un fluorocromo específico para cada uno de los nucleótidos terminadores de la cadena (ddNTP). A partir de una única reacción, era posible obtener los fragmentos de longitud variable acabados con los cuatro nucleótidos, y después se separaban a través de un gel de agarosa. La lectura se realizaba de forma automática, ya que el gel estaba conectado a un detector de fluorescencia CCD (*Charge-Coupled Devices*) que traducía la señal de la fluorescencia por la identidad del nucleótido en cada posición. Esta información se recogía en forma de cromatograma directamente en un ordenador.

- No obstante, el gran motor del desarrollo de las tecnologías de secuenciación del ADN fue el Proyecto Genoma Humano iniciado en 1990. El genoma humano publicado en 2001, costó unos 3 mil millones de dólares americanos y se consiguió mediante la colaboración de diversos grupos de investigación, tanto de carácter público (Consorcio Internacional) como privado (Celera Genomics Corporation) (Lander et al. 2001; Venter et al. 2001). Se realizó mediante la combinación del uso de las plataformas ABI PRISM 3700 y de la técnica *shotgun sequencing*, la cual se basaba en fragmentar al azar la molécula de DNA y después se insertan los fragmentos obtenidos en un vector de clonaje bacteriano (*Bacterial Artificial Chromosome-BAC*) para ser amplificados (Anderson 1981). Pero esta técnica era muy laboriosa y a partir de esta época empiezan a surgir diferentes técnicas con el fin de obtener secuencias más largas de forma automatizada (eliminando el paso de clonaje en bacterias), reduciendo costes y tiempo.

2.2 Desarrollo de las tecnologías de NGS

Las tecnologías de secuenciación de nueva generación (*Next Generation Sequencing* - NGS), también conocidas como tecnologías de secuenciación de alto rendimiento (*High Throughput Sequencing* - HTS) han evolucionado vertiginosamente desde la secuenciación del genoma humano, ofreciendo cada vez un mayor número de bases secuenciadas a un coste menor.

- En 2006 aparecen las primeras tecnologías de NGS de segunda generación (*Second-Generation Sequencing* - SGS) que se caracterizan por requerir librerías de fragmentos de DNA, para ser amplificados de forma individual previamente a la secuenciación, y poder realizar millones de reacciones de polimerización en paralelo (Metzker 2010). Las principales plataformas eran: GS FLX de 454 Life Sciences (Roche) (Margulies et al. 2005); Genome Analyzer, HiSeq, MiSeq y NextSeq de Illumina (Bentley et al. 2008); SOLiD de ABI (Ruparel et al. 2005); Ion Torrent de Life Technologies (Flusberg et al. 2010). Actualmente, algunas de estas plataformas han desaparecido como 454 y SOLID, en cambio Illumina es una de las plataformas más populares, debido a su alto rendimiento y bajo coste. Este tipo de tecnologías, se basan en fragmentar el ADN y después se amplifican los extremos del fragmento en múltiples reacciones en paralelo, obteniendo lecturas (*reads*) cortas, entre 100 y 300 pb. Dependiendo de la librería, es posible secuenciar sólo un extremo del fragmento, lecturas únicas (*single end*) o los dos extremos. Si secuenciamos los dos extremos, encontramos dos clases diferentes de *reads* dependiendo de la distancia genómica a la cual se encuentran las parejas: *paired end* o *mate paired* (Figura 2). La distancia que hay entre las parejas de los *reads* se denomina *insert size*, y en los *mate paired* es mucho mayor (2-5 kb) que en los *paired end* (<1kb).

Figura 2. Secuenciación de reads "emparejados". Los reads tipo *Paired End* y *Mate Paired* se obtienen a partir de secuenciar ambos extremos de un fragmento de DNA. Esta tecnología ofrece una mayor precisión para ensamblar regiones con repeticiones, ya que permite conocer la distancia entre las parejas de los reads secuenciados. Fuente: www.illumina.com.

- A partir de 2013, surgen las técnicas NGS de tercera generación, denominadas también secuenciación de molécula única (*Single Molecule Sequencing* - SMS). Estas técnicas no necesitan ningún paso de amplificación de la librería y son capaces de secuenciar directamente ("leer") una única molécula de DNA, sin aplicar ningún proceso enzimático ni de marcaje, copia o hibridación. Las principales plataformas

de tercera generación son: Pacific Biosciences (PacBio) (Roberts, Carneiro, y Schatz 2013) y Oxford Nanopore Technologies (ONT) (Jain et al. 2016). Estas plataformas generan *reads* más largos que las anteriores (5-50 kb) pero tiene una tasa de error bastante mayor. Principalmente, se utilizan para completar la secuencia genómica (*gap filling*), ya que las regiones repetitivas son difíciles de ensamblar, y para unir los *contigs* (secuencia contigua obtenida por el ensamblado de los *reads*) (*scaffolding*) previamente ensamblados con tecnologías de segunda generación (Lee et al. 2016).

- Por último, cabe destacar la aparición en 2016 de una nueva tecnología de tercera generación de mapeo, Dovetail Genomics, LLC (Santa Cruz, CA, USA), que permite realizar *scaffolding* a nivel cromosómico con lecturas de alta calidad a partir de un ensamblaje genómico fragmentado. Esta tecnología se basa en la secuenciación de dos tipos de librerías generadas a partir del proceso de ligación por proximidad de la cromatina, Chicago® (Putnam et al. 2016) y Dovetail™ Hi-C (Lieberman-Aiden et al. 2009). La conformación tridimensional de los cromosomas del genoma nos aporta información de la proximidad espacial de elementos distantes en el cromosoma, pero relacionados funcionalmente en procesos como la transcripción y la replicación, tales como promotores y "enhancers" (activadores de la transcripción). La metodología de estas dos librerías es bastante similar, únicamente difieren en el sustrato de partida. En la librería Hi-C, se fija la cromatina con formaldehído en el mismo núcleo celular a partir de tejido, sangre o cultivos celulares. En cambio, para generar la librería Chicago, la cromatina es reconstituida *in vitro*, combinando el DNA genómico extraído (DNAg) con histonas purificadas y factores de ensamblaje de la cromatina. A continuación, los pasos a seguir en ambas librerías son prácticamente iguales. Tras fijar la estructura conformacional con formaldehído de la cromatina, se fragmenta la cromatina usando enzimas de restricción, se marcan los extremos cohesivos con Biotina y se vuelven a unir los extremos romos al azar mediante ligación. Los fragmentos que tienen el marcaje de biotina en el interior del fragmento son seleccionados con Estreptavidina y se amplifican mediante PCR. Para secuenciar estas librerías se utilizan plataformas de Illumina para generar parejas de *reads* cortos tipo *Paired end* (100-300 bp), pero que están separados a una distancia mayor (~100 kb) que las lecturas únicas obtenidas por PacBio y ONT. Mediante esta metodología, tenemos información para poder reconstruir haplotipos (*phasing*) de gran longitud e identificar variaciones estructurales (SV) (Edge, Bafna, y Bansal 2017).

No obstante, las tecnologías de NGS también presentan limitaciones, la mayoría relacionadas con la capacidad analítica de los datos generados. El rápido progreso de las metodologías experimentales de NGS (Figura 3) ha forzado el continuo desarrollo de herramientas bioinformáticas y el uso de infraestructuras informáticas de altas prestaciones para poder almacenar y analizar grandes volúmenes de datos. Además, el aprendizaje de lenguajes de programación para poder procesar y analizar los datos de NGS, han provocado que la Bioinformática se convierta en una disciplina indispensable para poder analizar los datos de NGS. A partir de la llegada de las tecnologías de NGS, se produce un cambio de paradigma analítico. Anteriormente los estudios se basaban en un número limitado de genes y por lo tanto los datos eran accesibles a los investigadores desde cualquier ordenador sencillo. Sin embargo, en esta nueva era conocida como la de las "ómicas", pasamos a disponer de datos de secuenciación masiva, lo que requiere la implementación de algoritmos y sistemas de computación complejos, como *clusters* de ordenadores, generalmente bajo sistemas operativos de tipo **`UNIX`**.

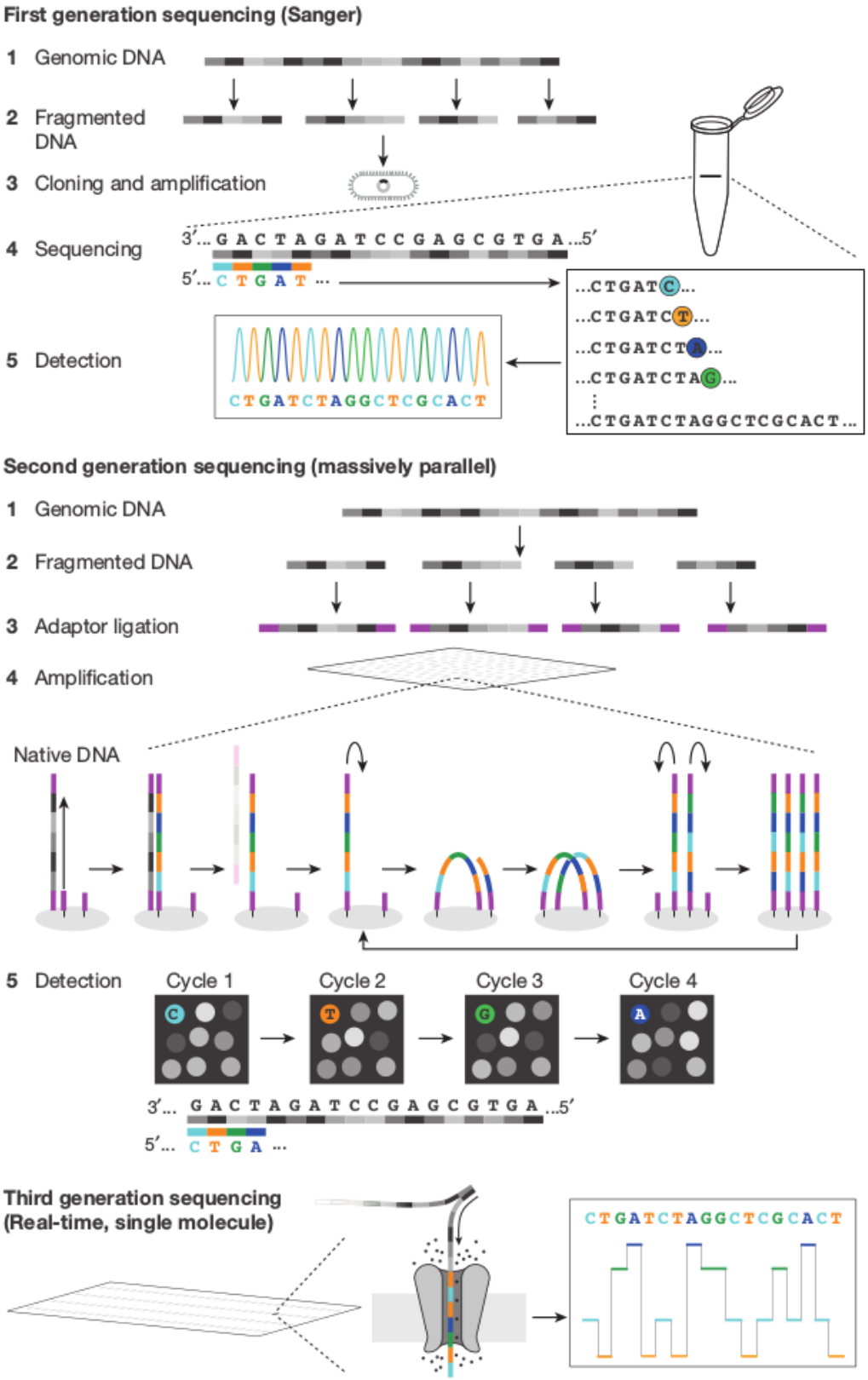


Figura 3. Resumen de las diferentes tecnologías de secuenciación. En esta ilustración se muestran ejemplos esquemáticos de las técnicas de secuenciación de primera, segunda y tercera generación. Fuente: (Shendure et al. 2017).

3. Análisis bioinformático de datos NGS

Básicamente, el proceso para obtener las secuencias genómicas a partir de datos de NGS se compone de dos fases. La primera fase (*in vitro*) está compuesta por todos los procesos experimentales relacionados con la secuenciación, como la extracción de los ácidos nucleicos y la preparación de la librería, en la cual podemos aplicar una metodología para enriquecer la secuenciación en nuestras regiones de interés y añadir adaptadores con secuencias conocidas para después poder identificar el origen de las secuencias. La segunda fase (*in silico*) comprende el pre-procesamiento, ensamblaje de los *reads* y el análisis de las secuencias ensambladas mediante herramientas bioinformáticas. La elección de la plataforma utilizada para secuenciar los datos NGS va a condicionar la fase de análisis, ya que cada tecnología presenta unas características específicas, como el tipo de *read* (*Single End*, *Paired End*, *Mate Paired*), errores de secuenciación o cobertura.

Una vez hemos obtenido los *reads* podemos dividir el proceso analítico en diferentes etapas. La primera etapa del procesamiento de los datos es común en todos los análisis, en cambio los últimos pasos dependen del tipo de cuestión que queramos responder. Todas estas etapas se pueden ejecutar de forma secuencial, formando un "pipeline" (tubería de procesos). Dado que intervienen varios programas, es necesario el uso de scripts (generalmente en **`Perl`**, **`Python`** o **`R`**) para automatizar el análisis, como cambiar el formato de los ficheros *output* según el formato *input* del programa siguiente, entre otros procesos. Normalmente, todos los comandos se suelen ejecutar en un único *script* (en lenguaje **`BASH`** o **`UNIX`**) para poder enviar los trabajos a un *cluster* de ordenadores.
A continuación, describiremos los pasos básicos de un análisis de datos de NGS.

3.1 Pre-procesamiento de los reads: Control de Calidad

El primer paso consiste en la evaluación de la calidad de los datos crudos (*raw data*) que recibimos del secuenciador. En general, este paso consiste en: eliminar la secuencia completa o parcial (*trimming*) de baja calidad de los *reads* y eliminar secuencias espurias y/o sobrerrepresentadas, que pueden ser restos de cebadores y/o adaptadores utilizados en la fase *in vitro* o contaminaciones de otras especies introducidas por una mala manipulación, como la presencia de DNA bacteriano en el caso de secuenciar genomas eucariotas.

El servicio de secuenciación entrega normalmente los datos NGS en formato FASTQ (Figura 4). Este formato es una extensión del formato FASTA, donde se muestra para cada posición del *read* el nucleótido y la calidad de la base secuenciada. En el caso de los *reads* tipo *Paired End* y *Mate Paired*, obtendremos un fichero para cada pareja (*forward* y *reverse*). En un fichero con formato FASTQ la información de cada secuencia se expresa en 4 líneas:

- la primera línea empieza por @ y a continuación encontramos el identificador de la secuencia e información de la carrera (run) de secuenciación.
- la segunda línea contiene la secuencia de nucleótidos (A, G, C, T).
- la tercera línea contiene un signo "+" para indicar el final de la secuencia anterior.
- la cuarta línea contiene los valores de calidad de cada base en código ASCII ("Q" o *Phred Quality Score*).

```
@HWI-ST1217:288:D2E5BACXX:4:1101:1330:1990 2:N:0:GTGGCC
AGCCACCAATAAACAAACCACACTCTTTGCCCTGCAGTTGCAAATCAAACAAAT
+
?@=*DDFFDFHHHIJJJJG<FHG@GHGGGHGHGYTYUTUIYUYIJKHKHYYTTIUY7
```

Figura 4. Ejemplo de un read completo con formato FASTQ. En este read la cuarta base secuenciada es una "C" con una probabilidad representada por el símbolo " * ". Este símbolo representa un Phred Quality Score (Q) igual a 9. El valor Q se obtiene a través de obtener el valor ASCII de " * ", el cual significa 42; y se le resta 33; obteniendo un valor de Q=9.

El valor de la calidad de la secuenciación (Q) nos indica la probabilidad de que la base secuenciada sea errónea. Actualmente, el formato de codificación más utilizado es el de Illumina 1.8+ (Phred-33), donde el valor numérico se representa por un carácter ASCII y se utiliza la ecuación de la figura 5 para calcular la probabilidad estimada de error (Cock et al. 2010).

Q = -10 log10 (Pe) o **Pe = 10^(-Q/10)**
Q = Valor de la calidad
Pe = Probabilidad estimada de error de la base secuenciada

Figura 5. En esta figura se muestra la fórmula para calcular la probabilidad estimada de error. Con los datos de la figura 4, donde Q=9 obtendremos una probabilidad estimada de error (Pe) de 0.12, la cual se calcula de la siguiente manera: P=10 ^ (-9/10); P =0.12589.

El filtro del valor de la calidad (Q) se seleccionará en base al análisis a realizar. Para un ensamblaje un valor de Q=20 (Q20) es suficiente, pero si estamos interesados en buscar posiciones variables lo ideal sería utilizar un valor más alto que el de la frecuencia de encontrar un polimorfismo de un sólo nucleótido (*Single Nucleotide Polymorphism* - SNP); por ejemplo, en el caso de humanos sabemos que la frecuencia media de encontrar SNP's es de 1 cada 1000 bp (Cooper et al. 1985), por lo tanto idealmente tendríamos que utilizar un *Phred Quality Scor*e mayor de Q30. A continuación, se muestra en la figura 6 la relación entre el valor de Q con su probabilidad estimada de error y el valor de precisión.

Q10 = 1 error cada 10 nucleótidos (**10% de error**)
Q20 = 1 error cada 100 nucleótidos (**1% de error**)
Q30 = 1 error cada 100 nucleótidos (**0.1% de error**)
Q40 = 1 error cada 10000 nucleótidos (**0.01% de error**)

Figura 6. Relaciones entre los valores Q con su probabilidad de error y el valor de la precisión expresado en porcentaje.

Para procesar los ficheros crudos (*raw reads*) los programas más utilizados son **PRINSEQ** (Schmieder y Edwards 2011), **NGS QC ToolKit** (Patel y Jain 2012), **FastP** (S. Chen et al. 2018) y para visualizar los datos, antes y después del pre-procesamiento, existen herramientas como **FastQC** (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/).

3.2 Ensamblaje (de novo / por referencia)

El siguiente paso consiste en obtener una secuencia contigua de mayor longitud a través de la unión por solapamiento de *reads* cortos (obtenidos con tecnologías de SGS*)*. La finalidad de este paso es reconstruir la secuencia original de donde proviene el *read.* En el caso de *reads* de DNAseq lo ideal sería obtener una secuencia de tamaño cromosómico, mientras que si son *reads* de RNAseq el objetivo es obtener el tránscrito completo.

Dependiendo de la existencia de una secuencia de referencia, encontramos dos tipos de aproximaciones para realizar el ensamblaje:

- Ensamblaje *de novo:* esta aproximación une los *reads* a través de sus regiones solapantes para obtener una secuencia contigua más larga (*contig*), sin utilizar una secuencia de referencia. Actualmente todos los softwares de ensamblaje *de novo* se basan en algoritmos de grafos (*graph-based algorithms*). Estos algoritmos representan a través de una relación binaria los *reads* como nodos y las regiones solapantes como conexiones. Actualmente, existen 4 estrategias para representar los solapamientos entre los grafos: *Overlap-Layout-Consensus* (OLC), grafos de *Bruijn* (DBG), grafos *greedy* que usan OLC o DBG, y el algoritmo híbrido (Khan et al. 2018; Miller, Koren, y Sutton 2010). Los programas más utilizados para el ensamblaje *de novo* para DNAseq son: **SPAdes** (Bankevich et al. 2012), **MaSuRCA** (Zimin et al. 2013), **SGA** (Simpson y Durbin 2012), **Abyss** (Simpson et al. 2009), **Ray** (Boisvert, Laviolette, y Corbeil 2010) y para RNAseq disponemos de **Trinity** (Grabherr et al. 2011) y **Bridger** (Chang et al. 2015). En general, el resultado de este proceso es un fichero en formato fasta con los contigs ensamblados. En caso de trabajar con RNAseq, es necesario el uso de programas que reduzcan el número de secuencias redundantes ensambladas, como **CD-HIT-EST** (W. Li y Godzik 2006).

- Ensamblaje por referencia (*mapping*): esta aproximación utiliza una secuencia de referencia como guía para alinear (mapear) los *reads* y de esta forma "colocarlos" en la posición correcta. Dependiendo de la naturaleza molecular de nuestras lecturas (DNA o RNA) encontramos dos tipos de softwares. Para alinear lecturas de DNA contra genoma se utilizan algoritmos *splice-unaware* y para alinear lecturas de RNA contra genoma los *splice-aware* (Engström et al. 2013). Los softwares *splice-unaware* son más apropiados para alinear datos de DNAseq, ya que alinean la lectura de forma contigua, sin considerar la información de los sitios de *splicing,* como **BWA** (Heng Li y Durbin 2010) y **Bowtie** (Langmead et al.

2009) (consultar revisión Mielczarek y Szyda 2016). En cambio, los softwares *splice-aware* consideran las coordenadas de las posiciones de los sitios de *splicing*, para identificar las uniones de empalme exon-exon y permite "partir" la lectura en las regiones intrónicas. Algunos de los softwares *splice-aware* más populares son **`TopHat`** (Trapnell, Pachter, y Salzberg 2009), **`STAR`** (Dobin et al. 2013), **`HISAT`** (D. Kim, Langmead, y Salzberg 2015), para más información consultar Baruzzo et al. 2017. El *pipeline* más popular en análisis de RNAseq es el paquete **`TUXEDO`**, donde la primera versión está formada por **`TopHat`** y **`Cufflinks`** (Trapnell et al. 2012) y la última versión por **`HISAT`** y **`Stringtie`** (Pertea et al. 2016). Además, estos *pipelines* contienen herramientas que permiten realizar análisis post-ensamblaje, como obtener la secuencia en formato fasta, identificar nuevas isoformas, análisis de expresión diferencial etc. (Griffith et al. 2015). En general, el resultado de este proceso está compuesto por un fichero que contiene los alineamientos en formato SAM/BAM y otros ficheros con las coordenadas de las estructuras genómicas mapeadas tipo GTF/GFF (los cuales serán descritos en la siguiente sección).

Debido a la ausencia de protocolos estandarizados, al desarrollo continuo de herramientas informáticas y las diferentes complejidades de cada genoma, es aconsejable utilizar varios softwares y comparar los ensamblajes producidos, convirtiendo en rutina el concepto de "hacer más pruebas" cada vez que abordamos un proyecto nuevo de secuenciación. Para evaluar la calidad y la cobertura media (número de veces que se secuencia una posición) alcanzada disponemos de varias aproximaciones. En el ensamblaje de genomas se suelen emplear medidas estadísticas que nos indican la distribución de la longitud de los *contigs*. El principal estadístico es el denominado N50 que indica la longitud mínima del *contig* que representa el 50% de las bases totales del ensamblaje. Pero, además es necesario una validación biológica, como la determinación de los genes identificados (completos o parciales) en el ensamblado. Para ello se utiliza, un conjunto de genes universales, genes ortólogos presentes en la mayoría de organismos, como los *housekeeping genes* (HKG). Algunas de las herramientas más populares para medir la proporción de los genes ortólogos altamente conservados, o para realizar una comparación contra un conjunto de genes referencia, son: **`BUSCO`** (Simão et al. 2015) y **`Detonate`** (B. Li et al. 2014).

3.3 Procesamiento del fichero de alineamiento (SAM/BAM)

Tras el proceso de mapeo, obtenemos un fichero SAM/BAM (S*equence Alignment Map/ Binary Alignment Map)* (H. Li et al. 2009). Este fichero describe los alineamientos de los *reads* que han mapeado contra la secuencia de referencia, pero también nos indica que *reads* no han mapeado. El formato SAM (*Sequence Alignment Map*) es un fichero de texto separado por tabuladores el cual se divide en dos secciones y el fichero BAM no es más que la forma binaria del SAM. En la primera sección, conocida como sección de encabezado, cada línea empieza por "@" y encontraremos información general sobre la secuencia de referencia, del proceso de secuenciación de las librerías de los *reads* y de los *softwares* utilizados para realizar el alineamiento. La segunda sección contiene la información de los alineamientos. Cada línea representa el alineamiento de un *read* y contiene 11 campos obligatorios que nos informan de la robustez del alineamiento a través de códigos (Figura 7). Además, en función del programa utilizado para realizar el alineamiento, podemos encontrar campos opcionales a partir de la columna 11.

A

```
@HD   VN: 1.0  SO: coordinate
@RG   ID: ANT
@PG   PN: Tophat     VN: 1.0.22
@SQ   SN: AFK010   LN: 382
HWI-ST1217:1839   145   AFK010   85   50   101M    =         1020   ACCGTGGTGCGT  HHI85JJJG<!B
HWI-ST1217:4171   97    AFK010   13   50   93M8N   AFK01   103    GGCCTCAGGTGA  GHIJHHF*+F(7
```

B

Col	Campo	Tipo	Descripción breve
1	QNAME	Cadena de caracteres	Identificador de la lectura
2	FLAG	Entero	Símbolo que indica el tipo de alineamiento
3	RNAME	Cadena de caracteres	Identificador de la secuencia mapeada
4	POS	Entero	Posición inicial del alineamiento
5	MAPQ	Entero	Calidad del alineamiento
6	CIGAR	Cadena de caracteres	Código que nos indica las regiones alineadas de la lectura y sus posiciones variables como (INDELs o MISMATCHs)
7	RNEXT	Cadena de caracteres	Identificador de la pareja de la lectura
8	PNEXT	Entero	Coordenada del inicio del alineamiento de la pareja de la lectura
9	TLEN	Entero	Distancia con la pareja de la lectura (Insert Size)
10	SEQ	Cadena de caracteres	Secuencia fasta de la región de la lectura alineada
11	QUAL	Cadena de caracteres	Calidad del alineamiento en código ASCII Phred-scaled base QUALITY + 33

Figura 7. Descripción fichero SAM/BAM. La figura A corresponde a un fichero SAM "simplificado", que contiene cuatro líneas en la sección de encabezado y dos líneas que corresponden a la información de los alineamientos de los reads HWI-ST1217:1839 y HWI-ST1217:4171; la figura B corresponde a la descripción de los campos del fichero SAM/BAM de la segunda sección, la cual contiene la descripción de los alineamientos.

Antes de utilizar el fichero SAM se deben eliminar los alineamientos de baja calidad y formatear el fichero según el software que vayamos a utilizar a continuación. Los alineamientos que se deben filtrar son los siguientes: *reads* que mapean en múltiples sitios, los *reads paired end* que tienen un *insert size* superior a la media o que su pareja no mapea, *reads* duplicados generados por PCR en el paso de amplificación de la librería, etc. Esta información se obtiene principalmente de los FLAGS, situados en la segunda columna de la sección del alineamiento. Para este proceso, encontramos una gran variedad de herramientas: **`SAMtools`** (H. Li et al. 2009), **`PicardTools`** (http://broadinstitute.github.io/picard/), **`BamTools`** (Barnett et al. 2011), etc. Todos estos programas permiten analizar ("parsear") los ficheros SAM/BAM, cambiar el formato del fichero, y algunos incluyen módulos para la detección de las posiciones variables, conocido como *variant calling* o *SNP calling* (Nielsen et al. 2011). El paquete más utilizado para procesar los ficheros SAM/BAM es **`SAMtools`**. Este programa ofrece diversas utilidades para filtrar las lecturas mal alineadas y cambiar el formato (ordenarlo e "indexarlo") para poder visualizar los datos o analizarlos con otros programas. Además, también tiene incorporada la función bcftools herramienta diseñada para calcular la variabilidad de las secuencias alineadas (H. Li 2011). En general los softwares suelen trabajar con el fichero BAM, ordenado e indexado, en lugar del fichero SAM, ya que es más rápido de procesar y analizar.

3.4 Análisis post-ensamblaje

Una vez obtenido el ensamblaje nos interesará descifrar la identidad y las características funcionales de las secuencias obtenidas, ya que los datos de NGS son fragmentos de secuencias de DNA con función desconocida (Abril y Castellano 2019). Para ello hay un paso fundamental, la anotación, mediante la cual se infiere la estructura y la función de las secuencias ensambladas.

3.4.1 Anotación estructural y funcional

Actualmente, para la anotación estructural (o predicción de genes) de un genoma, existen varias herramientas, que se pueden dividir en dos grupos según la aproximación metodológica utilizada: los que aplican métodos *ab-initio* (con evidencia intrínseca) y los que usan métodos comparativos basados en conocimientos previos (con evidencia extrínseca).

Los métodos *ab-initio* se basan en herramientas de predicción automática que únicamente emplean la información que se puede extraer de la propia secuencia genómica ensamblada y de modelos probabilísticos predefinidos (aunque estos modelos se pueden entrenar con los resultados de la predicción automática para mejorar la anotación). Esta aproximación utiliza algoritmos que analizan la secuencia de DNA/RNA en búsqueda de "señales" de elementos funcionales, como codones de inicio y terminación, sitios de empalme (*splicing*), etc. Para ello, usan perfiles probabilísticos, como los modelos ocultos de Markow (HMM en inglés *Hidden Markov Models*) o las matrices de peso posicional (PWM en inglés *Position Weight Matrices*), previamente creados a partir de evidencias curadas (Burge y Karlin 1997). Algunos de los programas más populares son: **`SNAP`** (Korf 2004), **`AUGUSTUS`** (Stanke y Waack 2003), **`GeneID`** (Guigó et al. 1992), **`GlimmerHMM`** (Majoros, Pertea, y Salzberg 2004), entre otros. En cambio, en la aproximación basada en evidencias externas a las propias del genoma ensamblado, se aplican métodos de búsqueda de similitud a nivel de secuencia, utilizando la información de los genomas anotados más cercanos y disponibles en las bases de datos, o de otro tipo de evidencia de la propia especie, como el transcriptoma. Los algoritmos usados en esta aproximación, se basan en métodos de alineamiento local, óptimos, como el de Smith-Waterman (T. F. Smith y Waterman 1981), o heurísticos, implementados en programas como **`BLAST`** (Altschul et al. 1990). Sin embargo, cabe destacar que el mejor enfoque es combinar las dos aproximaciones, con el fin de validar las predicciones probabilísticas con datos reales de la propia especie o de otros genomas, como por ejemplo alinear lecturas de secuencias de RNAseq con la secuencia genómica, método que se está implementando ya en algunos programas como **`Ipred`** (Zickmann y Renard 2015), **`MAKER2`** (Holt y Yandell 2011) o **`BRAKER1`** (Hoff et al. 2016).

Tras identificar las regiones codificadoras de proteínas putativas, la información de los ficheros de anotación obtenidos, en formato GTF/GFF (*General Transfer Format/General Features Format*) y/o los resultados de **`BLAST`** podemos inferir la pauta de lectura correcta (ORF en inglés *Open Reading Frames*) y traducir las secuencias de nucleótidos a aminoácidos utilizando por ej. **`Transdecoder`** (Haas et al. 2013). De esta manera, obtendremos el set de proteínas predichas codificadas por el genoma de estudio. A partir de ahí, podemos emplear herramientas como **`BLAST`** (Camacho et al. 2009), **`HMMER`** (Eddy 2009) o **`InterProScan`** (P. Jones et al. 2014) para anotar funcionalmente dichas proteínas. Los dos últimos son especialmente útiles cuando la divergencia entre nuestro organismo y los organismos con información incluida en las bases de datos es muy alta, o cuando queremos identificar genes específicos del linaje que estamos estudiando. **`HMMER`** (Eddy 2009) permite buscar patrones en las secuencias predichas a partir de diversas bases de datos de modelos de dominios proteicos, como la popular base de datos de Pfam (http://pfam.xfam.org/). **`InterProScan`** (P. Jones et al. 2014) nos aporta información de los procesos funcionales (términos GO, identificadores InterPro, etc.) y de las

rutas metabólicas (KEGG). También existen algunas plataformas *user-friendly* para realizar la anotación funcional de forma automática, aunque una de las más populares como **Blast2GO** (Conesa et al. 2005) sólo es gratuita si se usa a través de la línea de comandos.

3.4.2 Análisis de variación genética

Una vez obtenida la anotación funcional y estructural podemos realizar el análisis de los elementos funcionales de interés para el estudio, como pueden ser posiciones polimórficas o genes con expresión diferencial, etc.

- **Detección de variación genética:** en el caso de disponer de una secuencia de referencia y haber secuenciado múltiples individuos, podemos comparar las secuencias para detectar posiciones variables (genómica de poblaciones) (Santosh Kumar, Banks, y Cloutier 2012). No obstante, si carecemos de secuencia de referencia, podemos obtenerla realizando un ensamblaje *de novo* para después obtener un alineamiento. Dependiendo del elemento genético podemos diferenciar dos grupos:

    - variación de un único nucleótido: conocidas también como polimorfismo de un sólo nucleótido o SNP (*Single Nucleotide Polymorphism*) (Sachidanandam et al. 2001), donde la variación afecta sólo a un único nucleótido. Sin embargo, dentro de este grupo también se incluyen las Inserciones/Delecciones (INDELs) de secuencias cortas (< 10 bp) (Mullaney et al. 2010).

    - variación estructural (SV): este tipo de variación se caracteriza por la presencia de polimorfismos estructurales que se clasifican en dos grupos. Según el número de copias de la variante podemos diferenciar los de copia única que provocan reordenamientos del DNA en el genoma (p. ej., inversión, translocación equilibrada, etc), o las variantes que afectan al número de copias de un locus (CNV) (delecciones, inserciones, duplicaciones, inserción de retroelementos, etc.) (Freeman et al. 2006). El tamaño puede variar desde 100 - 200 pb hasta millones de pares de bases (Mb) de ADN (Lupski 2015).

    El software más popular para realizar la detección de variantes es **GATK** (*Genome Analysis Toolkit*) (DePristo et al. 2011). Esta herramienta dispone de un gran abanico de funciones para filtrar alineamientos erróneos, recalibrar los alineamientos y realinear las regiones cercanas a los INDELS. Después de filtrar el alineamiento podemos realizar la detección de variantes (*variant calling*) y obtendremos un fichero en formato VCF (*Variant Calling Format*) (Danecek et al. 2011).

- **Detección de expresión diferencial:** los análisis de expresión diferencial se emplean para identificar los genes que presentan cambios en el nivel de regulación entre diferentes condiciones, como tejidos, individuos, tratamientos, etc. En general, se alinean datos de RNAseq contra una secuencia de referencia (genoma o transcriptoma). Tras el proceso de alineamiento, a través del fichero SAM/BAM filtrado, podemos obtener una matriz de contajes utilizando programas como **HTSeq** (Anders, Pyl, y Huber 2015), Rsubread (Liao, Smyth, y Shi 2014), etc. A continuación, esta matriz se analizará para determinar qué genes presentan expresión genética diferencial estadísticamente significativa. Para este análisis encontramos una gran diversidad de paquetes de **R**, como **EdgeR** (Robinson, McCarthy, y Smyth 2010), **DESeq2** (Love, Huber, y Anders 2014), **Limma** (Law et al. 2014), etc., que difieren en el modelo estadístico usado en el estudio de la distribución de los genes.

4. Principales organismos utilizados en este estudio

En esta Tesis hemos generado y analizado datos de NGS de diferentes artrópodos, en concreto especies del subfilo de los quelicerados (Arthropoda, Chelicerata).

4. 1 Evolución y Filogenia de los Artrópodos

Los artrópodos son el filo animal con el mayor número de especies descritas. Actualmente, los artrópodos se clasifican en cuatro grandes subfilos: quelicerados, miriápodos, crustáceos y hexápodos (Gonzalo Giribet 2000). La mayoría de las hipótesis actuales proponen que los artrópodos son un grupo monofilético, y que los cuatro subfilos divergieron de un ancestro común marino (Daley et al. 2018), adaptándose posteriormente al medio terrestre en múltiples eventos independientes de terrestrialización. Es difícil inferir con exactitud cuándo se produjeron estas colonizaciones del medio terrestre. Algunos estudios, a partir de la estimación del tiempo de divergencia de los diferentes grupos, soportan que los quelicerados colonizaron la tierra hace unos 520 Ma, aunque según los registros fósiles, dicha colonización pudo ocurrir algo más tarde en los arácnidos, hace unos 495 Ma, similar al tiempo estimado para el subfilo de los hexápodos (485-445 Ma) (Lozano-Fernandez et al. 2019). En cambio, la evidencia fósil y la estima más antigua de terrestralización, es la de los miriápodos, datada hace unos 554 Ma (Lozano-Fernandez et al. 2019). En el caso de los crustáceos existe una gran incertidumbre acerca de cómo y cuándo se produjo el proceso. Los fósiles más antiguos sitúan la aparición de los primeros crustáceos durante el Mesozoico (250 Ma), sin embargo, existen grandes controversias ya que se han producido diferentes radiaciones e incluso transiciones reversibles del medio terrestre al medio acuático (Dunlop, Scholtz, y Selden 2013).

4.2 La araña *Macrothele calpeiana*

La araña *Macrothele calpeiana* (Walckenaer, 1805) (Chelicerata, Araneae, Macrothelidae) fue descrita por primera vez a partir de especímenes capturados en Ceuta. También se la conoce como araña de "tela de embudo" (*funnel web*) debido a que construye galerías subterráneas

recubiertas de seda. Pertenece al infraorden de arañas Mygalomorphae, que incluye alrededor de 3.000 especies, como las arañas de "trampilla" (*trap-door*, diversas familias) y las tarántulas (Theraphosidae) (World Spider Catalog 2019 - Natural History Museum Bern). Esta especie es probablemente la araña más grande de Europa (longitud total 40-60 mm) y es la única araña protegida por la legislación europea (Convenio de Berna (1979, apéndice II) (Collins, N. M., y Wells 1987). Es endémica del sur de la Península Ibérica y fue considerada inicialmente una especie vulnerable debido a la deforestación de su hábitat natural, los bosques de alcornoques (Collins, N. M., y Wells 1987). Sin embargo, estudios posteriores demostraron que la especie tiene una distribución mucho más amplia y podría encontrarse con frecuencia en áreas altamente deforestadas. En los últimos años, *M. calpeiana* se ha introducido en países europeos fuera de su área de distribución natural, probablemente mediante la exportación comercial de olivos españoles, lo que suscita algunas preocupaciones sobre su posible impacto en los ecosistemas invadidos (Bellvert y Arnedo 2016; Jiménez-Valverde, Decae, y Arnedo 2011).

M. calpeiana es también un organismo de particular interés en estudios biogeográficos. El género *Macrothele* muestra una distribución altamente fragmentada, con la mayor parte de su diversidad en el sudeste asiático (21 especies). Unas pocas especies habitan en el África tropical (4 especies) y sólo existen dos especies conocidas en Europa, *M. calpeiana* y *M. cretica (*Kulczynski, 1903), una araña endémica de Creta que también es motivo de preocupación para la conservación. Un estudio filogenético reciente (Opatova y Arnedo 2014) reveló que las dos especies europeas de *Macrothele* no son taxones hermanos, y sugiere que la colonización de Asia a Europa se produjo en eventos independientes. Otra característica interesante de este género es la potencia del veneno que producen. De hecho, estudios sobre la estructura molecular y las propiedades químicas de las toxinas venenosas de *Macrothele* (Satake et al. 2004; Yamaji et al. 2009; X.-Z. Zeng, Xiao, y Liang 2003) han revelado que algunas moléculas del veneno se pueden aplicar como inhibidores del crecimiento celular en terapias contra el cáncer (Gao et al. 2005; Z. Liu et al. 2012).

5. Estudio del origen y evolución del sistema quimiosensorial en artrópodos

5. 1 El sistema quimiosensorial

Todos los organismos tienen órganos sensoriales especializados que detectan estímulos del ambiente, incluidas señales visuales, acústicas, táctiles y químicas. El sistema quimiosensorial (SQ) es el responsable de detectar las señales químicas del exterior y generar una respuesta comportamental de los individuos, produciendo respuestas tanto de atracción como de repulsión (Anholt y Mackay 2001). Este sistema, esencial para la supervivencia y la reproducción de los organismos, está presente en todos los organismos desde procariotas a eucariotas, siendo uno de los más primitivos (Hildebrand y Shepherd 1997). Dependiendo de la naturaleza, concentración y rango espacial de la señal química, podemos clasificar este sistema en dos modalidades sensoriales fisiológicamente relacionadas, aunque en algunos casos anatómicamente independientes (ver más abajo): los sentidos del gusto y el olfato. En términos generales, estos dos procesos se diferencian por el índice de solubilidad de las moléculas que detectan, es decir, si son hidrofóbicas o hidrofílicas. Además otro factor importante que se ha usado para distinguir entre ambos es la distancia a la cual se pueden detectar dichas moléculas. El gusto normalmente detecta moléculas semi-solubles y en estado sólido a través del contacto físico directo con las estructuras quimiorreceptoras, mientras que el olfato suele especializarse en moléculas volátiles emitidas a distancias largas y en concentraciones mucho menores (E Mollo et al. 2017; E Mollo et al. 2014).

5.2 El SQ de los artrópodos

Como en el resto de los animales, el SQ de los artrópodos participa en la detección de alimento, pareja, depredadores, sitios de ovoposición, compuestos tóxicos e incluso en la comunicación social (Krieger y Ross 2002; Matsuo et al. 2007; Whiteman y Pierce 2008). Este gran filo de invertebrados es tan numeroso y diverso debido en gran parte a su habilidad de colonizar hábitats muy diferentes, incluyendo múltiples eventos independientes de colonización al medio terrestre, donde muchos de estos factores biológicos que acabamos de comentar (y por lo tanto el SQ) tuvieron seguro un

papel muy relevante. Todos estos factores convierten al SQ de los artrópodos y a sus componentes en un buen modelo de estudio de la adaptación ecológica y un sustrato ideal para detectar la huella de la selección natural a nivel molecular. Hasta la fecha, el SQ se ha estudiado principalmente en insectos, debido a la gran disponibilidad de recursos genómicos en las bases de datos de especies de hexápodos (Vieira y Rozas 2011) y de herramientas de disección genética funcional en *Drosophila* (Devaud 2003). No obstante, durante la última década, debido al avance de las tecnologías de NGS, se están obteniendo muchos datos genómicos y transcriptómicos de otros grupos de artrópodos, impulsando la investigación sobre el origen y evolución del SQ en estos animales tan diversos.

5.2.1 Componentes moleculares del SQ periférico de artrópodos

En los insectos (y en la mayoría de hexápodos estudiados hasta ahora), el proceso de quimiorrecepción se inicia con la entrada de las moléculas que actuarán como señales químicas a través de los poros de unas estructuras especializadas, similares a pelos, pero recubiertos de cutícula, denominados sensilios quimiosensoriales. Podemos distinguir dos grandes tipos de sensilios según su morfología y estructura: los olfativos y los gustativos (Figura 8). Los sensilios gustativos presentan un único poro en el ápice (0.5–2 um de diámetro) y se encuentran distribuidos por todo el cuerpo, aunque por lo general los podemos encontrar en aparato bucal, patas, alas e incluso la placa vaginal (Fiala 2007). En cambio, los sensilios olfativos presentan múltiples poros (50-200 nm de diámetro) y se encuentran distribuidos principalmente en el tercer segmento de la antena y de forma secundaria en otros apéndices como el palpo maxilar (Joseph y Carlson 2015a). La morfología y distribución de los sensilios quimiosensoriales de los otros tres mayores grupos del filo de los artrópodos presentan una gran diversidad. Mediante estudios de microscopía y tinciones del sistema nervioso, se han descubierto sensilios quimiosensoriales principalmente en las antenas y anténulas de crustáceos (Derby et al. 2016; Harzsch y Krieger 2018) , en las antenas y patas de los miriápodos (Ernst y Rosenberg 2003; Kenning, Müller, y Sombke 2017; Sombke et al. 2011) , y en el órgano de Haller y en los órganos tarsales de patas y pedipalpos (Carr et al. 2017; Foelix 1970; Ganske y Uhl 2018; Harris y Mill 1973) en quelicerados (Figura 9).

Una vez la molécula señalizadora entra al espacio sensiliar acuoso (linfa sensiliar), esta es transportada por difusión (en caso de ser hidrofílica), o unida a un transportador (si es una molécula hidrofóbica) que la solubiliza y la transporta hasta las proximidades del receptor ubicado en la dendrita de la membrana de las neuronas receptoras (RNs) (Pelosi et al. 2014a). Tras la interacción de la molécula con los receptores en la membrana de las RNs, estas convierten la señal química en eléctrica (transducción de la señal). Los receptores generan gradientes iónicos entre el interior (iones negativos) y el exterior (iones positivos) de la membrana de la RNs. Este diferencial de potencial despolariza la membrana y provoca unos potenciales de acción eléctricos. Estos potenciales son transmitidos por la membrana axonal hasta las estructuras del sistema nervioso central donde se interpreta la información (Joseph y Carlson 2015).

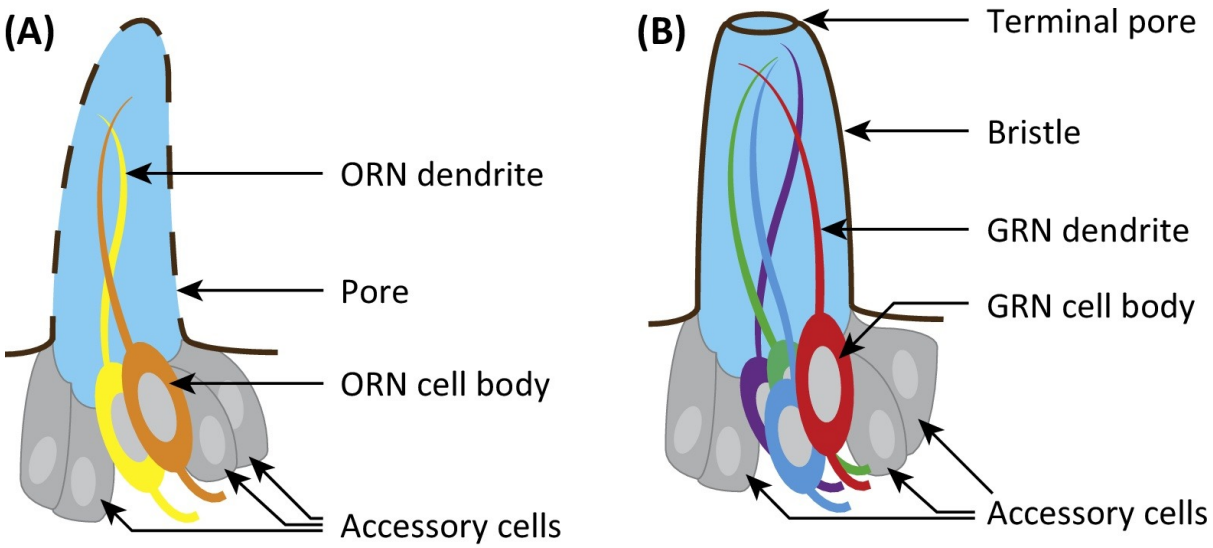


Figura 8. Esquema general de la estructura los sensilios quimiosensoriales de insectos. (A) Sensilio olfativo. (B) Sensilio gustativo. ORN, olfactory receptor neuron. GRN, gustatory receptor neuron. Las dendritas de las neuronas receptoras están bañadas en linfa sensiliar (en azul claro). Fuente: (Joseph y Carlson 2015a).

5.3 Las familias multigénicas del SQ de artrópodos

Las proteínas más importantes implicadas en el SQ de artrópodos están codificadas por genes que forman familias multigénicas de tamaño medio-grande (de diez hasta cientos o incluso miles de genes por familia). En hexápodos se han identificado varias familias de proteínas solubles transportadoras (Pelosi et al. 2014a, 2018), entre las que destacan las odorant-binding proteins (OBP), las chemosensory proteins (CSPs) o las Niemann-Pick protein type C2 (NPC2), y de receptores anclados a la membrana de las RN (Benton 2015; Joseph y Carlson 2015b), principalmente los olfactory (OR), gustatory (GR) y los ionotropic (IR) receptors. Además, otros receptores y proteínas de membrana han sido también implicados en el SQ de algunas especies, como por ejemplo los degenerin/epithelial sodium channels (DEG/ENaC), las neuron membrane proteins (SNMPs) y los transient receptor potential (TRP) (Figura 9). El conocimiento de todas estas proteínas en el resto de linajes de artrópodos es mucho más pobre, aunque se sabe que estos carecen de ORs y OBPs (Vieira y Rozas 2011).

Los estudios de genómica comparada muestran que todas estas estas familias evolucionan bajo el modelo denominado de "nacimiento y muerte", incluyendo cambios adaptativos en muchas de sus copias en diferentes linajes (Almeida et al. 2014; Sánchez-Gracia, Vieira, y Rozas 2009). Los nuevos componentes de estas familias son por lo tanto el resultado estocástico de ganancia y pérdida de genes y de la divergencia funcional en algunos de sus miembros guiada por la selección natural, aunque con la limitación de un conjunto mínimo de genes imprescindibles para realizar la actividad quimiorreceptora (Sánchez-Gracia, Vieira, y Rozas 2009).

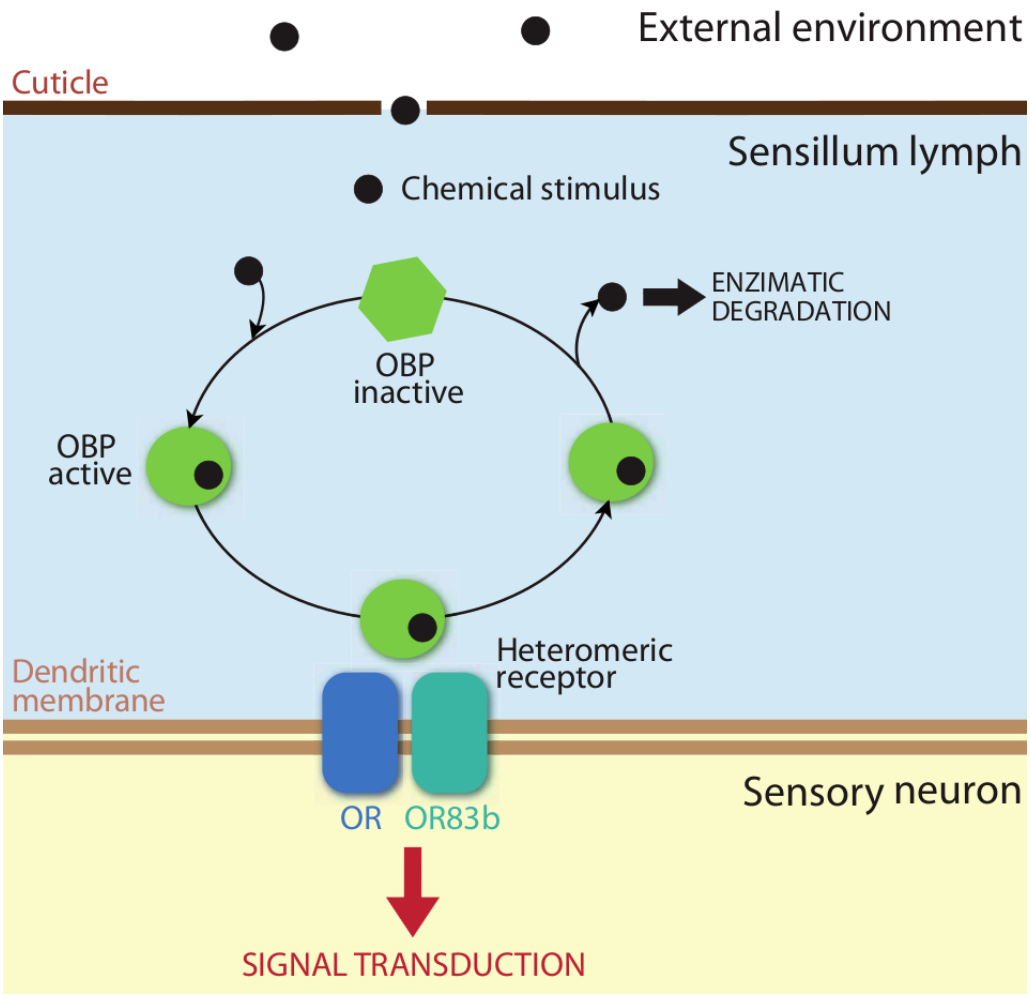


Figura 9. Representación esquemática de un modelo de los eventos que ocurren en la quimiorrecepción periférica en insectos. Esta figura representa un esquema funcional general simplificado. Se han propuesto esquemas alternativos para la actividad OBP (ver (R.G. Vogt 2005)). Fuente: (Sánchez-Gracia, Vieira, y Rozas 2009).

5.3.1 Proteínas solubles transportadoras

Las proteínas de unión a ligando son las responsables de transportar las señales volátiles (hidrofóbicas) hacia las proximidades del receptor o de ayudar en la degradación de la señal tras la interacción con el receptor (Scheuermann y Smith 2019; Richard G. Vogt y Riddiford 1981). Suelen tener un patrón conservado de cisteínas, el cual es necesario para dar estabilidad al plegamiento globular, presentando los residuos hidrofílicos en el exterior y los más hidrofóbicos alrededor del bolsillo donde se une el ligando (Pelosi et al. 2014b). Estas proteínas suelen tener una longitud de unos 140 aminoácidos y se encuentran de forma abundante en la linfa de los sensilios quimiosensoriales, ya que son secretadas por las células accesorias de los sensilios. Aunque, también se han identificado algunos genes de estas familias, como los Obp Minus-C, expresados en otros tejidos no relacionados directamente con el SQ, como la cabeza, la carcasa adulta, los testículos, las glándulas accesorias masculinas, la espermateca y algunos tejidos larvarios (datos del proyecto FlyAtlas; (Chintapalli, Wang, y Dow 2007). Esto sugiere, que también pueden estar involucradas en el transporte de otras sustancias no relacionadas con el SQ.

La familia de las OBP

Las OBPs presentan un patrón conservado de seis cisteínas que forman tres puentes disulfuro y no son homólogas a las odorant-binding proteins de vertebrados (Leal, Nikonova, y Peng 1999). Son específicas de hexápodos, lo que sugiere que se originaron después de la división de este grupo y del resto de pancrustáceos (~470 MA) (Vieira y Rozas 2011). La historia aparentemente paralela de OBP y OR (ver el apartado de las OR más abajo) se ha descrito como un escenario de coevolución entre estas familias después de colonización del medio terrestre por parte de de los insectos. No obstante, en los últimos años han surgido otras hipótesis sugiriendo que su origen podía ser diacrónico (Missbach et al. 2015).

La familia de las CSP

La secuencia típica de aminoácidos de una CSP varía entre 100 y 120 residuos, presentando un patrón conservado de cuatro cisteínas que forman dos puentes disulfuro en la proteína madura (Richard G. Vogt y Riddiford 1981; Angeli et al. 1999). Las CSPs están presentes en los órganos olfativos y gustativos de todos los artrópodos aunque cabe remarcar que en quelicerados sólo se ha identificado una única CSP en una especie de ácaro (Pelosi et al. 2018).

La familia de las NPC2

Esta familia se había relacionado principalmente con el transporte de lípidos en otras especies de animales (Liou et al. 2006). En estudios posteriores se han descubierto genes de esta familia altamente expresados en la linfa de algunos sensilios antenales en la hormiga *Camponotus japonicus* (Ishida et al. 2014). Esto sugiere que estas proteínas transportadoras podrían estar involucradas en el transporte de señales químicas implicadas la comunicación social (que en insectos suelen ser ácidos grasos de cadena larga, alcoholes y acetatos). Además, la estructura tridimensional de las NPC2 es similar al denominado barril β de las odorant-binding proteins de vertebrados (que estructuralmente son lipocalinas) y presenta un patrón conservado de seis cisteínas que se asemeja al de las OBP de hexápodos (Pelosi et al. 2014a).

5.3.2 Receptores de membrana

La primera clase de proteínas quimiorreceptoras que se descubrió fue la familia de los receptores acoplados a proteína G (GPCR) de vertebrados (Buck y Axel 1991). Este tipo de receptores están presentes en todos los vertebrados, pero también en nemátodos y moluscos (Bargmann 2006). Sin embargo, en artrópodos no hay evidencias firmes de su relación con el SQ (Spehr y Munger 2009). En este grupo, en cambio, las principales familias de quimiorreceptores, los GR, OR e IR, presentan dominios y configuraciones transmembrana muy diferentes, siendo claramente evolutivamente independientes a los receptores de vertebrados (Yarmolinsky, Zuker, y Ryba 2009). No obstante, durante los últimos años se han descubierto miembros de otras familias, previamente caracterizadas en procesos mecanosensoriales, como receptores de los SNMPs, los ENaCs y los TRPs, que también podrían estar involucradas en la quimiorrecepción en algunas especies de artrópodos (Joseph y Carlson 2015b) (Figura 10).

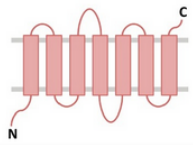
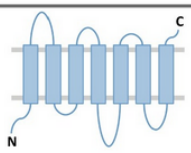
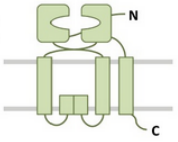
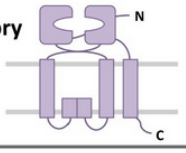
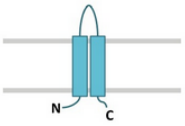
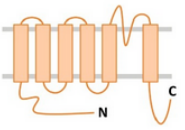
Receptors	Examples	Ligands	Behavioral output
Ors 	Or56a Or47b Or19a Or67d (♂) Or67d (♀)	Geosmin (toxic microbes) Methyl laurate (pheromone) Terpenes (citrus fruits) cis-Vaccenyl acetate cis-Vaccenyl acetate	Activates an aversion circuit Promotes male courtship of females Activates oviposition circuit Inhibits male-male courtship Promotes female receptivity
Grs 	Gr5a Gr93a Gr39a Gr63a	Trehalose (sugar) Caffeine (bitter) Presumptive female pheromone CO_2, stress pheromone	Promotes feeding Inhibits feeding Promotes male courtship of females Promotes avoidance
Antennal IRs 	IR64a IR92a IR84a IR40a	Acids Ammonia and amines Phenylacetic acid (food odor) DEET	Activates avoidance responses Activates attraction Stimulates courtship on food sources Activates aversion circuit
Gustatory IRs 	IR52c	Presumptive pheromone	Promotes male courtship behavior
Ppks 	Ppk23 Ppk25 Ppk28	7,11-Heptacosadiene Multiple pheromones Water	Acts in male courtship behavior Stimulates receptivity to courtship Promotes fluid intake
TRPs 	TRPA1 Painless TRPL	Aristolochic acid Isothiocyanate (wasabi) Camphor (bitter)	Inhibits feeding Activates aversion circuit Inhibits feeding

Figura 10. Principales familias de quimiorreceptores en artrópodos. En la primera columna, se muestran las topologías predichas para los diferentes tipos de receptores. Algunos ejemplos de receptores de cada clase, sus posibles estímulos y las posibles respuestas comportamentales asociadas se presentan en la segunda, tercera y cuarta columnas, respectivamente. Fuente: (Joseph y Carlson 2015a).

La superfamilia de quimiorreceptores (OR y GR)

Los primeros receptores quimiosensoriales descubiertos en artrópodos mediante el análisis del genoma de *D. melanogaster* fueron los OR (Clyne et al. 1999) y los GR (Craig Montell 2009). En general, estos receptores presentan una longitud de entre 300 y 500 aminoácidos y siete dominios transmembrana (7TM), aunque se disponen en la membrana con una topología inversa con respecto a los receptores de vertebrados. El extremo carboxi-terminal (C-terminal) está orientado al exterior de la célula mientras que el extremo amino-terminal (N-terminal) se localiza en el interior de la misma (Robertson, Warr, y Carlson 2003). Además, el mecanismo de señalización parece ser también diferente, las GPCRs olfativas de vertebrados actúan a través de mecanismos metabotrópicos, donde es necesario la activación de la señalización a partir de un segundo mensajero. En cambio, en insectos, los propios receptores constituyen complejos heteromultiméricos, como el dímero formado por el GR21a y GR63a (W. D. Jones et al. 2007), o el que componen el co-receptor OR83b (también conocido como ORCO) y otros OR específicos, donde ORCO actúa como canal iónico y la otra subunidad determina la especificidad de ligando (Stengl y Funk 2013).

A diferencia de los OR, que son exclusivos de los hexápodos (Missbach et al. 2014; Vieira y Rozas 2011), se han encontrado miembros de la família de los GR en todos los otros artrópodos, ver (Chipman et al. 2014; Ngoc et al. 2016; Peñalva-Arana, Lynch, y Robertson 2009) para algunos ejemplos. A nivel de secuencia proteica, presentan una divergencia muy alta entre las copias de la misma familia, con tan solo un 8% de similitud entre copias de la familia de diferentes subfilos de artrópodos. En *Drosophila* se han identificado miembros de la familia de los GR relacionados con la detección de sabores dulces y amargos (Fujii et al. 2015), y con la detección de CO_2 (W. D. Jones et al. 2007).

Las familias de los IR

En el 2009 se descubrió en *D. melanogaster* una nueva subfamilia de receptores ionotrópicos de glutamato (iGluR), los denominados IR (Benton et al. 2009). A diferencia de los iGluR, que están presentes en un amplio rango de especies (bacterias, plantas y animales), los IR no se han encontrado en ninguna especie del grupo de los deuteróstomos, pero sí en todas las especies de protóstomos analizadas, como nemátodos, artrópodos y moluscos (Croset et al. 2010). La subfamilia de IR presentan algunos miembros que se han relacionados con el SQ en insectos, como por ejemplo los IR "antenales", relacionados con el olfato en *Drosophila*, o algunos IR no antenales asociados con el gusto en esta misma especie (Rytz, Croset, y Benton 2013). En general, los iGluR presentan una longitud de alrededor de 1000 aminoácidos, tienen tres dominios transmembrana (3-TM) y tres dominios funcionales conocidos: un dominio N-terminal extracelular (ATD), involucrado en el ensamblaje de canales y un dominio extracelular de unión a ligando bipartido (LBD), el cual se divide en dos lóbulos (S1 y S2) que están separados por un dominio de canal iónico (ICD) (Croset et al. 2010). Aparte de los IR, la familia de los iGluR está formada por otras tres subfamilias de receptores, los AMPA, Kainato y NMDA (Armstrong et al. 1998). Estos receptores intervienen en el proceso de la transmisión sináptica a través de la detección del glutamato y en procesos relacionados con el aprendizaje (Littleton y Ganetzky 2000).

Interesantemente, los IR han perdido la capacidad de unir glutamato (no tienen el dominio ATD o lo presentan extraordinariamente divergente). Los IR antenales se han identificado en ORN que no expresan OR y se cree que funcionan a través de un mecanismo de acción similar al de estos últimos, con IR25a e IR8a actuando como correceptores (Abuin et al. 2011) y formando complejos heterodiméricos con otros IR específico de ligando (Rytz, Croset, y Benton 2013). Se han encontrado miembros de esta subfamilia en todas las especies de artrópodos, no obstante, presentan una baja similitud a nivel de secuencia proteica (8.5%) (Croset et al. 2010).

La familia de las SNMP

La familia de los receptores CD36 se caracterizan por la presencia de dos dominios transmembrana (2-TM), y se ha demostrado que participan en el metabolismo de lípidos y en procesos del sistema inmunológico, en mamíferos, peces y artrópodos (Fink et al. 2015; Neculai et al. 2013). Sin embargo, se encontraron algunos genes de esta familia, los SNMP, expresados en las ORN de insectos, en las mismas células que los OR que reconocen feromonas en esta especie. Parece ser que los receptores SNMP1 y el SNMP2, intervienen en el proceso de detección de feromonas, actuando como cofactores de los OR (van der Goes van Naters y Carlson 2007). En *D. melanogaster*, son esenciales para la detección de la feromona volátil, conocida como *11-cis-vaccenyl acetate* (cVA) (Jin, Ha, y Smith 2008).

La familia de las DEG/ENaC

Las DEG/ENaC presentan dos dominios transmembrana (2-TM) y un dominio conservado extracelular rico en cisteínas. Se identificaron por primera vez en el nemátodo modelo, *Caenorhabditis elegans* (Maupas, 1900) dónde participan en varios procesos sensoriales y mecanosensoriales, como la propiocepción, la quimiorecepción o el transporte de Na+ (Mano y Driscoll 1999). En *D. melanogaster* esta familia génica está compuesta por 31 genes llamados pickpocket (Zelle et al. 2013). Se han encontrado algunos miembros expresados en las neuronas gustativas, y hay estudios que soportan que están relacionadas con la detección de feromonas por contacto (PPK23) (Toda, Zhao, y Dickson 2012) o la detección de sales (L. Liu et al. 2003) y agua (PPK28) (Z. Chen, Wang, y Wang 2010).

La familia de los TRP

Los TRP forman parte de un grupo de receptores de canales iónicos no dependientes de voltaje, y son proteínas altamente conservadas que están presentes en todas las especies desde la levadura hasta los mamíferos (Nilius 2003; Vassort y Fauconnier 2008). Están relacionados con una amplia variedad de modalidades sensoriales, principalmente con la termo- y mecanorrecepción (Craig Montell 2005; Pedersen, Owsianik, y Nilius 2005)y en menor medida con el SQ, ya que en algunos hexápodos se ha demostrado que participa en procesos gustativos de compuestos nocivos (S. H. Kim et al. 2010).

Esta familia se caracteriza por la presencia de seis dominios transmembrana (6-TM), y está compuesta por siete subfamilias que se dividen en dos grupos. En el grupo 1, encontramos los TRPC (*canonic*), los TRPV (*vanilloid*), los TRPM (*melastatin*), los TRPP (*polycystin*), los TRPA (*ankyrin*) y los TRPN (*no mechanoreceptor potential C NOMPC*); en el grupo 2, tenemos los TRPP (*polycystin*) y los TRPML (*mucolipin*) (C. Montell 2005). El primer receptor de esta familia, se descubrió en *D. melanogaster* y estaba involucrado con procesos de fotodetección (Craig Montell 2005). Posteriormente, se descubrieron miembros de otras cuatro subfamilias (TRPA, TRPV, TRPC y TRPM) expresados en las neuronas gustativas y olfativas (Cattaneo et al. 2016; Kozma et al. 2018). Además, se constató su participación en procesos quimiosensoriales como mediadores en la detección de compuestos tóxicos. En *D. melanogaster*, los canales TRPA1 (TRPA) y el TRPL (TRPC) están implicados en la detección de compuestos tóxicos, como la citronela (Kwon et al. 2010) y el wasabi (Al-Anzi, Tracey, y Benzer 2006). También se han encontrado receptores TRPA1 expresados en la misma neurona que la GR66a, que detecta la cafeína, y junto otras GR relacionadas con el gusto amargo, como la GR32a y GR47 (S. H. Kim et al. 2010). En general, estos receptores se activan a través de la detección de compuestos nocivos y alteran la termorregulación. Por ejemplo, el receptor TRPV1 (TRPV) está involucrado en la detección de la capsaicina y produce un aumento de la temperatura (Caterina et al. 1997), y en cambio el TRPM8 (TRPM) con la identificación del mentol y un descenso de la temperatura (Peier et al. 2002). Debido a estas funciones, estas familias se estudian para desarrollar pesticidas comerciales (Salgado 2017).

Para profundizar en el conocimiento del origen y evolución de todas estas familias del SQ en artrópodos, y más concretamente en quelicerados, en esta tesis hemos secuenciado el transcriptoma de los principales apéndices quimiosensoriales de una especie no modelo de arácnido, *Macrothele calpeiana* (Chelicerata, Araneae, Macrothelidae).

6. Desarrollo de herramientas bioinformáticas para desarrollar nuevos marcadores moleculares

Una de las aplicaciones importantes de las tecnologías de NGS en biología evolutiva y sistemática es la de obtener y desarrollar marcadores con objeto de ser aplicados tanto para determinar los patrones e inferir los procesos evolutivos subyacentes al origen y mantenimiento de la diversidad biológica.

Tanto la filogenética como la filogeografía, aunque pueden diferir en los objetivos específicos de estudio, comparten metodologías y herramientas para analizar la biodiversidad. Durante la última década, el uso de datos de NGS ha revolucionado los estudios en el campo de la biología evolutiva y la sistemática al proporcionar una gran cantidad de datos moleculares, de organismos no modelo. Estos avances han mejorado la capacidad de identificación de linajes nuevos y delimitación de especies crípticas, así como su contextualización filogenética, información básica para avanzar en la investigación de otras disciplinas afines como la ecología, conservación, biogeografía y evolución.

6.1 Antecedentes de la biología evolutiva contemporánea

En 1859, en su publicación "El Origen de las especies" (Darwin et al. 1859), Darwin describe la primera teoría que describe un mecanismo sencillo para explicar cómo se genera la biodiversidad en la naturaleza y propone que los seres vivos son entidades que sufren modificaciones que son transmitidas a sus descendientes. A partir de entonces, se empieza a emplear el término de "Árbol de la Vida" para agrupar y representar las relaciones de entre los organismos en función de su parentesco evolutivo. No obstante, hasta casi un siglo después, no se desarrolla la primera metodología para realizar inferencias filogenéticas reproducibles de forma rigurosa y sistemática (Hennig 1966). El entomólogo alemán Willi Hennig, propone ese año el uso de homologías derivadas y compartidas entre organismos como base para establecer sus relaciones evolutivas, método que se conoce como como Sistemática Cladística o Filogenética.

En paralelo, durante esos años se estaban realizando una serie de importantes hallazgos

que pondrían las bases de la biología molecular moderna. En este sentido, cabe destacar, el trabajo de Hershey y Chase (Hershey y Chase 1952), que confirmó que el DNA era el material hereditario, la caracterización de la estructura de la molécula del DNA (Watson y Crick 1953b), y, sobre todo, la publicación del "Dogma central de la biología molecular" que describe los mecanismos moleculares involucrados en la transmisión de la información del DNA al RNA y de su traducción a proteínas. Estos y otros descubrimientos, como la secuenciación de proteínas, culminaría con el desarrollo de lo que se conoce como el concepto del reloj molecular (Zuckerkandl y Pauling 1964). Tras comparar la secuencia de la misma proteína en diferentes especies, Zuckerandl y Pauling constataron que el número de diferencias era aproximadamente proporcional al tiempo transcurrido desde que las especies habían compartido el último ancestro común. Por lo tanto, si aceptamos que las proteínas tienen una tasa de sustitución de aminoácidos constante a lo largo del tiempo, podemos inferir el tiempo de divergencia de dos especies en unidades de número de sustituciones. Si además, disponemos de alguna información independiente (por ejemplo de un fósil o de un evento geológico) que nos permita establecer a qué tiempo se corresponde un determinado número de sustituciones de aminoácido en una proteína determinada, podremos estimar ese tiempo de divergencia en tiempo de reloj (ej. Ma).

Las relaciones filogenéticas se representan gráficamente en forma de árbol, donde los nodos representan los taxones y las ramas definen las relaciones entre estos taxones. Los nodos terminales representan los taxones actuales o "OTUs" (operational taxonomic units), mientras que los nodos internos representan los diferentes ancestros de estos. Según el tipo de árbol filogenético, se puede mostrar sólo la topología (cladograma; muestra únicamente las relaciones de parentesco evolutivo entre los taxones) o también se puede añadir la longitud de las ramas de forma proporcional al tiempo evolutivo transcurrido desde los taxones ancestros a sus descendientes (filogramas).

Por otro lado, el concepto de filogeografía aparece a finales de los años ochenta del siglo pasado (John C. Avise et al. 1987) y se define como el estudio de los mecanismos, tanto evolutivos como históricos y/o demográficos, que establecen la distribución geográfica de los linajes genealógicos (a nivel poblacional o especies cercanas). Por lo tanto, la filogeografía es una disciplina que integra campos de la biología evolutiva a nivel macroevolutivo, como la filogenética, la biogeografía, la geoclimatología y la paleontología, y microevolutivo, como la evolución molecular, la demografía y en especial la genética de poblaciones (John C. Avise 1994, 2000).

Inicialmente, la inferencia de las relaciones filogenéticas se realizó a partir de la comparación de caracteres morfológicos (Hennig 1950). No obstante, a partir de los años 60 del siglo XX, se empieza a utilizar la información derivada de los análisis cromosómicos (citogenéticos), bioquímicos (como los cambios de movilidad electroforética de proteínas como las isoenzimas), de las secuencias aminoacídicas y finalmente las secuencias de los ácidos nucleicos (DNA/RNA) (John C. Avise 1994). Debido a su desarrollo posterior, los estudios de filogeografía estuvieron desde su inicio basados en marcadores moleculares, fundamentalmente en las técnicas de digestión con enzimas de restricción como los RFLPs (fragmentos de restricción de longitud polimórfica) (Botstein et al. 1980) y en secuencias del DNA mitocondrial (Moritz, Dowling, y Brown 1987) y ya posteriormente en marcadores nucleares como los AFLPs, los microsatélites y secuenciación directa de la secuencia de DNA de genes o fragmentos de genes concretos.

6.2 Obtención de marcadores moleculares mediante técnicas de secuenciación dirigida Sanger

A partir de 1990 se desarrollaron una gran diversidad de técnicas para obtener marcadores moleculares de DNA a través del uso de la reacción de PCR (Schlötterer 2004). Un marcador molecular de DNA puede ser un gen o un fragmento de DNA, heredable que presenta variabilidad entre individuos (polimorfismos) o entre especies (divergencia) y se puede identificar en una región específica del genoma nuclear o mitocondrial (Hedin 2001; Satish Kumar et al. 2009). La mayor limitación de estos marcadores estaba relacionada con el gran coste que suponía obtener un número suficiente de marcadores para resolver múltiples cuestiones evolutivas. Además, la mayoría de las metodologías implican un conocimiento previo de las secuencias genómicas de los taxones de interés para poder desarrollar cebadores y/o descartar loci parálogos (Brito y Edwards 2009). A continuación, comentaremos brevemente los marcadores moleculares obtenidos mediante la técnica de Sanger más populares en estudios evolutivos:

- **Microsatélites de DNA**: son regiones repetitivas del genoma (de entre 2-6 nucleótidos repetidas en tándem con una longitud de 100 pb). Se obtienen a través de amplificación por PCR y por lo tanto es necesario el uso de cebadores. Se suelen utilizar en estudios a escala poblacional ya que suelen ser altamente polimórficos (Sauné et al. 2015). Algunas de las limitaciones principales de estos marcadores son la facilidad de obtener errores durante la amplificación por PCR y la complejidad de desarrollar modelos adecuados de mutación (fórmula que estima la probabilidad de cambio de una secuencia) (Dieringer y Schlötterer 2003).

- **Polimorfismos de longitud de fragmentos amplificados (AFLPs):** son fragmentos de DNA polimórficos que se pueden obtener sin tener información previa de genoma. Esta técnica se basa en amplificar por PCR fragmentos de DNA obtenidos tras realizar una digestión completa del genoma con enzimas de restricción. La amplificación del fragmento digerido se lleva a cabo mediante la ligación de adaptadores de secuencia conocida a los extremos de los fragmentos (Williams et al. 1990). Se suelen utilizar para estudios de estructura poblacional. La mayor limitación es la asignación errónea de homología de los fragmentos y la baja reproducibilidad (Schierwater y Ender 1993).

- **Marcadores derivados de la secuencia de DNA mitocondrial**: el genoma mitocondrial es un genoma haploide circular que mide ~16.000 pb en metazoos. Se caracteriza por tener una tasa de evolución alta (entre 5 y 10 veces superior a los genes nucleares), carece de intrones y de recombinación y es de herencia uniparental (línea materna). Los genes mitocondriales más utilizados como marcadores en estudios de biogeografía y en delimitación de especies son el Citocromo c Oxidasa I (COI o *cox1*), el NADH deshidrogenasa 1 (NAD1) y el 16S rRNA (Gillespie, Croom, y Palumbi 1994; Johannesen et al. 2002). También se utilizan para estudios filogenéticos de nivel taxonómico medio y bajo, ya que a escalas muy profundas pueden aparecer posiciones saturadas debido a su rápida evolución (Brewer et al. 2013). Por otro lado, al ser de herencia materna, en casos de hibridación, obtendremos inferencias erróneas al obtener solo la historia evolutiva de las hembras (Rubinoff y Holland 2005). Por estos motivos, en los estudios filogenéticos y filogeográficos se empezaron a incluir información de diversos loci no ligados del genoma nuclear (Brito y Edwards 2009).

- **DNA nuclear**: son marcadores obtenidos a través de amplificar regiones del genoma nuclear mediante PCR, la cual implica el diseño de cebadores específicos. En general, el DNA nuclear presenta unas tasas de mutación inferiores a las del ADN mitocondrial. Tradicionalmente, los genes nucleares más utilizados en reconstrucciones filogenéticas son los genes de las histonas (H3) y los genes ribosomales (18S y 28S) (Giribet y Edgecombe 2013). A nivel poblacional se suelen utilizar intrones o regiones no codificantes ya que la tasa evolutiva es mayor y no están sujetas a presiones selectivas. Para resolver filogenias poco profundas, las regiones ITS (*internal transcribed spacer*) ofrecen suficiente variabilidad, aunque a veces los productos de eventos de paralogía y recombinación son difíciles de identificar (Hormiga, Arnedo, y Gillespie 2003).

6.3 Obtención de marcadores moleculares a través de metodologías NGS

La mayoría de los problemas asociados al tipo de marcador se pueden solventar combinando marcadores de diferentes procedencias, por ej. combinando genes nucleares con mitocondriales y aplicando el modelo más apropiado para analizar los datos. Sin embargo, como hemos comentado previamente, la principal limitación de los marcadores secuenciados mediante la tecnología de Sanger era la dificultad de obtener un número suficiente de marcadores para poder abordar estudios filogenéticos y filogeográficos con rigor y precisión (Brito y Edwards 2009).

Sin embargo, la llegada de las tecnologías de NGS han facilitado la identificación de un gran número de marcadores de DNA distribuidos aleatoriamente por el genoma, haciendo más asequible el uso de datos multi-locus (múltiples loci concatenados) en los análisis. Asimismo, se han desarrollado algoritmos que implementan diferentes modelos evolutivos y permiten particionar el análisis en función de la tasa evolutiva de los diferentes loci (o incluso diferentes regiones de un mismo gen). Pero dado que obtener un genoma completo sigue siendo un proceso caro y computacionalmente complejo, se han desarrollado metodologías que permiten aislar y secuenciar un número reducido de regiones concretas del genoma. Estas aproximaciones se conocen como técnicas de partición genómica (Mamanova et al. 2010; Turner et al. 2009) y permiten la amplificación de regiones homólogas entre individuos mediante el uso de sondas o enzimas de restricción que "capturan" las regiones de DNA. Estas técnicas han favorecido el desarrollo de marcadores moleculares universales para un amplio rango taxonómico ya que al reducir el volumen a secuenciar por individuo y utilizar adaptadores para identificarlos, es posible aumentar el número de taxones en el estudio (Zavodna, Grueber, y Gemmell 2013). Además, las tecnologías de NGS han permitido el uso de organismos no modelo e individuos de poblaciones naturales, ya que en general no es necesario disponer de información genómica previamente. Las metodologías más populares de reducción genómica para aplicaciones en estudios evolutivos son las siguientes:

- **Secuenciación del transcriptoma (RNAseq)**: esta técnica se basa en obtener sólo las secuencias de las regiones codificantes, la cual suele representa aprox. un 1.5 % del total del genoma, dependiendo de la especie. Estos datos son apropiados para detectar genes con selección, aumentar la potencia de estudios de estructura poblacional (Roulin et al. 2012) o para análisis filogenéticos (Fernández et al. 2018). Algunas de las limitaciones están relacionadas con la dificultad de identificar genes parálogos e isoformas.

- **Secuenciación dirigida de amplicones (TAS) con el uso de cebadores específicos:** esta metodología consta de dos pasos de amplificación por PCR. En el primer paso mediante el uso de cebadores específicos se amplifica una región génica dirigida (amplicón) y en la segunda PCR se ligan unos adaptadores, secuencias conocidas de 10 pb, conocidos como *MID* o *barcodes* (código de barras), que permitirán identificar los diferentes individuos secuenciados (Bybee et al. 2011). La técnica más popular consiste en amplificar múltiples loci de varios individuos utilizando una combinación de cebadores. Sin embargo, si trabajamos con rangos taxonómicos muy lejanos, pueden aparecer mutaciones en la región diana del cebador o pueden variar las condiciones de la PCR como la temperatura (Edwards y Gibbs 1994).

- **Enriquecimiento híbrido con el uso de sondas:** el enriquecimiento híbrido (o captura de secuencia) se basa en capturar regiones concretas de DNA o RNA a través del uso de sondas. En un principio, se utilizaban soportes físicos multipocillos, como en los microarrays, pero la aproximación más popular es la hibridación en fase líquida a través de sondas marcadas con biotina. Las sondas de captura, son secuencias de oligonucleótidos de una longitud comprendida entre 60-120 pb, que "capturan" las regiones del genoma a secuenciar, ya que la librería de secuenciación se prepara con las regiones que se hibridan con la sonda (Gnirke et al. 2009). Las técnicas más populares para estudios filogenéticos son el enriquecimiento anclado (AE) (A. R. Lemmon, Emme, y Lemmon 2012) y el enriquecimiento de elementos ultraconservados (UCE) (Faircloth et al. 2012). Estas dos técnicas son muy similares, implican el uso de sondas de regiones conservadas para amplificar regiones muy variables. Para diseñar las sondas es necesario identificar las regiones conservadas a través de comparar los genomas de las especies más distantes incluidas en el análisis. De esta forma, se aumenta la probabilidad de que estas regiones estén conservadas también en el resto de las especies del estudio (organismos de los nodos internos).

- **Preparación de librerías reducidas mediante el uso de enzimas de restricción:** la base de esta aproximación es el uso de enzimas de restricción para digerir el genoma y seleccionar los fragmentos digeridos en base a una longitud determinada para su posterior secuenciación (reduced-representation library - RRL) (Altshuler et al. 2000). A partir de esta metodología se han desarrollado diversas técnicas, las más populares son: la secuenciación de ADN asociada al sitio de restricción (*restriction-site-Associated DNA sequencing* - RAD-seq) (J. W. Davey et al. 2010) y la genotipificación por secuenciación (*Genotyping by sequencing* – GBS) (Elshire et al. 2011). Dependiendo de la técnica utilizada podemos aplicar estos datos a diferentes estudios (Kadlec et al. 2017), sobretodo en estudios de genética de poblaciones y filogeográficos (Emerson et al. 2010).

No obstante, como hemos visto, no todos los marcadores moleculares son apropiados para resolver cualquier cuestión filogenética o filogeográfica (Figura 11). Por este motivo, es muy importante la selección de la técnica de reducción genómica en base a la aplicación posterior que queremos realizar. Además, una vez obtenidos los datos de NGS, debemos aplicar rigurosos análisis bioinformáticos para la correcta selección de los marcadores de interés. Al utilizar grandes conjuntos de datos reduciremos el error de muestreo, pero en presencia de sesgos sistemáticos, también podríamos obtener respuestas incorrectas con un fuerte apoyo estadístico (S. Kumar et al. 2012). Por estos motivos, para obtener marcadores informativos debemos tener en cuenta los siguientes pasos bioinformáticos:

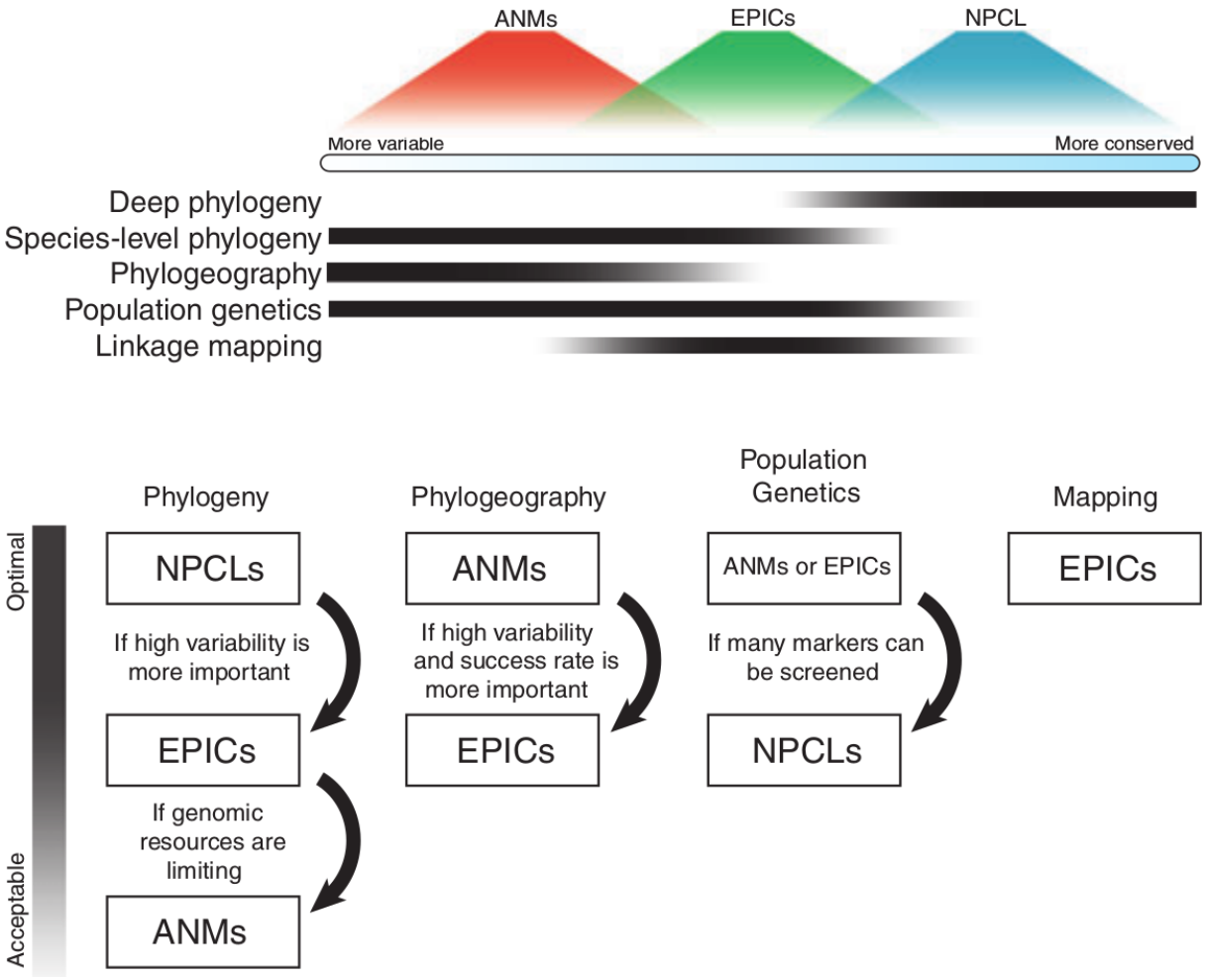


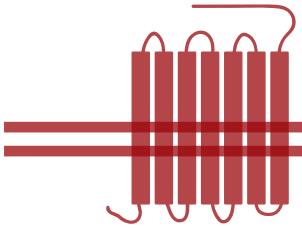
Figura 11. (A) Esquema de la variabilidad genética relativa entre las clases de marcadores y la cantidad de variación generalmente necesaria para diferentes tipos de preguntas de la biología evolutiva y la sistemática. (B) Resumen de las clases de marcadores, según su procedencia del genoma, y su nivel de eficiencia para resolver cuestiones concretas. Fuente: (Thomsom, Wang, y Jonhson 2010).

- **Procesamiento de las secuencias:** en esta fase además de realizar los análisis de control de la calidad de los *reads* descritos en la sección 3, es necesario eliminar posiciones variables (polimórficas) con baja cobertura (para evitar falsos positivos generados por errores de secuenciación), seleccionar los loci de copia única (ortólogos) y aquellos que estén presentes en la mayoría de los individuos incluidos en la preparación de la librería (marcadores universales).

- **Selección de los marcadores en función de la cuestión evolutiva a responder y según la clase de marcador molecular obtenido:** dado que a través de las tecnologías de NGS obtenemos *a priori* un gran número de marcadores con características diferentes, es apropiado seleccionar aquellos que mejor respondan nuestra cuestión de interés. Dependiendo del análisis podemos diferenciar entre las propiedades inherentes de las regiones codificadoras y las no codificadoras, como puede ser su tasa de mutación, si está sometido a presiones selectivas, nivel de variabilidad, etc. Las regiones codificadoras (nuclear protein coding loci - NPCL) son más apropiadas para estudios de genética de poblaciones (especiación y divergencia, extinción, introgresión, etc.) y para resolver

filogenias distantes, ya que suelen estar más conservadas que las regiones no codificantes. Por otro lado, podemos obtener marcadores moleculares que provienen de regiones no codificadoras como los EPICs (*Exon- primed Intron-crossing Markers*) (Backström, Fagerberg, y Ellegren 2007) y los ANMs (Anonymous Nuclear Markers) (Jennings y Edwards 2005). La técnica EPICs se basa en desarrollar cebadores de las regiones exónicas para amplificar los intrones, los cuales suelen ser regiones que acumulan más variabilidad que las regiones codificadoras. En cambio los ANMs son marcadores que se generan por azar a través de digestiones enzimáticas, y por lo tanto la gran mayoría son regiones no-codificantes (E. M. Lemmon y Lemmon 2013; McCormack et al. 2013), pero es necesario confirmar si son o no codificantes a través de procesos informáticos. Para aplicaciones filogenéticas el uso de los ANMs se ha hecho más popular, ya que en principio son regiones que con mayor probabilidad pueden presentar una evolución neutra y reflejar de una forma más precisa el tiempo de divergencia entre los taxones.

Por lo tanto, dependiendo del rango taxonómico y del tipo de estudio que queremos resolver es necesario el desarrollo de *scripts ad hoc* para seleccionar los marcadores más apropiados. En esta tesis, en el capítulo 3, presentamos una herramienta bioinformática desarrollada para identificar y seleccionar los marcadores moleculares más informativos para responder las preguntas evolutivas concretas que nos estemos planteando, tanto para aplicaciones filogenéticas como para estudios de genética de poblaciones y/o filogeografía.



OBJETIVOS

Objetivos

Comprender la base molecular de los mecanismos implicados en la generación y mantenimiento de la biodiversidad es una de las grandes cuestiones de la biología evolutiva. El Sistema Quimiosensorial (SQ) al ser el responsable de percibir las señales químicas del exterior, está íntimamente relacionado con la supervivencia, viabilidad y reproducción de los individuos. Por lo tanto, es un buen modelo para identificar los mecanismos evolutivos implicados con las adaptaciones a nuevos hábitats e identificar la huella molecular de la selección natural. Por otro lado, para poder inferir las relaciones filogenéticas y caracterizar los patrones de variabilidad bajo un enfoque filogeográfico (estructuras) es necesario el desarrollo de estrategias para la generación de marcadores moleculares informativos para responder las cuestiones biológicas de estudio. En la era "ómica" las principales limitaciones ya no consisten en obtener suficientes marcadores, sino en la capacidad de seleccionar los más adecuados y en la disponibilidad de herramientas bioinformáticas eficientes. En este contexto, hoy en día existen pocos protocolos informáticos estandarizados para poder procesar y analizar los datos.

En esta tesis, hemos abordado estas dos cuestiones mediante el uso de datos de NGS. Los objetivos específicos han sido:

1. Identificar y caracterizar los genes del SQ en artrópodos a partir de datos de RNAseq de un quelicerado, *Macrothele calpeiana*, con el fin de inferir el origen y evolución del SQ.

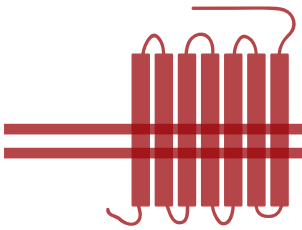
2. Desarrollar una herramienta bioinformática para analizar y procesar datos de RNAseq fundamentalmente en organismos no modelo.

3. Diseñar una herramienta bioinformática para generar marcadores moleculares para su uso en filogenética y genética de poblaciones:

3.1 Desarrollar un método que permita identificar nuevos marcadores para su aplicación mediante amplificación por PCR, o para diseñar sondas de captura.

3.2 Implementar métodos que permitan seleccionar marcadores informativos a partir de marcadores existentes, o de ficheros con alineamientos múltiples de secuencias

3.3 Validación experimental de la metodología implementada en la herramienta bioinformática mediante técnicas de reducción genómica



INFORME DE LOS DIRECTORES

UNIVERSITAT DE BARCELONA

Dr. Julio Rozas
Catedràtic de Genètica

Departament de Genètica,
Microbiologia i Estadística
Facultat de Biologia

Diagonal 643
Edifici Prevosti
08028 Barcelona
Spain

Tel. +34 93 4021495
Fax. +34 93 4034420
jrozas@ub.edu
www.ub.edu/molevol/julio

Informe signat del director de tesi del factor d'impacte dels articles publicats. En cas que es presenti algun treball en coautoria, caldrà incloure també un informe del director de la tesi signat, en què s'especifiqui exhaustivament quina ha estat la participació del doctorant/a en cada article, i si algun dels coautors d'algun dels treballs presentats en la tesi doctoral ha utilitzat, implícitament o explícitament aquests treballs per a la l'elaboració de la tesi doctoral

---

El Drs. **Julio Rozas i Miquel A. Arnedo,** directors de la Tesi Doctoral elaborada pel Sra. Cristina Frías López, amb el títol "**Desarrollo de técnicas bioinformáticas para el análisis de datos de secuenciación masiva en sistemática y genómica evolutiva: Aplicación en el análisis del sistema quimiosensorial en artrópodos**"

INFORMEN

Que la tesi doctoral està elaborada com a compendi de 3 publicacions amb dades originals (publicacions 1-3 en el cos central de la tesi), i 6 més (publicacions 4-9) a l'apèndix:

Publicacions (cos central de la tesi):

1. Frías-López, C., Almeida, F. C., Guirao-Rico, S., Vizueta, J., Sánchez-Gracia, A., Arnedo, M. A., Rozas, J. 2015. Comparative analysis of tissue-specific transcriptomes in the funnel-web spider *Macrothele calpeiana* (Araneae, Hexathelidae). *PeerJ* **3:** e1064.
   Factor d'impacte (5 Year Impact Factor): **[IF = 2.183; Q1].** Categoria de Multidisciplinary Sciences.

2. Frías-López, C.*, Sánchez-Herrero, J. F.*, Guirao-Rico, S., Mora, E., Arnedo, M. A., Sánchez-Gracia, A. and Rozas, J. 2016. DOMINO: Development of informative molecular markers for phylogenetic and genome-wide population genetic studies in non-model organisms. *Bioinformatics* **32:** 3753-3759.
   Factor d'impacte (5 Year Impact Factor): **[IF = 8.044; Q1 i D1].** Categoria de Mathematical and Computational Biology.
   *, la mateixa contribució

3. Kornobis, E., Cabellos, L., Aguilar, F., Frías-López, C., Rozas, J., Marco, J. and Zardoya, R. 2015. TRUFA: a USer-friendly Web server for de novo RNA-seq analysis using cluster computing. *Evolutionary Bioinformatics.* **11:** 97-104.
   Factor d'impacte (5 Year Impact Factor): **[IF = 1.580].** Categoria de Mathematical and Computational Biology.

UNIVERSITAT DE BARCELONA

Dr. Julio Rozas
Catedràtic de Genètica

Departament de Genètica,
Microbiologia i Estadística
Facultat de Biologia

Diagonal 643
Edifici Prevosti
08028 Barcelona
Spain

Tel. +34 93 4021495
Fax. +34 93 4034420
jrozas@ub.edu
www.ub.edu/molevol/julio

A la publicació 1, la doctoranda va fer la part més important del treball experimental, computacional i d'anàlisi de dades, i va redactar el primer esborrany dels manuscrits. A la publicació 2 (full article a la revista *Bioinformatics*), que també es presenta com a part de la tesi doctoral de José F. Sánchez-Herrero, va dissenyar els primers scripts que integren la *pipeline* del software DOMINO, la validació experimental (amb dades de seqüenciació de DNA de 4 especies, amb la plataforma 454), i va contribuir a la redacció del primer esborrany del manuscrit. A la publicació 3 (col·laboració amb membres del grup de recerca de R. Zardoya y J. Marco) va participar en el disseny de la *pipeline* del software TRUFA i va testar la integració dels diferents softwares en la seva GUI.

Publicacions en el apèndix:
Aquestes publicacions son resultats de col·laboracions científiques on el doctorant, fent servir eines computacionals o analítiques desenvolupats en la seva tesi doctoral, ha realitzat una part dels anàlisis experimentals o bioinformàtics.

4. Solà, E., Alvarez-Presas, M., Frías-López, C., Littlewood, T. J., Rozas, J., Riutort, M. 2015. Evolutionary Analysis of Mitogenomes from Parasitic and Free-Living Flatworms.
   PLoS ONE **10:** e0120081.
   Factor d'impacte (5 Year Impact Factor): **[IF = 3.535; Q1].** Categoria de Multidisciplinary Sciences.

5. Vizueta, J., Frías-López, C., Macías-Hernández, N., Arnedo, M. A., Sánchez-Gracia, A. and Rozas, J. 2017. Evolution of chemosensory gene families in arthropods: Insight from the first inclusive comparative transcriptome analysis across spider appendages.
   Genome Biol. Evol. **9:** 178-196.
   Factor d'impacte (5 Year Impact Factor): **[IF = 4.171; Q1].** Categoria de Genetics & Heredity.

6. Sousa, P., Frías-López, C., Harris, D. J., Rozas, J. & Arnedo, M. A.: Development of anonymous nuclear markers for Buthus scorpions (Scorpiones: Buthidae) using massive parallel sequencing, with an overview of nuclear markers used in Scorpions phylogenetics.
   Preparat per enviar a publicar.

7. Cerda-Mejía, L., Valenzuela, S. V., Frías, C., Diaz, P., & Pastor, F. J. 2017. A bacterial GH6 cellobiohydrolase with a novel modular structure.
   Applied microbiology and biotechnology, **101:** 2943-2952.
   Factor d'impacte (5 Year Impact Factor): **[IF = 3.602].** Categoria de Biotechnology & Applied Microbiology.

8. Crespo, L. C., Domènech, M., Enguídanos, A., Malumbres-Olarte, J., Cardoso, P., Moya-Laraño, Frías-López, C., Macías-Hernández, N., De Mas, E., Mazzuca, P., Mora, E., Opatova, V., Planas, E., Ribera, C., Roca-Cusachs, M., Ruiz, D., Sousa, P., Tonzo, V. & Arnedo, M. A. 2018. A DNA barcode-assisted annotated checklist of the spider (Arachnida, Araneae) communities associated to white oak woodlands in Spanish National Parks.
   Biodiversity data journal **6:** e29443.
   Factor d'impacte: **[IF = 1.029].** Categoria de Biodiversity Conservation.

UNIVERSITAT DE BARCELONA

Dr. Julio Rozas
Catedràtic de Genètica

Departament de Genètica,
Microbiologia i Estadística

Facultat de Biologia

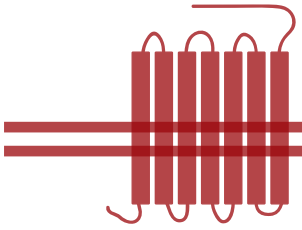
Diagonal 643
Edifici Prevosti
08028 Barcelona
Spain

Tel. +34 93 4021495
Fax. +34 93 4034420
jrozas@ub.edu
www.ub.edu/molevol/julio

9. Sánchez-Herrero, J. F., Frías-López, C., Escuer, P., Hinojosa-Alvarez, S., Arnedo, M. A., Sánchez-Gracia, A. and Rozas, J. 2019. The draft genome sequence of the spider *Dysdera silvatica* (Araneae, Dysderidae): A valuable resource for functional and evolutionary genomic studies in chelicerates.
GigaScience **In press** doi: 10.1093/gigascience/giz099.
Factor d'impacte (5 Year Impact Factor): **[IF = 7.441; Q1].** Categoria de Multidisciplinary Sciences.

Dr. Julio Rozas
Catedràtic de Genètica
Universitat de Barcelona

Miquel A. Arnedo
Catedràtic de Zoologia
Universitat de Barcelona



PUBLICACIONES

Artículo 1

Comparative analysis of tissue-specific transcriptomes in the funnel-web spider *Macrothele calpeiana* (Araneae, Hexathelidae)

Cristina Frías-López, Francisca C. Almeida, Sara Guirao-Rico, Joel Vizueta, Alejandro Sánchez-Gracia, Miquel A. Arnedo and Julio Rozas

2015, PeerJ, 3: e1064.

PeerJ

Comparative analysis of tissue-specific transcriptomes in the funnel-web spider *Macrothele calpeiana* (Araneae, Hexathelidae)

Cristina Frías-López[1,2], Francisca C. Almeida[1,*], Sara Guirao-Rico[1,**], Joel Vizueta[1], Alejandro Sánchez-Gracia[1], Miquel A. Arnedo[2] and Julio Rozas[1]

[1] Departament de Genètica and Institut de Recerca de la Biodiversitat (IRBio), Universitat de Barcelona, Barcelona, Spain
[2] Departament de Biologia Animal and Institut de Recerca de la Biodiversitat (IRBio), Universitat de Barcelona, Barcelona, Spain
[*] Current affiliation: Consejo Nacional de Investigaciones Científicas y Tecnológicas (CONICET), Departamento de Ecología, Genética y Evolución, Universidad de Buenos Aires, Intendente Güiraldes y Costanera Norte s/n, Pabellón II—Ciudad Universitaria, Capital Federal, Argentina
[**] Current affiliation: Centre for Research in Agricultural Genomics (CRAG) CSIC-IRTA-UAB-UB, Barcelona, Spain

ABSTRACT

The funnel-web spider *Macrothele calpeiana* is a charismatic Mygalomorph with a great interest in basic, applied and translational research. Nevertheless, current scarcity of genomic and transcriptomic data of this species clearly limits the research in this non-model organism. To overcome this limitation, we launched the first tissue-specific enriched RNA-seq analysis in this species using a subtractive hybridization approach, with two main objectives, to characterize the specific transcriptome of the putative chemosensory appendages (palps and first pair of legs), and to provide a new set of DNA markers for further phylogenetic studies. We have characterized the set of transcripts specifically expressed in putative chemosensory tissues of this species, much of them showing features shared by chemosensory system genes. Among specific candidates, we have identified some members of the iGluR and NPC2 families. Moreover, we have demonstrated the utility of these newly generated data as molecular markers by inferring the phylogenetic position *M. calpeina* in the phylogenetic tree of Mygalomorphs. Our results provide novel resources for researchers interested in spider molecular biology and systematics, which can help to expand our knowledge on the evolutionary processes underlying fundamental biological questions, as species invasion or biodiversity origin and maintenance.

Subjects Evolutionary Studies, Genetics, Genomics, Zoology
Keywords *De novo* transcriptome assembly, Molecular markers, Chemosensory system, RNA-seq, Mygalomorphae Phylogeny

Submitted 19 May 2015
Accepted 9 June 2015
Published 30 June 2015

Corresponding author
Julio Rozas, jrozas@ub.edu

Academic editor
Kimberly Bishop-Lilly

Additional Information and Declarations can be found on page 17

DOI **10.7717/peerj.1064**

INTRODUCTION

The funnel-web spider *Macrothele calpeiana* (family *Hexathelidae*) is a charismatic component of the European arthropod fauna. It belongs to the spider infraorder

How to cite this article Frías-López et al. (2015), Comparative analysis of tissue-specific transcriptomes in the funnel-web spider *Macrothele calpeiana* (Araneae, Hexathelidae). **PeerJ 3:e1064**; **DOI 10.7717/peerj.1064**

Mygalomorphae, which includes about 3,000 species of, among others, trap-door spiders, funnel-web spiders, and tarantulas (*Platnick, 2006*). *M. calpeiana* is a hairy, large spider that constructs extended and conspicuous funnel-web sheets close to the ground, and it is the only spider protected under European legislation (*Collins & Wells, 1987*). This spider is endemic to the southern Iberian Peninsula and was initially considered to be particularly vulnerable due to its close association with the highly threatened cork-oak forests found in the region (*Collins & Wells, 1987*). Subsequent studies, however, demonstrated that the species has a much wider distribution and could be frequently found in highly disturbed areas. In the last years, *M. calpeiana* has been introduced in European countries outside its natural range, probably associated with the commercial export of Spanish olive trees, raising some concerns about their possible impact on the invaded ecosystems (*Jiménez-Valverde, Decae & Arnedo, 2011*).

M. calpeiana is also an organism of particular interest in biogeographic studies. The *Macrothele* genus shows a highly disjointed distribution, with the bulk of its diversity in South-East Asia (21 species), a few species inhabiting tropical Africa (4 species) and only two known species in Europe, *M. calpeiana* itself and *M. cretica*, a Cretan endemic spider that is also of conservation concern. A recent phylogenetic study (*Opatova & Arnedo, 2014*) has revealed that the two European *Macrothele* species are not sister taxa, and that they most likely colonized independently Europe from Asia. Another interest in the genus relates to the venom toxins of some *Macrothele* spiders, which can be strong enough to cause envenomation, as in the case of some large Taiwanese *Macrothele* spiders (*Hung & Wang, 2004*). In fact, studies on the molecular structure and chemical properties of venom toxins (*Zeng, Xiao & Liang, 2003*; *Corzo et al., 2003*; *Satake et al., 2004*; *Yamaji et al., 2009*) have established the utility of *Macrothele* venom as cell growth inhibitors in cancer research (*Gao et al., 2005*; *Liu et al., 2012*).

The scarcity of genomic and transcriptomic data in chelicerates, which just cover a few species (*Grbić et al., 2011*; *Mattila et al., 2012*; *Cao et al., 2013*; *Clarke et al., 2014*; *Sanggaard et al., 2014*; *Posnien et al., 2014*) and the lack of tissue-specific transcript data in mygalomorphs, clearly limit the research on the molecular determinants of fundamental biological processes in this group of species. Within this context, with the aim of shedding light on the composition of Mygalomorph transcriptomes, we conducted the first RNA-seq study in one species of this group, *M. calpeiana*, including several tissues, and using a 454GS-FLX-based technology (*Prosdocimi et al., 2011*). The new sequence data will be an important, initial contribution to further basic, applied, and translational research in this non-model organism. Here we address two specific objectives: (i) to identify possible candidate chemosensory transcripts for future studies, and (ii) to provide new markers for further phylogenetic and evolutionary genomic-based studies in this group. As an example, we used some of the new generated transcripts to clarify the phylogenetic position of *M. calpeiana* in the Mygalomorph phylogeny.

The chemosensory system plays a key role in fundamental vital processes, including the localization of food, hosts, or predators and social communication; nevertheless, there are very few studies focused in non-insect species results (*Vieira & Rozas, 2011*;

Montagné et al., 2015), and almost unknown in mygalomorphs. In insects, the main molecular components of the chemosensory system are encoded by two main groups of gene families (*Sánchez-Gracia, Vieira & Rozas, 2009*; *Vieira & Rozas, 2011*; *Almeida et al., 2014*) the chemoreceptors and the secreted ligand-binding proteins. The first include the gustatory (GR), olfactory (OR), and ionotropic (IR) receptors, while the second group, known as ligand-binding families, are the odorant-binding protein (OBP), chemosensory protein (CSP), chemosensory type A and B (CheA/B), and probably some members of the Niemann-Pick disease type C2-related (NPC2) family (*Pelosi et al., 2014*). The preliminary analyses of the genomic sequences of the chelicerates *I. scapularis* (M Gulia-Nuss et al., 2015, unpublished data), *Stegodyphus mimosarum, Acanthoscurria geniculata*, (*Sanggaard et al., 2014*), *Mesobuthus martensii* (*Cao et al., 2013*), and *Tetranychus urticae* (*Grbić et al., 2011*), as well as in other arthropods, like the centipede *Strigamia maritima* (*Chipman et al., 2014*), revealed the absence of the typical insect OR and OBP gene families in these species.

Several experimental studies of chelicerates have identified the presence of specialised chemosensory hairs predominantly in the distal segment of the first pair of legs and in palps (*Foelix, 1970*; *Foelix & Chu-Wang, 1973*; *Kronestedt, 1979*; *Cerveira & Jackson, 2012*). In order to investigate the presence of transcripts related to the chemosensory system in spiders, we sequenced the specific transcriptomes of these two structures in *M. calpeiana*. To enrich our samples in tissue-specific transcripts, we built subtractive normalized cDNA libraries for each of these tissues separately. Additionally, for comparative purposes, we also analysed the ovary RNA-seq data. In this way, this study represents a starting-point to characterize the gene expression in the putative chelicerate chemosensory system structures.

Because of their low vagility and restricted distributions, mygalomorph spiders are well-suited for monitoring the ecological and evolutionary conservation status of terrestrial ecosystems (*Bond et al., 2006*), while at the same time are also highly threatened by habitat destruction (*Harvey, 2002*). To date, however, the lack of informative nuclear markers has limited research on these organisms and has hampered the assessment of their conservation or invasive species status. The method we employed here provides useful data for developing nuclear molecular markers to be used in other evolutionary genomic, phylogenetic, and phylogeographic studies of *Mygalomorphae*.

METHODS

Sample collection and preparation

Four adult females of the spider *Macrothele calpeiana* were collected (Junta de Andalucía, Spain; permission: SGYB-AFR-CMM) in two different localities in the southern Iberian Peninsula, namely Iznalloz (Granada, N37.36468 W3.47183, 1,011 m) (individuals MAC-GR1, MAC-GR2, MAC-GR3) and Finca de los Helechales, rd. Cabeza la Vaca (Huelva, N38.09032 W6.46621, 749 m) (individual CRBAMM000991). For each individual, palps, distal segments of the first pair of legs (denoted as legs), ovaries, brains and muscle tissues (from the rest of legs) were dissected and stabilized in RNA later (Applied Biosystems/Ambion).

Total RNA extraction and cDNA preparation

Each tissue was disrupted and homogenized separately using a rotor-stator homogenizer. Total RNA was extracted with the RNeasy midi kit (Qiagen, Hilden, Germany). For all dissected tissues, except the ovary, the protocol included a proteinase K digestion step in order to digest contaminant proteins. All samples were enriched in poly(A) mRNA prior to library preparation using the Oligotex RNA midi kit (Qiagen, Hilden, Germany).

The purified mRNA was used as a template for synthesizing the first cDNA strand using the SMARter PCR cDNA Synthesis Kit (Clontech, Mountain View, California, USA). In this protocol, a poly(A)-specific primer initiates the first strand synthesis of cDNA, thus selecting for polyadenylated RNA while simultaneously keeping the concentration of ribosomal RNA low. The resulting single stranded cDNA was amplified with the Advantage2 PCR kit (Clontech, Mountain View, California, USA), using 23 (brain, leg and muscle) and 20 (palp and ovary) amplification cycles. Double stranded cDNA was purified using CHROMA SPIN-1000 columns (Clontech, Mountain View, California, USA) and subsequently cleaved with *Rsa1* to generate shorter, blunt-ended cDNA fragments, which are necessary for adaptor ligation and subtraction. The digested cDNA were then purified using a standard phenol:chloroform:isoamyl extraction.

Subtractive hybridization and RNA sequencing

Transcripts expressed specifically in the palps, legs, and ovaries were enriched using the PCR-Select cDNA Subtraction Kit (Clontech, Mountain View, California, USA). This technique is based on a method of selective amplification of differently expressed sequences. We used leg, palp, and ovary cDNA as tester (samples of interest) and brain and muscle cDNAs samples as driver (transcripts exclusively for subtraction purposes) samples. According to the kit's protocol, the tester samples are subdivided into two aliquots that receive different adaptors. These aliquots are mixed to driver cDNA (in a higher concentration), denatured, and allowed to reanneal to form double chain cDNA. The process in repeated once, but with the two aliquots of tester cDNA mixed together and some more tester cDNA added. Then a PCR is done in a way that only double chain cDNA formed by fragments with different adaptors at each end will be amplified (i.e., cDNA formed by the hybridization of single chain cDNA from different tester aliquots). In this way, the sample is enriched with cDNA specific to the tester tissue since the tester cDNA that hybridizes with driver cDNA does not get amplified. The subtraction process also normalizes the library so that the frequencies of each unique cDNA became less unequal, increasing the chances of sequencing a large number of unique cDNAs. The subtracted cDNA products were treated with RNase (Qiagen, Hilden, Germany) and purified with QIAquick PCR Purification Kit (Qiagen, Hilden, Germany).

Two micrograms of subtracted cDNA from each tester tissue was prepared for sequencing on a 454/ Roche GS-FLX Titanium sequencer using three different MID tags, one for each tissue. Double-stranded cDNA was nebulized to generate 500-kb fragments and a shotgun library prepared for GS-FLX sequencing as per the manufacturer's instructions (Roche, Basel, Switzerland), which was run on a 1/4 picotitre plate region.

Read processing, handling, and *de novo* transcriptome assembly

We used *sffinfo* script (Roche's Newbler package; 454 SFF Tools) to extract the DNA sequences (FASTA format) and quality scores (FastQ format) independently for each MID tag from the SFF file. We removed adapters and putative contaminant sequences (upon the UniVec database and the *E. coli* genome sequence data) with SeqClean script (http://compbio.dfci.harvard.edu/tgi/software/), with parameters: -v <sequence of adapters> -c 8 -l 40 -x 95 -y 11 -M -L -s <database of contaminant sequences>. We trimmed low-quality bases at the ends of the reads and removed those shorter than 100bp or with a mean quality score (Q) below 20 using the NGS QC Toolkit (*Patel & Jain, 2012*).

First, we conducted a complete *de novo* assembly using all reads from the three tissues altogether in Newbler v2.6 GS (454 life Sciences, Roche Diagnostics) with parameters -urt -cDNA -Denovo -mol 100 -moi 95 -url. Subsequently, we used the contigs and the non-assembled reads (i.e., singletons) from this first step as input for a second assembly round in CAP3 (*Huang, 1999*), with parameters –o 60 –p 95. Redundant transcripts and putative isoforms were removed using cd-hit-est program, to generate a list of unique transcripts (*Fu et al., 2012*). We then used the gsMapper program (included in Newbler package) to map original (after filtering) reads (from the 3 tissues) to the unique transcripts, discarding all reads exhibiting hard clipping (more than 10% of read length) with an in-house Perl script.

Functional annotation

We carried out most of the functional annotation of the assembled transcripts with blast (v. 2.2.29) (*Altschul, 1997*; *Camacho et al., 2009*), Blast2GO (*Conesa et al., 2005*), InterProScan (*Jones et al., 2014*) and TRUFA (*Kornobis et al., 2015*). We first conducted a series of similarity-based searches with blastx (E-value cut-off 10^{-3}) against the NCBI non-redundant (NCBI-nr) database, retrieving the 5 hits with the lowest E-value for each query transcript. We then used Blast2GO and TRUFA to: (i) assign the Gene Ontology (GO) terms to each of these transcripts and determine the involved KEGG pathways (*Kanehisa & Goto, 2000*), (ii) identify particular protein domain structures in the sequenced transcripts using the InterProScan search engine, and (iii) determine which GO terms, InterPro domains, and KEGG pathways were significantly enriched in particular tissues by applying the Fisher's exact test and controlling by the False Discovery Rate (FDR) (*Benjamini & Hochberg, 1995*).

To determine the efficiency of the subtractive approach employed here to enrich samples with tissue specific transcripts, we estimated the fraction of assembled transcripts encoding for putative housekeeping (HK) genes (i.e., transcripts expected to be expressed across different tissues). For the analysis, we considered that a *M. calpeiana* transcript encodes a HK gene if we obtained a significant blastx hit (E-value cut-off 10^{-3}) against a database that includes all HK genes shared between humans (data set from *Eisenberg & Levanon, 2013*) and *Drosophila melanogaster* (data set from *Lam et al., 2012*) (which correspond to the 80% and 94% of the human and *Drosophila* HK genes, respectively; Table S1A).

Furthermore, we also estimated the number of transcripts that encode genes included in the CEG (Cluster of Essential Genes) database (a set of 458 Eukaryotic Orthologous Groups proteins identified by the Core Eukaryotic Genes Mapping Approach, CEGMA) (*Parra, Bradnam & Korf, 2007*; *Parra et al., 2009*). CEG proteins are highly conserved and present in a wide range of eukaryotic organisms, being therefore a good dataset to assess the reliability of our RNA sequencing and transcript annotation. *VennDiagram* R package was used to obtain all graphic representations of the logical relations (http://cran.r-project.org/web/packages/VennDiagram/index.html).

In order to identify putative *M. calpeiana* chemosensory related transcripts, we carried out an additional specific and customized search. We first built a protein database (CheDB) with vertebrate and insect sequences that match against the InterPro protein family signatures associated with chemosensory function (Table S1B). Then, we conducted a blastx search (E-value of 10^{-3}) using the assembled contigs as query against the CheDB database. To minimize the percentage of false positive results, we checked whether the candidate chemosensory transcripts from the blast searches truly encoded the Pfam HMM core profiles corresponding to chemosensory protein domains, using the programs HMMER (*Eddy, 2009*) (E-value of 10^{-3}) and InterProScan. Only *M. calpeiana* transcripts with positive hits in this second search step were unequivocally annotated as putative chemosensory genes. Finally, we also ran an additional tblastn search (E-value of 10^{-3}) of a set of proteins annotated as chemosensory in currently available chelicerate genomes—the common house spider *Parasteatoda tepidariorum* (https://www.hgsc.bcm.edu/arthropods/common-house-spider-genome-project), the social spider *Stegodyphus mimosarum* (*Sanggaard et al., 2014*), the mygalomorph spider *Acanthoscurria geniculate* (*Sanggaard et al., 2014*), the scorpion *Mesobuthus martensii* (*Cao et al., 2013*), and the tick *Ixodes scapularis* (https://www.vectorbase.org/) against *M. calpeiana* transcripts. In this last search, we also included as queries the translated sequences of the transcripts already identified as candidate *M. calpeiana* chemosensory genes in the first searches. In order to exclude spurious homologs caused by short-length false-positive hits, we only considered for further analyses those transcripts whose blast alignments span either at least 2/3 of the total number of amino acids of the query proteins or those covering at least 80% of the transcript length.

Phylogenetic analysis

To determine the utility of the newly sequenced transcripts as markers for molecular phylogenetics, we applied them to study the phylogenetic position of *M. calpeiana* in the tree of Mygalomorphs, a currently unresolved question. As a starting point, we used the phylogenetic analysis reported in *Bond et al. (2014)*. In particular, we first retrieved the amino acid data of all 16 mygalomorph and 3 non-mygalomorph outgroup species (*Stegodyphus*, *Hypochilus* and *Liphistius*) from the matrix d327 (44 taxa; 327 genes; 110,808 amino acid positions). Then, we searched for putative homologs of these 327 genes in *M. calpeiana* transcripts using the blastp program. For this analysis, we obtained the conceptual translation of the transcript sequences (in all six frames) using TransDecoder (version r20140704) as implemented in the Trinity software (*Haas et al., 2013*). We selected

all *Macrothele* translated amino acid sequences that produced a positive blast hit with an E-value $< 10^{-15}$ and with local alignment length >80 amino acids (i.e., in order to maximize the probability of using 1:1 orthologues). Then, we aligned each of these selected translated sequences of *M. calpeiana* with their corresponding homologs in the 19 chosen species (a single multiple sequence alignment, MSA, per gene) using MAFFT (option–merge) (*Katoh & Standley, 2013*). Finally, we concatenated all individual MSA with amino acid data in at least 50% of the species.

We also built family specific MSA with amino acid sequences of NMDA-ionotropic glutamate receptors (NMDA-iGluR) and with members of the Niemann-Pick C disease 2 (NPC2) family, to investigate the phylogenetic relationships between the candidate *M. calpeiana* transcripts and some representatives of these two families in arthropods. We included in these MSA the proteins already annotated in *D. melanogaster* (hexapod), *S. maritima* (myriapod) and *I. scapularis* (chelicerate), as well as the NPC2 genes expressed in *Apis melifera* and *Camponotus japonicus* antenna (*Pelosi et al., 2014*). For iGluR (including IR8a/IR25a proteins) we prepared two different MSA, one for each functional domain. We used HMMER and the Pfam profiles of these two domains (PF01094 "ANF_receptor," and PF00060 "Lig_chan") to identify and trim separately the extracellular amino-terminal and the ligand-gated ion channel domains, which were used to build two separate MSA (and separate trees) with HMMERALIGN.

We conducted all phylogenetic reconstructions by maximum likelihood (ML) using the PROTGAMMAWAG model in the program RAxML version 8 (*Stamatakis, 2014*). We carried out a multiple non-parametric bootstrap analysis (500 bootstrap runs) to obtain node support values.

RESULTS AND DISCUSSION

RNA-seq of *Macrothele calpeiana*

We sequenced a total of 164,111 raw reads across the three tester samples (i.e., leg, palp, and ovary), with a N50 value of 409bp (Table 1). After trimming, cleaning and removing very short reads (less than 100bp), we obtained a final set of 128,816 reads, which was used for further analyses. Our two-step *de novo* assembly strategy (applying Newbler v 2.6, and subsequently CAP3) yielded a total of 3,705 contigs (N50 of 647bp), composed by more than one read, plus 3,560 singletons. After running the cd-hit-est and gsMapper software these contigs clustered into 6,696 unique sequences (i.e., putative *M. calpeiana* individual coding genes), of which 3,467 corresponds to contigs assembled by more than one read (i.e., excluding singletons) (Table 1; Table S2). Table 2 and Table S3 show the distribution of these 6,696 (and also the 3,467) unique sequences across tissues. *M. calpeiana* reads data are available at the Sequence Read Archive (SRA) database under the accession numbers SRA: SRS951615, SRA: SRS951616 and SRA: SRS951618 (Bioproject number: PRJNA285862).

RNA-seq quality and functional annotation

We investigated the quality of our tissue specific transcriptome by a series of similarity-based searches of our transcripts against sequences in the NCBI-nr database. As expected,

Table 1 **Summary of RNA-seq data and assembly.**

Raw number of reads	164,111
N50	409
Reads used in the Newbler assembly[a]	128,818
Assembled reads	122,183
Isotigs (number of singletons)	3,635 (6,614)
N50 (Isotigs)	601
CAP3 assembly	
Contigs (number of singletons >100nuc)	3,705 (3,560)
N50 (Contigs)	647
Unique sequences[b]	
Total number of sequences (transcripts)	6,696
N50	455
Coverage	14.33X
Reads mapped	95,250
Sequences (excluding singletons)	3,467
N50	613
Coverage	22.94X
Reads mapped	90,267

Notes.

[a] Number of reads after trimming, cleaning and excluding short reads.

[b] Number of reads after clustering and mapping filtering.

Table 2 **Summary of RNA-seq data and assembly per tissue.**

	Leg	Palp	Ovary	Total
Driver[a]	Muscle	Muscle	Brain	
Raw number of reads	59,232	54,321	50,558	164,111
N50	404	405	419	409
Reads used in assembly[b]	46,474	41,545	40,799	128,818
N50	362	364	378	368
Unique sequences (transcripts)[c]	2,705	3,798	1,796	6,696
Longest transcript (in nucleotides)	3,053	3,057	4,116	4,116
HK, housekeeping sequences	426	638	328	1,005
CEG sequences	385	547	236	789
Sequences excluding HK-CEG genes	2,139	2,952	1,369	5,390
Sequences with GO annotation	1,147	1,612	816	2,619
Sequences within Interpro	1,464	1,966	988	3,353
Sequences within KEGG	389	509	173	776
Sequences with functional annotation[d]	1,704	2,363	1,152	3,970
Sequences with annotation[e]	2,060	2,915	1,428	4,978

Notes.

[a] Driver of subtractive cDNA library.

[b] Number of reads after trimming, cleaning and excluding short reads.

[c] Considering the total ($n = 6,696$) data set.

[d] GO, Interpro or KEGG hits.

[e] GO, Interpro, KEGG or blast hits.

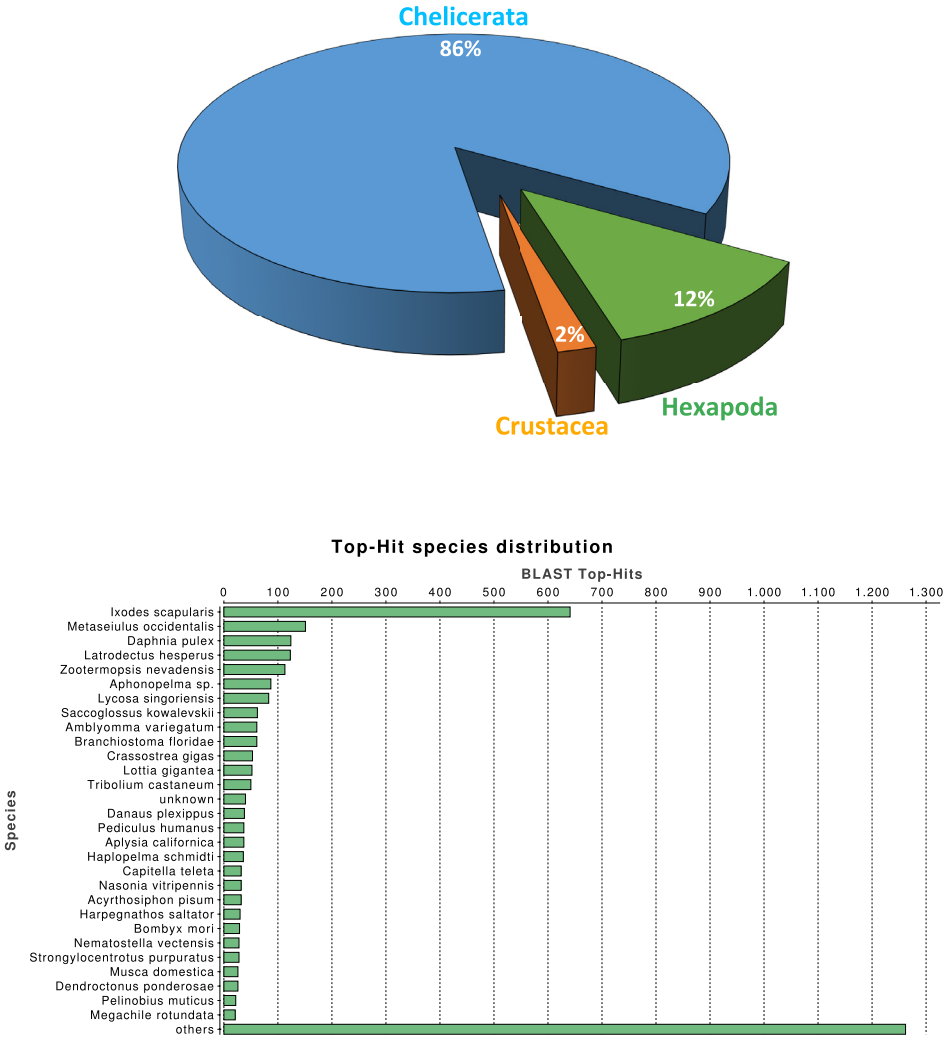


Figure 1 Macrothele taxonomic distribution. Taxonomic distribution of the 6,696 transcripts with significant blast hits against the NCBI-nr data base (using the top-hit; cut-off E-value of 10^{-3}) by means of the Blast2GO package (4,399 transcripts with blast hit). (A) Distribution of the top-hits across arthropod groups (29.4% of the transcripts with blast hit). (B) Top-hit species distribution.

the single largest category of top blast hits (blastx E-value cut-off 10^{-3}), corresponding to 25.3% of top blast hits, was to chelicerate protein coding genes, followed by hits to other arthropod species (4.1%). Within the Arthropoda, hits within Hexapoda represents about 12% (Fig. 1A), while *Ixodes scapularis* is the species receiving the majority of hits (Fig. 1B).

Overall, 2,619, 3,353 and 776 out of the 6,696 identified transcripts have a GO, InterPro, or KEGG associated term, respectively (Table 2); in total 4,978 of them (74.3%) have some functional annotation information. We analysed the distribution of GO terms (at GO level 2) across the 2,619 *M. calpeiana* transcripts sequences with GO annotation. We found that

the most frequent GO terms present in this sample are "metabolic" and "cellular processes" within the biological process domain (BP), and "binding" and "catalytic activities" within molecular function domain (MF). The distribution of GO terms in the complete data set (2,619 GO terms; Fig. 2) and in the data set excluding singleton sequences (1,734 GO terms; Fig. S1) is not significantly different (two tailed FET, P-value $= 0.592$ and 0.757 for BP and MF, respectively). Hence, we used the complete dataset for further functional annotation analyses.

Tissue-specific expression

With our subtractive approach we aimed to enrich a number of tissue-specific transcripts. We detected 1,005 transcripts annotated as housekeeping genes (Table 2) and 789 transcripts with putative homology to 290 of 458 CEG members of the CEGs dataset. Out of the 789 transcripts with CEG homologs, 488 are also annotated as HK genes (Fig. S2 and Tables S3–S5). Despite the finding of about 15% of HK and CEG genes, the largest proportion of them are located at the intersection of the Venn diagram (Figs. 3C and 3D), indicating that tissue-specific transcripts should reliably represent tissue-specific functions. After excluding these likely ubiquitously expressed genes, the remaining sample ($n = 5,390$ transcripts; 1,523 with GO annotation) exhibits the desired tissue-specific expression profile. In fact, the distributions of GO terms including (2,619 transcripts) or not (1,523 transcripts) HK/CEG genes are significantly different from each other (two tailed P-value < 0.018 for the most frequent GO categories within BP and MP) (Fig. 2).

To gain further insight into transcript function, we compared transcript expression across legs, palps, and ovaries (Fig. 3; Fig. S3). We found a high proportion of transcripts shared between leg and palp (1,112 and 848, including or not HK and CEG genes, respectively), and a few between these tissues and ovary (Figs. 3A and 3B). This result was expected given the ontogenetic similarities of legs and palps.

The overrepresentation analysis of the GO terms across the different Venn diagram sections (Table S3; see also Fig. 3E) detected 26 significant overrepresented GO terms in legs-palps (sections I, II and IV) or ovary transcripts (sections III, V, VI and VII) after the FDR (Fig. 4; Table S6A and Fig. S4). For instance, the GO terms "cation binding," "metal ion binding," and "oxidation–reduction process" are clearly overrepresented in legs-palps specific transcripts (P-value $< 6.9 \times 10^{-8}$). These significant differences are also found in comparisons involving only section III (i.e., considering only ovary-specific transcripts instead of all ovary-transcripts), or only section IV (considering only specific transcripts shared between leg and palp) (results not shown). Indeed, the major over- or underrepresentation effect appears in individual sections III and IV (Table S6).

To investigate the biological pathways that are differently expressed among the studied tissues, we analysed the distribution of transcripts associated with different KEGG terms (Tables S3 and S7). Again, we found significant differences between transcripts expressed exclusively in legs and/or palps (sections I, II, and IV) and the ovary-expressed transcripts (sections III, V, VI, and VII) (two tailed FET, P-value of 2.6×10^{-3}). For instance, we detected 3 KEGG pathways (Tropane, piperidine and pyridine alkaloid biosynthesis;

PeerJ

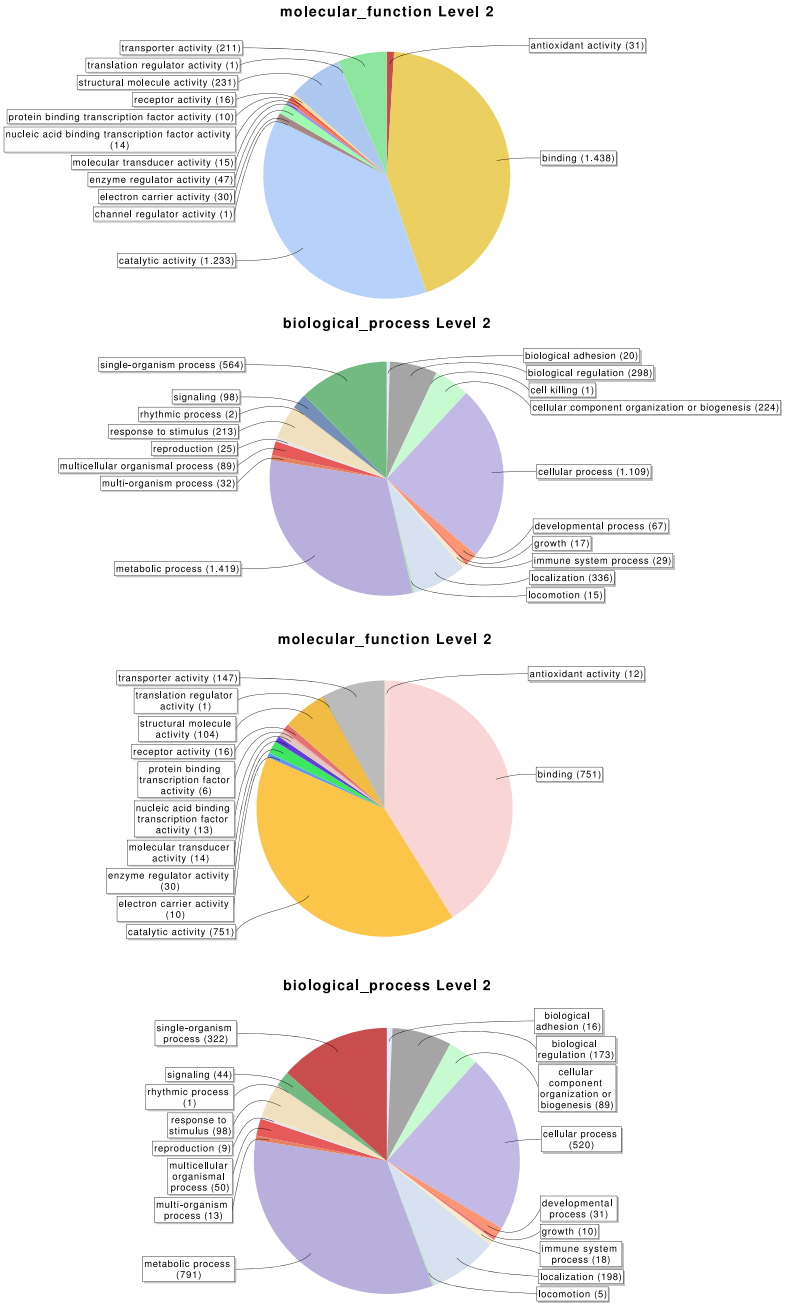


Figure 2 Distribution of the Gene Ontology (GO) terms associated with the complete set of *M. calpeiana* transcripts (2,619 transcripts with GO annotations over 6,696 sequences). (A) MF, molecular function. (B) BP, Biological process. Distribution GO terms excluding transcripts encoding HK or CEG genes (1,523 transcripts with GO annotations over 5,390 sequences). (C) MF, molecular function. (D) BP, Biological process.

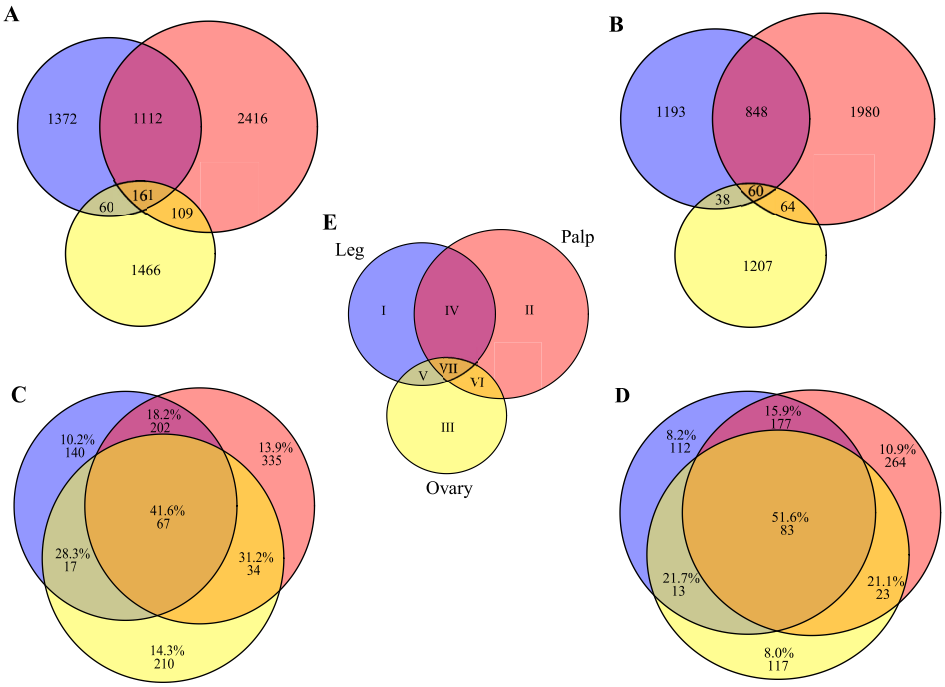


Figure 3 Transcript distribution across tissues. Venn diagrams showing the number of sequences expressed specifically in each tissue or in their intersections (blue, ochre and yellow indicate leg, palp and ovary, respectively). (A) All transcripts ($n = 6{,}696$). (B) Transcripts excluding putative housekeeping or CEG genes ($n = 5{,}390$). (C) Number and percentage of transcripts encoded by housekeeping genes ($n = 1{,}005$). (D) Number and percentage of transcripts with homologs included in the CEG database ($n = 789$). The area of each Venn diagram section is approximately proportional to the number of transcripts (A and B), or to the particular fraction value (C and D). (E) Roman numerals used to designate the different sections.

Tryptophan metabolism; and Tyrosine metabolism) specifically expressed in sections I, II and IV; none of the 11 detected transcripts of these three pathways had ovary expression (Table S7). Actually, these pathways are not directly related to chemosensory function. It has been shown that the golden orb web spider *Nephila antipodiana* (Walckenaer) coats its web with an alkaloid (2-pyrrolidinone), which apparently provides protection against ant invasion (*Zhang et al., 2012*). *Macrothele* large funnel-webs are equally exposed to predators, both insects and small vertebrates, and hence the use of a chemical defense against invaders would be highly advantageous. Further studies on the presence of these chemical clues on the funnel-webs are needed to confirm this hypothesis.

Chemosensory-related genes

As a starting point for the identification of chemosensory organs in *M. calpeiana*, we studied two features commonly present in the chemosensory-related proteins, the existence of a signal peptide (characteristic of soluble binding proteins such as insect and vertebrate OBP, and the NPC2, CSP, and CheA/B), and the presence of a transmembrane domain (characteristic of all chemosensory receptors, such as insect and vertebrate ORs,

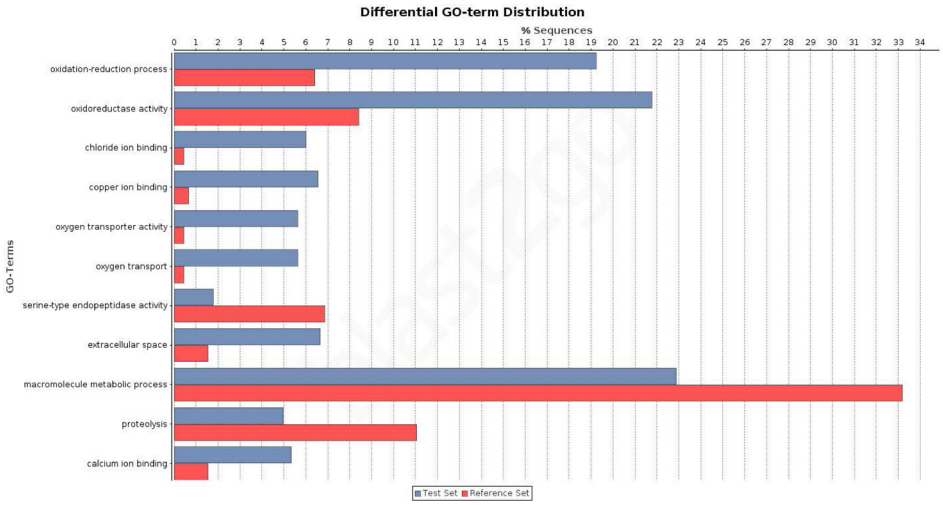


Figure 4 Differential distribution of GO terms across tissues. Differential distribution of the GO terms of the transcripts from leg and palp (Venn sections I, II and IV; in blue) and ovary (sections III, V, VI and VII; in red). Analysis conducted excluding HK and CEG encoding genes (1,523 transcripts over 5,390).

GRs and IRs). For that, we searched for a putative tissue-specific overrepresentation of such features in legs and palps (the candidate chemosensory structures in spiders) among the 3,353 transcripts with InterPro annotation. We found a significant over-representation of the signal peptide-encoding transcripts in legs-palps specific transcripts (Venn sections I, II and IV against the rest) (two tailed FET, *P*-value of 6.9×10^{-3}), being especially evident for transcripts shared between palps and legs tissues (Venn section IV; two tailed FET, *P*-value of 9.7×10^{-7}). Remarkably, the percentage of transcripts with signal peptide in section IV of the Venn diagram (transcripts expressed in both legs and palps, but not in ovary) is 27.8% (Fig. 5A), while the 40.6% of leg-specific transcripts have at least one transmembrane domain (Fig. 5B). Given that these features are not completely exclusive of chemosensory genes it is difficult to clearly assess whether these differences may reflect true differences in the chemosensory role of these tissues (see also Fig. S5).

The specific blast searches for chemosensory genes against the CheDB database detected several candidate transcripts. Nevertheless, the examination of the conceptual translation of these transcripts using HMM profiles showed that only seven candidates (two IR and five NPC2; Table S3) have the specific molecular signature of a chemosensory protein domain. Almost all the other candidates either exhibit non-chemosensory domain signatures or yielded no significant results in the search against HMM profiles. The two putative IR transcripts are specifically expressed in palps and each of them encodes a different Pfam domain characteristic of these receptors (*Croset et al., 2010*), the extracellular amino-terminal domain (PF01094; transcript Mcal_4794) and the ligand-gated ion channel domain (PF00060; transcript Mcal_5646). The closest related proteins of the *M. calpeiana* transcripts in the CheDB database correspond with two *S. mimosarum* predicted proteins annotated as "Glutamate receptor, ionotropic kainate 2" products (GenBank accessions

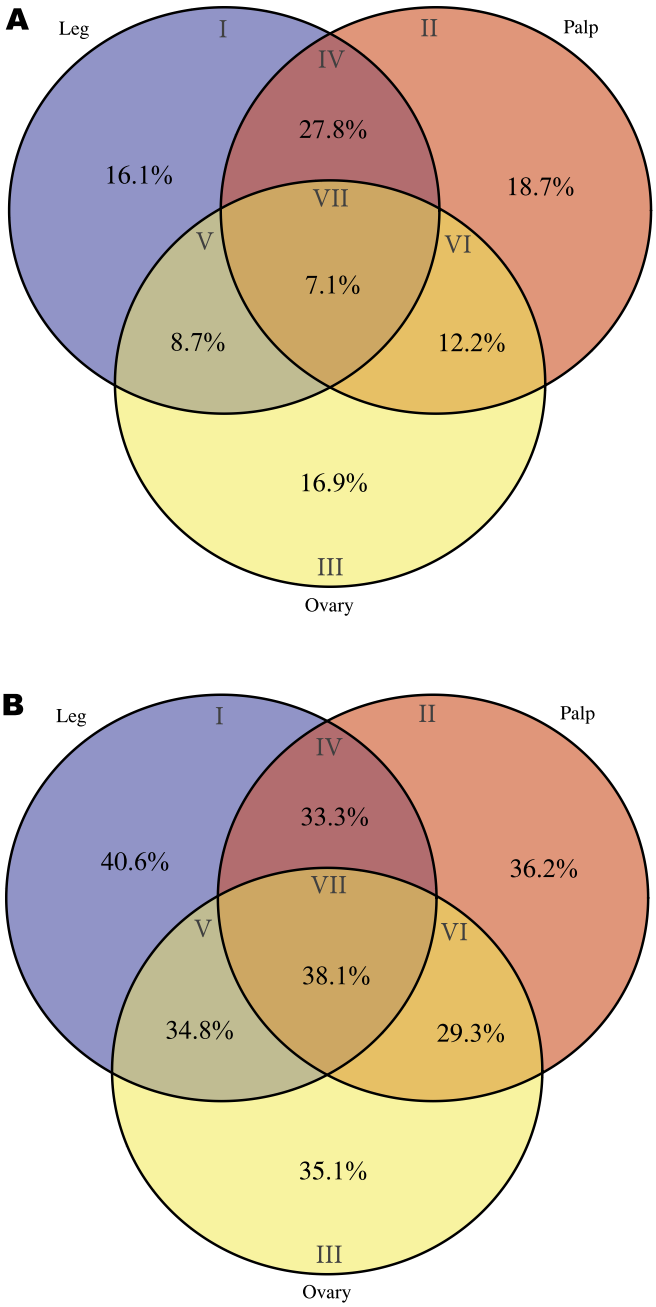


Figure 5 Distribution of specific interpro domains across tissues. Venn diagrams showing the percentage of specific interpro domains across tissues (the different Venn sections are indicated in roman numbers). Analysis conducted excluding HK and CEG encoding genes (2,364 transcripts with Interpro annotation over 5,390). (A) Signal peptide domain. (B) Transmembrane domain.

KFM81344 and KFM59881, 48% and 67% of identity, with Mcal_4794 and Mcal_5646, respectively). Nevertheless, we cannot rule out that the two *M. calpeiana* transcripts were in fact two fragments of the same iGluR gene since KFM59881 is also a partial product that only includes the "Lig_chan" domain. Besides, the rest of best-hits in blast searches using these two *M. calpeiana* transcripts as queries correspond to kainate (KA) receptors followed by α-amino-3-hydroxy-5-methyl-4-isoxazole propionate (AMPA) members in other arthropod species. The phylogenetic trees of the members of these subfamilies in arthropods (built separately for each protein domain; see 'Methods') show that the translated proteins of Mcal_4794 and Mcal_5646 group in the same clade with some KA receptors of insects, centipedes or ticks (Figs. S6A and S6B), again suggesting their putative role in synaptic transmission and regulation (i.e., it would not be a chemosensory receptor).

The products of three of the five putative NPC2 encoding transcripts constitute a *M. calpeiana* specific monophyletic clade in the NPC2 family tree (Fig. S6C) and are specifically expressed in ovary, which is suggestive of a non-chemosensory function. The other two NPC2 are expressed in palp and legs (Mcal_1484) or palp-specific (Mcal_6333). Both encoding proteins are relatively distant to the *Apis mellifera* and *Camponotus japonicus* antennal expressed NPC2, being more related to some *I. scapularis* and *S. maritima* members as well as with the ovarian clade of NPC2. In light of these results, the possible chemosensory function of these proteins in palps and legs remains to be elucidated. These results strongly encourage further functional analyses to determine the putative chemosensory role of these NPC2 genes specifically expressed in palps and legs.

Recent genome sequencing projects have revealed that chelicerate genomes contain numerous copies of ionotropic (IR) and insect-like gustatory (GR) receptors, which are the principal candidates to perform chemoreceptor functions in these species. The apparent absence of genes belonging to these families specifically expressed in *M. calpeiana* palp/leg tissues might be explained by low sequence coverage. Many of these receptors are probably encoded by low expressed genes, and their detection might need more extensive sequencing. However, to date, there is no other study of the specific expression of either these receptors or other chemosensory family members in different tissues of a chelicerate. Given the life-style of *M. calpeiana*, i.e., it builds funnel-shaped webs, which it uses to trap prey, we cannot rule out the possibility of a residual role of a chemoreceptor system in favour of mechanoreception in this species. New deep sequencing transcriptomic data from other spider species are needed to answer this question. In fact, our preliminary results from tissue specific transcriptomes in *Dysdera silvatica* (Araneae, Haplogynae) (J Vizueta et al., 2015, unpublished data) indicate that members IRs and GRs families are specifically expressed in leg and palp tissues, suggesting their putative role in chemoreception in nocturnal running hunter spiders.

Mygalomorph phylogeny

From the data matrix d327 of *Bond et al. (2014)*, we built a new MSA with information of *M. calpeiana* obtained from our transcriptome analysis. We have filtered the data in order to include high quality homologous data with high coverage per taxon. Our final MSA

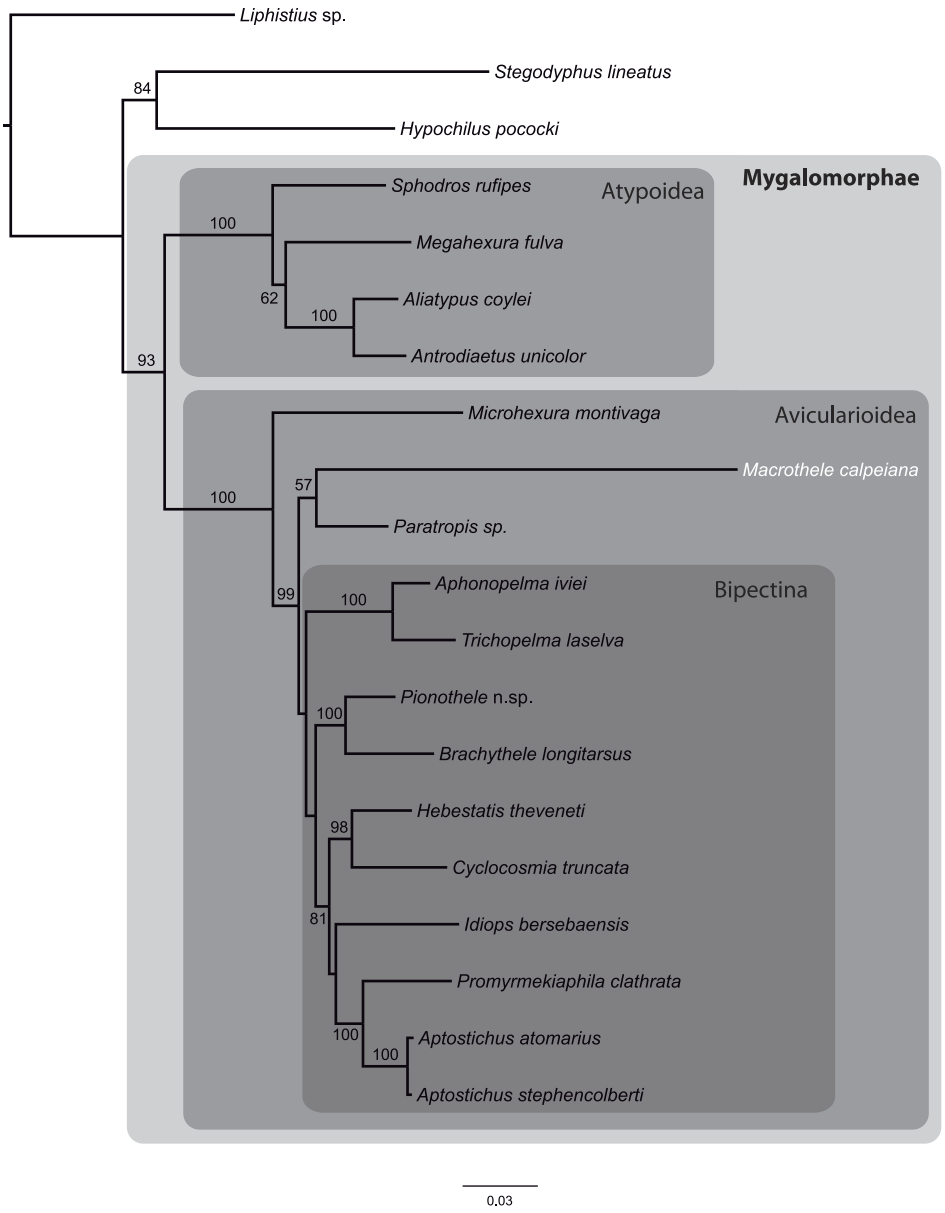


Figure 6 Phylogenetic relationships of major Mygalomorphae lineages sampled. ML tree showing the phylogenetic relationships of major Mygalomorphae lineages sampled. The analysis is based on a supermatrix of 35 putative orthologs (4,531 amino acids). Numbers indicate bootstrap support values >50%.

comprises 17 Mygalomorph species (including *M. calpeiana*) and 3 non-mygalomorph outgroups (20 taxa; 35 genes; 4,531 amino acids; Table S8), with an average taxa coverage of 17.1. Our ML phylogenetic tree, rooted using *Liphistus* as an outgroup, mirrors those reported in *Bond et al. (2014)* and shows *M. calpeiana* as the sister lineage of the genus *Paratropis* (Fig. 6), albeit with low node support (57%), as part of the non-Bipectina

Avicularioidea. Interestingly, in a recent study focused on the phylogenetic relationship and biogeographic origins of the genus *Macrothele* (*Opatova & Arnedo, 2014*) based on a denser taxonomic sampling but lower gene coverage (3 genes), a similar position of *Macrothele*, within the Avicularioidea but outside the Bipectina lineage, was also recovered.

CONCLUSIONS

The tissue specific transcriptome presented here provides a novel resource for *Macrothele* researchers, and for people interested in spider systematics and molecular biology. Having ovary and non-ovary expressed transcripts-based markers, which may potentially differ in their evolutionary rates, can become instrumental for further studies aiming to understand the evolutionary processes acting at different time-scales, such as biological invasions, secondary gene flow or speciation, and to implement successful conservation polices; in particular, we have demonstrated the utility of these newly generated data by inferring the phylogenetic position of *M. calpeiana* in the Mygalomorphae tree. Moreover, our tissue-specific gene expression study represents a starting point to understanding the chemosensory system in spiders and, in general, in chelicerates.

ACKNOWLEDGEMENT

We thank Centres Científics i Tecnòlogics de la Universitat de Barcelona for the sequencing facilities.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

Grants from the Ministerio de Educación y Ciencia of Spain (BFU2010-15484 and CGL2013-45211 to JR, and CGL2012-36863 to MAA), and from the Comissió Interdepartamental de Recerca i Innovació Tecnològica of Spain (2009SGR-1287; 2014SGR1055; 2014SGR1604). JR and MAA were partially supported by ICREA Academia (Generalitat de Catalunya). CF-L was supported by an IRBio fellowship (Universitat de Barcelona), FCA by a Juan de la Cierva postdoctoral fellowship (Spanish Ministerio de Economía y Competitividad; JCI-2008-3456), and SG-R and AS-G by a grant under the program Beatriu de Pinós (Generalitat de Catalunya, 2010BP-A 00438 and 2010BP-B 00175, respectively). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
Ministerio de Educación y Ciencia of Spain: BFU2010-15484, CGL2013-45211, CGL2012-36863.
Comissió Interdepartamental de Recerca i Innovació Tecnològica of Spain: 2009SGR-1287, 2014SGR1055, 2014SGR1604.
ICREA Academia (Generalitat de Catalunya).
IRBio fellowship (Universitat de Barcelona).

Juan de la Cierva postdoctoral fellowship (Spanish Ministerio de Economía y Competitividad): JCI-2008-3456.
Beatriu de Pinós postdoctoral fellowships (Generalitat de Catalunya): 2010BP-A 00438, 2010BP-B 00175.

Competing Interests

Julio Rozas is an Academic Editor for PeerJ.

Author Contributions

- Cristina Frías-López performed the experiments, analyzed the data, wrote the paper, prepared figures and/or tables, reviewed drafts of the paper.
- Francisca C. Almeida and Sara Guirao-Rico performed the experiments, analyzed the data, reviewed drafts of the paper.
- Joel Vizueta analyzed the data, prepared figures and/or tables, reviewed drafts of the paper.
- Alejandro Sánchez-Gracia analyzed the data, wrote the paper, prepared figures and/or tables, reviewed drafts of the paper.
- Miquel A. Arnedo conceived and designed the experiments, analyzed the data, contributed reagents/materials/analysis tools, reviewed drafts of the paper.
- Julio Rozas conceived and designed the experiments, analyzed the data, wrote the paper, prepared figures and/or tables, reviewed drafts of the paper.

Field Study Permissions

The following information was supplied relating to field study approvals (i.e., approving body and any reference numbers):

Field permission from the Junta de Andalucia (Spain); reference: SGYB-AFR-CMM.

DNA Deposition

The following information was supplied regarding the deposition of DNA sequences:

http://www.ncbi.nlm.nih.gov/bioproject/PRJNA285862.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.1064#supplemental-information.

REFERENCES

Almeida FC, Sánchez-Gracia A, Campos JL, Rozas J. 2014. Family size evolution in *Drosophila* chemosensory gene families: a comparative analysis with a critical appraisal of methods. *Genome Biology and Evolution* **6**:1669–1682 DOI 10.1093/gbe/evu130.

Altschul S. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25**:3389–3402 DOI 10.1093/nar/25.17.3389.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)* **57**:289–300 DOI 10.2307/2346101.

Bond JE, Beamer DA, Lamb T, Hedin M. 2006. Combining genetic and geospatial analyses to infer population extinction in mygalomorph spiders endemic to the Los Angeles region. *Animal Conservation* **9**:145–157 DOI 10.1111/j.1469-1795.2006.00024.x.

Bond JE, Garrison NL, Hamilton CA, Godwin RL, Hedin M, Agnarsson I. 2014. Phylogenomics resolves a spider backbone phylogeny and rejects a prevailing paradigm for orb web evolution. *Current Biology: CB* **24**:1765–1771 DOI 10.1016/j.cub.2014.06.034.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* **10**:421 DOI 10.1186/1471-2105-10-421.

Cao Z, Yu Y, Wu Y, Hao P, Di Z, He Y, Chen Z, Yang W, Shen Z, He X, Sheng J, Xu X, Pan B, Feng J, Yang X, Hong W, Zhao W, Li Z, Huang K, Li T, Kong Y, Liu H, Jiang D, Zhang B, Hu J, Hu Y, Wang B, Dai J, Yuan B, Feng Y, Huang W, Xing X, Zhao G, Li X, Li Y, Li W. 2013. The genome of *Mesobuthus martensii* reveals a unique adaptation model of arthropods. *Nature Communications* **4**:Article 2602 DOI 10.1038/ncomms3602.

Cerveira AM, Jackson RR. 2012. Love is in the air: olfaction-based mate-odour identification by jumping spiders from the genus Cyrba. *Journal of Ethology* **31**:29–34 DOI 10.1007/s10164-012-0345-x.

Chipman AD, Ferrier DEK, Brena C, Qu J, Hughes DST, Schröder R, Torres-Oliva M, Znassi N, Jiang H, Almeida FC, Alonso CR, Apostolou Z, Aqrawi P, Arthur W, Barna JCJ, Blankenburg KP, Brites D, Capella-Gutiérrez S, Coyle M, Dearden PK, Du Pasquier L, Duncan EJ, Ebert D, Eibner C, Erikson G, Evans PD, Extavour CG, Francisco L, Gabaldón T, Gillis WJ, Goodwin-Horn EA, Green JE, Griffiths-Jones S, Grimmelikhuijzen CJP, Gubbala S, Guigó R, Han Y, Hauser F, Havlak P, Hayden L, Helbing S, Holder M, Hui JHL, Hunn JP, Hunnekuhl VS, Jackson L, Javaid M, Jhangiani SN, Jiggins FM, Jones TE, Kaiser TS, Kalra D, Kenny NJ, Korchina V, Kovar CL, Kraus FB, Lapraz F, Lee SL, Lv J, Mandapat C, Manning G, Mariotti M, Mata R, Mathew T, Neumann T, Newsham I, Ngo DN, Ninova M, Okwuonu G, Ongeri F, Palmer WJ, Patil S, Patraquim P, Pham C, Pu L-L, Putman NH, Rabouille C, Ramos OM, Rhodes AC, Robertson HE, Robertson HM, Ronshaugen M, Rozas J, Saada N, Sánchez-Gracia A, Scherer SE, Schurko AM, Siggens KW, Simmons D, Stief A, Stolle E, Telford MJ, Tessmar-Raible K, Thornton R, Van der Zee M, von Haeseler A, Williams JM, Willis JH, Wu Y, Zou X, Lawson D, Muzny DM, Worley KC, Gibbs RA, Akam M, Richards S. 2014. The first myriapod genome sequence reveals conservative arthropod gene content and genome organisation in the centipede *Strigamia maritima*. *PLoS Biology* **12**:e1002005 DOI 10.1371/journal.pbio.1002005.

Clarke TH, Garb JE, Hayashi CY, Haney RA, Lancaster AK, Corbett S, Ayoub NA. 2014. Multi-tissue transcriptomics of the black widow spider reveals expansions, co-options, and functional processes of the silk gland gene toolkit. *BMC Genomics* **15**:365 DOI 10.1186/1471-2164-15-365.

Collins NM, Wells SM. 1987. *Invertebrates in need of special protection in Europe. Augier H. Nature & Environment Series No. 35*. Strasbourg: Council of Europe, pp. 162.

Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M. 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **21**:3674–3676 DOI 10.1093/bioinformatics/bti610.

Corzo G, Gilles N, Satake H, Villegas E, Dai L, Nakajima T, Haupt J. 2003. Distinct primary structures of the major peptide toxins from the venom of the spider *Macrothele gigas* that bind to sites 3 and 4 in the sodium channel. *FEBS Letters* **547**:43–50 DOI 10.1016/S0014-5793(03)00666-5.

Croset V, Rytz R, Cummins SF, Budd A, Brawand D, Kaessmann H, Gibson TJ, Benton R. 2010. Ancient protostome origin of chemosensory ionotropic glutamate receptors and the evolution of insect taste and olfaction. *PLoS Genetics* **6**:e1001064 DOI 10.1371/journal.pgen.1001064.

Eddy SR. 2009. A new generation of homology search tools based on probabilistic inference. *Genome Informatics. International Conference on Genome Informatics* **23**:205–211.

Eisenberg E, Levanon EY. 2013. Human housekeeping genes, revisited. *Trends in Genetics* **29**:569–574 DOI 10.1016/j.tig.2013.05.010.

Foelix RF. 1970. Chemosensitive hairs in spiders. *Journal of Morphology* **132**:313–333 DOI 10.1002/jmor.1051320306.

Foelix RF, Chu-Wang IW. 1973. The morphology of spider sensilla. II. Chemoreceptors. *Tissue & Cell* **5**:461–478 DOI 10.1016/S0040-8166(73)80038-2.

Fu L, Niu B, Zhu Z, Wu S, Li W. 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**:3150–3152 DOI 10.1093/bioinformatics/bts565.

Gao L, Shan B, Chen J, Liu J, Song D, Zhu B. 2005. Effects of spider *Macrothele raven* venom on cell proliferation and cytotoxicity in HeLa cells. *Acta Pharmacologica Sinica* **26**:369–376 DOI 10.1111/j.1745-7254.2005.00052.x.

Grbić M, Van Leeuwen T, Clark RM, Rombauts S, Rouzé P, Grbić V, Osborne EJ, Dermauw W, Thi Ngoc PC, Ortego F, Hernández-Crespo P, Diaz I, Martinez M, Navajas M, Sucena É, Magalhães S, Nagy L, Pace RM, Djuranović S, Smagghe G, Iga M, Christiaens O, Veenstra JA, Ewer J, Villalobos RM, Hutter JL, Hudson SD, Velez M, Yi S V, Zeng J, Pires-daSilva A, Roch F, Cazaux M, Navarro M, Zhurov V, Acevedo G, Bjelica A, Fawcett JA, Bonnet E, Martens C, Baele G, Wissler L, Sanchez-Rodriguez A, Tirry L, Blais C, Demeestere K, Henz SR, Gregory TR, Mathieu J, Verdon L, Farinelli L, Schmutz J, Lindquist E, Feyereisen R, Van de Peer Y. 2011. The genome of *Tetranychus urticae* reveals herbivorous pest adaptations. *Nature* **479**:487–492 DOI 10.1038/nature10640.

Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, Macmanes MD, Ott M, Orvis J, Pochet N, Strozzi F, Weeks N, Westerman R, William T, Dewey CN, Henschel R, Leduc RD, Friedman N, Regev A. 2013. *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols* **8**:1494–1512 DOI 10.1038/nprot.2013.084.

Harvey MS. 2002. Short-range endemism amongst the Australian fauna: some examples from non-marine environments. *Invertebrate Systematics* **16**:555–570 DOI 10.1071/IS02009.

Huang X. 1999. CAP3: a DNA sequence assembly program. *Genome Research* **9**:868–877 DOI 10.1101/gr.9.9.868.

Hung S-W, Wang T-L. 2004. Arachnid envenomation in Taiwan. *Ann Disaster Med* **3**:12–17.

Jiménez-Valverde A, Decae AE, Arnedo MA. 2011. Environmental suitability of new reported localities of the funnel-web spider *Macrothele calpeiana*: an assessment using potential distribution modelling with presence-only techniques. *Journal of Biogeography* **38**:1213–1223 DOI 10.1111/j.1365-2699.2010.02465.x.

Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, Pesseat S, Quinn AF, Sangrador-Vegas A, Scheremetjew M, Yong S-Y, Lopez R, Hunter S. 2014. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**:1236–1240 DOI 10.1093/bioinformatics/btu031.

Kanehisa M, Goto S. 2000. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research* **28**:27–30 DOI 10.1093/nar/28.1.27.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution* **30**:772–780 DOI 10.1093/molbev/mst010.

Kornobis E, Cabellos L, Aguilar F, Frías-López C, Rozas J, Marco J, Zardoya R. 2015. TRUFA: a user-friendly web server for *de novo* RNA-seq analysis using cluster computing. *Evolutionary Bioinformatics* **11**:97–104 DOI 10.4137/EBO.S23873.

Kronestedt T. 1979. Study on chemosensitive hairs in wolf spiders (Araneae, Lycosidae) by scanning electron microscopy. *Zoologica Scripta* **8**:279–285 DOI 10.1111/j.1463-6409.1979.tb00639.x.

Lam KC, Mühlpfordt F, Vaquerizas JM, Raja SJ, Holz H, Luscombe NM, Manke T, Akhtar A. 2012. The NSL complex regulates housekeeping genes in *Drosophila*. *PLoS Genetics* **8**:e1002736 DOI 10.1371/journal.pgen.1002736.

Liu Z, Zhao Y, Li J, Xu S, Liu C, Zhu Y, Liang S. 2012. The venom of the spider *Macrothele raveni* induces apoptosis in the myelogenous leukemia K562 cell line. *Leukemia Research* **36**:1063–1066 DOI 10.1016/j.leukres.2012.02.025.

Mattila TM, Bechsgaard JS, Hansen TT, Schierup MH, Bilde T. 2012. Orthologous genes identified by transcriptome sequencing in the spider genus *Stegodyphus*. *BMC Genomics* **13**:70 DOI 10.1186/1471-2164-13-70.

Montagné N, de Fouchier A, Newcomb RD, Jacquin-Joly E. 2015. Advances in the identification and characterization of olfactory receptors in insects. *Progress in Molecular Biology and Translational Science* **130**:55–80 DOI 10.1016/bs.pmbts.2014.11.003.

Opatova V, Arnedo MA. 2014. From Gondwana to Europe: inferring the origins of Mediterranean *Macrothele* spiders (Araneae: Hexathelidae) and the limits of the family Hexathelidae. *Invertebrate Systematics* **28**:361–374 DOI 10.1071/IS14004.

Parra G, Bradnam K, Korf I. 2007. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23**:1061–1067 DOI 10.1093/bioinformatics/btm071.

Parra G, Bradnam K, Ning Z, Keane T, Korf I. 2009. Assessing the gene space in draft genomes. *Nucleic Acids Research* **37**:289–297 DOI 10.1093/nar/gkn916.

Patel RK, Jain M. 2012. NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. *PLoS ONE* **7**:e30619 DOI 10.1371/journal.pone.0030619.

Pelosi P, Iovinella I, Felicioli A, Dani FR. 2014. Soluble proteins of chemical communication: an overview across arthropods. *Frontiers in Physiology* **5**:Article 320 DOI 10.3389/fphys.2014.00320.

Platnick NI. 2006. The world spider catalog, V6.5 by N. I. Platnick. AMNH. *Available at https://research.amnh.org/iz/spiders/catalog_15.0/index.html*.

Posnien N, Zeng V, Schwager EE, Pechmann M, Hilbrant M, Keefe JD, Damen WGM, Prpic N-M, McGregor AP, Extavour CG. 2014. A comprehensive reference transcriptome resource for the common house spider *Parasteatoda tepidariorum*. *PLoS ONE* **9**:e104885 DOI 10.1371/journal.pone.0104885.

Prosdocimi F, Bittencourt D, da Silva FR, Kirst M, Motta PC, Rech EL. 2011. Spinning gland transcriptomics from two main clades of spiders (order: Araneae)–insights on their molecular, anatomical and behavioral evolution. *PLoS ONE* **6**:e21634 DOI 10.1371/journal.pone.0021634.

Sánchez-Gracia A, Vieira FG, Rozas J. 2009. Molecular evolution of the major chemosensory gene families in insects. *Heredity* **103**:208–216 DOI 10.1038/hdy.2009.55.

Sanggaard KW, Bechsgaard JS, Fang X, Duan J, Dyrlund TF, Gupta V, Jiang X, Cheng L, Fan D, Feng Y, Han L, Huang Z, Wu Z, Liao L, Settepani V, Thøgersen IB, Vanthournout B, Wang T, Zhu Y, Funch P, Enghild JJ, Schauser L, Andersen SU, Villesen P, Schierup MH, Bilde T, Wang J. 2014. Spider genomes provide insight into composition and evolution of venom and silk. *Nature Communications* **5**:Article 3765 DOI 10.1038/ncomms4765.

Satake H, Villegas E, Oshiro N, Terada K, Shinada T, Corzo G. 2004. Rapid and efficient identification of cysteine-rich peptides by random screening of a venom gland cDNA library from the hexathelid spider *Macrothele gigas*. *Toxicon* **44**:149–156 DOI 10.1016/j.toxicon.2004.05.012.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**:1312–1313 DOI 10.1093/bioinformatics/btu033.

Vieira FG, Rozas J. 2011. Comparative genomics of the odorant-binding and chemosensory protein gene families across the Arthropoda: origin and evolutionary history of the chemosensory system. *Genome Biology and Evolution* **3**:476–490 DOI 10.1093/gbe/evr033.

Yamaji N, Little MJ, Nishio H, Billen B, Villegas E, Nishiuchi Y, Tytgat J, Nicholson GM, Corzo G. 2009. Synthesis, solution structure, and phylum selectivity of a spider delta-toxin that slows inactivation of specific voltage-gated sodium channel subtypes. *The Journal of Biological Chemistry* **284**:24568–24582 DOI 10.1074/jbc.M109.030841.

Zeng X-Z, Xiao Q-B, Liang S-P. 2003. Purification and characterization of raventoxin-I and raventoxin-III, two neurotoxic peptides from the venom of the spider *Macrothele raveni*. *Toxicon* **41**:651–656 DOI 10.1016/S0041-0101(02)00361-6.

Zhang S, Koh TH, Seah WK, Lai YH, Elgar MA, Li D. 2012. A novel property of spider silk: chemical defence against ants. *Proceedings Biological Sciences of the Royal Society* **279**:1824–1830 DOI 10.1098/rspb.2011.2193.

Artículo 1

Supplemental Information

Figure S1.

Distribution of the Gene Ontology (GO) terms associated with the complete set of *M. calpeiana* transcripts excluding singletons (transcripts formed by a single read). Panels A (for MF, molecular function) and B (for BP, Biological process) include information from the 1,734 transcripts with GO annotations over 3,467 sequences. Panels C (for MF, molecular function) and D (for BP, Biological process) include information from the 973 transcripts with GO annotations over 2,573 sequences (i.e. the 3,467 sequences after excluding HK or CEG genes).

Figure S2.

Number of *Macrothele* transcripts encoding HK or CEG genes.

Figure S3.

Venn diagrams showing the number of sequences excluding singletons expressed specifically in each tissue or in their intersections (blue, ochre and yellow indicate leg, palp and ovary, respectively). A) All transcripts (n = 3,467); B) Number of transcripts excluding those coding for housekeeping and CEG genes (n = 2,589); C) Number and percentage of transcripts coding for housekeeping genes (n = 688). D) Number and percentage of transcripts including in the CEG database (n = 533). In panels A and B, The area of each Venn diagram section is

approximately proportional to the number of transcripts (panels A and B), or to the particular fraction value (panels C and D).

Figure S4.

Differential distribution of the GO terms of the transcripts from leg or palp (Venn sections I, II and IV; in blue) and ovary (sections III, V, VI and VII; in red). Analysis comprising all transcripts with GO terms (2,619 transcripts over 6,696).

Figure S5.

Venn diagrams showing the percentage of specific interpro domains across tissues (the different Venn sections are indicated in roman numbers). Analysis conducted including HK and CEG encoding genes (3,353 transcripts with Interpro annotation over 6,696). A) Signal peptide domain. B) Transmembrane domain.

Figure S6

Phylogenetic relationships of a representative subset of arthropod iGluR (including IR8a/IR25a) and NPC2 family members. A) iGluR family tree based on "ANF_receptor" domain. Since we fail to detect the highly divergent "ANF_receptor" domain of IR8a and IR25a proteins in our HMM-search using the PF01094 profile (but see Croset et al., 2010) we did not include these sequences in the tree of this domain. B) iGluR family tree based on "Lig_chan" domain. C) NPC2 family tree, using the same protein identifiers as in Pelosi et al. (2014). NPC2 proteins expressed in the antennae of *A. melifera* and *C. japonicus* are indicated with an asterisk. Hexapods (*A. melifera*, *C. japonicum* and *D. melanogaster*), myriapods (*S. maritima*) and chelicerates (*I. scapularis*) sequences are shown in green, red and blue, respectively, while *M. calpeiana* are represented in shaded boxes. Numbers indicate node support values (percentage over 500 bootstrap replicates).

Figure S1A

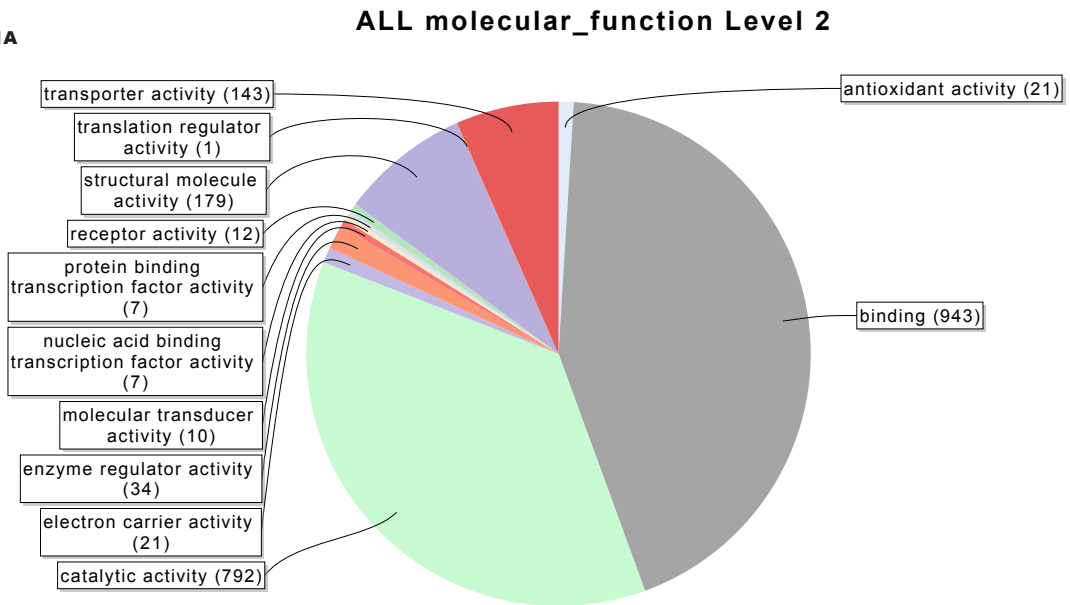


Figure S1B

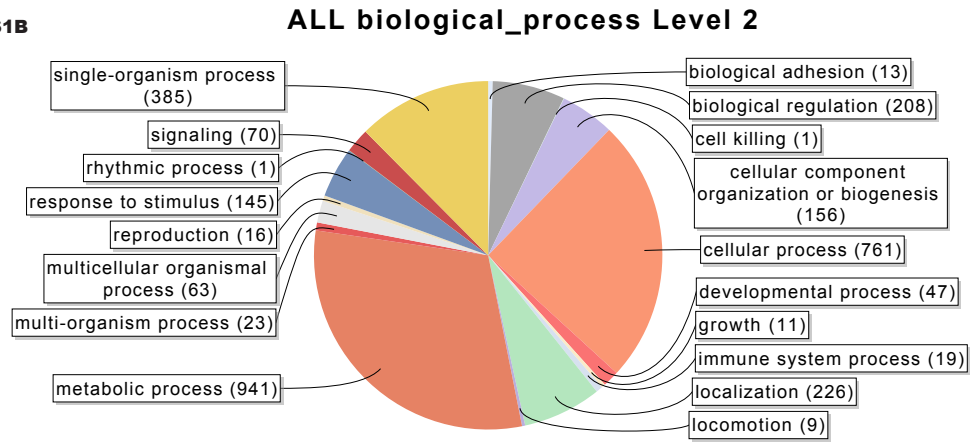


Figure S1C

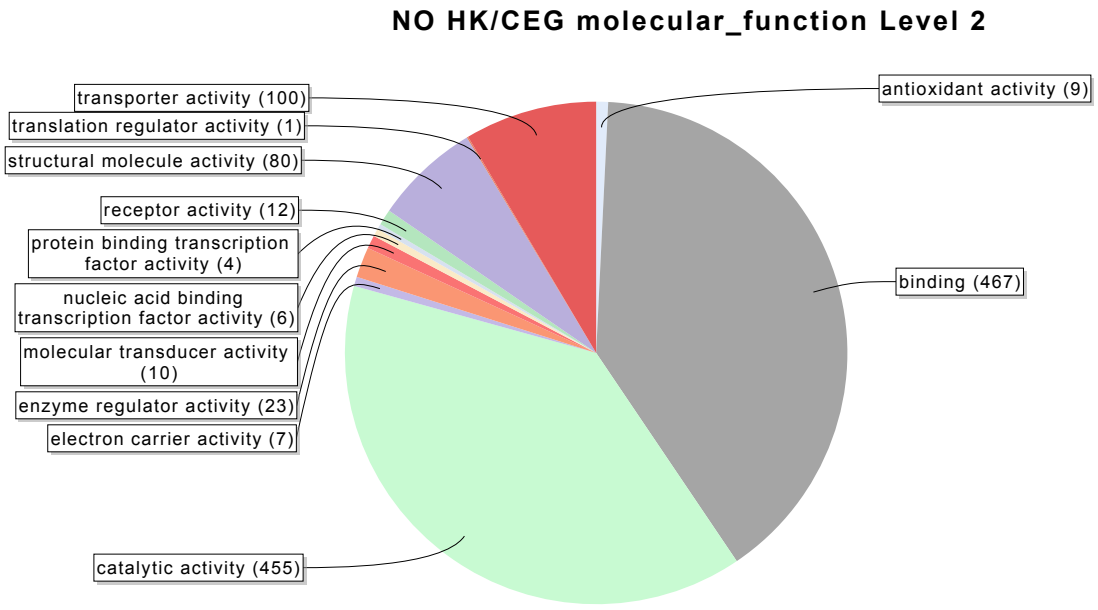


Figure S1D

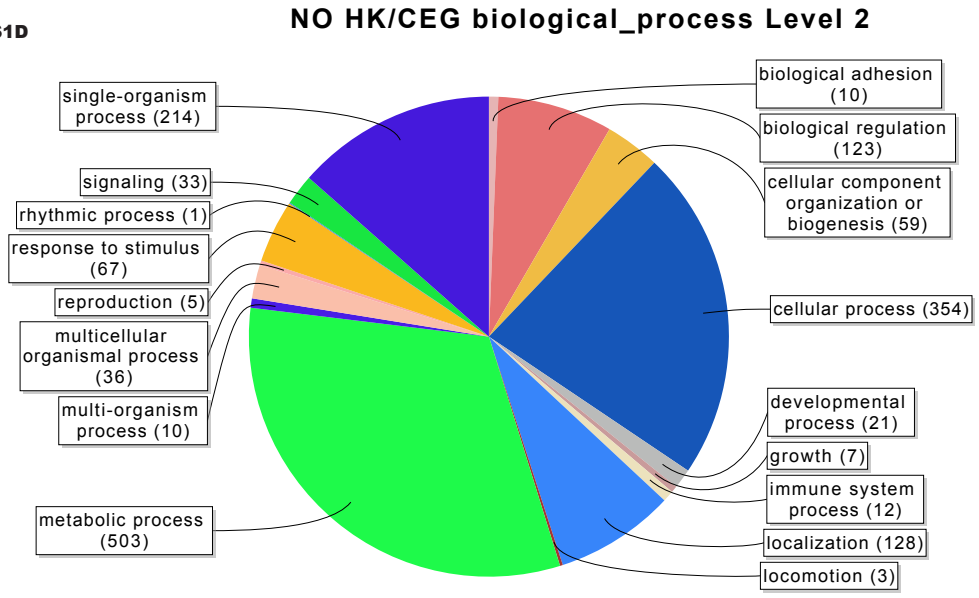


Figure S2

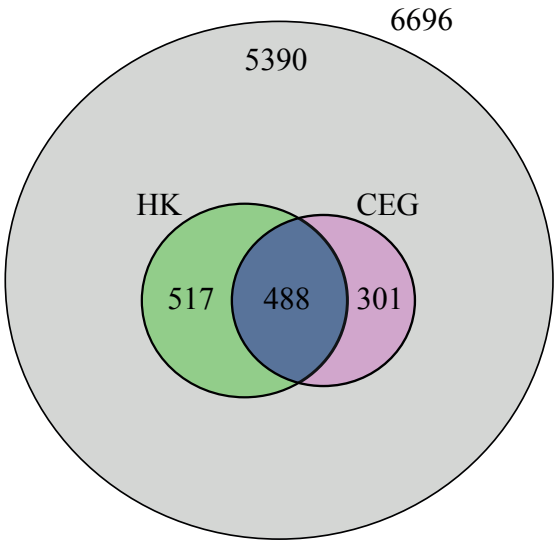
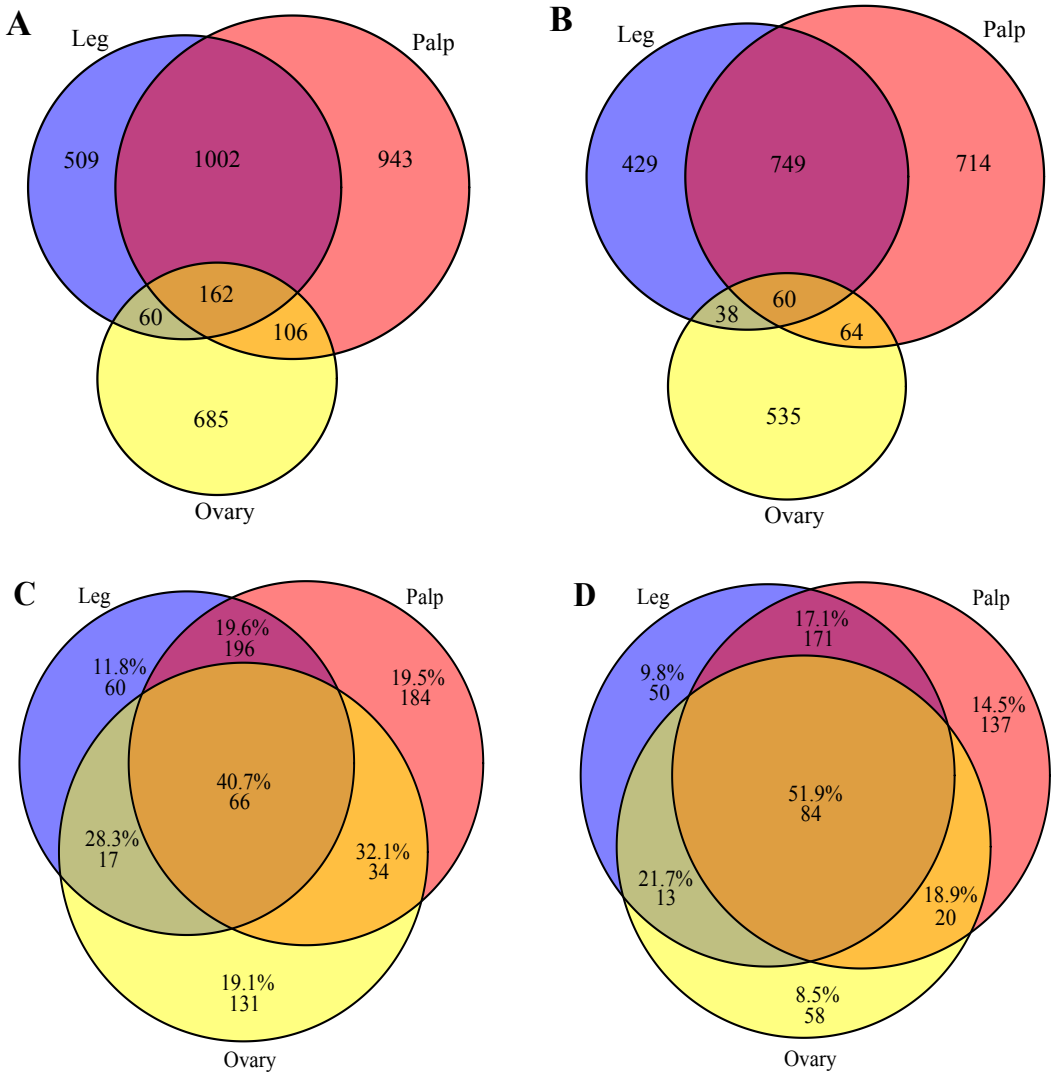


Figure S3



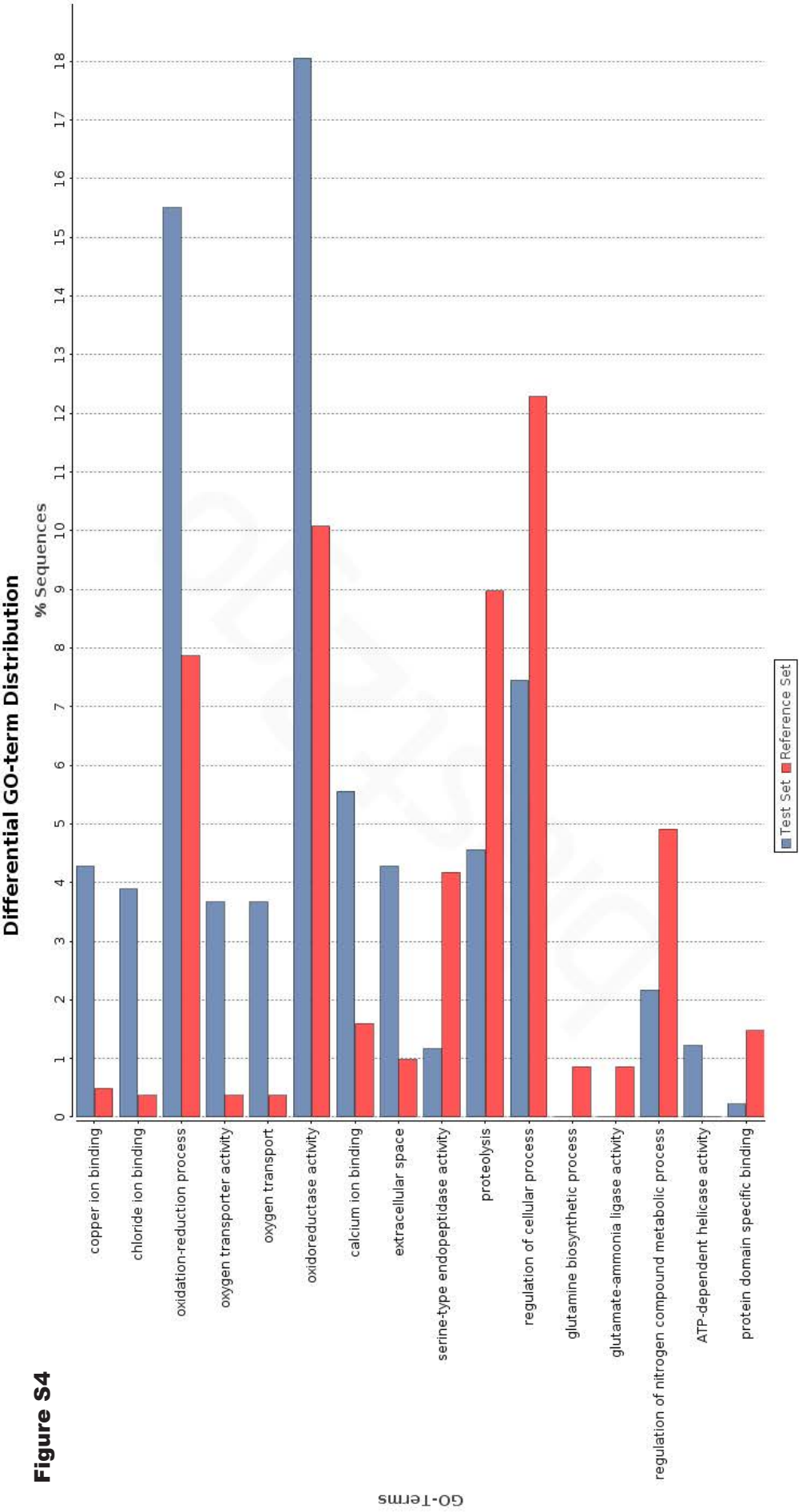

Differential GO-term Distribution
% Sequences
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
copper ion binding
chloride ion binding
oxidation-reduction process
oxygen transporter activity
oxygen transport
oxidoreductase activity
calcium ion binding
extracellular space
serine-type endopeptidase activity
proteolysis
regulation of cellular process
glutamine biosynthetic process
glutamate-ammonia ligase activity
regulation of nitrogen compound metabolic process
ATP-dependent helicase activity
protein domain specific binding
GO-Terms
Test Set
Reference Set

Figure S4

Figure S5A

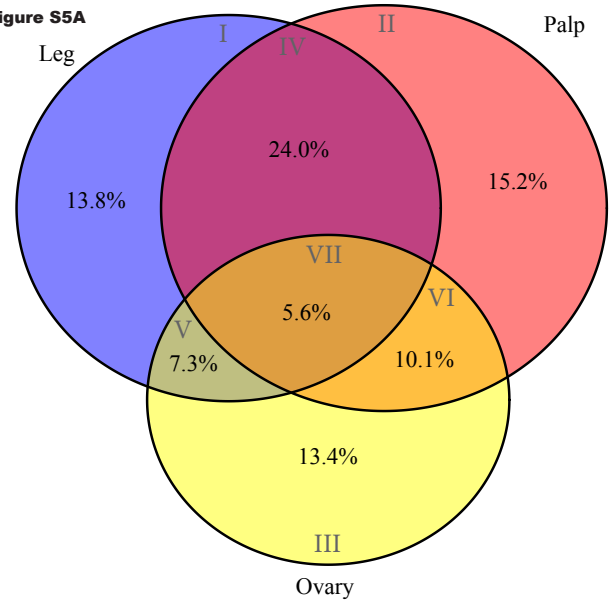


Figure S5B

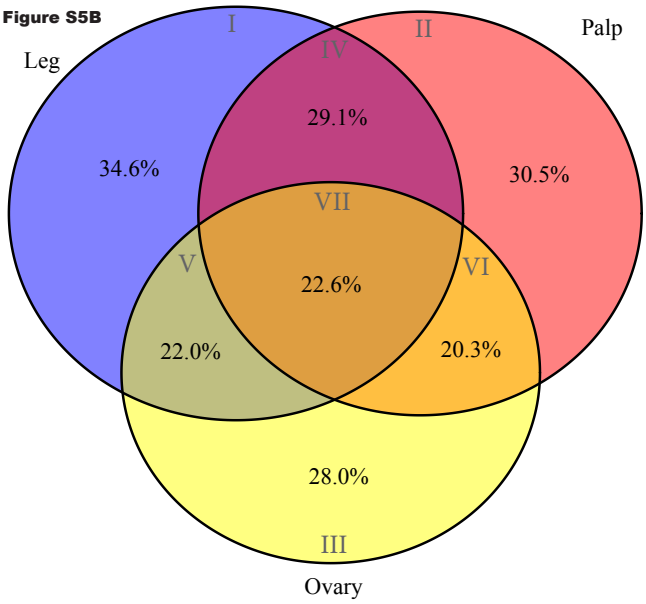


Figure 6A

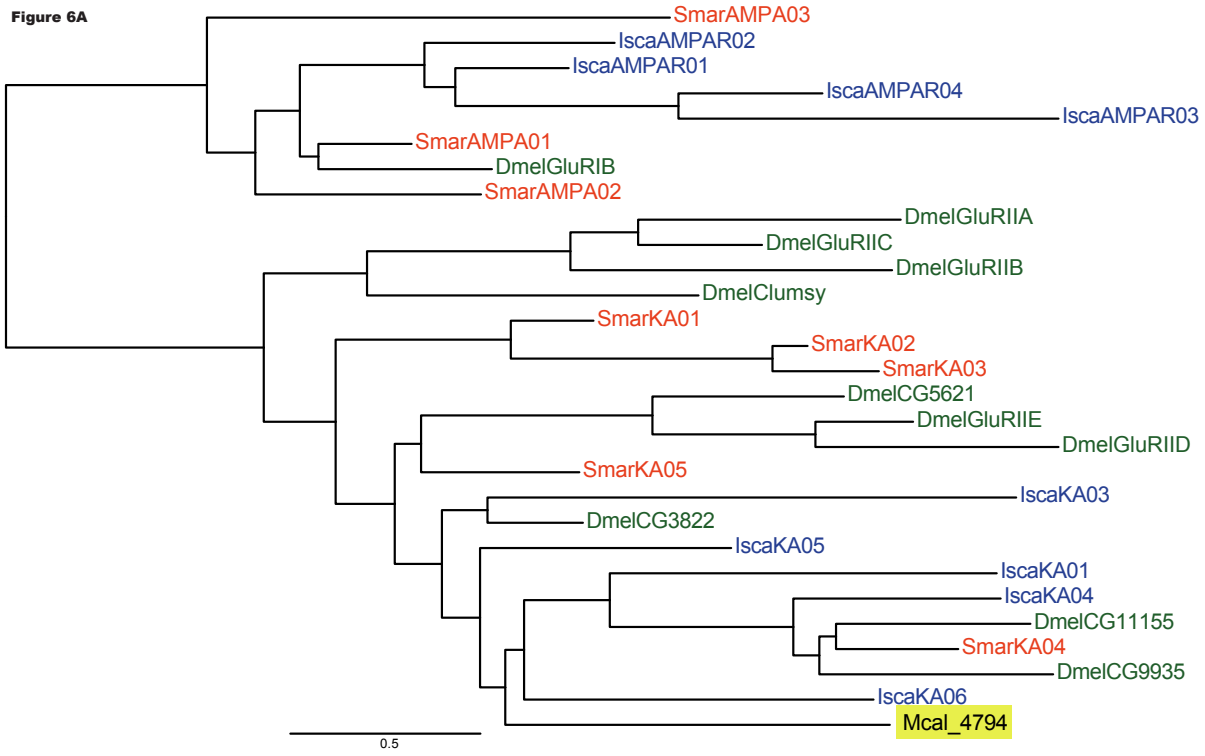
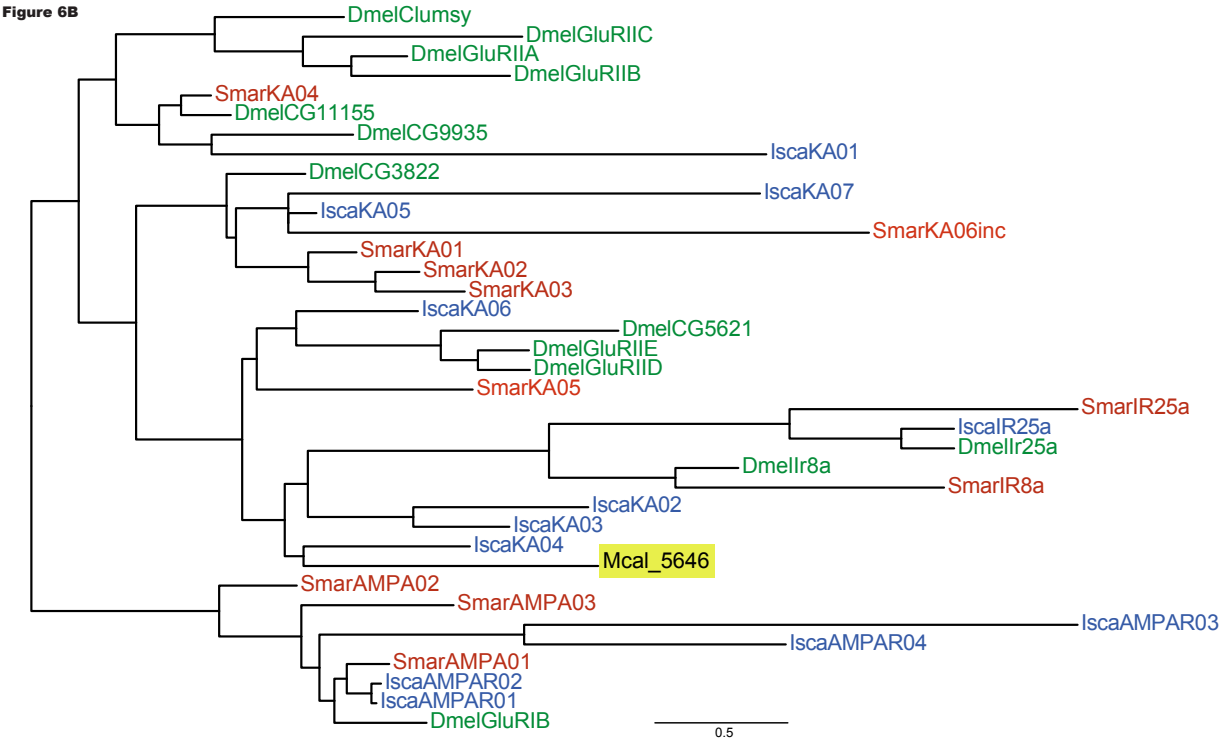


Figure 6B



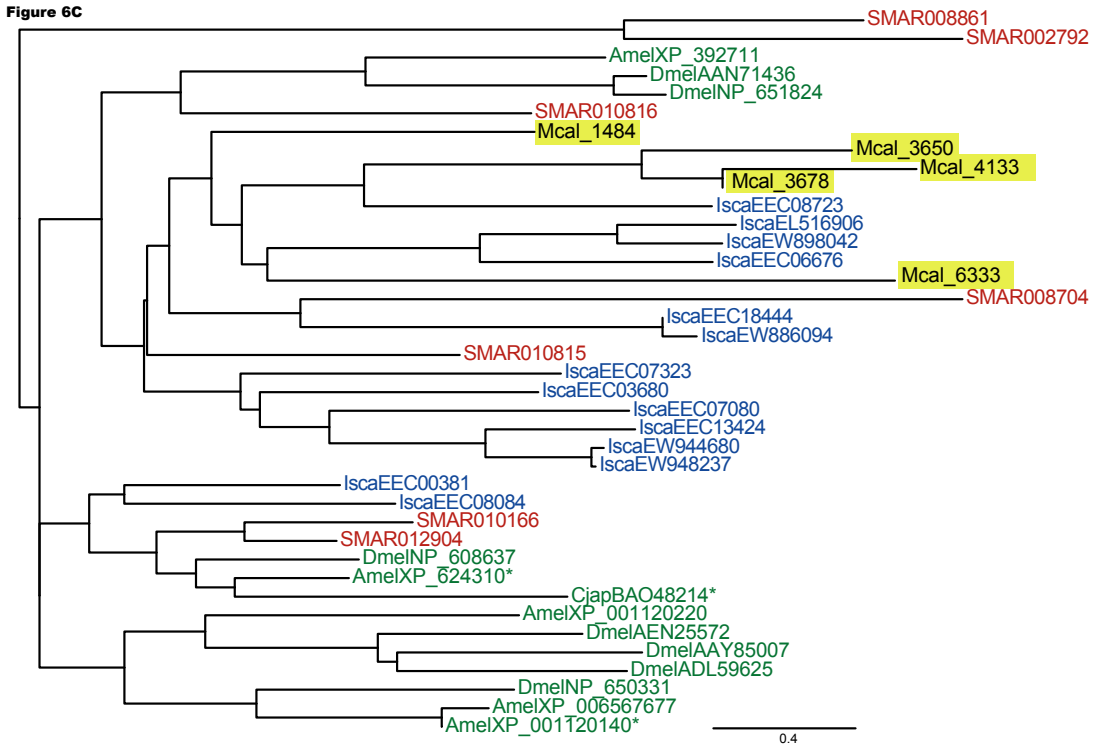


Figure 6C
SMAR008861
SMAR002792
AmelXP_392711
DmelAAN71436
DmelNP_651824
SMAR010816
Mcal_1484
Mcal_3650
Mcal_4133
Mcal_3678
IscaEEC08723
IscaEL516906
IscaEW898042
IscaEEC06676
Mcal_6333
SMAR008704
IscaEEC18444
IscaEW886094
SMAR010815
IscaEEC07323
IscaEEC03680
IscaEEC07080
IscaEEC13424
IscaEW944680
IscaEW948237
IscaEEC00381
IscaEEC08084
SMAR010166
SMAR012904
DmelNP_608637
AmelXP_624310*
CjapBAO48214*
AmelXP_001120220
DmelAEN25572
DmelAAY85007
DmelADL59625
DmelNP_650331
AmelXP_006567677
AmelXP_001120140*
0.4

Table S1
HK and CEG genes (A) List of the 5,752 Drosophila melanogaster and 3,786 human housekeeping genes used this study. (B) List of the Interpro domain associated with the chemosensory function.
peerj-03-1064-s001.xlsx (165K)
DOI: 10.7717/peerj.1064/supp-1

Table S2
DNA sequence of Macrothele transcripts DNA sequence of the 6,696 transcripts identified in M. calpeiana.
peerj-03-1064-s002.xlsx (1.1M)
DOI: 10.7717/peerj.1064/supp-2

Table S3
Functional annotation of the M. calpeiana transcripts Summary of the functional annotation of the M. calpeiana transcripts.
peerj-03-1064-s003.xlsx (1.0M)
DOI: 10.7717/peerj.1064/supp-3

Table S4
CEG genes identified in Macrothele CEG homologous genes identified in M. calpeiana transcripts.
peerj-03-1064-s004.xlsx (19K)
DOI: 10.7717/peerj.1064/supp-4

Table S5
Macrothele transcripts with significant HK or CEG blast hits M. calpeiana transcripts with significant HK or CEG blast hits
peerj-03-1064-s005.xlsx (253K)
DOI: 10.7717/peerj.1064/supp-5

Table S6
Gene Ontology terms over- and under-represented among M. calpeiana tissues List of the GO terms over- and under-represented among M. calpeiana tissues. (B) Analysis conducted excluding HK and CEG encoding genes (1,523 transcripts with GO annotation over 5,390 transcripts). (B) Analysis conducted using all data (2,619 transcripts with GO annotation over 6,619 transcripts).
peerj-03-1064-s006.xlsx (122K)
DOI: 10.7717/peerj.1064/supp-6

Table S7
KEGG pathways identified in Macrothele List of KEGG pathways identified in M. calpeiana transcripts.
peerj-03-1064-s007.xlsx (124K)
DOI: 10.7717/peerj.1064/supp-7

Table S8
Genes used in the molecular phylogenetics analyses List of the 35 genes used to infer the phylogenetic relationships among of major Mygalomorphae lineages sampled.
peerj-03-1064-s008.xlsx (15K)
DOI: 10.7717/peerj.1064/supp-8

Artículo 2

DOMINO: Development of informative molecular markers for phylogenetic and genome-wide population genetic studies in non-model organisms

Cristina Frías-López, José F. Sánchez-Herrero, Sara Guirao-Rico, Elisa Mora, Miquel A. Arnedo, Alejandro Sánchez-Gracia and Julio Rozas.

2016, Bioinformatics, 32: 3753-3759.

Bioinformatics, 32(24), 2016, 3753–3759
doi: 10.1093/bioinformatics/btw534
Advance Access Publication Date: 24 August 2016
Original Paper

OXFORD

Phylogenetics

DOMINO: development of informative molecular markers for phylogenetic and genome-wide population genetic studies in non-model organisms

Cristina Frías-López[1,2,‡], José F. Sánchez-Herrero[1,‡], Sara Guirao-Rico[1,†], Elisa Mora[2], Miquel A. Arnedo[2], Alejandro Sánchez-Gracia[1,*,§] and Julio Rozas[1,*,§]

[1]Departament de Genètica, Microbiologia i Estadística, and Institut de Recerca de la Biodiversitat (IRBio), Universitat de Barcelona, Barcelona, Spain and [2]Departament de Biologia Evolutiva, Ecologia i Ciències Ambientals, and Institut de Recerca de la Biodiversitat (IRBio), Universitat de Barcelona, Barcelona 08028, Spain

*To whom correspondence should be addressed.
[†]Present address: Centre for Research in Agricultural Genomics (CRAG) CSIC-IRTA-UAB-UB, Barcelona, Spain

[‡]The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

§The authors wish it to be known that, in their opinion, the last two authors should be regarded as joint Last Authors.

Associate Editor: Alfonso Valencia

Received on May 11, 2016; revised on July 7, 2016; accepted on August 9, 2016

Abstract

Motivation: The development of molecular markers is one of the most important challenges in phylogenetic and genome wide population genetics studies, especially in studies with non-model organisms. A highly promising approach for obtaining suitable markers is the utilization of genomic partitioning strategies for the simultaneous discovery and genotyping of a large number of markers. Unfortunately, not all markers obtained from these strategies provide enough information for solving multiple evolutionary questions at a reasonable taxonomic resolution.

Results: We have developed Development Of Molecular markers In Non-model Organisms (DOMINO), a bioinformatics tool for informative marker development from both next generation sequencing (NGS) data and pre-computed sequence alignments. The application implements popular NGS tools with new utilities in a highly versatile pipeline specifically designed to discover or select personalized markers at different levels of taxonomic resolution. These markers can be directly used to study the taxa surveyed for their design, utilized for further downstream PCR amplification in a broader set taxonomic scope, or exploited as suitable templates to bait design for target DNA enrichment techniques. We conducted an exhaustive evaluation of the performance of DOMINO via computer simulations and illustrate its utility to find informative markers in an empirical dataset.

Availability and Implementation: DOMINO is freely available from www.ub.edu/softevol/domino.

Contact: elsanchez@ub.edu or jrozas@ub.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

It is well known that phylogenetic inferences based on a single or very few genetic markers can lead to systematic errors and reach invalid conclusions (Brito and Edwards, 2009; Maddison *et al.*, 1997). Next generation sequencing (NGS) has become a feasible and cost-effective way of obtaining large amounts of genetic markers suitable for addressing ecological and evolutionary questions. Among current methodologies, the hybrid enrichment and the reduction representation sequencing methods (for a review see Lemmon and Lemmon, 2013) are particularly promising approaches for studies in non-model organisms. Markers developed with these methodologies, however, may not be informative enough to resolve multiple evolutionary questions across a reasonable taxonomic range; indeed, some markers may be inefficient for a particular study in a specific taxonomic group, or can be useful only for limited phylogenetic ranges. These problems make often necessary to accomplish various cost-intensive enrichment or reduction representation experiments to further obtain makers suitable to be applicable across a wide range of species.

Recently, some optimizing approaches have been developed to try to overcome this limited marker informativeness. For instance, the *MarkerMiner* 1.0 pipeline (Chamala *et al.*, 2015), outputs different types of multiple sequence alignments (MSA) files, some of them including reference coding sequences containing introns, which facilitates the downstream evaluation of the phylogenetic utility of each marker or the prediction of intron–exon boundaries and intron sizes, very useful for primer or probe of development. Nevertheless, the pipeline does not perform these assessments by itself and the application is specifically devised to work only with transcriptome assemblies and with a set of plant reference genomes. Indeed, the possibility of selecting particular markers with a specific number of samples has been recently implemented in the RAD-Seq data processing pipeline *RADIS* (Cruaud *et al.*). However, this application does not include other key options and parameter combinations, such as the selection of a specific nucleotide variation range across a set of pre-defined taxa, options that can be very useful for a plethora of studies. *BaitFisher* (Mayer *et al.*, 2016) also implements a novel approach to optimize the design of target enrichment baits to be applicable across a wide range of taxa. This software includes an algorithm to infer target DNA enrichment baits from multiple taxa by exploiting user-provided nucleotide sequence information of target loci in a representative set of species and can handle both genomic and cDNA data. Nevertheless, this software works on the basis of MSA of already known target loci that directly serves as templates for bait design (i.e. it cannot be used with raw NGS data or for *de novo* marker discovery).

Here we present Development Of Molecular markers In Non-model Organisms (DOMINO) a new bioinformatics tool that facilitates the development of highly informative markers from different data sources, including raw NGS reads and pre-computed MSA in various formats (such as those from RAD data). DOMINO efficiently process NGS data or pre-computed MSA and identifies (i.e. *de novo* discovery) or selects the sequence regions or alignments that meet user-defined criteria. Customizable features include the length of variable and conserved regions (when requested), the minimum levels (or a preferred range) of nucleotide variation, how to manage polymorphic variants, or which taxa (or what fraction of them) should be covered by the marker. All these criteria can be easily defined in a user-friendly graphical user interface (GUI) or under a command-line version that implements some extended options and that it is particularly useful for working with large NGS datasets in high performance computers (Supplementary Fig. S2; see also the DOMINO documentation). The regions identified or selected in DOMINO can be (i) directly used as markers with a particular depth of taxonomic resolution, (ii) utilized for their downstream PCR amplification in a broader taxonomic scope or (iii) used as suitable templates to optimized bait design for target DNA enrichment techniques.

2 Methods and implementation

2.1 DOMINO workflow

The DOMINO workflow consists of four main phases (Fig. 1) that can be run either using the DOMINO GUI or the extended command-line version (see the DOMINO manual in the DOMINO Web page). In both cases, the most relevant results from each phase are conveniently reported in the appropriate output files.

2.1.1 Input data and pre-processing phase

DOMINO accepts input sequence data files in two different formats, the 454 Pyrosequencing Standard Flowgram Format (SFF), and FASTQ format (Cock *et al.*, 2010). These input files can contain 454 or Illumina (single- or paired-end) raw reads from *m* taxa (the 'taxa panel'). The sequences from each taxon should be properly identified with a specific barcode (aka, tag, MID or index), or loaded in separate files, also appropriately named (see the DOMINO manual in the DOMINO Web site for details). DOMINO is designed to filter low quality, low complexity, contaminant and very short reads using either default or user-specified filtering parameters. Mothur, PRINSEQ, NGS QC toolkit, BLAST, as well as new Perl functions specifically written for DOMINO (DM scripts) are used to perform these tasks (Supplementary Table S1). DOMINO uses Mothur v1.32.0 (Schloss *et al.*, 2009) to extract reads from SFF files and store them in FASTQ format, which are subsequently converted to FASTA and QUAL files. Low quality or very short reads are trimmed, or definitely removed, using NGS QC Toolkit v2.3.1 (Patel and Jain, 2012). PRINSEQ v0.20.3 package (Schmieder and Edwards, 2011) is used to eliminate low complexity reads using the implemented DUST algorithm. Putative contaminant sequences, such as bacterial DNA frequently found in genomic samples (Leese *et al.* 2012), cloning vectors, adaptors, linkers and high-throughput library preparation primers, can also be removed using a DOMINO function that performs a BLAST search (BLAST v2.2.28) (Altschul *et al.*, 1990) against UniVec database (http://www.ncbi.nlm.nih.gov/tools/vecscreen/univec/) and/or against a user-supplied contaminant database (see the DOMINO manual).

2.1.2 Assembly phase

When working with just NGS reads, the program first applies an assembly-based approach; the pipeline is therefore optimized to work with genome partitioning methods in which the length of the size-selected (or enriched) fragments and the sequencing depth are enough to permit the assembly of a set of homologous fragments. For data from restriction-site associated DNA (RAD) sequencing and related methods see the Mapping/Alignment phase section. DOMINO performs separate assemblies, one for each panel taxon, using MIRA v4.0.2 (Chevreux, 1999), either with the pre-processed reads from the previous step or with those supplied by the user. Although the default parameter values vary in function of the particular sequencing technology, the majority of them are shared (see the DOMINO manual). In order to avoid including repetitive and chimeric regions, all contigs (and the corresponding reads) identified as repeats in the MIRA algorithm are discarded from the mapping/

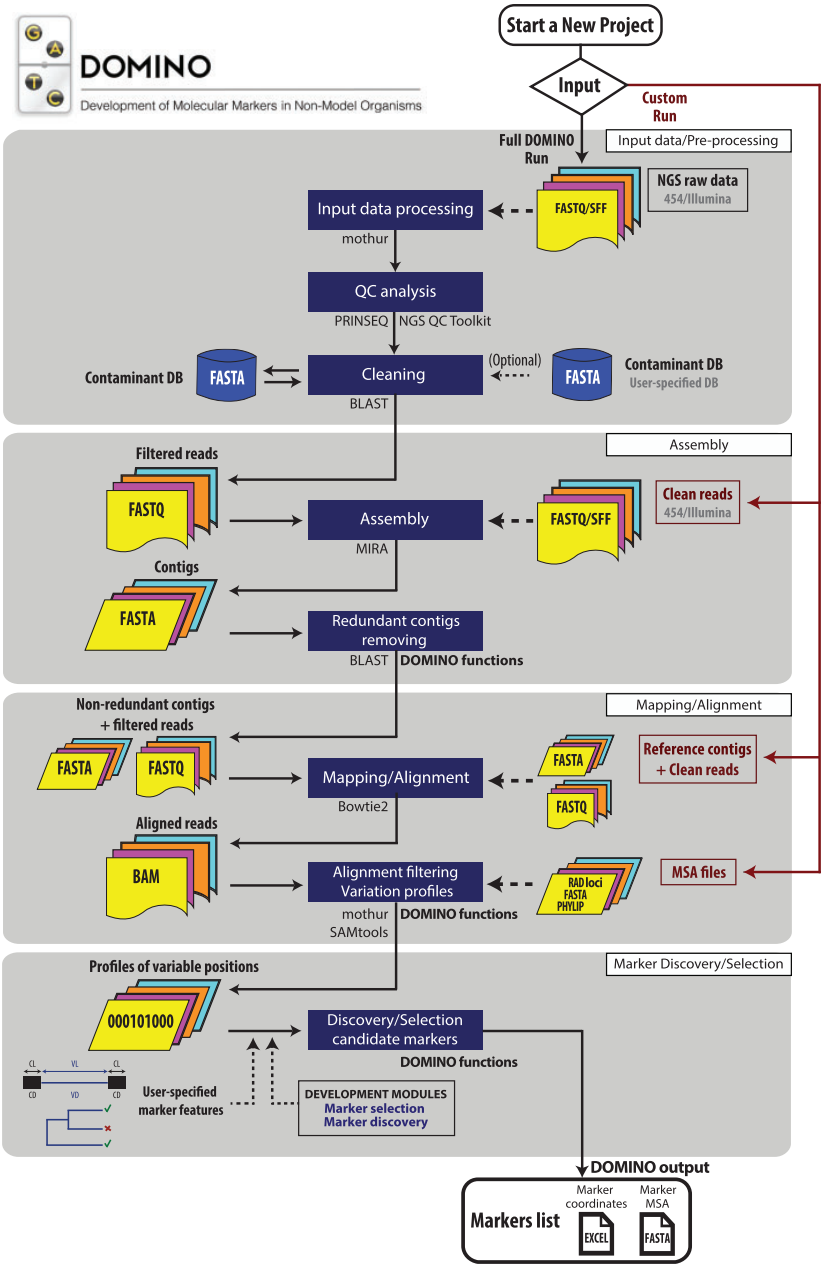


Fig. 1. Workflow showing the basic steps used to discover or select molecular markers with the DOMINO software

alignment phase (Chevreux, 1999). Since MIRA can generate redundant contigs because of polymorphic and paralogous regions, we have implemented a specific DOMINO function that performs a clustering of all contigs based on an all versus all contigs BLAST search to identify and remove such redundancies. The DOMINO command line version (see below) also includes an option to perform a second iterative assembly step using the software CAP3 (Huang, 1999). If selected, this option uses MIRA output sequences (contigs and singletons) as input for CAP3 under a relaxed parameter scheme.

2.1.3 Mapping/alignment phase

DOMINO uses *Bowtie2* (Langmead and Salzberg, 2012) to map the pre-processed reads from each taxon to the assembled contigs of the other $m-1$ taxa from the panel. Thus, in this step, DOMINO builds $m(m-1)$ sequence alignment/map files (SAM/BAM files. In the case of a panel of $m=4$ taxa, e.g. DOMINO will build $4 \times 3 = 12$ SAM/BAM files during this step. The reason behind this particular mapping strategy lies in the dissimilar performance of alignment/mapping algorithms depending on the divergence between the reads and the reference sequences. Immediately after generating BAM files, DOMINO

Fig. 2. Screenshot of Marker Discovery/Selection TAB included in the DOMINO GUI

removes all unmapped contigs and multi-mapping reads. This step is critical to avoid alignment artifacts, which can create false positive markers (i.e. sequence regions with misleading high levels of nucleotide diversity). The contigs with an unusually large number of aligned reads, which can correspond to repetitive regions, are also removed (they are not suitable for designing single copy markers) Later, DOMINO will build one pileup file per each BAM/SAM file using the SAMtools v0.1.19 suite (*mpileup* option) (Li *et al.*, 2009).

Since sequencing errors might have a great effect on the marker selection, DOMINO incorporates their own functions for detecting and masking putative sequencing errors, which apply a very conservative criterion for variant calling. First, to avoid the calling of spurious nucleotide variants in low sequencing coverage experiments (i.e. erroneously assigned variants fixed between the taxa from the panel), DOMINO mask the information from positions with only one read mapped to the reference. Furthermore, sequencing errors may also inflate the number of called polymorphisms under the Polymorphic Variants option in the marker identification/selection phase. To avoid such undesirable effect, DOMINO incorporates a similar conservative criterion to use only highly credible polymorphisms. Under the Polymorphic Variant option, DOMINO will assume that each taxon represents a diploid individual; for positions with eight or more reads mapped, DOMINO discards those polymorphic variants in which the frequency of the minor allele is significantly lower than the expected under error free data (hence, in absence of sequencing errors the distribution of observed nucleotide counts at each position would follow a binomial distribution). For lower coverage values, DOMINO will use the information of a polymorphic variant only if the allele with the minor frequency is present in two or more reads. This testing procedure, applied independently for each position within each species, will likely discard some true polymorphic sites; this variant calling approach, however, makes DOMINO highly conservative in detecting true markers when including polymorphisms in the analysis (i.e. DOMINO will use only highly confident within-species segregating variants for the marker Discovery/Selection phase). Ambiguity codes, either introduced by MIRA assembler in contig sequences or present in user-supplied reference sequences or MSA, are also considered by DOMINO to decide whether a position is or not variable.

After applying all the above-mentioned post-mapping filters, DOMINO combines the variation profiles (arrays with the information about the state of each position, conserved or variable between taxa pairs) obtained from each of the $m-1$ pileup files including the same reference sequence (i.e. the same taxon), into a single multiple taxa variation profile (MTVP). Since each of these references will be likely fragmented in i contigs, DOMINO will build $i \times m$ MTVP per taxon. Each of these MTVP will be independently scanned for regions containing candidate markers in the next phase. If the user provides reference sequences from a single taxon (e.g. a genome draft), plus the reads from the m different taxa, the program builds only one MTVP set (one per contig or scaffold in the supplied reference). On the other hand, if the input includes a single or multiple pre-computed MSA instead of NGS data, DOMINO skips the alignment/mapping phase and directly generates the single MTVP set (one per aligned region). In this point, the program accepts MSA files in FASTA (multiple FASTA files, one per linked region), PHYLIP (multiple PHYLIP files, one per linked region, or one multi PHYLIP file with the alignment of all regions) and pyRAD LOCI (*.loci files generated by the program pyRAD; Eaton, 2014) and STACKS fasta (batch_X.fa output files generated from the population analyses in the program STACKS; Catchen *et al.*, 2011) output files.

2.1.4 Marker discovery/selection phase

Each MTVP generated in the previous step is either scanned for the presence of candidate marker regions using a sliding window approach (DOMINO marker discovery module), or used to select markers (with the desired features) among the MSA loaded in the previous tab (DOMINO marker selection module). In the first case, a specific DOMINO function searches for sequence regions of desired length (Variable region Length, VL), showing the minimum level of variation indicated by the user (Variable region Divergence, VD). DOMINO can also restrict that this variable region was flanked (or not) by highly conserved regions (Conserved region Divergence, CD) of a predefined length (Conserved region Length, CL); an information useful to further design PCR primers. Moreover, DOMINO can strictly restrict the search to a particular set of taxa (from the panel), or just specify the minimum number of taxa required to be covered by the marker (by changing the Minimum number of Covering Taxa parameter; MCT $< m$). As indicated, DOMINO can use or not the information from polymorphic sites. An appropriated combination of selected taxa and MCT and VD parameter values will allow the user select a large set of informative markers suitable to be applicable across a wide range of taxa. In the second case, the DOMINO selection module allows directly selecting the most informative markers among the loaded by the user in the same way and with the same personalized features described above. For RAD loci, a particular range of variable positions (VP) between the closest taxa (instead of the VD parameter) must be specified. This option allows selecting informative RAD loci while excluding those exhibiting anomalous high levels of variation, which might reflect RAD tag clustering errors. The specific selection of a set of loci/MSA that meet some specific phylogenetic criteria using the DOMINO selection module can be very helpful to further design probes for different target enrichment techniques, including the enrichment of specific RAD segments using hyRAD (Suchan *et al.*, 2016).

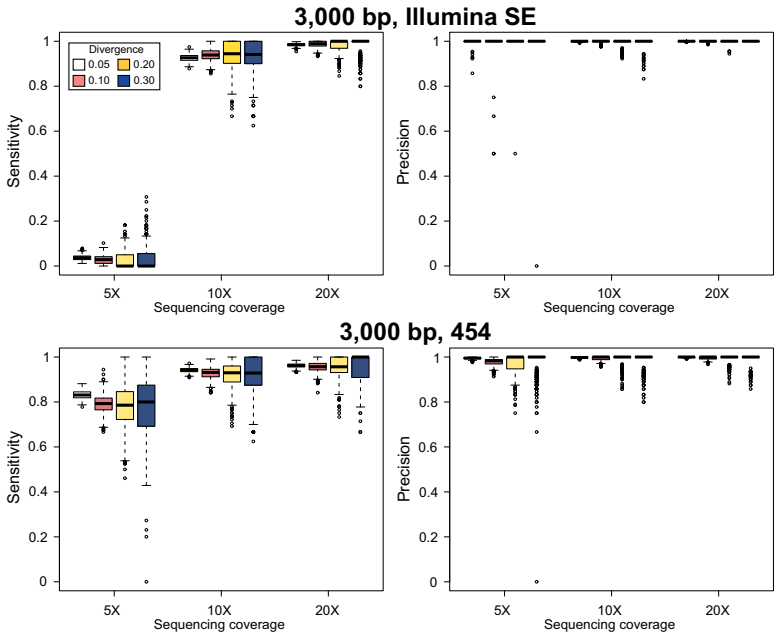


Fig. 3. Sensitivity and precision estimates for simulated datasets of 100 fragments of 3 kb after their *in silico* sequencing with Illumina and Roche-454 technologies

After the last phase, DOMINO reports the list the genomic regions (and their coordinates) or MSA that meet the selection criteria, along with the corresponding MSA of these regions for the selected taxa. Since DOMINO can work with more than one MTVP set (*m* in a full DOMINO run), some of the markers found in MTVP based on different reference taxa may be redundant (they can cover the same genomic region, although with different coordinates; see Mapping/Alignment phase section), while other can be found only in one particular profile. To avoid reporting redundant information, we have implemented a BLAST-based function to collapse these maker sequences, only reporting unique markers. To maximize the probability of finding informative markers, the final list of candidates under the DOMINO marker discovery module can include overlapped regions that fulfill the specified characteristics. Operationally, all regions that meet the criteria for being considered a candidate marker (after moving the scanning window five or more base pairs) are listed as different markers in the final output. In this way, users can choose the best marker to be used directly for further analyses or the more appropriated region of each contig to be PCR amplified and sequenced in additional focal species (i.e. the best marker from each linked block).

2.2 DOMINO GUI

DOMINO can be run either in the command prompt, by setting a large set of command line options, or using the GUI specifically developed to facilitate its use to non-experts in NGS bioinformatics tools (Fig. 2; see also the DOMINO manual for details). The DOMINO GUI is a cross-platform application that allows the user to interactively set marker selection criteria by tuning the most important parameters and options available in the command prompt version. It should be noted that for huge NGS datasets (which require substantial amounts of computational resources) a full DOMINO run using the GUI version is not recommendable. In this case, the user can either run DOMINO under the command line version using high performance computer clusters or, take advantage of the custom run options available in the GUI version to enter in DOMINO partially processed data, e.g. pre-processed reads, assemblies or alignment files (SAM/BAM) obtained with other memory-efficient software (Supplementary Table S2).

2.3 System and availability

The GUI was built using the cross-platform library and user interface framework Qt (https://www.qt-project.org/) based on C++ scripting language. Since most of the functions specifically developed for this work are implemented in Perl scripting language, users need to install first a recent version of Perl (version 5.12 or higher; http://www.perl.org/). The source code, the documentation and some example data files are freely distributed under the GNU GPL software license at: http://www.ub.edu/softevol/domino.

3 Results and conclusions

3.1 Computer simulations

We conducted an exhaustive computer simulation study to assess the performance of DOMINO in detecting informative markers (i.e. simulated regions that meet specific marker selection criteria) from NGS data. For that, we emulated an RRL-like experiment of four closely related species exhibiting different levels of nucleotide divergence among them and incorporating substitution rate heterogeneity across sites to create genuine informative markers. The topology of the species tree used for the simulations was fixed (Supplementary Fig. S1). In each replicate, we generated an independent RRL-like dataset of 100 fragments, of different length (3 or 10 kb) each. The nucleotide sequences were simulated with the program evolver, included in the PAML v4.7 package (Yang, 1997, 2007), using 0.1, 0.15, 0.20 or 0.30 substitutions per site between the two most

divergent sequences, under the Jukes and Cantor (1969) substitution model with substitution rate heterogeneity across sites (modeled as a discrete gamma with 10 categories and $\alpha = 0.01$). For each replicate, we simulated a complete NGS experiment in the Roche-454 (reads with an average length of ~400 bp), and the Illumina HiSeq2000 platforms (average length of 101 bp; single and paired-ends) using the ART v2.5.8 program (Huang *et al.*, 2012) with default parameters and three different sequencing coverage values (5×, 10× and 20×). We generated 500 simulation replicates for each of the 48 possible scenarios (i.e. for each combination RRLs fragment length, divergence, sequencing platform and coverage values), resulting in a total 27 000 DOMINO runs, which took roughly 80 000 CPU h.

Using the DOMINO marker discovery module under the command line version, we first traced the number and the location of the regions that meet the selection criteria present in each simulated fragment previous to emulate their NGS sequencing (true markers; TNM). Subsequently, for each dataset, we execute a full run of our program using the simulated NGS reads to obtain the list of candidate markers (detected markers; DNM) for each scenario. For this experiment, we define an informative marker as a variable region of 600 bp (VL = 600), present in all four species (MCT = 4), showing at least 0.01 nucleotide substitutions per site between any pair of species (VD = 0.01), and flanked by two highly conserved regions of 60 or more bp long (CL = 60; only one substitution across species was permitted; CD = 1). We assessed the performance of DOMINO in detecting the TNM by measuring the sensitivity and precision in each replicate and plotting their distribution across the 500 replicates (Fig. 3; Supplementary Material).

We found that DOMINO pipeline has a high sensitivity in detecting the existing TNM, yielding averages of true positive rates values >0.9 for Illumina reads and when coverage values are equal or higher than 10× (Fig. 3). As expected, lower coverage values (5×) result in a reduction of the sensitivity estimates; in this case, DOMINO runs using 454 long reads outperforms those using Illumina short reads (e.g. average sensitivities close to 0.8 for the 454 under all tested nucleotide divergences in 3 kb fragments). Noticeably, we found that DOMINO show high sensitivities even for relative high divergence levels (up to 0.3 substitutions per site between the two more diverged taxa); in this case, the program performs slightly better when using short reads as input. In the light of this high sensitivity, precision becomes a critical aspect to be considered for further successful marker discovery. We found that DOMINO also detects TNM regions with high precision (most values are close to 1 regardless of the condition), yielding very few number of false positives. The performance of DOMINO when using reads from larger library fragments (10 kb) is very similar to that of the observed for 3 kb (Supplementary Fig. S2).

3.2 Application to empirical data

To illustrate the utility of DOMINO on real biological data, we performed a RRL sequencing experiment (using 454 reads; see Supplementary Material for details), which allow running all phases of the application and the DOMINO marker identification module, from raw reads to marker selection. We used four individuals (panel with four taxa) belonging to the spider family *Nemesiidae* (Araneae) for this analysis (Supplementary Material, Fig. S3). We identified many candidate regions that fulfill the requested marker characteristics (Supplementary Tables S3–S6), and tested the suitability of six of them by PCR amplification and Sanger sequencing in a larger panel that also included other 14 phylogenetically related species (focal species). The obtained phylogenetic tree not only recovered the expected relationships among the taxa from the panel but also demonstrates that the sequenced markers are useful to establish the phylogenetic relationships of the focal ones (Supplementary Fig. S4).

3.3 Conclusions

DOMINO will assist researches working with non-model organisms in the development of molecular markers for DNA variation studies. First, it allows obtaining a list of 'personalized' markers that meet user specific criteria without the mandatory need of a reference genome, which will improve their application from highly specific taxonomic scopes to more wide phylogenetic ranges. Second, its output alignment files, jointly with the information about markers coordinates and features provided by the program, can be either directly utilized in variation studies, or used as a templates for further downstream PCR amplification or target DNA enrichment probe design. Third, the DOMINO GUI makes this application accessible and easy-to-use to non-experts in the bioinformatics of NGS data handling and analysis. Finally, DOMINO is open cross-platform software that can be straightforwardly adapted to other pipelines or used in high performance computers. Although current version of the program works with raw reads of a limited number of reduction representation schemes (e.g. DOMINO cannot process raw reads from RAD- or RNA-Seq approaches) and sequencing platforms (Illumina short and 454 long reads), the modular structure of DOMINO will allow easily expanding the software to accept NGS data from other sources.

Funding

This work was supported by the Ministerio de Educación y Ciencia of Spain (No. BFU2010-15484 and No. CGL2013-45211 to J.R. and No. CGL2012-36863 to M.A.A.).

References

Altschul,S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Brito,P.H. and Edwards,S.V. (2009) Multilocus phylogeography and phylogenetics using sequence-based markers. *Genetica*, **135**, 439–455.

Catchen,J.M. *et al.* (2011) Stacks: building and genotyping Loci de novo from short-read sequences. *G3 (Bethesda)*, **1**, 171–182.

Chamala,S. *et al.* (2015) MarkerMiner 1.0: A new application for phylogenetic marker development using angiosperm transcriptomes. *Appl. Plant Sci.*, **3**, 1400115.

Chevreux,B. *et al.* (1999) Genome sequence assembly using trace signals and additional sequence information. *Comput. Sci. Biol. Proc. German Conf. Bioinform.*, **99**, 45–56.

Cock,P.J.A. *et al.* (2010) The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res.*, **38**, 1767–1771.

Cruaud,A. *et al.* (2016) *RADIS*: analysis of *RAD*-seq data for interspecific phylogeny. *Bioinformatics*, doi:10.1093/bioinformatics/btw352.

Eaton,D.A.R. (2014) PyRAD: assembly of de novo RADseq loci for phylogenetic analyses. *Bioinformatics*, **30**, 1844–1849.

Huang,W. *et al.* (2012) ART: a next-generation sequencing read simulator. *Bioinformatics*, **28**, 593–594.

Huang,X. (1999) CAP3: a DNA sequence assembly program. *Genome Res.*, **9**, 868–877.

Jukes,T.H. and Cantor,C.R. (1969). Evolution of protein molecules. In: Munro, H.N. (ed.) *Mammalian Protein Metabolism*. Academic Press, New York, pp. 21–132.

Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.

Leese,F. *et al.* (2012) Exploring Pandora's box: potential and pitfalls of low coverage genome surveys for evolutionary biology. *PLoS One*, **7**, e49202.

Lemmon,E.M. and Lemmon,A.R. (2013) High-throughput genomic data in systematics and phylogenetics. *Annu. Rev. Ecol. Evol. Syst.*, **44**, 99–121.

Li,H. *et al*. (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Maddison,W.P. *et al*. (1997) Gene trees in species trees. *Syst. Biol*., **46**, 523–536.

Mayer,C. *et al*. (2016) BaitFisher: a software package for multispecies target DNA enrichment probe Design. *Mol. Biol. Evol*., **33**, 1875–1886.

Patel,R.K. and Jain,M. (2012) NGS QC toolkit: a toolkit for quality control of next generation sequencing data. *PLoS One*, **7**, e30619.

Schloss,P.D. *et al*. (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol*., **75**, 7537–7541.

Schmieder,R. and Edwards,R. (2011) Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, **27**, 863–864.

Suchan,T. *et al*. (2016) Hybridization Capture Using RAD Probes (hyRAD), a new tool for performing genomic analyses on collection specimens. *PLoS One*, **11**, e0151651.

Yang,Z. (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci*., **13**, 555–556.

Yang,Z. (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol*. **24**, 1586–1591.

Artículo 2

Supplemental Information

Computer Simulations

We assessed the performance of DOMINO in detecting the genuine makers present in the simulated sequences (i.e., the regions in these sequences that meet the marker selection criteria to be further specified in DOMINO) by measuring the sensitivity and precision in each independent replicate and by plotting their distribution in the 500 replicates (Figure 3; Supplementary Fig. S2), being:

Sensitivity = TP / (TP+FN) and
Precision = TP / (TP+FP),

where, TP (true positives) indicates the number of simulated regions meeting the criteria that are correctly identified, FN (false negatives) indicates the number of these genuine marker regions that were not detected by the DOMINO discovery module after read assembly and mapping and FP (false positives) indicates the number of simulated regions incorrectly identified as markers.

Empirical Data

As an example of the application of DOMINO to empirical NGS data, we used the program to search for highly informative markers suitable to be used in phylogenetic studies in the spider family *Nemesiidae* (Araneae, Mygalomorphae). *Nemesia* and. For that, we chose a taxa panel of four species, consisting in three *Nemesia* (Audouin, 1826) and one *Iberesia* (its putative sister group, *Iberesia* Decae & Cardoso, 2006) samples (Supplementary Fig. S3). Specifically, we included in the panel two individuals from two different populations of *Nemesia raripilia* (Simon, 1914; *Nemesia raripilia* populations 061 and 079; collected in Coll de les Tres Creus, Sant Llorenç del Munt i Serra de l'Obac Natural Park, Barcelona, Spain), one individual of a different unidentified *Nemesia* species (*Nemesia* sp population 043 from Cabrera de Mar - Barcelona, Spain) and one individual of *Iberesia brauni* (L. Koch, 1882) (locality 098, collected in Port de Soller, Majorca, Spain).

We digested the genomic DNA of these four individuals with the eight-cutter restriction enzyme *Not*I (restriction site 5' GC/GGCCGC 3'). Fragments ranging from 2.5 to 3 kb were selected to construct the representation libraries by excising the corresponding bands from

the agarose gel, following by purification with the Qiagen Gel Extraction Kit (Qiagen). The Illustra GenomiPhi V2 Amplification Kit (GE Healthcare) was used to increase the amount of recovered DNA following the manufacturer's specifications. The amplified DNA was treated with RNase (Qiagen), and subsequently purified using the Qiagen PCR Purification Kit (Qiagen). The purified DNA sample was quantified with the Qubit® 2.0 Fluorometer (QBIT Assays, Invitrogen). The sequencing was conducted on a 454/Roche GS-FLX Titanium, with each sample individually tagged using the Roche 454 Pyrosequencing MID (Multiplex Identifier DNA) tags, and using 1/2 picotitre plate. Assuming that the restriction sites are randomly distributed across the genome, we estimated that the libraries represent ~0.007 of each genome (~21 Mbp, assuming a ~3Gbp genome). We obtained ~425,000 reads, and used DOMINO to pre-process this raw data, and to the *de novo* assembly each RRL fragment Supplementary Table S3). We searched, applying different parameter settings, for regions candidate to encompass suitable makers (Supplementary Table S4). We tested the suitability of six of the identified candidate markers by PCR amplification in individuals from the same four species of the panel along with other 14 phylogenetically related species (focal species). After the Sanger sequencing of each fragment, we built an MSA per each marker region using the program Mafft (Katoh *et al.*, 2002; Katoh and Standley, 2013), which were further concatenated to obtain the final MSA used for the phylogenetic analysis in RAxML version 8 (Stamatakis, 2014) (Supplementary Fig. S4).

DNA Deposition

The data have been deposited in the NCBI BioProject database (https://www.ncbi.nlm.nih.gov/bioproject/), with number: PRJNA327555

References

Katoh,K. et al. (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res., **30,** 3059–66.

Katoh,K. and Standley,D.M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol., **30,** 772–80.

Stamatakis,A. (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics, **30,** 1312–3.

Supplementary Figures

Figure S1. Topology and relative branch lengths of the tree used to simulate sequence data. In this example, we show the tree used to simulate sequences with 0.2 nucleotide substitutions per site between the two more distant taxa.

Figure S2. Sensitivity and precision estimates for data sets of 100 fragments of 10 kb after their in silico sequencing of simulated fragments with Illumina and Roche-454 technologies.

Figure S3. Maximum likelihood phylogenetic tree showing the relationships among the four species included in the taxa panel. This tree was built using a multiple sequence alignment of COI sequences. Branch lengths are in nucleotide substitutions per site.

Figure S4. Maximum likelihood phylogenetic tree showing the relationships among the four taxa included in the panel and other 11 focal species. The tree was built using a concatenated multiple sequence alignment with the sequence information of six of the markers identified by DOMINO.

Figure S1

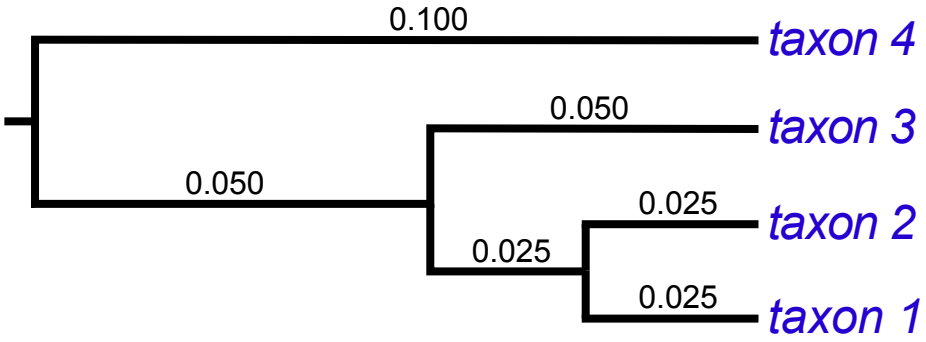


Figure S2

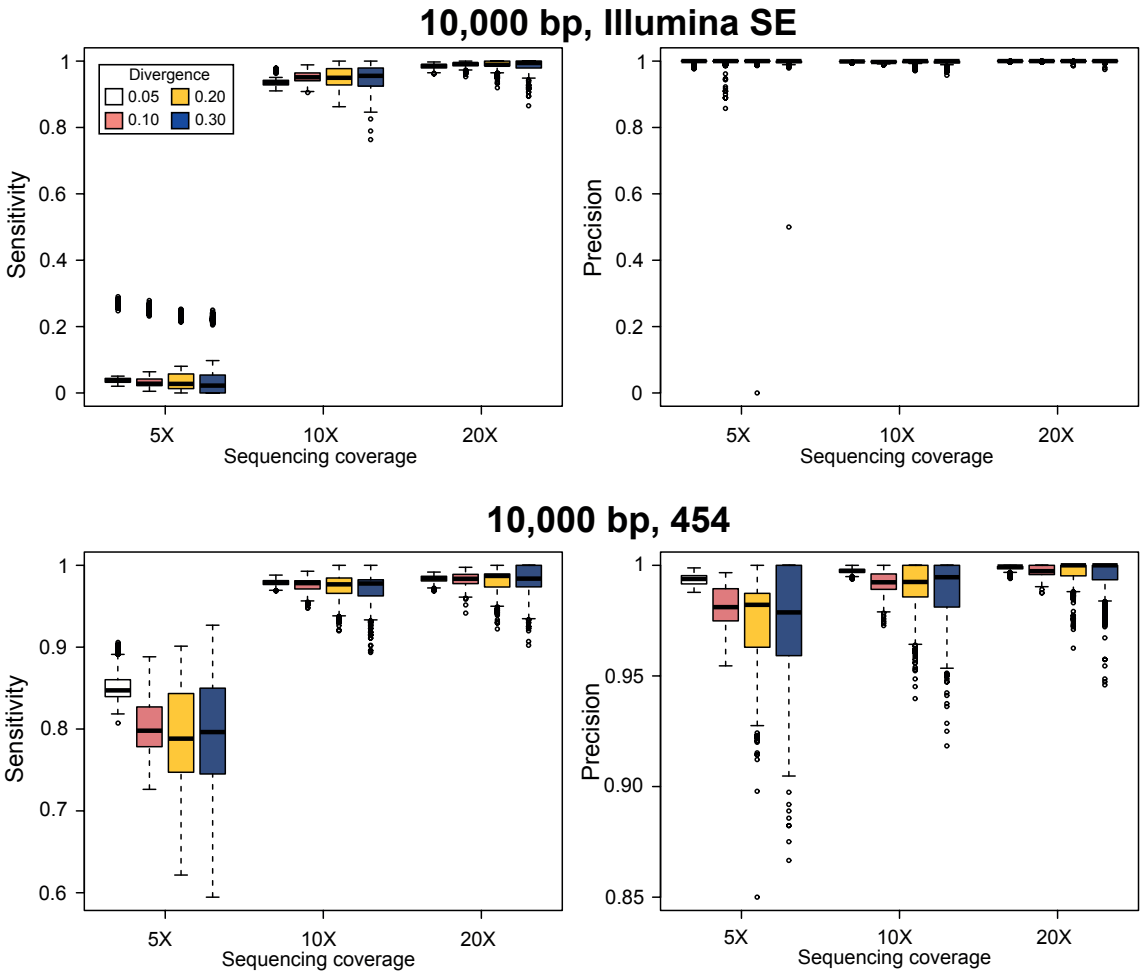


Figure S3

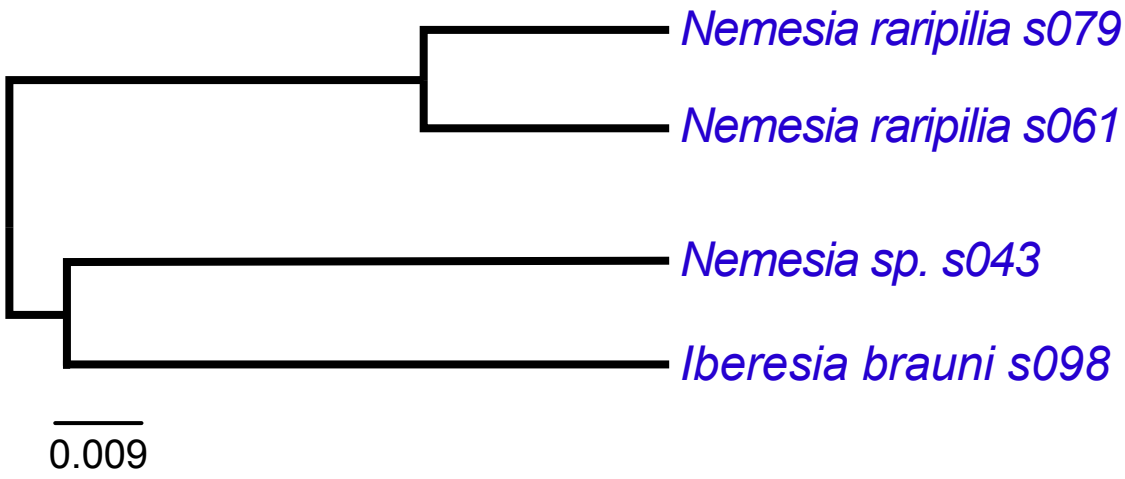
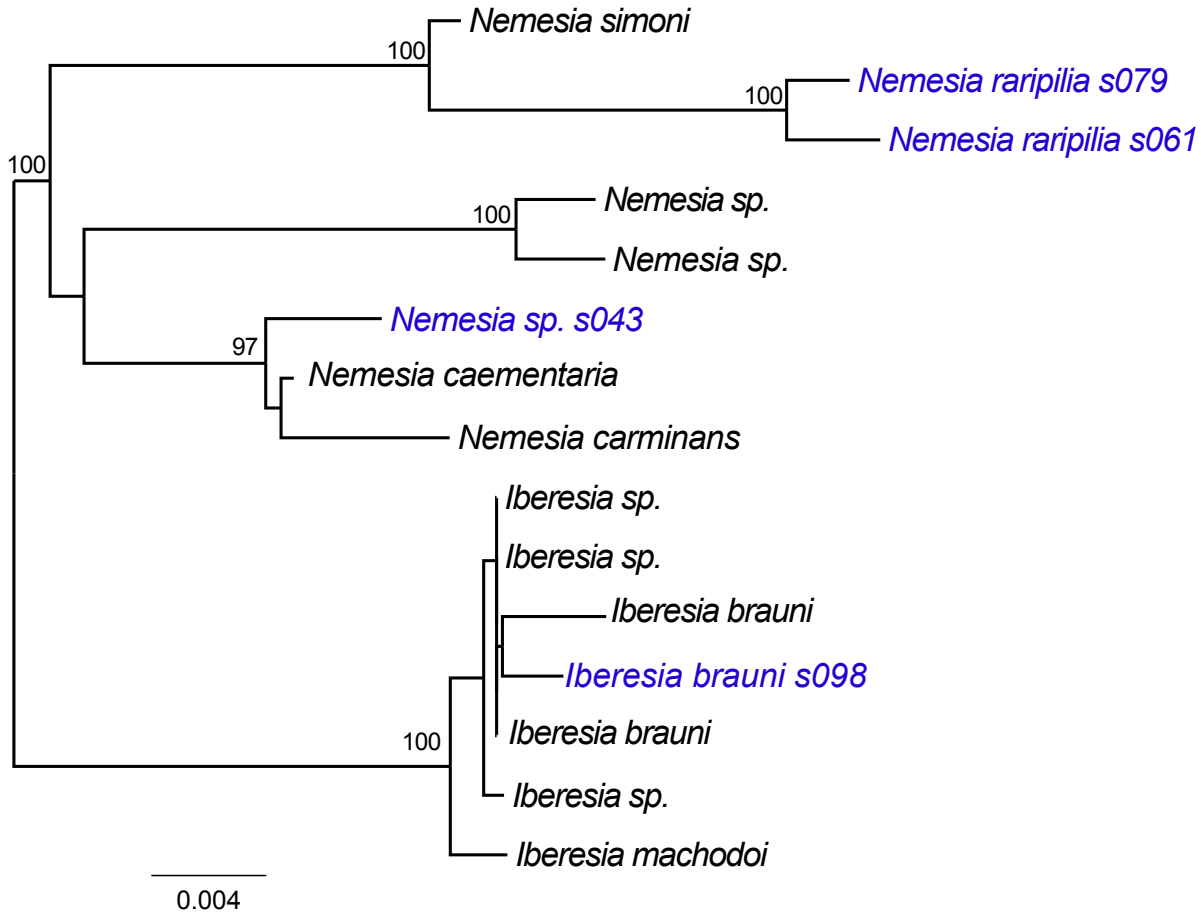


Figure S4



Supplementary tables

Table S1. Summary of software, algorithms and Perl functions included in DOMINO.

Table S2. RAM, CPU and execution times of different DOMINO runs and data sets.

Table S3. Summary statistics of the analysis of the NGS data from the four *Nemesiidae* species

Table S4. Results of the DOMINO maker discovery module using the NGS data from the four *Nemesiidae* species

Table S5. Summary statistics of the analysis of a subset of 4,000 reads (NGS data from the four *Nemesiidae* species).

Table S6. Results of the DOMINO maker discovery module using the subset of 4,000 reads (NGS data from the four *Nemesiidae* species).

Documentation

http://www.ub.edu/softevol/domino/

Artículo 3

TRUFA: a User-friendly Web server for *de novo* RNA-seq analysis using cluster computing.

Etienne Kornobis, Luis Cabellos, Fernando Aguilar, Cristina Frías-López, Julio Rozas, Jesús Marco and Rafael Zardoya

2015, Evolutionary Bioinformatics, 11: 97-104

TRUFA: A User-Friendly Web Server for *de novo* RNA-seq Analysis Using Cluster Computing

Libertas Academica
FREEDOM TO RESEARCH

Etienne Kornobis[1], Luis Cabellos[2], Fernando Aguilar[2], Cristina Frías-López[3], Julio Rozas[3], Jesús Marco[2] and Rafael Zardoya[1]

[1]Departamento de biodiversidad y biología evolutiva, Museo Nacional de Ciencias Naturales MNCN (CSIC), Madrid, Spain. [2]Instituto de Física de Cantabria, IFCA (CSIC-UC), Edificio Juan Jordá, Santander, Spain. [3]Departament de Genètica and Institut de Recerca de la Biodiversitat (IRBio), Universitat de Barcelona, Barcelona, Spain.

ABSTRACT: Application of next-generation sequencing (NGS) methods for transcriptome analysis (RNA-seq) has become increasingly accessible in recent years and are of great interest to many biological disciplines including, eg, evolutionary biology, ecology, biomedicine, and computational biology. Although virtually any research group can now obtain RNA-seq data, only a few have the bioinformatics knowledge and computation facilities required for transcriptome analysis. Here, we present TRUFA (TRanscriptome User-Friendly Analysis), an open informatics platform offering a web-based interface that generates the outputs commonly used in *de novo* RNA-seq analysis and comparative transcriptomics. TRUFA provides a comprehensive service that allows performing dynamically raw read cleaning, transcript assembly, annotation, and expression quantification. Due to the computationally intensive nature of such analyses, TRUFA is highly parallelized and benefits from accessing high-performance computing resources. The complete TRUFA pipeline was validated using four previously published transcriptomic data sets. TRUFA's results for the example datasets showed globally similar results when comparing with the original studies, and performed particularly better when analyzing the green tea dataset. The platform permits analyzing RNA-seq data in a fast, robust, and user-friendly manner. Accounts on TRUFA are provided freely upon request at https://trufa.ifca.es.

KEYWORDS: transcriptomics, RNA-seq, *de novo* assembly, read cleaning, annotation, expression quantification

CITATION: Kornobis et al. TRUFA: A User-Friendly Web Server for *de novo* RNA-seq Analysis Using Cluster Computing. *Evolutionary Bioinformatics* 2015:11 97–104 doi: 10.4137/EBO.S23873.

RECEIVED: January 09, 2015. **RESUBMITTED:** March 09, 2015. **ACCEPTED FOR PUBLICATION:** March 16, 2015.

ACADEMIC EDITOR: Jike Cui, Associate Editor

TYPE: Technical Advance

FUNDING: This work was partially funded with Spanish Ministry of Science and Innovation grants CGL2010–18216 and CGL2013–45211-C2–2-P to RZ and CGL2013–45211-C2–1-P to JR. JR was partially supported by ICREA Academia (Generalitat de Catalunya; Spain). The authors confirm that the funder had no influence over the study design, content of the article, or selection of this journal.

COMPETING INTERESTS: Authors disclose no potential conflicts of interest.

CORRESPONDENCE: ekornobis@gmail.com

Paper subject to independent expert blind peer review by minimum of two reviewers. All editorial decisions made by independent academic editor. Upon submission manuscript was subject to anti-plagiarism scanning. Prior to publication all authors have given signed confirmation of agreement to article publication and compliance with all applicable ethical and legal requirements, including the accuracy of author and contributor information, disclosure of competing interests and funding sources, compliance with ethical requirements relating to human and animal study participants, and compliance with any copyright requirements of third parties. This journal is a member of the Committee on Publication Ethics (COPE).

Published by Libertas Academica. Learn more about this journal.

Introduction

Since the introduction of the RNA-seq methodology around 2006,[1–6] studies based on whole transcriptomes of both model and non-model species have been flourishing. RNA-seq data are widely used for discovering novel transcripts and splice variants, finding candidate genes, or comparing differential gene expression patterns. The applications of this technology in many fields are vast,[1,7] including researches on, eg, splicing signatures of breast cancer,[8] host–pathogen interactions,[9] the evolution of the frog immunome,[10] the plasticity of butterfly wing patterns,[11] the study of conotoxin diversity in *Conus tribblei*,[12] and the optimization of trimming parameters for *de novo* assemblies.[13]

Despite the tremendous decrease in sequencing costs, which allows virtually any laboratory to obtain RNA-seq data, transcriptome analyses are still challenging and remain the main bottleneck for the widespread use of this technology. User-friendly applications are scarce,[14] and the post-analysis of generated sequence data demands appropriate bioinformatics know-how and suitable computing infrastructures.

When a reference genome is available, which is normally the case for model system species, a reference-guided assembly is preferable to a *de novo* assembly. However, an increasing number of RNA-seq studies are performed on non-model organisms with no available reference genome for read mapping (particularly those studies focused on comparative transcriptomics above the species level), and thus require a *de novo* assembly approach. Moreover, when a reference genome is available, combining both *de novo* and reference-based approaches can lead to better assemblies.[15,16] Analysis pipelines encompassing *de novo* assemblies are varied, and generally include steps such as cleaning and assembly of the reads, annotation of transcripts, and gene expression quantification.[16] A variety of software programs have been developed to perform different steps of the RNA-seq analysis,[17–19] but most of them are computationally intensive. The vast majority of these programs run solely with command lines. Processing the data to connect one step to the next in RNA-seq pipelines can be cumbersome in many instances, mainly due to the variety of output formats produced and the postprocessing needed to accept them further as input.

Moreover, as soon as a large computing effort is required, interactive execution is usually not feasible and an interface with the underlying batch systems used in clusters or supercomputers is needed. In order to provide users with such a bioinformatics tool that solve the above-mentioned problems, we have developed TRUFA (TRanscriptomes User-Friendly Analysis), an informatics platform for RNA-seq data analysis, which runs on the ALTAMIRA supercomputer at the Instituto de Fisica de Cantabria (IFCA), Spain.[20] The platform is highly parallelized both at the pipeline and program level. It can access up to 256 cores per execution instance for certain components of the pipeline. On top of allowing the user to obtain results in a relatively short time thanks to HPC (high-performance computing) resources, TRUFA is an integrative and graphical web tool for performing the main and most computationally demanding steps of a *de novo* RNA-seq analysis.

The first step of a *de novo* RNA-seq analysis consists in assessing data quality and cleaning raw reads. The output of a next-generation sequencing (NGS) reaction contains traces of polymerase chain reaction (PCR) primers and sequencing adapters as well as poor-quality bases/reads. Hence, it is advised to perform read trimming, which has been shown to have a positive effect on the rest of the RNA-seq analysis,[21] although parameter values for such trimming have to be optimized.[13]

Once reads have been cleaned, they are assembled into transcripts, which are subsequently categorized into functional classes in order to understand their biological meaning. Finally, it is possible to perform expression quantification analyses by estimating the amount of reads sequenced per assembled transcript and taking into account that the number of reads sequenced theoretically correlates with the number of copies of the corresponding mRNA *in vivo*.[6] All the above-mentioned steps in the RNA-seq analysis pipeline are included in TRUFA and correspond to distinct sections in the web-based user interface (see Figs. 1 and 2). For each step, the options available are those that are either critical to the analysis or, to our knowledge, the most widely used in the literature.

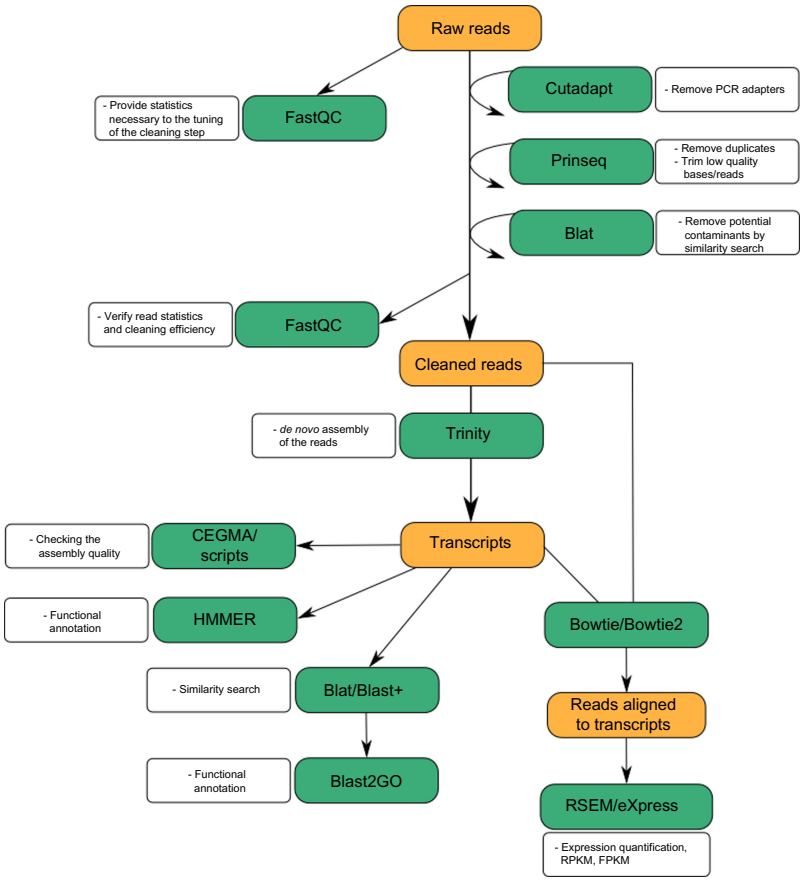


Figure 1. Overview of the TRUFA pipeline.

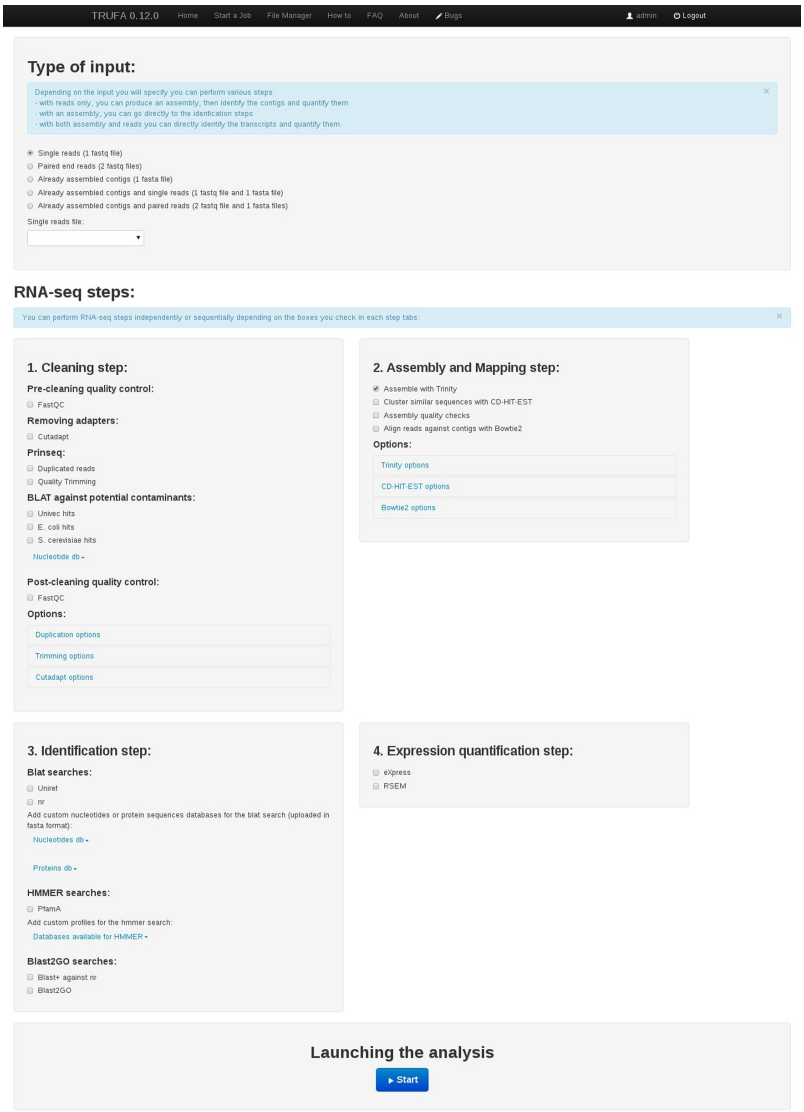


Figure 2. Snapshot of the TRUFA web page for running RNA-seq analysis.

There are several online platforms already available to perform different parts of a RNA-seq analysis. For example, Galaxy (https://usegalaxy.org/)[22] allows analyzing RNA-seq data with a reference genome (using Tophat[23] and Cufflinks[24]), whereas GigaGalaxy (http://galaxy.cbiit.cuhk.edu.hk/) can produce *de novo* assemblies using SOAPdenovo.[25] Another transcriptome analysis package integrated in Galaxy, Oqtans,[26] provides numerous features including *de novo* assembly with Trinity, read mapping, and differential expression. Nonetheless, to our knowledge, GigaGalaxy or Oqtans do not perform *de novo* annotations. Conversely, Fastannotator[27] is a platform specialized in transcript annotations using Blast2GO,[28] PRIAM,[29] and domain identification pipelines, but does not perform other steps of the RNA-seq analysis.

The TRUFA platform has been designed to be interactive, user-friendly, and to cover a large part of a RNA-seq analysis pipeline. Users can launch the pipeline from raw or cleaned Illumina reads as well as from already assembled transcripts. Each of the implemented programs (Table 1) can be easily integrated into the analysis and tuned depending on the needs of the user. TRUFA provides a comprehensive output, including read quality reports, cleaned read files, assembled transcript

Table 1. List of available software on TRUFA.

RNA-SEQ STEPS	AVAILABLE PROGRAMS	VERSIONS
Read cleaning	PRINSEQ	0.20.3
	CUTADAPT	1.3
	BLAT	v.35
Assembly and mapping	Trinity	r2012–06–08
	CD-HIT	4.5.4
	CEGMA	2.4
	Bowtie	0.12.8
	Bowtie2	2.0.2
Annotation	BLAT	v.35
	HMMER	3.0
	Blast+	2.2.28
	Blast2GO	2.5.0
Expression quantification	RSEM	1.2.8
	eXpress	1.5.1

files, assembly quality statistics, Blast, Blat, and HMMER search results, read alignment files (BAM files), and expression quantification files (including values of read counts, expected counts, and TPM, ie, transcripts per million[30]). Some outputs can be directly visualized from the web server, and all outputs can be downloaded in order to locally perform further analyses such as single nucleotide polymorphisms (SNPs) calling and differential expression quantification. The platform is mainly written in Javascript, Python, and Bash. The source code is available at Github (https://github.com/TRUFA-rnaseq). The long-term availability of the TRUFA web server (and further developed versions) is ensured given that it is currently installed in the ALTAMIRA supercomputer, a facility integrated in the Spanish Supercomputing Network (RES). The number of users is currently not limited and accounts are freely provided upon request.

Implementation

The overall workflow of TRUFA is shown in Figure 1. The input, output, and different components of the pipeline are the following:

Input. Currently, the input data accepted by TRUFA includes Illumina read files and/or reads already assembled into contigs. Read files should be in FASTQ format and can be uploaded as gzip compressed files (reducing uploading times). Reads from the NCBI SRA databases can be used but should be first formatted into FASTQ format using, eg, the SRA toolkit.[31] Already assembled contigs should be uploaded as FASTA files. Other FASTA files and HMM profiles can be uploaded as well for custom blast-like and protein profile-based transcript annotation steps, respectively. Thus far, no data size limitation is set.

Pipeline. Several programs can be called during the cleaning step (Table 1). The program FASTQC (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc) has been implemented to assess the quality of raw reads and give the statistics necessary to tune cleaning parameters (Fig. 1). After the quality of the data is determined, CUTADAPT[32] and PRINSEQ[33] allow, among other functionalities, the removal of adapters as well as low quality bases/reads. In particular, PRINSEQ has been chosen for its ability to treat both single and paired-end reads and to perform read quality trimming as well as duplicate removal. Using the BLAT fast similarity search tool, reads can be compared against databases of potential contaminants such as, eg, UniVec (which allows identifying sequences of vector origin; http://www.ncbi.nlm.nih.gov/VecScreen/UniVec.html) or user-specified databases. TRUFA's scripts will automatically remove those reads, giving hits with such queried databases.

Cleaned reads, after passing an optional second quality control with FASTQC to verify the overall efficiency of the first cleaning step, are ready for assembly. TRUFA implements the software Trinity,[34] which is an extensively used *de novo* assembler and has been shown to perform better than other single k-mer assemblers.[35] After the assembly, an in-house script provides basic statistics describing transcripts lengths distribution, total bases incorporated in the assembly, N50, and GC content. In addition, to evaluate the completeness of the assembly, a Blast+[36] similarity search is performed against the UniProtKB/Swiss-Prot database, and a Trinity script evaluates whether those assembled transcripts with hits are full-length or nearly full-length. The CEGMA software can also provide a measure of the completeness of the assembly by comparing the transcripts to a set of 248 core eukaryotic genes, which are conserved in highly divergent eukaryotic taxa.[37] Both the number of recovered genes from the total of 248 and their completeness have been used for *de novo* assembly quality assessments.[38,39]

The newly assembled transcripts can be used as query for similarity searches with BLAT[40] or Blast+ against the NCBI nr and UniRef90 databases. In parallel, HMMER[41] searches can be performed applying hidden markov models (HMM) against the PFAM-A database. Both analysis can be run as well with user-specified databases or models respectively. Further annotation and assignation of gene ontology (GO) terms can be obtained with Blast2GO[28] for the transcripts with blast hits against the nr database.

For expression quantification, Bowtie2[42] is used to produce alignments of the reads against the assembled transcripts. Alignments are then properly formatted using SAMtools[43] and Picard (http://broadinstitute.github.io/picard/).[43] Using these alignments, eXpress[44] can be used to quantify the expression of all isoforms. Additionally, the script "run_RSEM_align_n_estimate" of the Trinity package implemented in TRUFA uses Bowtie[45] and RSEM[46] to provide an alternative procedure for expression quantification

of both genes and isoforms. Moreover, the percentage of reads mapping back to the assembled transcripts (obtained with Bowtie and Bowtie2) can be used as another indication of the assembly quality.[35,38]

Output. TRUFA generates a large amount of output information from the different programs used in the customized pipeline. Briefly, a user should be able to download FastQC html reports, FASTQ files with cleaned reads (without duplicated reads and/or trimmed), Trinity-assembled transcripts (FASTA), read alignments against the transcripts (BAM files), GO annotations (.txt and.dat files which can be imported into the Blast2GO java application), and read counts (text files providing read counts and TPM). Various statistics are computed at each step and are reported in text files, such as the percentage of duplicated/trimmed reads, CEGMA completeness report, assembly sequence composition, percentage of mapped reads, and read count distributions.

Results and Discussion

We have built an informatics platform that performs a nearly complete *de novo* RNA-seq analysis in a user-friendly manner (amenable to the nonexpert user, avoiding command lines, and providing a lightweight visual interface), and tested its performance using four publicly available transcriptome datasets. A small dataset of the fission yeast, *Schizosaccharomyces pombe*, which is provided in a published Trinity tutorial,[47] was used to test the correct functioning of the assembly process on TRUFA. Two previously well-characterized datasets from the green tea, *Camelia sinensis* (SRX020193), and the fruit fly, *Drosophila melanogaster* (SRR023199, SRR023502, SRR023504, SRR023538, SRR023539, SRR023540, SRR023600, SRR023602, SRR023604, SRR027109, SRR027110, SRR027114 and SRR035403), were used to compare assembly and read mapping statistics with the results from Zhao et al.[35] Finally, TRUFA was tested using a rice (*Oryza sativa*) dataset[48,49] (SRX017630, SRX017631, SRX017632, SRX017633). When applicable, reads corresponding to each end of a pair-ended reaction were concatenated separately into two files, and all files were compressed with gzip before uploading to the platform. Each of the compressed read files was uploaded to TRUFA in less than a day (typical uploading times from a personal computer anywhere ranging from 30 seconds to 12 hours for files ranging from 200 MB to 12 GB, ie, between 0.25 and 25 Gbp).

The results of a first run performing only a FASTQC analysis were used to set the parameters (see Supplementary Table 1) for the cleaning process, except for the yeast dataset, which was assembled without preprocessing. Read cleaning, assembly, mapping, and annotation statistics are shown in Tables 2 and 3. The yeast dataset showed highly similar results to the original analysis, validating the TRUFA assembly. The difference observed in the number of transcripts is most likely due to the not fully deterministic nature of the Trinity algorithm.[47] However, the percentage of reads mapped back to the transcripts was slightly higher in the original study.[47] For the other three datasets, TRUFA showed globally comparable results. Except for the mean transcript length for the *C. sinensis* assembly, all other statistics for both *C. sinensis* and *D. melanogaster* assemblies were higher in the present analyses with respect to the original ones (Table 2). Remarkably, the percentage of reads mapping back to the transcripts was significantly higher for the green tea dataset using TRUFA. This could be due to a more efficient read-cleaning step or to differences between Bowtie2 (used in TRUFA) and Bowtie used by Zhao et al (2011) mappings. CEGMA analysis showed that more than 80% (range 85.5%–98.39%) of the core eukaryotic genes are fully recovered and more than 98% (range 98.8%–100%) are partially recovered in all dataset assemblies (Fig. 3). This indicates an overall high completeness of the assemblies performed herein with TRUFA. In addition to the assembly and the mapping of the reads, TRUFA was able to annotate *de novo* 25%–42% of the transcripts using

Table 2. Comparison of outputs between original and TRUFA analyses.

NO. OF RAW BASES	*S. pombe*		*C. sinensis*		*O. sativa*		*D. melanogaster*	
	PESS		PE		PE		PE	
	544M		2320M		5983M		24740M	
Pipeline	Trufa	Haas et al (2013)	Trufa	Zhao et al (2011)	Trufa	Xie et al (2014)	Trufa	Zhao et al (2011)
No. of bases after cleaning	No cleaning	No cleaning	2,017M	NA	5,342M	NA	5,028M	NA
No. of transcripts	9,370	9,299	201,892	188,950	166,512	170,880	80,999	70,906
Mean transcript length	1,014	NA	319	332	480	552	847	751
No. of bases in the assembly	9M	NA	64M	63M	80M	94M	69M	53M
N50	1,585	1,585	542	525	1,205	1,392	2,960	2,499
No. of transcripts >1000 nt	3,680	NA	13,276	12,495	22,317	28,578	17,251	12,511
Total alignment rate	94.98%	99.93%	88.84%	61.04%	94.76%	NA	92.39%	89.9
Concordant pairs	92.21%	93.12%	74.45%	NA	87.51%	NA	84.73%	NA

Note: Concordant pairs are considered when they report at least one alignment.
Abbreviations: PE, Paired-end; SS, strand-specific; M, million; NA, data not available.

Table 3. Summary of the *de novo* annotation step for the four assembled transcriptomes.

	S. pombe	*C. sinensis*	*O. sativa*	*D. melanogaster*
# transcripts	9,370	201,892	166,512	80,999
# Blast Hits	8,257	72,559	66,129	29,924
# Annotations	3,922	51,272	50,721	22,534
% of annotated transcripts	42%	25%	30%	28%
# HMMER hits	5,588	34,689	28,736	16,552
User time	11 h	3 d 19 h	6 d 8 h	4 d 15 h

Notes: # Transcripts, number of transcripts assembled by Trinity; # Blast hits, number of transcripts with at least one hit against the NCBI nr database (e-value $<10^{-6}$); # Annotation, number of transcripts with at least one annotation after Blast2GO analysis; # HMMER hits, number of transcripts with at least one hit against the Pfam A database (e-value $<10^{-6}$); User time, time needed to perform the complete pipeline (cleaning, assembly, annotation, and expression quantification).

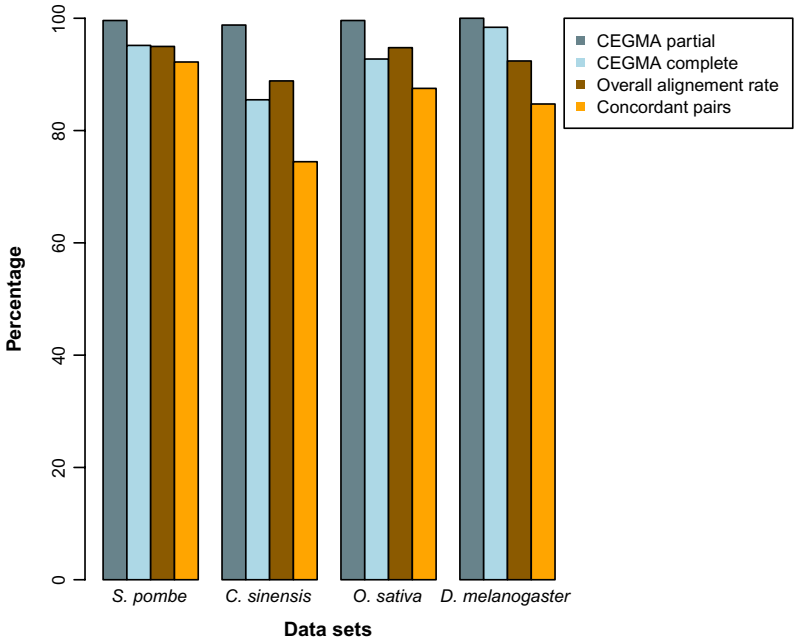


Figure 3. Measures of completeness and read usage for the assemblies produced with TRUFA. CEGMA results represent the percentage of completely and partially recovered genes in the assemblies for a subset of 248 highly conserved core eukaryotic genes. Overall alignment rate and concordant pairs (providing at least one alignment) were computed with Bowtie2.

the Blat, Blast+, and Blast2GO pipeline with an e-value of $<10^{-6}$ (Table 3). HMMER searches identified 17%–60% of the transcripts with at least one hit with an e-value $<10^{-6}$. The expression of each transcript was quantified using RSEM and eXpress, although no data were available for comparison with the original studies.

Considering the entire pipeline, each testing dataset was analyzed by TRUFA in less than a week (Table 3), confirming a good time efficiency of the platform. According to Macmanes[13] on the effect of read trimming for RNA-seq analysis, optimizing trimming parameters leads to better assembly results. This optimization should take no longer than 3 days of computation for datasets such as the ones used here and can be easily done with TRUFA by producing in parallel various assemblies and their quality statistics with different sets of trimming parameters and parameter values.

In Prospect

To complete the RNA-seq analysis pipeline available in TRUFA, we plan to expand the platform by incorporating programs for differential expression analysis and SNP calling. Other programs, especially for assembly (eg, SOAPdenovo-Trans, Velvet-Oases) and visualization (eg, GBrowse) of the data, are planned to be also included in the future. In addition, integrating GO terms for each annotated transcripts would permit the user to browse sequences of interest directly from

the web server without the need to download large quantities of output. We also plan to complete the platform by providing features for read mapping against a reference genome (such as, eg, STAR,[50] Tophat, and Cufflinks). A cloud version of TRUFA, which would increase considerably its global capabilities, is also envisioned to be run in the EGI.eu Federated Cloud (see https://www.egi.eu/infrastructure/cloud/) in the near future.

Conclusion

We presented TRUFA, a bioinformatics platform offering a web interface for *de novo* RNA-seq analysis. It is intended for scientists analyzing transcriptome data who do not have either bioinformatics skills or access to fast computing services (or both). TRUFA is essentially a wrapper of various widely used RNA-seq analysis tools, allowing the generation of RNA-seq outputs in an efficient, consistent, and user-friendly manner, based on a pipeline approach.

The trimming and assembly steps are guided by the integration of widely used quality control programs toward the optimization of the assembly process. Moreover, the implementation of HMMER, BLAST+, and Blast2GO to the platform for *de novo* annotation is, to our knowledge, a feature not available in other RNA-seq analysis web servers such as GigaGalaxy or Oqtans. This step is the most computationally demanding among all RNA-seq analysis steps (including SNPs calling and differential expression), and TRUFA uses highly parallelized steps to obtain annotations in a relatively short time frame. Although annotations can be performed in other platforms such as FastAnnotator, having all these steps from cleaning to annotations and expression quantification in the same pipeline reduces unnecessary transfer of large outputs and provides an advantage to the nonexpert user.

Data Accessibility

TRUFA platform, user manual, example data sets and tutorial videos are accessible at the web page https://trufa.ifca.es/web. Accession numbers to the read files used in this study are provided in the Results and Discussion section and can be obtained from http://www.ncbi.nlm.nih.gov/sra/.

Abbreviations

TRUFA: transcriptome user-friendly analysis
TPM: transcripts per million
SNP: single nucleotide polymorphism
HPC: high performance computing

Acknowledgments

We would like to thank Beatríz Ranz for her help during the validation process and Iván Cabrillo for his help managing the accounts on the Altamira supercomputer. We are grateful to Federico Abascal for his input in the review process, his suggestions, and his help during the beta testing. Thanks to all other beta testers: Anna María Addamo, Carlos Canchaya, Michel Domínguez, Iván Gómez-Mestre, Iker Irisarri, Nathan Kenny, Diushi Keri, David Osca, Snæbjörn Pálsson, Sara Rocha, Diego San Mauro, Maria Torres, Juan E. Uribe, Ignacio Varela, and Joel Vizueta. We thank five anonymous reviewers for their valuable feedback on a previous version of the manuscript. The platform was developed thanks to the use of bootstrap (http://getbootstrap.com/), jquery (http://jquery.com/), sqlite (https://sqlite.org/), webpy (http://webpy.org/), and filemanager (https://github.com/simogeo/Filemanager). The Altamira supercomputer is member of the Spanish Supercomputing Network.

Author Contributions

Conceived the study: RZ, EK. Constructed the pipeline: EK. Performed testing runs: EK, CFL. Implemented the web version: LC, FA, EK, JM. Tuned parts of the pipeline: CFL, JR. All authors contributed to the writing and improving of the manuscript, and read and approved the final version.

Supplementary Material

Supplementary Table 1. List of the main command lines used for the analysis of each data sets. Datasets: 1, *S. pombe*; 2, *C. sinensis*; 3, *O. sativa*; 4, *D. melanogaster*.

REFERENCES

1. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*. 2009;10(1):57–63.
2. Marguerat S, Wilhelm BT, Bähler J. Next-generation sequencing: applications beyond genomes. *Biochem Soc Trans*. 2008;36(pt 5):1091–6.
3. Lister R, Gregory BD, Ecker JR. Next is now: new technologies for sequencing of genomes, transcriptomes, and beyond. *Curr Opin Plant Biol*. 2009;12(2):107–18.
4. Wilhelm BT, Landry J-R. RNA-Seq-quantitative measurement of expression through massively parallel RNA-sequencing. *Methods*. 2009;48(3):249–57.
5. Bainbridge MN, Warren RL, Hirst M, et al. Analysis of the prostate cancer cell line LNCaP transcriptome using a sequencing-by-synthesis approach. *BMC Genomics*. 2006;7:246.
6. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods*. 2008;5(7):621–8.
7. Marguerat S, Bähler J. RNA-seq: from technology to biology. *Cell Mol Life Sci*. 2010;67(4):569–79.
8. Eswaran J, Horvath A, Godbole S, et al. RNA sequencing of cancer reveals novel splicing alterations. *Sci Rep*. 2013;3:1689.
9. Westermann AJ, Gorski SA, Vogel J. Dual RNA-seq of pathogen and host. *Nat Rev Microbiol*. 2012;10(9):618–30.
10. Savage AE, Kiemnec-Tyburczy KM, Ellison AR, Fleischer RC, Zamudio KR. Conservation and divergence in the frog immunome: pyrosequencing and de novo assembly of immune tissue transcriptomes. *Gene*. 2014;542(2):98–108.
11. Daniels EV, Murad R, Mortazavi A, Reed RD. Extensive transcriptional response associated with seasonal plasticity of butterfly wing patterns. *Mol Ecol*. 2014;23(24):6123–34.
12. Barghi N, Concepcion GP, Olivera BM, Lluisma AO. High conopeptide diversity in *Conus tribblei* revealed through analysis of venom duct transcriptome using two high-throughput sequencing platforms. *Mar Biotechnol*. 2014;17(1):81–98.
13. Macmanes MD. On the optimal trimming of high-throughput mRNA sequence data. *Front Genet*. 2014;5:13.
14. Smith DR. The battle for user-friendly bioinformatics. *Front Genet*. 2013;4:187.
15. Jain P, Krishnan NM, Panda B. Augmenting transcriptome assembly by combining de novo and genome-guided tools. *PeerJ*. 2013;1:e133.
16. Martin JA, Wang Z. Next-generation transcriptome assembly. *Nat Rev Genet*. 2011;12(10):671–82.
17. Bao S, Jiang R, Kwan W, Wang B, Ma X, Song Y-Q. Evaluation of next-generation sequencing software in mapping and assembly. *J Hum Genet*. 2011;56(6):406–14.

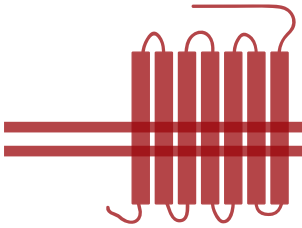
18. Garber M, Grabherr MG, Guttman M, Trapnell C. Computational methods for transcriptome annotation and quantification using RNA-seq. *Nat Methods*. 2011;8(6):469–77.
19. Lee HC, Lai K, Lorenc MT, Imelfort M, Duran C, Edwards D. Bioinformatics tools and databases for analysis of next-generation sequence data. *Brief Funct Genomics*. 2012;11(1):12–24.
20. Cabrillo I, Cabellos L, Marco J, Fernandez J, Gonzalez I. Direct exploitation of a top 500 supercomputer for analysis of CMS data. *J Phys Conf Ser*. 2014;513(3):032014.
21. Del Fabbro C, Scalabrin S, Morgante M, Giorgi FM. An extensive evaluation of read trimming effects on Illumina NGS data analysis. *PLoS One*. 2013;8(12):e85024.
22. Goecks J, Nekrutenko A, Taylor J. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol*. 2010;11(8):R86.
23. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*. 2009;25(9):1105–11.
24. Trapnell C, Williams BA, Pertea G, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol*. 2010;28(5):511–5.
25. Luo R, Liu B, Xie Y, et al. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience*. 2012;1(1):18.
26. Sreedharan VT, Schultheiss SJ, Jean G, et al. Oqtans: the RNA-seq workbench in the cloud for complete and reproducible quantitative transcriptome analysis. *Bioinformatics*. 2014;30(9):1300–1.
27. Chen TW, Gan RC, Wu TH, et al. FastAnnotator – an efficient transcript annotation web tool. *BMC Genomics*. 2012;13(suppl 7):S9.
28. Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*. 2005;21(18):3674–6.
29. Claudel-Renard C. Enzyme-specific profiles for genome annotation: PRIAM. *Nucleic Acids Res*. 2003;31(22):6633–9.
30. Wagner GP, Kin K, Lynch VJ. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory Biosci*. 2012;131(4):281–5.
31. Leinonen R, Sugawara H, Shumway M. The sequence read archive. *Nucleic Acids Res*. 2011;39(Database issue):D19–21.
32. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J*. 2011;17(1):10.
33. Schmieder R, Edwards R. Quality control and preprocessing of metagenomic datasets. *Bioinformatics*. 2011;27(6):863–4.
34. Grabherr MG, Haas BJ, Yassour M, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol*. 2011;29(7):644–52.
35. Zhao Q-Y, Wang Y, Kong Y-M, Luo D, Li X, Hao P. Optimizing de novo transcriptome assembly from short-read RNA-Seq data: a comparative study. *BMC Bioinformatics*. 2011;12(suppl 14):S2.
36. Camacho C, Coulouris G, Avagyan V, et al. BLAST+: architecture and applications. *BMC Bioinformatics*. 2009;10:421.
37. Parra G, Bradnam K, Korf I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics*. 2007;23(9):1061–7.
38. Moreton J, Dunham SP, Emes RD. A consensus approach to vertebrate de novo transcriptome assembly from RNA-seq data: assembly of the duck (*Anas platyrhynchos*) transcriptome. *Front Genet*. 2014;5:190.
39. Nakasugi K, Crowhurst RN, Bally J, Wood CC, Hellens RP, Waterhouse PM. De novo transcriptome sequence assembly and analysis of RNA silencing genes of *Nicotiana benthamiana*. *PLoS One*. 2013;8(3):e59534.
40. Kent WJ. BLAT – The BLAST-like alignment tool. *Genome Res*. 2002;12(4):656–64.
41. Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res*. 2011;39(Web Server issue):W29–37.
42. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods*. 2012;9(4):357–9.
43. Li H, Handsaker B, Wysoker A, et al; 1000 Genome Project Data Processing Subgroup. The sequence alignment/map format and SAMtools. *Bioinformatics*. 2009;25(16):2078–9.
44. Roberts A, Pachter L. Streaming fragment assignment for real-time analysis of sequencing experiments. *Nat Methods*. 2013;10(1):71–3.
45. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*. 2009;10(3):R25.
46. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*. 2011;12:323.
47. Haas BJ, Papanicolaou A, Yassour M, et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc*. 2013;8(8):1494–512.
48. Xie Y, Wu G, Tang J, et al. SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads. *Bioinformatics*. 2014;30(12):1660–6.
49. Zhang G, Guo G, Hu X, et al. Deep RNA sequencing at single base-pair resolution reveals high complexity of the rice transcriptome. *Genome Res*. 2010;20(5):646–54.
50. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013;29(1):15–21.

Artículo 3

Supplemental Information

Supplementary Table 1
List of the main command lines used for the analysis of each data sets. Datasets: 1, *S. pombe*; 2, *C. sinensis;* 3, *O. sativa*; 4, *D. melanogaster*.
EBO-11-2015-097-s001.csv (1.6K)
GUID: 86058EDD-0765-4A4A-9309-8700643F2562

Data Availability Statement
TRUFA platform, user manual, example data sets and tutorial videos are accessible at the web page https://trufa.ifca.es/web. Accession numbers to the read files used in this study are provided in the Results and Discussion section and can be obtained from http://www.ncbi.nlm.nih.gov/sra/.



DISCUSIÓN

1. Evolución del Sistema Quimiosensorial en Macrothele calpeiana

A diferencia de los hexápodos en los quelicerados y miriápodos, el conocimiento actual de las moléculas específicas involucradas en la quimiorrecepción proviene exclusivamente del análisis comparativo de secuencias genómicas (Vizueta, Rozas, y Sánchez-Gracia 2018). De hecho, los genomas de quelicerados y miriápodos contienen varios genes que codifican algunos quimioreceptores homólogos de insectos y pequeñas proteínas solubles quimiosensoriales (Chipman et al. 2014; Gulia-Nuss et al. 2016; Hoy et al. 2016). Como muchos otros organismos no modelo, estos grupos son difíciles de mantener en el laboratorio y no se disponen de herramientas moleculares apropiadas para su manipulación genética. Como una primera aproximación para abordar la identificación y caracterización de las familias multigénicas (FM) involucradas en el sistema quimiosensorial (SQ) hemos generado el transcriptoma de los potenciales órganos quimiosensoriales de la araña *Macrothele calpeiana* (Frías-López et al. 2015), un quelicerado, y el geofilomorfo *Strigamia maritima* (Leach, 1817), un miriápodo (resultados no publicados).

1.1 Estrategia diseñada para identificar genes involucrados en el SQ en *M. calpeiana*

Dado que *M. calpeiana* no dispone de un genoma de referencia, obtuvimos el transcriptoma mediante la tecnología más potente disponible en aquel momento, la plataforma de GS-FLX Titanium de 454 (Ekblom y Galindo 2011; Sujai Kumar y Blaxter 2010; Wheat 2010). Esta tecnología generaba *reads* razonablemente largos (~600 pb) pero, con un rendimiento más bajo por carrera que otras tecnologías como Illumina. Por ello, empleamos una técnica para enriquecer la muestra en los tránscritos de interés. En concreto, aplicamos una técnica de hibridación sustractiva (*Suppression subtractive hybridization* - SSH) la cual mediante un proceso de hibridación elimina las secuencias de cDNA repetidas entre dos muestras y únicamente se secuencian los tránscritos que están presentes en un único tejido.

Para construir las librerías SSH de los principales órganos quimiosensoriales, pata y palpo (*drivers*) utilizamos tejido muscular (*tester*) con el fin de enriquecer la secuenciación de

tránscritos de genes implicados en el SQ y descartar aquellos que no tuvieran relación, como tránscritos relacionados con procesos involucrados en la locomoción u otros no deseados. Además, la técnica de SSH nos permitía también normalizar el nivel de expresión y aumentar la probabilidad de secuenciar aquellos genes con bajos niveles de expresión, como son la mayoría de los genes que codifican quimioreceptores (H. Guo et al. 2017; J. Zhang et al. 2015).

1.2 Validación de la librería sustractiva y anotación funcional

Con el objetivo de determinar la calidad del transcriptoma ensamblado (6.696 contigs), se utilizó la herramienta blastx para buscar los tránscritos potencialmente homólogos con los genes *Housekeeping* (HKG) y con los genes de la base de datos CEGMA (*Cluster of Essential Genes en inglés* - CEG), un conjunto de genes altamente conservados en organismos eucariotas. La gran mayoría de los genes HKG y CEG están presentes en los tres tejidos y sólo representan un 15% de los tránscritos secuenciados. Esto demuestra que la hibridación se produjo de manera efectiva, ya que en condiciones normales suelen representar entre el 50-60% de los tránscritos secuenciados (J. Zeng et al. 2016). A pesar del uso de SSH, es de esperar que se hayan secuenciado algunos genes HKG, concretamente aquellos que no mantienen un nivel de expresión constante entre tejidos (Pfaffl et al. 2004; Ponton et al. 2011). De hecho, en estudios mediante técnicas de RT-PCR de muestras de cDNA de librerías SSH, se ha demostrado que algunos genes HKG pueden aparecer en un ciclo tardío, es decir, se reduce su nivel de expresión pero no son eliminados totalmente (Fang et al. 2011). Como no existe una lista de genes HKG específicos para quelicerados, algunos de los genes identificados como HKG a partir de la similitud de secuencia con genes de *Drosophila melanogaster* (Lam et al. 2012) y de *Homo sapiens* (Eisenberg y Levanon 2013), pueden que en realidad no sean HKG en arañas. Estos genes al no presentar una expresión estable entre tejidos no se han podido eliminar completamente mediante el proceso de hibridación de la librería SSH. Existe actualmente un gran debate alrededor de los genes definidos como HKG que pueden ocasionar interpretaciones erróneas en los estudios de expresión génica. Por eso, en la actualidad se están validando experimentalmente el conjunto de genes establecidos como HKG en los organismos no modelo en estudios de expresión génica mediante RT-PCR (Leelatanawit et al. 2012; Yang et al. 2018). En cualquier caso, en nuestro estudio los pocos tránscritos detectados como HKG fueron eliminados *in silico* de los análisis.

Como era de esperar, mediante la búsqueda de homología por similitud de secuencia mediante blastx los homólogos de los tránscritos se identificaron en un bajo porcentaje (65%), debido a la carencia de genomas cercanos y la ausencia de transcriptomas específicos de pata, palpo y ovario en otros quelicerados. De este 65% un 85% de los *hits* corresponden a proteínas de quelicerados, el 12% a hexápodos y un 2% a crustáceos, siendo la garrapata *Ixodes scapularis* Say, 1821 la especie con más *hits*. A través de la búsqueda de patrones mediante la herramienta HMMER (Eddy 2009), obtuvimos un porcentaje mayor de tránscritos anotados, en total un 75% de los tránscritos. Los términos GO más frecuentes de nivel 2 en pata y palpo son "metabólico" y "procesos celulares" (dentro del grupo de proceso biológico (BP), y "unión" y "actividades catalíticas", dentro de la categoría de función molecular (MF). Estos términos podrían estar involucrados con el SQ, pero también se ha visto que los procesos “metabólicos” tienen un papel clave para inducir cambios de expresión de los receptores en procesos de comunicación social. En concreto, se ha observado que en insectos que forman agregados, como las langostas (*Locusta*) los cambios de expresión de las CSPs pueden estar relacionados con el cambio de

fase gregaria (W. Guo et al. 2011; Martín-Blázquez et al. 2017) . Por otro lado, los términos GO de “unión” y “actividades catalíticas” se pueden relacionar con los procesos de detección del ligando quimiosensorial y la señalización inducida por los receptores de canal iónico, procesos implicados en la quimiorecepción, como se ha demostrado en otros RNAseq de antena de insecto (Su-fang Zhang et al. 2018). Sin embargo, estos términos son muy generales y no son exclusivos del SQ.

Tras la caracterización funcional del transcriptoma, comparamos los tránscritos de cada órgano para identificar los genes con expresión específica de tejido. Como era de esperar encontramos una alta proporción de tránscritos compartidos entre pata y palpo. Este resultado, apoyaría nuestra hipótesis situando estos dos tejidos como posibles órganos quimiosensoriales. De hecho, en hexápodos, se ha caracterizado la presencia de sensilios gustativos y olfativos en el primer par de patas, mientras que en el resto de patas principalmente sólo gustativos. En arañas, en cambio, se ha observado en la superficie de las patas por imágenes de microscopía electrónica de barrido (SEM) la presencia de sensilios con un único poro en el ápice (Ganske y Uhl 2018; Jiao et al. 2011), similar al gustativo de hexápodos. Debido a la ausencia de la identificación de sensilios multi-poros similares a los olfativos de hexápodos (Figura 12), se postula que este tipo de sensilio además de estar implicado en procesos gustativos podría estar involucrado en la detección de substancias volátiles (Ganske y Uhl 2018). Sin embargo, sería necesario completar estos resultados con una aproximación experimental, como por ejemplo la tinción de las neuronas sensoriales para comprobar si éstas efectivamente están inervadas en el deutocerebro (Chahda et al. 2019; Jefferis et al. 2001).

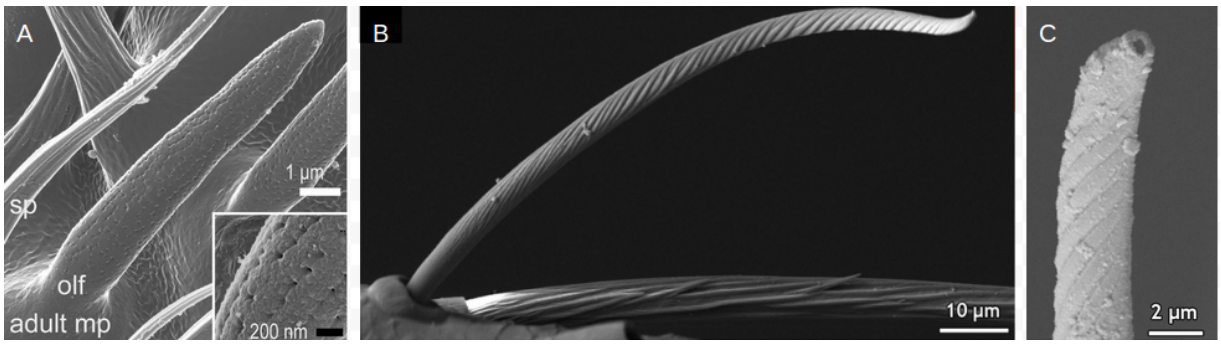


Figura 12. A) Imagen de la superficie del palpo maxilar de *Drosophila melanogaster* que muestra un sensilio olfatoria (olf) y una espínola no sensorial (sp). El recuadro muestra poros dispuestos regularmente. B) y C) Imágenes de un sensilio quimiosensorial con un único poro en el ápice de la araña *Argiope bruennichi*. Escala de las imágenes: A) 1 µm, B) 10 µm , C) 2 µm. Fuente A) (Ando et al. 2019), B) y C) (Ganske y Uhl 2018)

El estudio de las categorías funcionales sobre- e infra- representadas comparando las tránscritos comunes en palpo y pata con los de ovario, mostraron varios términos GO sobrerrepresentados. Así, los términos GO de "unión de cationes", "unión de iones metálicos" y "proceso de oxidación-reducción" están claramente sobrerrepresentados en tránscritos específicos de pata y palpo. Estos términos podrían estar involucrados en los procesos de señalización originados por los quimiorreceptores que actúan como canales iónicos y en la despolarización de la neurona sensorial durante la transducción de la señal química (Sparks et al. 2018; Wicher 2015). No obstante, estos procesos no son exclusivos del SQ.

También analizamos mediante el estudio de los términos KEGG las rutas biológicas que se utilizan de forma diferencial en los tejidos estudiados. En este caso, también hemos obtenido diferencias significativas entre las rutas de palpo y pata respecto a las de ovario. En concreto, tres de las rutas específicas de pata y palpo son "biosíntesis de alcaloides del tropano, piperidina y piridina", "metabolismo de triptófano" y "metabolismo de tirosina". Sin embargo, no hay evidencias de que ninguna de estas vías esté directamente relacionada con la función quimiosensorial. Los derivados de alcaloides como el tropano, la piperidina y la piridina son sintetizados por algunas plantas como mecanismos de defensa (Fürstenberg-Hägg, Zagrobelny, y Bak 2013) y se utilizan para desarrollar pesticidas (Chowański et al. 2016) (Application EP-2099455-A1 s. f.). También se ha demostrado que la araña *Nephila antipodiana* (Walckenaer, 1841) recubre su tela con otro tipo de alcaloide (2-pirrolidinona), que aparentemente proporciona protección contra la invasión de hormigas (Shichang Zhang et al. 2012). Las grandes telas de embudo que construye *M. calpeiana* para atrapar a sus presas están igualmente expuestas a los depredadores y, por lo tanto, podrían utilizar también derivados de estos alcaloides como una defensa química. Sin embargo, sería necesario realizar más estudios sobre la presencia de estas substancias químicas en las telas de embudo para confirmar esta hipótesis.

1.3 Transcriptoma quimiosensorial de *M. calpeiana*

Con el objetivo de identificar tránscritos de las FM del SQ, realizamos una aproximación bioinformática basada en la detección de las características principales de estas proteínas: (i) la presencia de un péptido señal (característico de las proteínas transportadoras como las OBPs de insectos y vertebrados, las NPC2, las CSPs y las ChesA/B), y (ii) la existencia de un dominio transmembrana (característico de todos los quimiorreceptores, como los ORs, los GRs y los IRs de insectos y vertebrados). Dado que la identificación de potenciales homólogos a través de perfiles de HMM fue más eficiente, realizamos la búsqueda de dominios proteicos mediante **`InterProScan`** (P. Jones et al. 2014). Encontramos una sobrerrepresentación del dominio péptido señal en los tránscritos de pata y palpo, siendo mucho más significativa para los tránscritos compartidos entre los dos tejidos. Además, un 40.6% de los tránscritos específicos de pata presentan al menos un dominio transmembrana. Estos resultados son coherentes con una función quimiosensorial, aunque pueden estar relacionados tanto con procesos gustativos como olfativos, los cuales se podrían dar en las patas de los quelicerados, de la misma manera que ocurre en las patas de los hexápodos (Joseph y Carlson 2015a). Hay que tener en cuenta que estas características no son exclusivas de los genes quimiosensoriales y por lo tanto no podemos asegurar con rotundidad que estas características sugieran un papel quimiosensorial de estos tejidos.

Para profundizar en el papel del SQ en los tejidos analizados, realizamos una serie de búsquedas recursivas, exhaustivas y excluyentes a dos niveles, uno en base al nivel de similitud de secuencia, y un segundo en base a la predicción de perfiles de HMM de dominios de proteínas del SQ. Para ello, construimos una base de datos que incluía sólo secuencias del SQ de vertebrados e insectos, detectadas previamente mediante la detección de dominios del SQ con el software InterProScan. A continuación, realizamos una búsqueda con perfiles HMM de los dominios proteicos quimiosensoriales y por último una búsqueda

basada en similitud de secuencia utilizando como base de datos las secuencias genómicas disponibles de 5 especies de quelicerados, las arañas *Parasteatoda tepidariorum* (C. L. Koch, 1841) (Theridiidae), *Stegodyphus mimosarum* Pavesi, 1883 (Eresidae) y *Acanthoscurria geniculate* (C. L. Koch, 1841) (Theraphosidae), el escorpión *Mesobuthus martensii* (Karsch, 1879) y la garrapata *Ixodes scapularis*. Esta aproximación bioinformática nos permitió detectar un número importante de tránscritos, aunque desgraciadamente escaso debido a la baja cobertura obtenida con la secuenciación de 454. En concreto, detectamos 7 candidatos quimiosensoriales, 2 *Irs* y 5 tránscritos de la familia de las *Npc2*. Encontramos que los dos tránscritos con similitud a los IRs (homólogos) se expresan específicamente en palpo y cada uno de ellos codifica para uno de los tres dominios Pfam característicos de estos receptores (Croset et al. 2010), el dominio extracelular amino-terminal y el dominio de canal iónico activado por ligando. Las secuencias identificadas más cercanas a estos dos tránscritos de *M. calpeiana* corresponden a sendas proteínas predichas en la araña *S. mimosarum* y anotadas como productos de "receptor de glutamato, kainato ionotrópico 2". A pesar de que probablemente corresponden a dos genes diferentes, no podemos descartar por completo que estos dos tránscritos de *M. calpeiana* que codifican IRs fueran en realidad dos fragmentos del mismo gen iGluR. Además, los siguientes mejores hits de estos dos tránscritos corresponden a receptores de kainato (KA) seguidos de miembros de los receptores de ácidos α-amino-3-hidroxi-5-metilo-4-isoxazol propiónico (AMPA) de otras especies de artrópodos. Por último, las relaciones filogenéticas de los miembros de estas subfamilias en artrópodos demuestran que las proteínas codificadas de los dos tránscritos de *M. calpeiana* se agrupan en el mismo clado con algunos receptores KA de insectos, ciempiés y garrapatas. Por consiguiente, estos dos tránscritos tendrían un papel en la transmisión y regulación sináptica y no serían un receptor quimiosensorial.

Las proteínas predichas de tres de los cinco tránscritos de los genes *Npc2* identificados en el transcriptoma constituyen un clado específico de *M. calpeian*a en el análisis filogenético, pero dado que se expresan específicamente en ovario, probablemente tampoco estarían involucradas en el SQ. Las otras dos *Npc2* detectadas se expresan en palpo y pata, y ambas tienen una cierta similitud con algunos miembros identificados en la garrapata *I. scapularis* y el miriápodo *S. maritima*, apareciendo relativamente distantes de las NPC2 antenales de la abeja *Apis mellifera* y la hormiga *Camponotus japonicus*. A la luz de estos resultados, queda por dilucidar la posible función quimiosensorial de estas proteínas en palpos y patas. Pelosi y col. (Pelosi et al. 2014a) propusieron que algunos miembros de la familia de las NPC2 podrían estar involucrados en el transporte y solubilización de semioquímicos en los diferentes linajes de artrópodos. Así esta familia podría ser la responsable de realizar la función biológica que efectúan las OBPs y las CSPs (procesos periféricos del olfato) en insectos, proteínas que están ausentes en el resto de artrópodos. No obstante, para poder determinar el papel quimiosensorial de los genes *Npc2* expresados específicamente en palpo y pata, sería necesaria su validación funcional, ya que las NPC2 también realizan otras funciones fisiológicas importantes, como la unión y el transporte de lípidos como el colesterol, que es la función conocida de estas proteínas en los vertebrados (Storch y Xu 2009).

Los datos disponibles al inicio de esta tesis doctoral (Vieira y Rozas 2011) sugerían la ausencia de genes de la família de las *Obps* y de los *Ors* en los genomas de quelicerados. De hecho, no los hemos identificado ni en *M. calpeiana* ni, más recientemente, en la también araña *Dysdera silvatica* (Schmidt, 1981, Dysderidae) (pero ver debajo el papel

de las OBP-like) (Vizueta et al. 2017). No obstante, tendríamos que haber encontrado receptores ionotrópicos (*Irs*) y gustativos (*Grs*), ya que recientemente se han identificado numerosas copias en otros genomas de quelicerados (Vizueta, Rozas, y Sánchez-Gracia 2018). De hecho, se ha postulado que los IRs y los GRs serían los principales candidatos para realizar funciones de quimiorrecepción en estas especies (Gulia-Nuss et al. 2016). No obstante, no hemos identificado ninguno de estos dos receptores, hecho que probablemente se explique por la baja expresión de estos genes y la poca cobertura ofrecida por el sistema de 454. De hecho, 454 no es una tecnología que permita detectar genes con baja expresión (Tarazona et al. 2011). Muchos de estos receptores probablemente están codificados por genes de baja expresión, y su detección podría necesitar una secuenciación de mayor profundidad. En algunos casos, se ha determinado que para identificar tránscritos poco abundantes es necesario secuenciar como mínimo 200 millones de *reads* (Sims et al. 2014).

En el momento de realizar el análisis del transcriptoma no pudimos comparar nuestros resultados con datos similares, ya que no existían estudios sobre la expresión específica de estos receptores u otros miembros del SQ en diferentes tejidos de un quelicerado. Sin embargo, en estudios posteriores realizados en nuestro grupo de investigación se ha demostrado la presencia de tránscritos de *Irs*, *Grs* y *Snmps* en la araña *D. silvatica* (Vizueta et al. 2017). No obstante, no se han encontrado genes codificantes de OBPs ni ORs, apoyando la teoría de que estas FM serían específicas del linaje de los insectos alados (Vieira y Rozas 2011). De hecho, en *D. silvatica* y otros quelicerados, así como también en las antenas de *S. maritima* (resultados no publicados), se han encontrado unas proteínas similares a las OBPs, las OBP-*like* que tienen un patrón similar a las OBP Minus-C, con sólo 4 cisteínas conservadas (Eliash et al. 2019; Vizueta et al. 2017). La diferencia en el número de genes de las FM detectadas entre el transcriptoma de *M. calpeiana* y el de *D. silvatica* se puede explicar con casi total certeza por las diferentes tecnologías de secuenciación utilizadas. Para secuenciar el transcriptoma de *M. calpeiana* utilizamos 454 y obtuvimos unos ~50.000 *reads* por tejido, en cambio el de *D. silvatica* fue secuenciado con Illumina y se obtuvieron unos 100 millones de *reads* por cada tejido.

Otro importante hallazgo de nuestro estudio está relacionado con la eficiencia de los perfiles de HMM para detectar dominios funcionales. Repitiendo la búsqueda de las FM en *M. calpeiana* utilizando perfiles creados con las secuencias genómicas disponibles actualmente en las bases de datos, hemos recuperado 36 tránscritos relacionados con el SQ en lugar de los 7 detectados en el estudio inicial. En concreto hemos identificado 10 ENaCs, 5 NPC2, 14 TRPs, dos IRs, un GR y una OBP-l. Cabe remarcar, que de estos 36 tránscritos sólo hemos recuperado 6 secuencias que cubran más del 70% de la longitud de la proteína, que corresponden a los 5 miembros de las NPC2 y a la única OBP-l. El resto de los candidatos corresponden a secuencias parciales con una longitud inferior del 50% respecto a la secuencia original. Esto demuestra, que la disponibilidad de datos de especies cercanas para hacer búsquedas es un factor clave en la construcción de los modelos de HMM, y en particular para identificar proteínas muy divergentes como puede ser el caso de las proteínas del SQ. En todo caso, los resultados obtenidos han proporcionado los primeros datos para comprender el origen y evolución del SQ en los quelicerados (Figura 13).

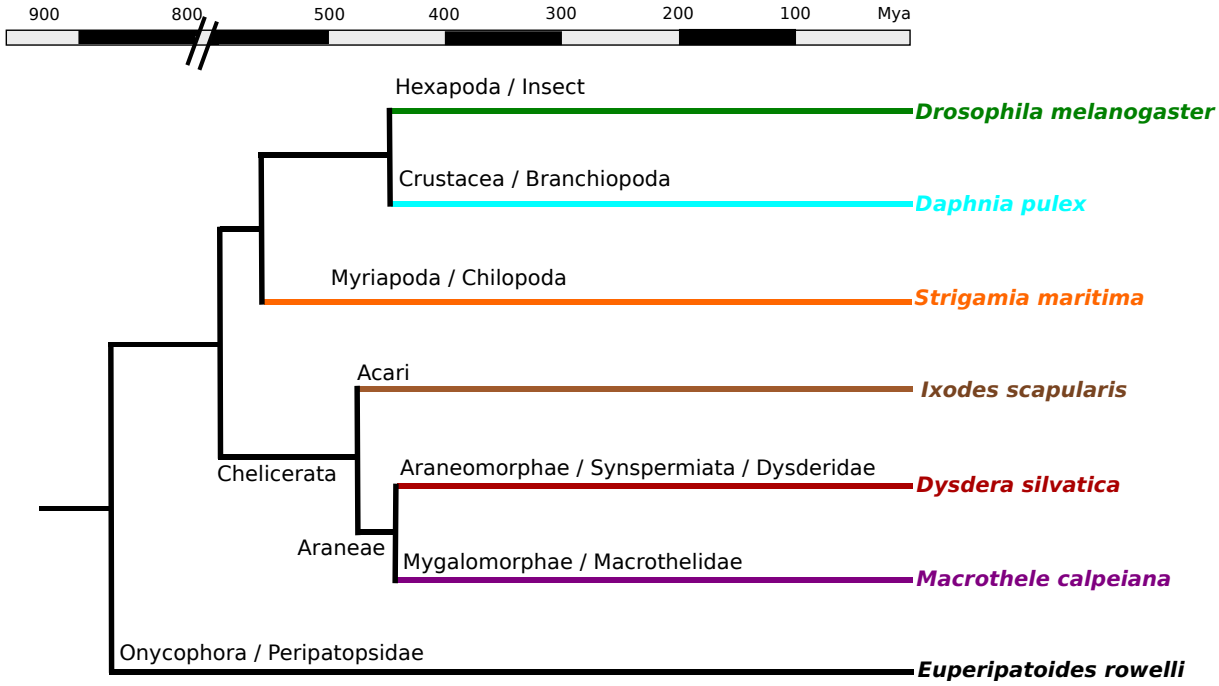


Figura 13. Posición filogenética de *M. calpeiana* en el filo de los artrópodos. Tiempos de divergencia obtenidos mediante TimeTree (Hedges et al. 2015).

2. Desarrollo de herramientas bioinformáticas para el análisis de datos NGS

Las tecnologías de NGS han facilitado la secuenciación del genoma de organismos no modelo; sin embargo, la obtención de la secuencia genómica completa, ensamblada y anotada, sigue siendo un proceso caro y laborioso, además de ser computacionalmente complejo. Para abaratar costes, se han desarrollado diversas técnicas de reducción genómica, es decir restringir la secuenciación a ciertas regiones del genoma, por ejemplo, la secuenciación de sólo las regiones codificantes del genoma mediante los tránscritos (p.ej. RNAseq) o el uso de cebadores, por ejemplo para secuenciar exomas (Puritz y Lotterhos 2018; Sulonen et al. 2011), enzimas de restricción, como los utilizados para la técnica RADseq de *screening* genómico (John W Davey et al. 2010) o el uso de sondas, como en el caso de la secuenciación de elementos ultraconservados (UCEs) (Faircloth et al. 2012), entre otras.

A diferencia de las secuencias obtenidas por la técnica de Sanger, el tratamiento de datos de NGS implica el conocimiento en herramientas bioinformáticas, la aplicación de lenguajes de programación y el uso de *pipelines* eficientes para obtener y analizar las secuencias. Además, dada la constante evolución de las tecnologías NGS, es necesario el constante desarrollo de herramientas bioinformáticas cada vez más eficientes,y que faciliten a los investigadores el análisis de los datos. Con objeto de facilitar al usuario los análisis de NGS, hemos desarrollado dos herramientas bioinformáticas, una para analizar datos de RNAseq, y otra para identificar y seleccionar marcadores moleculares a partir de datos de NGS para su aplicación en estudios de biología evolutiva.

2.1 Herramienta bioinformática para análisis de datos de RNAseq en organismos no modelo

Durante la última década, la secuenciación del transcriptoma (RNAseq) se ha convertido en una metodología indispensable y casi rutinaria para el análisis funcional del genoma, identificar genes específicos de tejido, detectar genes con expresión diferencial y la identificación de isoformas poco frecuentes (Stark, Grzelak, y Hadfield 2019). Cuando el

organismo de estudio no dispone de un genoma de referencia, no es posible reconstruir los tránscritos a través del alineamiento (mapeo) de los *reads* sobre el genoma, y se debe realizar un ensamblaje *de novo,* que es un proceso computacionalmente más complejo (Martin y Wang 2011). Dado el heterogéneo y denso rango de estudios que se pueden realizar con datos de RNAseq, existe una gran diversidad de herramientas computacionales y estadísticas para las distintas etapas de procesamiento y análisis de los datos. La gran mayoría de estas herramientas se suelen ejecutar a través de la línea de comandos bajo un entorno **UNIX** (como bash o shell), y para conectar un programa con otros, suele ser necesario aplicar algún post-procesamiento, o cambiar el formato de los ficheros de salida. En general, para automatizar los análisis se desarrollan los denominados *pipelines* (tubería de procesos) que se suelen ejecutar mediante *scripts*, ficheros de texto que ejecutan de manera secuencial los comandos de los distintos programas que intervienen en el análisis, y procesan el formato de los ficheros intermedios. Por lo tanto, es necesario unos conocimientos mínimos de programación para poder ejecutar de una manera eficiente el análisis completo, además de disponer de un sistema de almacenamiento de datos y un clúster de computación.

Como miembros de la red de investigación AdapNET (financiada por el MINECO, CGL2015-71726-REDT), hemos participado en el intento de unificar metodologías de trabajo e impulsar sinergias para potenciar la transferencia del conocimiento en el ámbito de la biología evolutiva y genómica de la adaptación. Uno de los resultados es la plataforma bionformática **TRUFA** (Kornobis et al. 2015) que permite realizar el análisis -casi completo- de datos de RNAseq de una manera accesible al usuario sin experiencia en bioinformática, evitando ejecutar programas por líneas de comando y proporcionando una interfaz visual. Entre las principales características de **TRUFA** destaca la implementación de diferentes softwares, para realizar un análisis de RNAseq, que incluye los siguientes procesos: i) el procesamiento de las lecturas crudas, ii) el ensamblaje *de novo* iii) el mapeo de los *reads* sobre el transcriptoma ensamblado, iv) la anotación funcional del transcriptoma (tanto mediante con algoritmos de búsqueda de similitud a nivel de secuencia como a través del uso de perfiles HMM para detectar dominios funcionales), y v) el análisis de expresión diferencial. Cabe remarcar que **TRUFA** está diseñado esencialmente para el ensamblaje de organismos no modelo, aquellos organismos que carecen de genoma de referencia y por lo tanto el ensamblaje se realiza *de novo*. Para este paso **TRUFA** implementa uno de los softwares disponibles más eficientes, **Trinity** (Grabherr et al. 2011). Además **TRUFA**, se ejecuta en la supercomputadora ALTAMIRA del Instituto de Física de Cantabria (IFCA), España. La plataforma permite el ensamblaje y el análisis en un tiempo relativamente corto gracias a los recursos de HPC (computación de alto rendimiento), que incluyen el acceso a hasta 256 núcleos para la ejecución de ciertos componentes del *pipeline* y un alto grado de paralelización de los programas.

Para evaluar la eficiencia de **TRUFA**, se reprodujo el análisis de cuatro conjuntos de transcriptomas publicados. Así, para validar el ensamblaje producido por **Trinity** implementado en **TRUFA**, se utilizaron los mismos datos distribuidos en la demo de **Trinity**. El ensamblaje obtenido demostró la correcta integración de **Trinity** en el flujo de trabajo implementado en **TRUFA**. No obstante, algunos de los ensamblajes mostraron ligeras diferencias en la métrica, como en la N50 y el número de tránscritos ensamblados. Estos hechos podrían estar relacionados con la estocasticidad de algunos de los pasos del ensamblaje. Así ocurre, entre otros casos, cuando varios *reads* tienen el mismo MAPQ y se selecciona uno de ellos de forma aleatoria (Langmead y Salzberg 2012; Sachdeva et al.

2014) . Además, pueden intervenir otros factores como, un proceso de limpieza de los *reads* más exhaustivo, o a las diferencias entre **Bowtie2** (Langmead y Salzberg 2012) (integrado en **TRUFA**) y **Bowtie** (Langmead et al. 2009) (utilizado en el trabajo original). Además, la compleción de los transcriptomas ensamblados fue posteriormente validada con la recuperación de un 85-98% de los 248 genes CEG utilizados.

En la actualidad existen varias herramientas con interfaz gráfica para el análisis de datos de RNAseq, no obstante, ninguna presenta un *pipeline* analítico tan completo como el de **TRUFA**. Algunas herramientas, como la utilidad de GigaGalaxy, **Oqtans** (Sreedharan et al. 2014) sólo realiza los procesos generales relacionados con el ensamblaje. Por el contrario, **Fastannotator** (T.-W. Chen et al. 2012) es una plataforma diseñada para realizar las anotaciones funcionales mediante **Blast2GO** (Conesa et al. 2005), pero no incluye ninguna herramienta para realizar los pasos generales que acompaña un estudio con datos de NGS.

TRUFA es de gran utilidad para el análisis automatizado de transcriptomas de organismos no modelo. En concreto, se ha utilizado para caracterizar las bases genómicas de la adaptación biológica de un grupo de anfibios (Torres-Sánchez et al. 2019), o para investigar los mecanismos evolutivos responsables de generar y mantener la diversidad de las conotoxinas (Abalde et al. 2018), entre otros estudios. Su utilidad radica en que permite realizar todas las etapas bioinformáticas comunes que implica trabajar con datos de NGS, como el pre-procesamiento de las lecturas (*raw reads*), el ensamblaje y mapeo, y además dispone de varias herramientas para realizar diversos análisis sobre el transcriptoma ensamblado, como la evaluación de la profundidad del transcriptoma secuenciado mediante el uso de los genes CEG, anotación funcional y detectar los genes con expresión diferencial significativa.

2.2 Herramienta bioinformática para generar y seleccionar marcadores moleculares a partir de datos NGS

El desarrollo de marcadores moleculares es uno de los desafíos más importantes en los estudios de genética de poblaciones, filogeografía y filogenética, especialmente en organismos no modelo. Con el desarrollo de estrategias de reducción genómica (E. M. Lemmon y Lemmon 2013), la obtención de grandes cantidades de marcadores genéticos adecuados para abordar cuestiones ecológicas y evolutivas se ha hecho asequible a muchos investigadores. Desgraciadamente, algunos de estos marcadores pueden no ser suficientemente informativos para resolver preguntas evolutivas específicas, por lo que es necesario el desarrollo de herramientas potentes que permitan al investigador seleccionar los marcadores más adecuados.

En esta tesis doctoral hemos desarrollado **DOMINO**, una herramienta bioinformática para la identificación de marcadores informativos a partir de datos de NGS. La aplicación implementa las herramientas más populares para realizar todos los pasos informáticos para el tratamiento de datos NGS, para *reads* producidos por las plataformas de 454 e Illumina, como i) el pre-procesamiento de las lecturas crudas (*raw reads*), ii) el ensamblaje de los *reads* filtrados, iii) y el alineamiento de los *reads* sobre las secuencias ensambladas. Además, hemos implementado nuevas utilidades para descubrir marcadores moleculares personalizados en función del nivel de variabilidad y de la resolución taxonómica de interés. En concreto,

destacan dos módulos uno i) para detectar regiones informativas para desarrollar cebadores que permitan la amplificación por PCR de la región variable en otros taxones (Figura 14), ii) y otro para seleccionar los marcadores más adecuados derivados de otras técnicas de reducción genómica como RAD-Seq (John W Davey et al. 2010), ddRAD (Peterson et al. 2012) o GBS (Elshire et al. 2011). Además, el *pipeline* es flexible y se pueden introducir los datos en distintas etapas de la ejecución, como utilizar directamente ficheros BAM o alineamientos múltiples de secuencia (MSA).

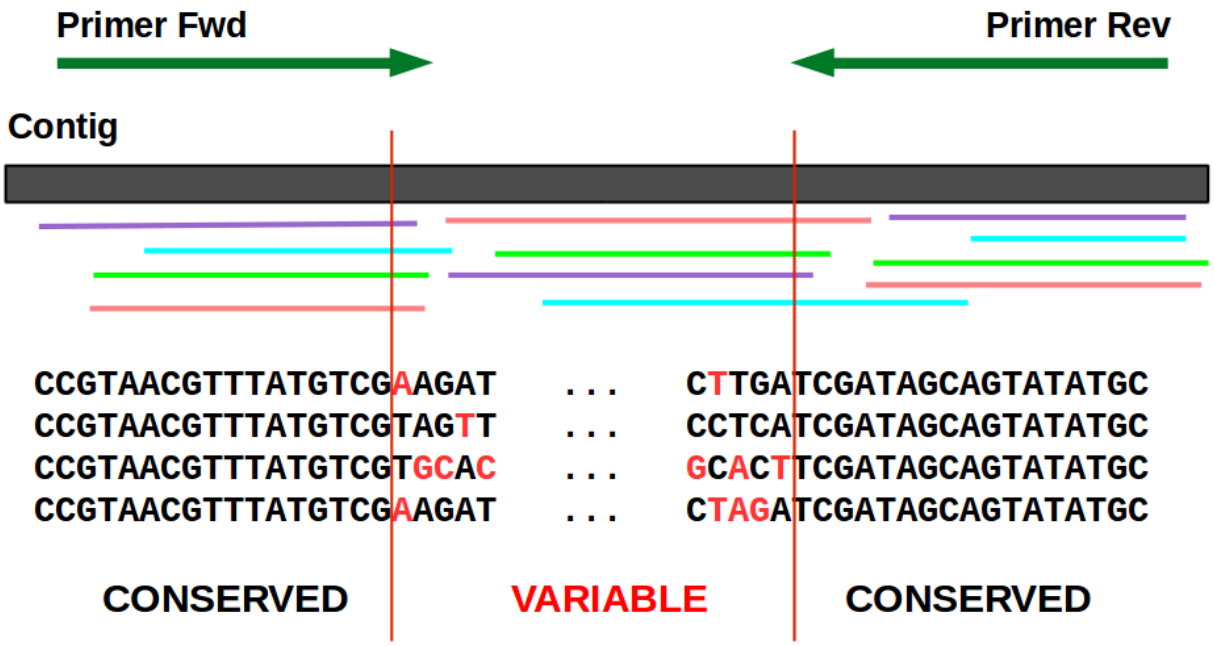


Figura 14. Diagrama de la detección de un candidato para desarrollar un marcador molecular.

Cabe destacar que **DOMINO** implementa una función para realizar el *SNP caller* basada en descartar posibles errores de secuenciación en las posiciones polimórficas derivadas. Para ello, se analizan las variantes que tengan una frecuencia más pequeña que no sea compatible con un verdadero polimorfismo. Esto se realiza bajo un marco estadístico que estima la probabilidad de que la posición variable sea realmente polimórfica a partir de la aplicación de un test binomial. Esta aproximación es una de las más eficientes para estudios de baja cobertura implementada en otros programas más recientes como **Heap** (Kobayashi et al. 2017), pero este software no incluye el proceso de ensamblaje.

Se hizo una validación por simulaciones computacionales, y se observó que los softwares que componen el *pipeline* son capaces de procesar eficientemente los *reads* obtenidos mediante diferentes tecnologías, como 454 e Illumina, y a diferentes niveles de cobertura. Como se esperaba, en el análisis de *SNP calling* los valores de cobertura más altos (10x y 15x) produjeron mejores resultados, en términos de sensibilidad y precisión (Song, Li, y Zhang 2016). En cambio, los valores más bajos (5x) producen una disminución de la sensibilidad; en concreto cuando se utilizan *reads* de Illumina. Esto puede ser debido a los artefactos ocurridos en la etapa de mapeo, ya que los *reads* cortos tiene mayor probabilidad de mapear en múltiples loci y el *read* sería eliminado, descartando del análisis las variantes que

presente ese *read*. En general, los buenos resultados obtenidos en los diferentes escenarios evaluados (excepto a coberturas bajas con datos de Illumina) demuestran una alta eficiencia para detectar marcadores moleculares, dado que la mayoría de los valores promedio de sensibilidad y precisión están alrededor de 0.9, independientemente de las condiciones.

También validamos el software mediante una aproximación experimental. En concreto, se secuenció mediante 454 (Roche) una librería reducida (RRLs) compuesta por 4 taxones de la familia de arañas migalomorfas de trampilla *Nemesiidae* y evaluamos la capacidad de **DOMINO** para identificar marcadores moleculares a bajas coberturas de secuenciación (2x). El objetivo era desarrollar marcadores útiles para resolver filogenias a nivel de género y especie, pero a su vez informativos para inferir las relaciones más profundas. Realizamos una búsqueda de marcadores de diferente longitud y con secuencias conservadas adecuadas para diseñar cebadores; para ello se fijó un valor mínimo de divergencia entre taxones y la longitud deseada de la región conservada. Utilizamos 6 marcadores para desarrollar cebadores que fueron amplificados en otras 14 especies filogenéticamente relacionadas, especies con una divergencia comprendida entre los taxones más distantes utilizados en nuestro panel de 4 especies. El análisis filogenético mostró que los marcadores seleccionados reconstruyeron las relaciones esperadas en los 14 taxones. Sin embargo, algunos marcadores no recuperaban la variabilidad detectada *in silico*.

Recientemente, y fruto de una colaboración con un grupo del Jardín Botánico de Madrid hemos podido testar la eficiencia de **DOMINO** con datos reales de Illumina (Garrido et al. datos no publicados). En concreto, mediante el uso de datos genómicos de 5 especies de hongos liquenizados del género *Ramalina* (Lecanorales, Ramalinaceae), hemos procesado, ensamblado *de novo* y mapeado un total de 30x10^6 de *reads* PE (125 bp) de Illumina por especie. Tras aplicar el módulo para detectar marcadores moleculares se identificaron 235 candidatos entre las 5 especies. De estos 235 se seleccionaron 11 marcadores que fueron validados mediante PCR en las especies del panel, y actualmente se están utilizando para obtener la filogenia de 100 especies del género *Ramalina*. Algunos problemas observados en la amplificación de los marcadores seleccionados, como la aparición de doble banda (secuenciación de parálogos recientes) o la ausencia de banda, se pudieron solventar mediante la modificación de la temperatura de hibridación de los cebadores.

En comparación con otros programas informáticos, **DOMINO** combina una alta flexibilidad -facilita el análisis de una gran diversidad de datos NGS- con el uso simple de una GUI (*Graphic User Interface*) que permite que el software sea utilizado por investigadores sin conocimientos bioinformáticos. **DOMINO** es una herramienta bioinformática útil en diversas áreas de la biología evolutiva que requieren marcadores moleculares como por ejemplo la filogenética, la fileogeografía y la genética de poblaciones.

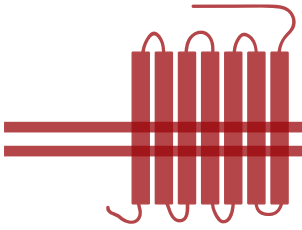
2.3 Validación de datos de RNAseq como marcadores moleculares para aplicaciones filogenéticas

Los datos de RNAseq de 454 en *M. calpeiana* nos han servido para validar el rango de aplicación de los mismos en aplicaciones filogenéticas. Para ello, utilizamos las proteínas que componen la matriz de datos d327 de Bond et al. 2014 (Bond et al. 2014) para buscar los potenciales ortólogos (*orthologous genes* - OGs) en *M. calpeiana*, e hicimos una

reconstrucción filogenética incluyendo taxones de los 5 mayores linajes de arañas. El árbol filogenético inferido por máxima verosimilitud, enraizado con *Liphistus*, perteneciente al suborden Mesothelae, grupo hermano del resto de arañas, migalomorfas y araneomorfas, (Bond et al. 2014) como grupo externo, recupera los mismos grupos filogenéticos focales obtenidos en trabajos anteriores (Bond et al. 2014), y muestra a *M. calpeiana* como el taxón hermano del género *Paratropis*, aunque con bajo soporte estadístico (*bootstrap* 57%) dentro del clado de Avicularioidea, pero fuera del subclado Bipectina. De hecho, en un estudio filogenético y biogeográfico reciente del género *Macrothele* (Opatova y Arnedo 2014), basado en un muestreo taxonómico más denso, pero con menor cobertura génica (3 genes), también recupera a *M. calpeiana* en una posición similar a la inferida por nuestros análisis filogenéticos. Más recientemente, un estudio incluyendo transcriptomas de un gran número de arañas, que representan cerca de un cuarto de las familias conocidas, y con datos generados con Illumina (incluyendo nueva secuenciación de *M calpeiana*) corrobora también nuestros resultados (Fernández et al. 2018). Es decir, los datos del transcriptoma se pueden utilizar fácilmente para realizar inferencias filogenéticas, y dado que suelen ser regiones más conservadas que las regiones no codificantes, son útiles para resolver filogenias a nivel de familia.

2.4 Integración de herramientas bioinformáticas

Los softwares desarrollados en esta tesis doctoral se siguen desarrollando activamente gracias a la implementación de nuevas utilidades que permitan realizar todo el análisis bioinformático relacionado con datos de RNAseq, tanto a nivel del tratamiento de lecturas crudas (pre-procesamiento, ensamblaje, mapeo, etc.), como el del desarrollo de marcadores a partir de datos de transcriptomas. Además, se está trabajando en la adaptación de la interfaz gráfica a diversos S.O para que se facilite su uso a usuarios inexpertos en entornos bioinformáticos complejos. En un futuro próximo, esperamos implementar todas estas herramientas bioinformáticas en un sistema único que permita procesar y analizar los datos de NGS de múltiples fuentes en su aplicación en estudios de biología evolutiva. Para ello se pretende integrar **DOMINO** con **TRUFA**, y migrar el conjunto del sistema para que pueda ser ejecutado desde una nube de cómputo.



CONCLUSIONES

Conclusiones

- Hemos identificado algunos genes candidatos del sistema quimiosensorial en el transcriptoma de la especie de araña migalomorfa *M. calpeiana*.

- Hemos identificado miembros de la familia de los *Irs* y las *Npc2*, a pesar de que la baja cobertura de la secuenciación de la tecnología por 454 no ofrecen la profundidad suficiente para detectar genes con baja expresión.

- Las búsquedas exhaustivas con modelos proteicos de HMM incluyendo un mayor número de secuencias genómicas de quelicerados, nos ha permitido identificar miembros de las familias quimiosensoriales de las *Obp-like*, *Snmps*, *Trps*, *Enacs* y *Grs*.

- Hemos participado en el desarrollo de **TRUFA,** una herramienta bioinformática para analizar transcriptomas de organismos no modelo, que realiza todos los procesos bioinformáticos comunes en los análisis de datos de RNAseq.

- Hemos validado la implementación de **TRUFA** con los datos distribuidos por lo softwares que integran el *pipeline* desarrollado, obteniendo resultados satisfactorios.

- Hemos desarrollado **DOMINO,** una herramienta bioinformática para generar marcadores moleculares para su uso en filogenética y genética de poblaciones. Esta herramienta permite realizar todas las etapas comunes para el tratamiento de secuencias de NGS. Sirve tanto para el análisis de datos genómicos como procedentes de técnicas de reducción genómica o incluso de alineamientos múltiples de secuencias (MSA).

- Hemos validado **DOMINO** a través de datos simulados, y hemos demostrado que es eficiente para detectar marcadores moleculares en un conjunto amplio de situaciones.

- Hemos validado **DOMINO** a través de una aproximación experimental; en particular en el desarrollo de marcadores en especies del género de arañas migalomorfas de trampilla *Nemesiidae*. Los resultados confirman que **DOMINO** es una herramienta apropiada para el desarrollo y la selección de marcadores.

- Los resultados de la inferencia de la posición filogenética de *M. calpeiana* en el árbol del suborden de arañas Mygalomorphae indican que los marcadores derivados del transcriptoma son adecuados para resolver estudios filogenéticos profundos.

BIBLIOGRAFÍA

Bibliografía

Abalde, Samuel, Manuel J Tenorio, Carlos M L Afonso, y Rafael Zardoya. 2018. «Conotoxin Diversity in Chelyconus ermineus (Born, 1778) and the Convergent Origin of Piscivory in the Atlantic and Indo-Pacific Cones» ed. Mandë Holford. *Genome Biology and Evolution* 10(10): 2643-62.

Abril, Josep F., y Sergi Castellano. 2019. «Genome Annotation». *Encyclopedia of Bioinformatics and Computational Biology*: 195-209.

Abuin, Liliane et al. 2011. «Functional Architecture of Olfactory Ionotropic Glutamate Receptors». *Neuron* 69(1): 44-60.

Al-Anzi, Bader, W. Daniel Tracey, y Seymour Benzer. 2006. «Response of Drosophila to Wasabi Is Mediated by painless, the Fly Homolog of Mammalian TRPA1/ANKTM1». *Current Biology* 16(10): 1034-40.

Almeida, Francisca C, Alejandro Sánchez-Gracia, Jose Luis Campos, y Julio Rozas. 2014. «Family size evolution in Drosophila chemosensory gene families: a comparative analysis with a critical appraisal of methods.» *Genome biology and evolution* 6(7): 1669-82.

Altschul, S F et al. 1990. «Basic local alignment search tool.» *Journal of molecular biology* 215(3): 403-10.

Altshuler, D et al. 2000. «An SNP map of the human genome generated by reduced representation shotgun sequencing.» *Nature* 407(6803): 513-16.

Anders, Simon, Paul Theodor Pyl, y Wolfgang Huber. 2015. «HTSeq--a Python framework to work with high-throughput sequencing data.» *Bioinformatics (Oxford, England)* 31(2): 166-69.

Anderson, S. 1981. «Shotgun DNA sequencing using cloned DNase I-generated fragments.» *Nucleic acids research* 9(13): 3015-27.

Ando, Toshiya et al. 2019. «Nanopore Formation in the Cuticle of an Insect Olfactory Sensillum». *Current Biology* 29(9): 1512-1520.e6.

Angeli, Sergio et al. 1999. «Purification, structural characterization, cloning and immunocytochemical localization of chemoreception proteins from Schistocerca gregaria». *European Journal of Biochemistry* 262(3): 745-54.

Anholt, Robert R. H., y Trudy F. C. Mackay. 2001. «The Genetic Architecture of Odor-Guided Behavior in Drosophila melanogaster». *Behavior Genetics* 31(1): 17-27.

Application EP-2099455-A1. «METHODS AND COMPOSITIONS FOR REPELLING ARTHROPODS - Dimensions».

Armstrong, Neali, Yu Sun, Guo-Qiang Chen, y Eric Gouaux. 1998. «Structure of a glutamate-receptor ligand-binding core in complex with kainate». *Nature* 395(6705): 913-17.

Avise, J C et al. 1987. «Intraspecific Phylogeography: The Mitochondrial DNA Bridge Between Population Genetics and Systematics». *Annual Review of Ecology and Systematics* 18(1): 489-522.

Avise, John C. 1994. *Molecular Markers, Natural History and Evolution*. Springer US.

Avise, John C. 2000. *Phylogeography : the history and formation of species*. Harvard University Press.

Backström, Niclas, Sofie Fagerberg, y Hans Ellegren. 2007. «Genomics of natural bird populations: a gene-based set of reference markers evenly spread across the avian genome». *Molecular Ecology* 17(4): 964-80.

Bankevich, Anton et al. 2012. «SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing.» *Journal of computational biology : a journal of computational molecular cell biology* 19(5): 455-77.

Bargmann, Cornelia. 2006. «Chemosensation in C. elegans». *WormBook*: 1-29.

Barnett, D. W. et al. 2011. «BamTools: a C++ API and toolkit for analyzing and managing BAM files». *Bioinformatics* 27(12): 1691-92.

Baruzzo, Giacomo et al. 2017. «Simulation-based comprehensive benchmarking of RNA-seq aligners». *Nature Methods* 14(2): 135-39.

Bellvert, Adrià, y Miquel A. Arnedo. 2016. «Threatened or Threatening? Evidence for Independent Introductions of *Macrothele calpeiana* () (Araneae: Hexathelidae) and First Observation of Reproduction Outside its Natural Distribution Range». *Arachnology* 17(3): 137-41.

Bentley, David R. et al. 2008. «Accurate whole human genome sequencing using reversible terminator chemistry». *Nature* 456(7218): 53-59.

Benton, Richard. 2015. «Multigene Family Evolution: Perspectives from Insect Chemoreceptors». *Trends in Ecology & Evolution* 30(10): 590-600.

Benton, Richard, Kirsten S. Vannice, Carolina Gomez-Diaz, y Leslie B. Vosshall. 2009. «Variant Ionotropic Glutamate Receptors as Chemosensory Receptors in Drosophila». *Cell* 136(1): 149-62.

Boisvert, Sébastien, François Laviolette, y Jacques Corbeil. 2010. «Ray: Simultaneous Assembly of Reads from a Mix of High-Throughput Sequencing Technologies». *Journal of Computational Biology* 17(11): 1519-33.

Bond, Jason E et al. 2014. «Phylogenomics resolves a spider backbone phylogeny and rejects a prevailing paradigm for orb web evolution.» *Current biology : CB* 24(15): 1765-71.

Botstein, D, R L White, M Skolnick, y R W Davis. 1980. «Construction of a genetic linkage map in man using restriction fragment length polymorphisms.» *American journal of human genetics* 32(3): 314-31.

Brewer, Michael S., Lynn Swafford, Chad L. Spruill, y Jason E. Bond. 2013. «Arthropod Phylogenetics in Light of Three Novel Millipede (Myriapoda: Diplopoda) Mitochondrial Genomes with Comments on the Appropriateness of Mitochondrial Genome Sequence Data for Inferring Deep Level Relationships» ed. Andreas Hejnol. *PLoS ONE* 8(7): e68005.

Brito, Patrícia H, y Scott V Edwards. 2009. «Multilocus phylogeography and phylogenetics using sequence-based markers.» *Genetica* 135(3): 439-55.

Buck, L, y R Axel. 1991. «A novel multigene family may encode odorant receptors: a molecular basis for odor recognition.» *Cell* 65(1): 175-87.

Burge, Chris, y Samuel Karlin. 1997. «Prediction of complete gene structures in human genomic DNA». *Journal of Molecular Biology* 268(1): 78-94.

Bybee, Seth M et al. 2011. «Targeted amplicon sequencing (TAS): a scalable next-gen approach to multilocus, multitaxa phylogenetics.» *Genome biology and evolution* 3: 1312-23.

Camacho, Christiam et al. 2009. «BLAST+: architecture and applications.» *BMC bioinformatics* 10(1): 421.

Carr, Ann L et al. 2017. «Tick Haller's Organ, a New Paradigm for Arthropod Olfaction: How Ticks Differ from Insects.» *International journal of molecular sciences* 18(7).

Caterina, Michael J. et al. 1997. «The capsaicin receptor: a heat-activated ion channel in the pain pathway». *Nature* 389(6653): 816-24.

Cattaneo, Alberto Maria et al. 2016. «TRPA5, an Ankyrin Subfamily Insect TRP Channel, is Expressed in Antennae of *Cydia pomonella* (Lepidoptera: Tortricidae) in Multiple Splice Variants». *Journal of Insect Science* 16(1): 83.

Chahda, J. Sebastian et al. 2019. «The molecular and cellular basis of olfactory response to tsetse fly attractants» ed. Gaiti Hasan. *PLOS Genetics* 15(3): e1008005.

Chang, Zheng et al. 2015. «Bridger: a new framework for de novo transcriptome assembly using RNA-seq data». *Genome Biology* 16(1): 30.

Chen, Shifu, Yanqing Zhou, Yaru Chen, y Jia Gu. 2018. «fastp: an ultra-fast all-in-one FASTQ preprocessor». *Bioinformatics* 34(17): i884-90.

Chen, Ting-Wen et al. 2012. «FastAnnotator--an efficient transcript annotation web tool.» *BMC genomics* 13 Suppl 7(Suppl 7): S9.

Chen, Z., Q. Wang, y Z. Wang. 2010. «The Amiloride-Sensitive Epithelial Na+ Channel PPK28 Is Essential for Drosophila Gustatory Water Reception». *Journal of Neuroscience* 30(18): 6247-52.

Chintapalli, Venkateswara R, Jing Wang, y Julian A T Dow. 2007. «Using FlyAtlas to identify better Drosophila melanogaster models of human disease». *Nature Genetics* 39(6): 715-20.

Chipman, Ariel D. et al. 2014. «The First Myriapod Genome Sequence Reveals Conservative Arthropod Gene Content and Genome Organisation in the Centipede Strigamia maritima» ed. Chris Tyler-Smith. *PLoS Biology* 12(11): e1002005.

Chowański, Szymon et al. 2016. «A Review of Bioinsecticidal Activity of Solanaceae Alkaloids». *Toxins* 8(3): 60.

Clyne, P J et al. 1999. «A novel family of divergent seven-transmembrane proteins: candidate odorant receptors in Drosophila.» *Neuron* 22(2): 327-38.

Cock, Peter J. A. et al. 2010. «The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants». *Nucleic Acids Research* 38(6): 1767-71.

Collins, N. M., and Wells, S. 1987. Augier H. Nature & Environment Series No. 35. Council of Europe, Strasbourg. pp. 162. *Invertebrates in need of special protection in Europe. Augier H. Nature & Environment Series No. 35. Council of Europe, Strasbourg. pp. 162.*

Conesa, Ana et al. 2005. «Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research.» *Bioinformatics (Oxford, England)* 21(18): 3674-76.

Cooper, D N et al. 1985. «An estimate of unique DNA sequence heterozygosity in the human genome.» *Human genetics* 69(3): 201-5.

Croset, Vincent et al. 2010. «Ancient Protostome Origin of Chemosensory Ionotropic Glutamate Receptors and the Evolution of Insect Taste and Olfaction» ed. David L. Stern. *PLoS Genetics* 6(8): e1001064.

Daley, Allison C, Jonathan B Antcliffe, Harriet B Drage, y Stephen Pates. 2018. «Early fossil record of Euarthropoda and the Cambrian Explosion.» *Proceedings of the National Academy of Sciences of the United States of America* 115(21): 5323-31.

Danecek, Petr et al. 2011. «The variant call format and VCFtools.» *Bioinformatics (Oxford, England)* 27(15): 2156-58.

Darwin, Charles et al. 1859. *On the origin of species by means of natural selection, or, The preservation of favoured races in the struggle for life /*. London : John Murray, Albemarle Street,.

Davey, J. W., M. L. Blaxter, Mark L Blaxter, y Mark W Blaxter. 2010. «RADSeq: next-generation population genetics». *Briefings in Functional Genomics* 9(5-6): 416-23.

Davey, John W, John L Davey, Mark L Blaxter, y Mark W Blaxter. 2010. «RADSeq: next-generation population genetics.» *Briefings in functional genomics* 9(5-6): 416-23.

DePristo, Mark A et al. 2011. «A framework for variation discovery and genotyping using next-generation DNA sequencing data». *Nature Genetics* 43(5): 491-98.

Derby, Charles D., Mihika T. Kozma, Adriano Senatore, y Manfred Schmidt. 2016. «Molecular Mechanisms of Reception and Perireception in Crustacean Chemoreception: A Comparative Review». *Chemical Senses* 41(5): 381-98.

Devaud, Jean Marc. 2003. «Experimental studies of adult Drosophila chemosensory behaviour.» *Behavioural processes* 64(2): 177-96.

Dieringer, Daniel, y Christian Schlötterer. 2003. «Two distinct modes of microsatellite mutation processes: evidence from the complete genomic sequences of nine species.» *Genome research* 13(10): 2242-51.

Dobin, Alexander et al. 2013. «STAR: ultrafast universal RNA-seq aligner». *Bioinformatics* 29(1): 15-21.

Dunlop, Jason A., Gerhard Scholtz, y Paul A. Selden. 2013. «Water-to-Land Transitions». En *Arthropod Biology and Evolution*, Berlin, Heidelberg: Springer Berlin Heidelberg, 417-39.

Eddy, Sean R. 2009. «A new generation of homology search tools based on probabilistic inference.» *Genome informatics. International Conference on Genome Informatics* 23(1): 205-11.

Edge, Peter, Vineet Bafna, y Vikas Bansal. 2017. «HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies». *Genome Research* 27(5): 801-12.

Edwards, Mary C., y Richard A. Gibbs. 1994. «Multiplex PCR: Advantages, development, and applications». *Genome Research*.

Eisenberg, Eli, y Erez Y. Levanon. 2013. «Human housekeeping genes, revisited». *Trends in Genetics* 29(10): 569-74.

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

Eliash, N. et al. 2019. «Varroa chemosensory proteins: some are conserved across Arthropoda but others are arachnid specific». *Insect Molecular Biology* 28(3): 321-41.

Ellegren, Hans. 2008. «Comparative genomics and the study of evolution by natural selection». *Molecular Ecology* 17(21): 4586-96.

Ellegren, Hans. 2014. «Genome sequencing and population genomics in non-model organisms». *Trends in Ecology & Evolution* 29(1): 51-63.

Elshire, Robert J. et al. 2011. «A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species» ed. Laszlo Orban. *PLoS ONE* 6(5): e19379.

Emerson, Kevin J et al. 2010. «Resolving postglacial phylogeography using high-throughput sequencing.» *Proceedings of the National Academy of Sciences of the United States of America* 107(37): 16196-200.

Engström, Pär G et al. 2013. «Systematic evaluation of spliced alignment programs for RNA-seq data». *Nature Methods* 10(12): 1185-91.

Ernst, Alfred, y Jörg Rosenberg. 2003. 44 African Invertebrates *African invertebrates: a journal of biodiversity research. SENSILLA COELOCONICA ON MAXILLIPEDES*. eds. South Africa). Council. Natal Museum (Pietermaritzburg y South Africa). Council. KwaZulu-Natal Museum (Pietermaritzburg. Council of the Natal Museum.

Faircloth, Brant C. et al. 2012. «Ultraconserved Elements Anchor Thousands of Genetic Markers Spanning Multiple Evolutionary Timescales». *Systematic Biology* 61(5): 717-26.

Fang, Dong et al. 2011. «Identification of Genes Directly Involved in Shell Formation and Their Functions in Pearl Oyster, Pinctada fucata» ed. Anna Mitraki. *PLoS ONE* 6(7): e21860.

Fernández, Rosa et al. 2018. «Phylogenomics, Diversification Dynamics, and Comparative Transcriptomics across the Spider Tree of Life». *Current Biology* 28(9): 1489-1497.e5.

Fiala, André. 2007. «Olfaction and olfactory learning in Drosophila: recent progress». *Current Opinion in Neurobiology* 17(6): 720-26.

Fink, Inge R. et al. 2015. «Molecular and functional characterization of the scavenger receptor CD36 in zebrafish and common carp». *Molecular Immunology* 63(2): 381-93.

Fleischmann, R. et al. 1995. «Whole-genome random sequencing and assembly of Haemophilus influenzae Rd». *Science* 269(5223): 496-512.

Flusberg, Benjamin A et al. 2010. «Direct detection of DNA methylation during single-molecule, real-time sequencing». *Nature Methods* 7(6): 461-65.

Foelix, Rainer F. 1970. «Chemosensitive hairs in spiders». *Journal of Morphology* 132(3): 313-33.

Freeman, J. L. et al. 2006. «Copy number variation: New insights in genome diversity». *Genome Research* 16(8): 949-61.

Frías-López, C. et al. 2015. «Comparative analysis of tissue-specific transcriptomes in the funnel-web spider Macrothele calpeiana (Araneae, Hexathelidae)». *PeerJ* 2015(6).

Fujii, Shinsuke et al. 2015. «Drosophila Sugar Receptors in Sweet Taste Perception, Olfaction, and Internal Nutrient Sensing». *Current Biology* 25(5): 621-27.

Fürstenberg-Hägg, Joel, Mika Zagrobelny, y Søren Bak. 2013. «Plant Defense against Insect Herbivores». *International Journal of Molecular Sciences* 14(5): 10242-97.

Galindo, J., J. W. Grahame, y R. K. Butlin. 2010. «An EST-based genome scan using 454 sequencing in the marine snail Littorina saxatilis». *Journal of Evolutionary Biology* 23(9): 2004-16.

Ganske, Anne-Sarah, y Gabriele Uhl. 2018. «The sensory equipment of a spider – A morphological survey of different types of sensillum in both sexes of Argiope bruennichi (Araneae, Araneidae)». *Arthropod Structure & Development* 47(2): 144-61.

Gao, Li et al. 2005. «Effects of spider Macrothele raven venom on cell proliferation and cytotoxicity in HeLa cells.» *Acta pharmacologica Sinica* 26(3): 369-76.

Gilad, Yoav, Jonathan K. Pritchard, y Kevin Thornton. 2009. «Characterizing natural variation using next-generation sequencing technologies». *Trends in Genetics* 25(10): 463-71.

Gilbert, W, y A Maxam. 1973. «The nucleotide sequence of the lac operator.» *Proceedings of the National Academy of Sciences of the United States of America* 70(12): 3581-84.

Gillespie, R G, H B Croom, y S R Palumbi. 1994. «Multiple origins of a spider radiation in Hawaii.» *Proceedings of the National Academy of Sciences of the United States of America* 91(6): 2290-94.

Giribet, Gonzalo, y Gregory Edgecombe. 2013. «Stable phylogenetic patterns in scutigeromorph centipedes (Myriapoda : Chilopoda : Scutigeromorpha): dating the diversification of an ancient lineage of terrestrial arthropods». *Invertebrate Systematics* 27.

Gnirke, Andreas et al. 2009. «Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing». *Nature Biotechnology* 27(2): 182-89.

Van der Goes van Naters, Wynand, y John R. Carlson. 2007. «Receptors and Neurons for Fly Odors in Drosophila». *Current Biology* 17(7): 606-12.

Gonzalo Giribet, Gregory D. Edgecombe y Ward C. Wheeler. 2000. «Sistemática y filogenia de Artrópodos: estado de la cuestión con énfasis en datos moleculares».

Grabherr, Manfred G et al. 2011. «Full-length transcriptome assembly from RNA-Seq data without a reference genome». *Nature Biotechnology* 29(7): 644-52.

Griffith, Malachi et al. 2015. «Informatics for RNA Sequencing: A Web Resource for Analysis on the Cloud» ed. Francis Ouellette. *PLOS Computational Biology* 11(8): e1004393.

Guigó, R, S Knudsen, N Drake, y T Smith. 1992. «Prediction of gene structure.» *Journal of molecular biology* 226(1): 141-57.

Gulia-Nuss, Monika et al. 2016. «Genomic insights into the Ixodes scapularis tick vector of Lyme disease». *Nature Communications*.

Guo, Huizhen et al. 2017. «Expression map of a complete set of gustatory receptor genes in chemosensory organs of Bombyx mori». *Insect Biochemistry and Molecular Biology* 82: 74-82.

Guo, Wei et al. 2011. «CSP and takeout genes modulate the switch between attraction and repulsion during behavioral phase change in the migratory locust». *PLoS Genetics*.

Haas, Brian J et al. 2013. «De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis.» *Nature protocols* 8(8): 1494-1512.

Harris, D J, y P J Mill. 1973. «The ultrastructure of chemoreceptor sensilla in Ciniflo (Araneida, Arachnida).» *Tissue & cell* 5(4): 679-89.

Harzsch, S., y J. Krieger. 2018. «Crustacean olfactory systems: A comparative review and a crustacean perspective on olfaction in insects». *Progress in Neurobiology* 161: 23-60.

Hedges, S Blair et al. 2015. «Tree of life reveals clock-like speciation and diversification.» *Molecular biology and evolution* 32(4): 835-45.

Hedin, Marshal C. 2001. «Molecular insights into species phylogeny, biogeography, and morphological stasis in the ancient spider genus Hypochilus (Araneae: Hypochilidae)». *Molecular Phylogenetics and Evolution* 18(2): 238-51.

Hennig, Willi. 1950. *Grundzüge einer Theorie der phylogenetischen Systematik*. ed. Berlin: Deutscher Zentralverlag.

Hennig, Willi. 1966. *Phylogenetic Systematics*. eds. D. Davis y Rainer Zangerl. Urbana: University of Illinois Press.

Hershey, A D, y M Chase. 1952. «Independent functions of viral protein and nucleic acid in growth of bacteriophage.» *The Journal of general physiology* 36(1): 39-56.

Hildebrand, John G., y Gordon M. Shepherd. 1997. «MECHANISMS OF OLFACTORY DISCRIMINATION:Converging Evidence for Common Principles Across Phyla». *Annual Review of Neuroscience* 20(1): 595-631.

Hoff, Katharina J. Et al. 2016. «BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS: Table 1.» *Bioinformatics* 32(5): 767-69.

Holt, Carson, y Mark Yandell. 2011. «MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects». *BMC Bioinformatics* 12(1): 491.

Hormiga, Gustavo, Miquel Arnedo, y Rosemary G. Gillespie. 2003. «Speciation on a Conveyor Belt: Sequential Colonization of the Hawaiian Islands by Orsonwelles Spiders (Araneae, Linyphiidae)» ed. Crandall Keith. *Systematic Biology* 52(1): 70-88.

Hoy, Marjorie A. et al. 2016. «Genome Sequencing of the Phytoseiid Predatory Mite Metaseiulus occidentalis Reveals Completely Atomized Hox Genes and Superdynamic Intron Evolution». *Genome Biology and Evolution*.

Hudson, Matthew E. 2008. «Sequencing breakthroughs for genomic ecology and evolutionary biology». *Molecular Ecology Resources* 8(1): 3-17.

Ishida, Yuko et al. 2014. «Niemann-Pick type C2 protein mediating chemical communication in the worker ant.» *Proceedings of the National Academy of Sciences of the United States of America* 111(10): 3847-52.

Jain, Miten, Hugh E. Olsen, Benedict Paten, y Mark Akeson. 2016. «The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community». *Genome Biology* 17(1): 239.

Jefferis, Gregory S. X. E., Elizabeth C. Marin, Reinhard F. Stocker, y Liqun Luo. 2001. «Target neuron prespecification in the olfactory map of Drosophila». *Nature* 414(6860): 204-8.

Jennings, W Bryan, y Scott V Edwards. 2005. «Speciational history of Australian grass finches (Poephila) inferred from thirty gene trees.» *Evolution; international journal of organic evolution* 59(9): 2033-47.

Jiao, Xiaoguo et al. 2011. «Chemoreceptors distribution and relative importance of male forelegs and palps in intersexual chemical communication of the wolf spider Pardosa astrigera». *Chemoecology*.

Jiménez-Valverde, Alberto, Arthur E. Decae, y Miquel A. Arnedo. 2011. «Environmental suitability of new reported localities of the funnelweb spider Macrothele calpeiana: an assessment using potential distribution modelling with presence-only techniques». *Journal of Biogeography* 38(6): 1213-23.

Jin, Xin, Tal Soo Ha, y Dean P Smith. 2008. «SNMP is a signaling component required for pheromone sensitivity in Drosophila.» *Proceedings of the National Academy of Sciences of the United States of America* 105(31): 10996-1.

Johannesen, Jes, Anna Hennig, Bianca Dommermuth, y Jutta M. Schneider. 2002. «Mitochondrial DNA distributions indicate colony propagation by single matri-lineages in the social spider Stegodyphus dumicola (Eresidae)». *Biological Journal of the Linnean Society* 76(4): 591-600.

Jones, Philip et al. 2014. «InterProScan 5: genome-scale protein function classification.» *Bioinformatics (Oxford, England)* 30(9): 1236-40.

Jones, Walton D., Pelin Cayirlioglu, Ilona Grunwald Kadow, y Leslie B. Vosshall. 2007. «Two chemosensory receptors together mediate carbon dioxide detection in Drosophila». *Nature* 445(7123): 86-90.

Joseph, Ryan M., y John R. Carlson. 2015a. «Drosophila Chemoreceptors: A Molecular Interface Between the Chemical World and the Brain». *Trends in Genetics*.

Joseph, Ryan M., y John R. Carlson. 2015b. «Drosophila Chemoreceptors: A Molecular Interface Between the Chemical World and the Brain». *Trends in Genetics* 31(12): 683-95.

Kadlec, Malvina, Dirk U Bellstedt, Nicholas C Le Maitre, y Michael D Pirie. 2017. «Targeted NGS for species level phylogenomics : ‘‘ made to measure ’’ or ‘‘ one size fits all ’’?» : 1-25.

Kenning, Matthes, Carsten H.G. Müller, y Andy Sombke. 2017. «The ultimate legs of Chilopoda (Myriapoda): a review on their morphological disparity and functional variability». *PeerJ* 5: e4023.

Khan, Abdul Rafay et al. 2018. «A Comprehensive Study of De Novo Genome Assemblers: Current Challenges and Future Prospective». *Evolutionary Bioinformatics* 14: 117693431875865.

Kim, Daehwan, Ben Langmead, y Steven L Salzberg. 2015. «HISAT: a fast spliced aligner with low memory requirements». *Nature Methods* 12(4): 357-60.

Kim, Sang Hoon et al. 2010. «Drosophila TRPA1 channel mediates chemical avoidance in gustatory receptor neurons.» *Proceedings of the National Academy of Sciences of the United States of America* 107(18): 8440-45.

Kobayashi, Masaaki et al. 2017. «Heap: a highly sensitive and accurate SNP detection tool for low-coverage high-throughput sequencing data». *DNA Research* 24(4): 397-405.

Korf, Ian. 2004. «Gene finding in novel genomes.» *BMC bioinformatics* 5: 59.

Kornobis, Etienne et al. 2015. «TRUFA: A User-Friendly Web Server for *de novo* RNA-seq Analysis Using Cluster Computing». *Evolutionary Bioinformatics* 11: EBO.S23873.

Kozma, Mihika T. et al. 2018. «Chemoreceptor proteins in the Caribbean spiny lobster, Panulirus argus: Expression of Ionotropic Receptors, Gustatory Receptors, and TRP channels in two chemosensory organs and brain» ed. Michel Renou. *PLOS ONE* 13(9): e0203935.

Krieger, M. J. B., y Kenneth G Ross. 2002. «Identification of a Major Gene Regulating Complex Social Behavior». *Science* 295(5553): 328-32.

Kumar, S. et al. 2012. «Statistics and Truth in Phylogenomics». *Molecular Biology and Evolution* 29(2): 457-72.

Kumar, Santosh, Travis W Banks, y Sylvie Cloutier. 2012. «SNP Discovery through Next-Generation Sequencing and Its Applications.» *International journal of plant genomics* 2012: 831460.

Kumar, Satish et al. 2009. «Reconstructing Indian-Australian phylogenetic link». *BMC Evolutionary Biology* 9(1): 173.

Kumar, Sujai, y Mark L Blaxter. 2010. «Comparing de novo assemblers for 454 transcriptome data». *BMC Genomics* 11(1): 571.

Kwon, Young et al. 2010. «Drosophila TRPA1 Channel Is Required to Avoid the Naturally Occurring Insect Repellent Citronellal». *Current Biology* 20(18): 1672-78.

Lam, Kin Chung et al. 2012. «The NSL complex regulates housekeeping genes in Drosophila.» *PLoS genetics* 8(6): e1002736.

Lander, E S et al. 2001. «Initial sequencing and analysis of the human genome». *Nature* 409(6822): 860-921.

Langmead, Ben, y Steven L Salzberg. 2012. «Fast gapped-read alignment with Bowtie 2». *Nature Methods* 9(4): 357-59.

Langmead, Ben, Cole Trapnell, Mihai Pop, y Steven L Salzberg. 2009. «Ultrafast and memory-efficient alignment of short DNA sequences to the human genome». *Genome Biology* 10(3): R25.

Law, Charity W, Yunshun Chen, Wei Shi, y Gordon K Smyth. 2014. «voom: precision weights unlock linear model analysis tools for RNA-seq read counts». *Genome Biology* 15(2): R29.

Leal, W S, L Nikonova, y G Peng. 1999. «Disulfide structure of the pheromone binding protein from the silkworm moth, Bombyx mori.» *FEBS letters* 464(1-2): 85-90.

Lee, Hayan et al. 2016. «Third-generation sequencing and the future of genomics». *bioRxiv*: 048603.

Leelatanawit, Rungnapa, Amornpan Klanchui, Umaporn Uawisetwathana, y Nitsara Karoonuthaisiri. 2012. «Validation of Reference Genes for Real-Time PCR of Reproductive System in the Black Tiger Shrimp» ed. Christian Schönbach. *PLoS ONE* 7(12): e52677.

Lemmon, Alan R., Sandra A. Emme, y Emily Moriarty Lemmon. 2012. «Anchored Hybrid Enrichment for Massively High-Throughput Phylogenomics». *Systematic Biology* 61(5): 727-44.

Lemmon, Emily Moriarty, y Alan R. Lemmon. 2013. «High-Throughput Genomic Data in Systematics and Phylogenetics». *Annual Review of Ecology, Evolution, and Systematics* 44(1): 99-121.

Li, Bo et al. 2014. «Evaluation of de novo transcriptome assemblies from RNA-Seq data». *Genome Biology* 15(12): 553.

Li, H. et al. 2009. «The Sequence Alignment/Map format and SAMtools». *Bioinformatics* 25(16): 2078-79.

Li, H. 2011. «A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data». *Bioinformatics* 27(21): 2987-93.

Li, Heng, y Richard Durbin. 2010. «Fast and accurate long-read alignment with Burrows-Wheeler transform.» *Bioinformatics (Oxford, England)* 26(5): 589-95.

Li, R. et al. 2009. «SNP detection for massively parallel whole-genome resequencing». *Genome Research* 19(6): 1124-32.

Li, Weizhong, y Adam Godzik. 2006. «Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences.» *Bioinformatics (Oxford, England)* 22(13): 1658-59.

Liao, Y., G. K. Smyth, y W. Shi. 2014. «featureCounts: an efficient general purpose program for assigning sequence reads to genomic features». *Bioinformatics* 30(7): 923-30.

Lieberman-Aiden, Erez et al. 2009. «Comprehensive mapping of long-range interactions reveals folding principles of the human genome.» *Science (New York, N.Y.)* 326(5950): 289-93.

Liou, Heng-Ling et al. 2006. «NPC2, the Protein Deficient in Niemann-Pick C2 Disease, Consists of Multiple Glycoforms That Bind a Variety of Sterols». *Journal of Biological Chemistry* 281(48): 36710-23.

Littleton, J T, y B Ganetzky. 2000. «Ion channels and synaptic organization: analysis of the Drosophila genome.» *Neuron* 26(1): 35-43.

Liu, Lei et al. 2003. «Contribution of Drosophila DEG/ENaC genes to salt taste.» *Neuron* 39(1): 133-46.

Liu, Zhonghua et al. 2012. «The Venom of the Spider Macrothele Raveni Induces Apoptosis in the Myelogenous Leukemia K562 Cell Line.» *Leukemia research* 36(8): 1063-66.

Love, Michael I, Wolfgang Huber, y Simon Anders. 2014. «Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2». *Genome Biology* 15(12): 550.

Lozano-Fernandez, Jesus et al. 2019. «Increasing species sampling in chelicerate genomic-scale datasets provides support for monophyly of Acari and Arachnida». *Nature Communications* 10(1): 2295.

Lupski, James R. 2015. «Structural variation mutagenesis of the human genome: Impact on disease and evolution.» *Environmental and molecular mutagenesis* 56(5): 419-36.

Majoros, W. H., M. Pertea, y S. L. Salzberg. 2004. «TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders». *Bioinformatics* 20(16): 2878-79.

Mamanova, Lira et al. 2010. «Target-enrichment strategies for next-generation sequencing». *Nature Methods* 7(2): 111-18.

Mano, Itzhak, y Monica Driscoll. 1999. «DEG/ENaC channels: A touchy superfamily that watches its salt». *BioEssays* 21(7): 568-78.

Margulies, Marcel et al. 2005. «Genome sequencing in microfabricated high-density picolitre reactors». *Nature* 437(7057): 376.

Martín-Blázquez, R., B. Chen, L. Kang, y M. Bakkali. 2017. «Evolution, expression and association of the chemosensory protein genes with the outbreak phase of the two main pest locusts». *Scientific Reports*.

Martin, Jeffrey A., y Zhong Wang. 2011. «Next-generation transcriptome assembly». *Nature Reviews Genetics* 12(10): 671-82.

Matsuo, Takashi et al. 2007. «Odorant-Binding Proteins OBP57d and OBP57e Affect Taste Perception and Host-Plant Preference in Drosophila sechellia» ed. Mohamed A. F Noor. *PLoS Biology* 5(5): e118.

Maxam, A. M., y W. Gilbert. 1977. «A new method for sequencing DNA.» *Proceedings of the National Academy of Sciences* 74(2): 560-64.

McCormack, John E. et al. 2013. «Applications of next-generation sequencing to phylogeography and phylogenetics». *Molecular Phylogenetics and Evolution* 66(2): 526-38.

Metzker, Michael L. 2010. «Sequencing technologies — the next generation». *Nature Reviews Genetics* 11(1): 31-46.

Mielczarek, M., y J. Szyda. 2016. «Review of alignment and SNP calling algorithms for next-generation sequencing data». *Journal of Applied Genetics* 57(1): 71-79.

Miller, Jason R, Sergey Koren, y Granger Sutton. 2010. «Assembly algorithms for next-generation sequencing data.» *Genomics* 95(6): 315-27.

Missbach, Christine et al. 2014. «Evolution of insect olfactory receptors». *eLife* 3: e02115.

Missbach, Christine, Heiko Vogel, Bill S. Hansson, y Ewald Groβe-Wilde. 2015. «Identification of Odorant Binding Proteins and Chemosensory Proteins in Antennal Transcriptomes of the Jumping Bristletail *Lepismachilis y-signata* and the Firebrat *Thermobia domestica:* Evidence for an Independent OBP–OR Origin». *Chemical Senses* 40(9): 615-26.

Mollo, Ernesto et al. 2017. «Taste and smell in aquatic and terrestrial environments.» *Natural product reports* 34(5): 496-513.

Mollo, Ernesto et al. 2014. «Sensing marine biomolecules: smell, taste, and the evolutionary transition from aquatic to terrestrial life». *Frontiers in Chemistry* 2: 92.

Montell, C. 2005. «The TRP Superfamily of Cation Channels». *Science Signaling* 2005(272): re3-re3.

Montell, Craig. 2005. «TRP channels in Drosophila photoreceptor cells.» *The Journal of physiology* 567(Pt 1): 45-51.

Montell C. 2009. «A taste of the Drosophila gustatory receptors». *Current Opinion in Neurobiology* 19(4): 345-53.

Moritz, C, T E Dowling, y W M Brown. 1987. «Evolution of Animal Mitochondrial DNA: Relevance for Population Biology and Systematics». *Annual Review of Ecology and Systematics* 18(1): 269-92.

Mullaney, Julienne M, Ryan E Mills, W Stephen Pittard, y Scott E Devine. 2010. «Small insertions and deletions (INDELs) in human genomes.» *Human molecular genetics* 19(R2): R131-6.

Mullis, Kary B., y Fred A. Faloona. 1987. «[21] Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction». En *Methods in enzymology*, , 335-50.

Natural History Museum Bern. 2019. «World Spider Catalog (2019) - Version 20.5.» https://wsc.nmbe.ch/ (25 de septiembre de 2019).

Neculai, Dante et al. 2013. «Structure of LIMP-2 provides functional insights with implications for SR-BI and CD36». *Nature* 504(7478): 172-76.

Ngoc, Phuong Cao Thi et al. 2016. «Complex Evolutionary Dynamics of Massively Expanded Chemosensory Receptor Families in an Extreme Generalist Chelicerate Herbivore». *Genome Biology and Evolution* 8(11): 3323-39.

Nielsen, Rasmus, Joshua S Paul, Anders Albrechtsen, y Yun S Song. 2011. «Genotype and SNP calling from next-generation sequencing data.» *Nature reviews. Genetics* 12(6): 443-51.

Nilius, Bernd. 2003. «From TRPs to SOCs, CCEs, and CRACs: consensus and controversies». *Cell Calcium* 33(5-6): 293-98.

Opatova, Vera, y Miquel A. Arnedo. 2014. «From Gondwana to Europe: inferring the origins of Mediterranean Macrothele spiders (Araneae : Hexathelidae) and the limits of the family Hexathelidae». *Invertebrate Systematics* 28(4): 361.

Patel, Ravi K, y Mukesh Jain. 2012. «NGS QC Toolkit: a toolkit for quality control of next generation sequencing data.» *PloS one* 7(2): e30619.

Pedersen, Stine Falsig, Grzegorz Owsianik, y Bernd Nilius. 2005. «TRP channels: An overview». *Cell Calcium* 38(3-4): 233-52.

Peier, Andrea M et al. 2002. «A TRP channel that senses cold stimuli and menthol.» *Cell* 108(5): 705-15.

Pelosi, Paolo et al. 2018. «Beyond chemoreception: diverse tasks of soluble olfactory proteins in insects». *Biological Reviews* 93(1): 184-200.

Pelosi, Paolo, Immacolata Iovinella, Antonio Felicioli, y Francesca R. Dani. 2014a. «Soluble proteins of chemical communication: an overview across arthropods». *Frontiers in Physiology* 5: 320.

Pelosi, Paolo et al. 2014b. «Soluble proteins of chemical communication: an overview across arthropods». *Frontiers in Physiology* 5: 320.

Peñalva-Arana, D Carolina, Michael Lynch, y Hugh M Robertson. 2009. «The chemoreceptor genes of the waterflea Daphnia pulex: many Grs but no Ors». *BMC Evolutionary Biology* 9(1): 79.

Pertea, Mihaela et al. 2016. «Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown». *Nature Protocols* 11(9): 1650-67.

Peterson, Brant K. et al. 2012. «Double Digest RADseq: An Inexpensive Method for De Novo SNP Discovery and Genotyping in Model and Non-Model Species» ed. Ludovic Orlando. *PLoS ONE* 7(5): e37135.

Pfaffl, Michael W, Ales Tichopad, Christian Prgomet, y Tanja P Neuvians. 2004. «Determination of stable housekeeping genes, differentially regulated target genes and sample integrity: BestKeeper--Excel-based tool using pair-wise correlations.» *Biotechnology letters* 26(6): 509-15.

Ponton, Fleur et al. 2011. «Evaluation of potential reference genes for reverse transcription-qPCR studies of physiological responses in Drosophila melanogaster». *Journal of Insect Physiology* 57(6): 840-50.

Puritz, Jonathan B., y Katie E. Lotterhos. 2018. «Expressed exome capture sequencing: A method for cost-effective exome sequencing for all organisms». *Molecular Ecology Resources* 18(6): 1209-22.

Putnam, Nicholas H. et al. 2016. «Chromosome-scale shotgun assembly using an in vitro method for long-range linkage». *Genome Research* 26(3): 342-50.

Roberts, Richard J, Mauricio O Carneiro, y Michael C Schatz. 2013. «The advantages of SMRT sequencing». *Genome Biology* 14(6): 405.

Robertson, H. M., C. G. Warr, y J. R. Carlson. 2003. «Molecular evolution of the insect chemoreceptor gene superfamily in Drosophila melanogaster». *Proceedings of the National Academy of Sciences* 100(Supplement 2): 14537-42.

Robinson, M. D., D. J. McCarthy, y G. K. Smyth. 2010. «edgeR: a Bioconductor package for differential expression analysis of digital gene expression data». *Bioinformatics* 26(1): 139-40.

Roulin, Anne et al. 2012. «The fate of duplicated genes in a polyploid plant genome.» *The Plant journal : for cell and molecular biology*: 143-53.

Rubinoff, Daniel, y Brenden S. Holland. 2005. «Between two extremes: mitochondrial DNA is neither the panacea nor the nemesis of phylogenetic and taxonomic inference.» *Systematic biology*.

Ruparel, H. et al. 2005. «Design and synthesis of a 3'-O-allyl photocleavable fluorescent nucleotide as a reversible terminator for DNA sequencing by synthesis». *Proceedings of the National Academy of Sciences* 102(17): 5932-37.

Rytz, Raphael, Vincent Croset, y Richard Benton. 2013. «Ionotropic Receptors (IRs): Chemosensory ionotropic glutamate receptors in Drosophila and beyond». *Insect Biochemistry and Molecular Biology* 43(9): 888-97.

Sachdeva, V., C.S. Kim, K.E. Jordan, y M.D. Winn. 2014. «Parallelization of the Trinity Pipeline for De Novo Transcriptome Assembly». En *2014 IEEE International Parallel & Distributed Processing Symposium Workshops*, IEEE, 566-75.

Sachidanandam, R et al. 2001. «A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms». *Nature* 409(6822): 928-33.

Saiki, R. et al. 1988. «Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase». *Science* 239(4839): 487-91.

Salgado, Vincent L. 2017. «Insect TRP channels as targets for insecticides and repellents.» *Journal of pesticide science* 42(1): 1-6.

Sánchez-Gracia, A, F G Vieira, y J Rozas. 2009. «Molecular evolution of the major chemosensory gene families in insects». *Heredity* 103(3): 208-16.

Sanger, F. et al. 1977. «Nucleotide sequence of bacteriophage φX174 DNA». *Nature* 265(5596): 687-95.

Sanger, F. et al. 1975. «The Croonian Lecture, 1975: Nucleotide Sequences in DNA». *Proceedings of the Royal Society of London. Series B, Biological Sciences* 191: 317-33.

Sanger, F., y A.R. Coulson. 1975. «A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase». *Journal of Molecular Biology* 94(3): 441-48.

Satake, H et al. 2004. «Rapid and efficient identification of cysteine-rich peptides by random screening of a venom gland cDNA library from the hexathelid spider Macrothele gigas». *Toxicon* 44(2): 149-56.

Sauné, Laure et al. 2015. «Isolation, characterization and PCR multiplexing of microsatellite loci for a mite crop pest, Tetranychus urticae (Acari: Tetranychidae).» *BMC research notes* 8: 247.

Scheuermann, Elizabeth A, y Dean P Smith. 2019. «Odor-Specific Deactivation Defects in a Drosophila Odorant Binding Protein Mutant.» *Genetics*: genetics.302629.2019.

Schierwater, B, y A Ender. 1993. «Different thermostable DNA polymerases may amplify different RAPD products.» *Nucleic acids research* 21(19): 4647-48.

Schlötterer, Christian. 2004. «The evolution of molecular markers — just a matter of fashion?» *Nature Reviews Genetics* 5(1): 63-69.

Schmieder, R., y R. Edwards. 2011. «Quality control and preprocessing of metagenomic datasets». *Bioinformatics* 27(6): 863-64.

Scholtz, Gerhard, y Gregory Edgecombe. 2005. «Heads, Hox and the phylogenetic position of trilobites». En , 139-65.

Shendure, Jay et al. 2017. «DNA sequencing at 40: past, present and future». *Nature* 550(7676): 345-53.

Simão, Felipe A. et al. 2015. «BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs». *Bioinformatics* 31(19): 3210-12.

Simpson, J. T. et al. 2009. «ABySS: A parallel assembler for short read sequence data». *Genome Research* 19(6): 1117-23.

Simpson, J. T., y R. Durbin. 2012. «Efficient de novo assembly of large genomes using compressed data structures». *Genome Research* 22(3): 549-56.

Sims, David et al. 2014. «Sequencing depth and coverage: key considerations in genomic analyses». *Nature Reviews Genetics* 15(2): 121-32.

Smith, Hamilton O., y K.W. Welcox. 1970. «A Restriction enzyme from Hemophilus influenzae: I. Purification and general properties». *Journal of Molecular Biology* 51(2): 379-91.

Smith, Lloyd M. et al. 1986. «Fluorescence detection in automated DNA sequence analysis». *Nature* 321(6071): 674-79.

Smith, T.F., y M.S. Waterman. 1981. «Identification of common molecular subsequences». *Journal of Molecular Biology* 147(1): 195-97.

Sombke, Andy et al. 2011. «The source of chilopod sensory information: External structure and distribution of antennal sensilla in Scutigera coleoptrata (chilopoda, Scutigeromorpha)». *Journal of Morphology* 272(11): 1376-87.

Song, Kai, Li Li, y Guofan Zhang. 2016. «Coverage recommendation for genotyping analysis of highly heterologous species using next-generation sequencing technology». *Scientific Reports* 6(1): 35736.

Sparks, Jackson T et al. 2018. «Membrane Proteins Mediating Reception and Transduction in Chemosensory Neurons in Mosquitoes.» *Frontiers in physiology* 9: 1309.

Spehr, Marc, y Steven D. Munger. 2009. «Olfactory receptors: G protein-coupled receptors and beyond». *Journal of Neurochemistry* 109(6): 1570-83.

Sreedharan, Vipin T. et al. 2014. «Oqtans: the RNA-seq workbench in the cloud for complete and reproducible quantitative transcriptome analysis». *Bioinformatics* 30(9): 1300-1301.

Stanke, M., y S. Waack. 2003. «Gene prediction with a hidden Markov model and a new intron submodel». *Bioinformatics* 19(Suppl 2): ii215-25.

Stapley, Jessica et al. 2010. «Adaptation genomics: the next generation». *Trends in Ecology & Evolution* 25(12): 705-12.

Stark, Rory, Marta Grzelak, y James Hadfield. 2019. «RNA sequencing: the teenage years». *Nature Reviews Genetics*.

Stengl, Monika, y Nico W. Funk. 2013. «The role of the coreceptor Orco in insect olfactory transduction». *Journal of Comparative Physiology A* 199(11): 897-909.

Storch, Judith, y Zhi Xu. 2009. «Niemann–Pick C2 (NPC2) and intracellular cholesterol trafficking». *Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids* 1791(7): 671-78.

Sulonen, Anna-Maija et al. 2011. «Comparison of solution-based exome capture methods for next generation sequencing». *Genome Biology* 12(9): R94.

Tarazona, S. et al. 2011. «Differential expression in RNA-seq: A matter of depth». *Genome Research* 21(12): 2213-23.

Thomsom, Robert C., Ian J. Wang, y Jarret R. Jonhson. 2010. «Genome-enabled development of DNA markers for ecology, evolution and conservation». *Molecular Ecology* 19(11): 2184-95.

Toda, Hirofumi, Xiaoliang Zhao, y Barry J. Dickson. 2012. «The Drosophila Female Aphrodisiac Pheromone Activates ppk23+ Sensory Neurons to Elicit Male Courtship Behavior». *Cell Reports* 1(6): 599-607.

Torres-Sánchez, María et al. 2019. «Multi-tissue transcriptomes of caecilian amphibians highlight incomplete knowledge of vertebrate gene families». *DNA Research* 26(1): 13-20.

Townsend, Ted M et al. 2008. «Rapid development of multiple nuclear loci for phylogenetic analysis using genomic resources: an example from squamate reptiles.» *Molecular phylogenetics and evolution* 47(1): 129-42.

Trapnell, Cole et al. 2012. «Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks». *Nature Protocols* 7(3): 562-78.

Trapnell, Cole, Lior Pachter, y Steven L. Salzberg. 2009. «TopHat: discovering splice junctions with RNA-Seq». *Bioinformatics* 25(9): 1105-11.

Turner, Emily H., Sarah B. Ng, Deborah A. Nickerson, y Jay Shendure. 2009. «Methods for Genomic Partitioning». *Annual Review of Genomics and Human Genetics* 10(1): 263-84.

Vassort, Guy, y Jérémy Fauconnier. 2008. «Les canaux TRP *(transient receptor potential)*». *médecine/sciences* 24(2): 163-68.

Venter, J. Craig et al. 2001. «The Sequence of the Human Genome». *Science* 291(5507): 1304-51.

Vieira, Filipe G, y Julio Rozas. 2011. «Comparative genomics of the odorant-binding and chemosensory protein gene families across the Arthropoda: origin and evolutionary history of the chemosensory system.» *Genome biology and evolution* 3: 476-90.

Vizueta, Joel et al. 2017. «Evolution of Chemosensory Gene Families in Arthropods: Insight from the First Inclusive Comparative Transcriptome Analysis across Spider Appendages.» *Genome biology and evolution* 9(1): 178-96.

Vizueta, Joel, Julio Rozas, y Alejandro Sánchez-Gracia. 2018. «Comparative Genomics Reveals Thousands of Novel Chemosensory Genes and Massive Changes in Chemoreceptor Repertories across Chelicerates.» *Genome biology and evolution* 10(5): 1221-36.

Vogt, R.G. 2005. «Molecular Basis of Pheromone Detection in Insects». En *Comprehensive Molecular Insect Science*, Elsevier, 753-803.

Vogt, Richard G., y Lynn M. Riddiford. 1981. «Pheromone binding and inactivation by moth antennae». *Nature* 293(5828): 161-63.

Wang, Zhong, Mark Gerstein, y Michael Snyder. 2009. «RNA-Seq: a revolutionary tool for transcriptomics». *Nature Reviews Genetics* 10(1): 57-63.

Watson, J D, y F H Crick. 1953a. «Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid». *Nature* 171(4356): 737-38.

Watson, J D, y F H Crick. 1953b. «The structure of DNA.» *Cold Spring Harbor symposia on quantitative biology* 18: 123-31.

Wheat, Christopher W. 2010. «Rapidly developing functional genomics in ecological model systems via 454 transcriptome sequencing». *Genetica* 138(4): 433-51.

Whiteman, Noah K, y Naomi E Pierce. 2008. «Delicious poison: genetics of Drosophila host plant preference.» *Trends in ecology & evolution* 23(9): 473-78.

Wicher, Dieter. 2015. «Olfactory Signaling in Insects». En , 37-54.

Williams, John G.K. et al. 1990. «DNA polymorphisms amplified by arbitrary primers are useful as genetic markers». *Nucleic Acids Research*.

Wu, Ray. 1972. «Nucleotide Sequence Analysis of DNA». *Nature New Biology* 236(68): 198-200.

Yamaji, Nahoko et al. 2009. «Synthesis, Solution Structure, and Phylum Selectivity of a Spider δ-Toxin That Slows Inactivation of Specific Voltage-gated Sodium Channel Subtypes». *Journal of Biological Chemistry* 284(36): 24568-82.

Yang, Xiaowei, Huipeng Pan, Ling Yuan, y Xuguo Zhou. 2018. «Reference gene selection for RT-qPCR analysis in Harmonia axyridis, a global invasive lady beetle.» *Scientific reports* 8(1): 2689.

Yarmolinsky, David A., Charles S. Zuker, y Nicholas J.P. Ryba. 2009. «Common Sense about Taste: From Mammals to Insects». *Cell* 139(2): 234-44.

Zavodna, Monika, Catherine E. Grueber, y Neil J. Gemmell. 2013. «Parallel Tagged Next-Generation Sequencing on Pooled Samples – A New Approach for Population Genetics in Ecology and Conservation» ed. Paul Sunnucks. *PLoS ONE* 8(4): e61471.

Zelle, Kathleen M, Beika Lu, Sarah C Pyfrom, y Yehuda Ben-Shahar. 2013. «The genetic architecture of degenerin/epithelial sodium channels in Drosophila.» *G3 (Bethesda, Md.)* 3(3): 441-50.

Zeng, Jingyao et al. 2016. «Identification and analysis of house-keeping and tissue-specific genes based on RNA-seq data sets across 15 mouse tissues». *Gene* 576(1): 560-70.

Zeng, Xiong-Zhi, Qiao-Bin Xiao, y Song-Ping Liang. 2003. «Purification and characterization of raventoxin-I and raventoxin-III, two neurotoxic peptides from the venom of the spider Macrothele raveni». *Toxicon* 41(6): 651-56.

Zhang, Jin et al. 2015. «Antennal Transcriptome Analysis and Comparison of Chemosensory Gene Families in Two Closely Related Noctuidae Moths, Helicoverpa armigera and H. assulta» ed. Walter S. Leal. *PLOS ONE* 10(2): e0117054.

Zhang, Shichang et al. 2012. «A novel property of spider silk: chemical defence against ants». *Proceedings of the Royal Society B: Biological Sciences* 279(1734): 1824-30.

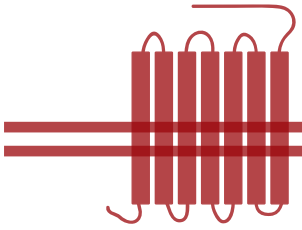
Zhang, Su-fang et al. 2018. «Dynamic Changes in Chemosensory Gene Expression during the Dendrolimus punctatus Mating Process». *Frontiers in Physiology* 8: 1127.

Zhou, Xiaofan et al. 2014. «Divergent and conserved elements comprise the chemoreceptive repertoire of the non-blood feeding mosquito Toxorhynchites amboinensis.» *Genome biology and evolution*.

Zickmann, Franziska, y Bernhard Y Renard. 2015. «IPred - integrating ab initio and evidence based gene predictions to improve prediction accuracy». *BMC Genomics* 16(1): 134.

Zimin, Aleksey V. et al. 2013. «The MaSuRCA genome assembler». *Bioinformatics* 29(21): 2669-77.

Zuckerkandl, Emile, y Linus Pauling. 1964. Evolving Genes and Proteins *Evolutionary Divergence and Convergence in Proteins*. ed. ed. Washburn SL. Perspectives in molecular anthropology. In: Classification and human evolution.



ANEXOS

A

Evolutionary Analysis of Mitogenomes from Parasitic and Free-Living Flatworms

Eduard Solà, Marta Álvarez-Presas, Cristina Frías-López,
D. Timothy J. Littlewood, Julio Rozas, Marta Riutort

2015, PLoS ONE 10(3): e0120081

PLOS ONE

RESEARCH ARTICLE

Evolutionary Analysis of Mitogenomes from Parasitic and Free-Living Flatworms

Eduard Solà[1☯], Marta Álvarez-Presas[1☯], Cristina Frías-López[1], D. Timothy J. Littlewood[2], Julio Rozas[1], Marta Riutort[1]*

1 Institut de Recerca de la Biodiversitat and Departament de Genètica, Facultat de Biologia, Universitat de Barcelona, Catalonia, Spain, **2** Department of Life Sciences, Natural History Museum, Cromwell Road, London, United Kingdom

☯ These authors contributed equally to this work.
* mriutort@ub.edu (MR)

OPEN ACCESS

Citation: Solà E, Álvarez-Presas M, Frías-López C, Littlewood DTJ, Rozas J, Riutort M (2015) Evolutionary Analysis of Mitogenomes from Parasitic and Free-Living Flatworms. PLoS ONE 10(3): e0120081. doi:10.1371/journal.pone.0120081

Academic Editor: Hector Escriva, Laboratoire Arago, FRANCE

Received: September 18, 2014

Accepted: January 19, 2015

Published: March 20, 2015

Data Availability Statement: All sequence data are available from GenBank under the accession numbers KP208776 and KP208777. Information on the localities where the animals sequenced were sampled is included in Table S1. Tables S3 and S4 list the specific primers the authors designed to reamplify the genomes of C. alpina and Obama sp., respectively.

Funding: Funding was received from Ministerio de Economía y competitividad (Spain): CGL 2008-00378 MR, CGL 2011-23466 MR, CGL2013-45211-C2-1-P (http://www.idi.mineco.gob.es/portal/site/MICINN/). The funders had no role in study design, data

Abstract

Mitochondrial genomes (mitogenomes) are useful and relatively accessible sources of molecular data to explore and understand the evolutionary history and relationships of eukaryotic organisms across diverse taxonomic levels. The availability of complete mitogenomes from Platyhelminthes is limited; of the 40 or so published most are from parasitic flatworms (Neodermata). Here, we present the mitogenomes of two free-living flatworms (Tricladida): the complete genome of the freshwater species *Crenobia alpina* (Planariidae) and a nearly complete genome of the land planarian *Obama* sp. (Geoplanidae). Moreover, we have reannotated the published mitogenome of the species *Dugesia japonica* (Dugesiidae). This contribution almost doubles the total number of mtDNAs published for Tricladida, a species-rich group including model organisms and economically important invasive species. We took the opportunity to conduct comparative mitogenomic analyses between available free-living and selected parasitic flatworms in order to gain insights into the putative effect of life cycle on nucleotide composition through mutation and natural selection. Unexpectedly, we did not find any molecular hallmark of a selective relaxation in mitogenomes of parasitic flatworms; on the contrary, three out of the four studied free-living triclad mitogenomes exhibit higher A+T content and selective relaxation levels. Additionally, we provide new and valuable molecular data to develop markers for future phylogenetic studies on planariids and geoplanids.

Introduction

Complete mitochondrial genomes (mitogenomes) provide a diversity of molecular markers suitable to study a variety of biological features, including the effects of different life habits (e.g. [1]) or the phylogenetic relationships among populations or species. This is because mitochondrial (mt) DNA does not usually recombine, commonly exhibits neutral evolution, and mt markers have smaller effective population sizes than their nuclear counterparts which result in

collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

shorter coalescent times [2,3]. These features make mtDNA to be especially appropriate for either phylogeographical or population genetic studies (e.g. [4]).

Currently, within the phylum Platyhelminthes (Lophotrochozoa) there is available mitogenome sequence information for up to 40 parasitic species of Neodermata, which includes the Trematoda, Cestoda and Monogenea [5,6]. In contrast, there are few available complete mitogenomes from free-living flatworms [7,8]: one complete mitogenome (*Dugesia japonica*; ~18 kb), another almost complete (*Dugesia ryukyuensis*; ~17 kb) and a fragment of 6.8 kb (*Microstomum lineare*), and also a complete mitogenome of *Schmidtea mediterranea* available in GenBank (Acc. N.: NC_022448.1). Three of these mitogenomes belong to the Tricladida (*Dugesia* and *Schmidtea*), a clade not distantly related to the parasitic flatworms (Fig. 1), although the two groups split possibly in the Paleozoic [9].

The free-living triclads (Tricladida) have been included recently in biogeographical, phylogeographical and conservation studies [10,11]. In particular land planarians have become convenient models for understanding the origins and maintenance of biological diversity because of their low vagility and extreme dependence on the continuity and stability of their habitats. To date, all these studies have been based on partial gene fragments (particularly *cox1*), due to limitations in amplifying other mitochondrial genes or regions.

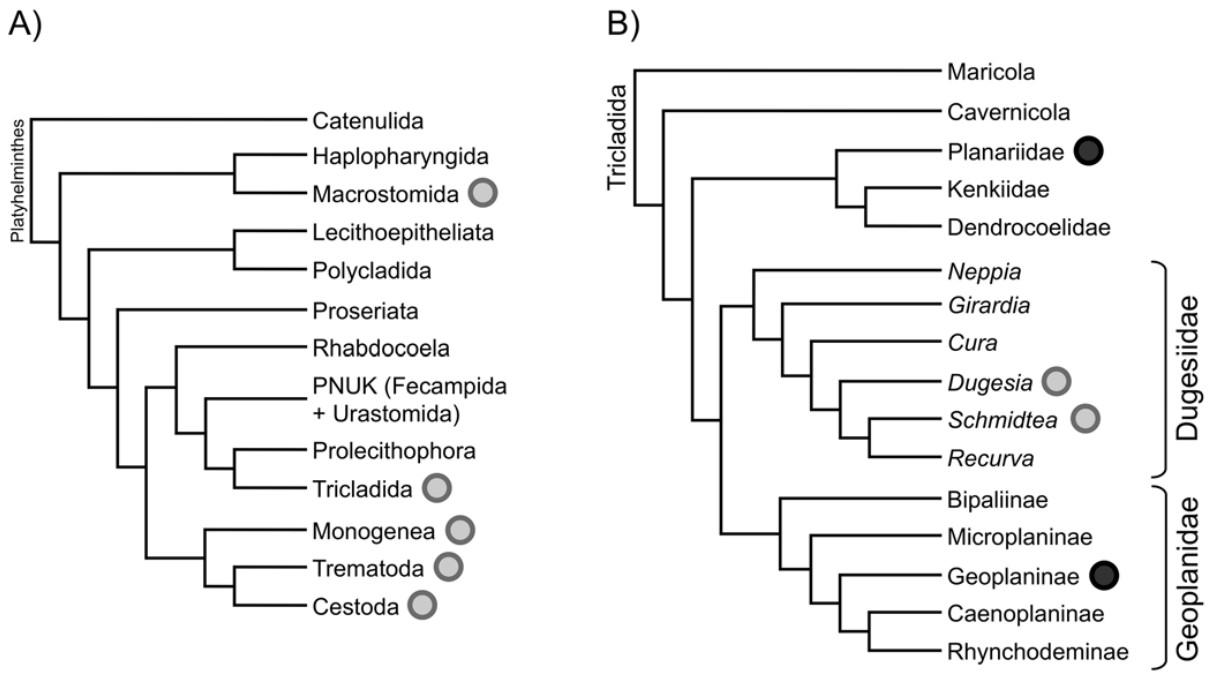


Fig 1. Phylogenetic schemes indicating the relationships of groups for which mitogenomes are available. A) Phylogeny of the Platyhelminthes according to Riutort *et al*., 2012 ([9]) and B) phylogeny of the Tricladida according to Riutort *et al*., 2012 and Sluys *et al*., 2013 ([52]). Monogenea, Trematoda and Cestoda constitute the Neodermata (parasitic flatworms) group. Grey circles indicate those groups for which mitogenomes are already available. Black circles indicate new obtained mitogenomes.

doi:10.1371/journal.pone.0120081.g001

Through denser taxon sampling the development of universal and specific primers within this group should be achievable. Additionally, this will provide gene order, nucleotide and amino acid data for phylogenetic studies across the phylum, confirming for example the use of the rhabditophoran mitochondrial genetic code for the whole group [12], the identity of initiation and stop codons, and composition skews. Finally, it will also allow a comparison between mitogenomes from free-living and parasitic taxa, providing insights as to whether these different lifestyles have left a molecular signature.

Here we have determined the mitochondrial genomes of two Tricladida species belonging to two different superfamilies (*Crenobia alpina*, Planarioidea; *Obama* sp., Geoplanoidea) with two major aims, (i) to study the molecular evolution of mitochondrial molecules in the platyhelminths and (ii) to determine the putative different impact of natural selection in free-living and parasitic species caused by their lifestyles. In order to achieve the first objective we have compared the sequence and gene annotations of the new mitogenomes together with those of available free-living species (*Dugesia*, [8]; *Schmidtea mediterranea*, Ross *et al*., Acc. N.: NC_022448.1). For the second objective, we used complete mitogenomic data to determine whether parasitic species exhibit higher evolutionary rates or a relaxation of natural selection as previously proposed [13–16]. For the study, we contrasted the impact of mutational and selective strengths on nucleotide composition and codon bias. Additionally, our new mitogenomic data will be useful to further conduct phylogenetic and phylogeographic-based analyses in triclads.

Material and Methods

Samples

None of the species used in this study are protected or endangered, and most sampling sites did not require permission for collecting. For *D. subtentaculata* locality in Sta. Fe del Montseny within the Parc Natural del Montseny, permission was provided by the Parc authorities. Four species of Tricladida from three different families (Dugesiidae, Geoplanidae, Planariidae) were targeted for complete mitochondrial genome characterization (Table 1). Live specimens of *Crenobia alpina* (Dana, 1766), *Polycelis felina* (Dalyell, 1844), *Dugesia subtentaculata* (Draparnaud, 1801) and *Obama* sp. (*Obama* sp. [17]) were collected from different localities within Catalonia. Sample locality data is shown in Table A in S1 Tables file. It was not possible to obtain the complete mitogenome for two of these species, owing to different reasons, hence the analyses and results from here on will only refer to the species *Crenobia alpina* and *Obama* sp. Information on the problems found and results obtained for the other two species can be found in the S1 File. The complete mitochondrial genomes of two triclads and eight neodermatans were also retrieved from GenBank (Table 1) to carry out a preliminary gene checking of the mitogenomes obtained in this study by means of 454 (Roche) pyrosequencing, and to perform analytical comparisons between triclads and parasitic flatworms.

Mitochondrial DNA extraction

We isolated mitochondrial DNA from about 100 animals for each species based on a modification of the protocol described in Bessho *et al*. (1997) [18]. We first removed the mucus from the planarians with a diluted cysteine chloride solution (pH 7.0) obtained from effervescent tablets (CINFA) and then dipped the animals in buffer 1 (0.1 M sucrose, 10 mM TrisHCl, pH 7.4) overnight at −80°C. Animals were next homogenized, transferred to two PPCO tubes and centrifuged at 600 g (Beckman JA-20 rotor) at 2°C during 10 minutes in order to remove nuclei. The supernatant was centrifuged in FEP tubes at 15,000 g at 2°C for 10 minutes in a Sorvall centrifuge (SS-34 rotor). The pellet was dissolved in 40 mL (20 mL in each tube) of 0.1 M

Table 1. List of all Platyhelminthes species included in the present work.

Species	Classification	Life cycle	Acc. Number	Analysis			References
				CG	PGS	SQ	
Crenobia alpina	Tricladida, Planariidae	FL	KP208776	X		X	This work
Dugesia japonica	Tricladida, Dugesiidae	FL	AB618487.1	X			[8]
Obama sp.	Tricladida, Geoplanidae	FL	KP208777	X		X	This work
Schmidtea mediterranea	Tricladida, Dugesiidae	FL	NC_022448.1	X			Not published
Benedenia hoshinai	Monogenea, Capsalidae	P	NC_014591.1	X			[53]
Diplogonoporus balaenopterae	Cestoda, Diphyllobothriidae	P	NC_017613.1	X			[54]
Fasciola hepatica	Trematoda, Fasciolidae	P	NC_002546.1	X	X		[23]
Schistosoma japonicum	Trematoda, Schistosomatidae	P	NC_002544.1	X			[23]
Taenia saginata	Cestoda, Taeniidae	P	NC_009938.1	X			[55]
Taenia solium	Cestoda, Taeniidae	P	AB086256.1		X		[21]
Tetrancistrum sigani	Monogenea, Ancyrocephalidae s.l.	P	NC_018031.1	X			[56]
Gyrodactylus derjavinoides	Monogenea, Gyrodactylidae	P	NC_010976.1	X			[22]

Acronyms indicating the different analyses: CG, Comparative genomics; PGS, Preliminary gene screening; SQ, Sequencing.
Acronyms indicating life cycle: FL, Free-living; P, Parasitic.

doi:10.1371/journal.pone.0120081.t001

sucrose solution containing 50 mM $MgCl_2$ (buffer 2). To remove any contamination of nuclear DNA from mitochondrial membranes, the solution was treated with 10 μl of 70 units/mL DNase. After inactivating the DNase (80°C for 10 minutes), 200 mL (100 mL per tube) of 0.6% SDS, 10 mM EDTA, 10 mM Tris-HCl (pH 8.0) (buffer 3) were added and incubated at 60°C for 10 minutes to disrupt mitochondrial membranes. Finally, an ordinary phenol chloroform extraction was applied to isolate mitochondrial DNA [19].

Mitochondrial DNA quantification and 454 sequencing

We quantified the DNA amount with a Qubit 2.0 fluorometer (Invitrogen) following manufacturer's instructions. After precipitating the DNA it was resuspended in TE to a final concentration of 20 ng/μL. The five DNA samples were multiplexed identifier (MID) tagged, and the 454 libraries prepared at the Centres Científics i Tecnològics de la Universitat de Barcelona (CCiTUB). The samples were run into a ¼ 454 plate of the GS FLX titanium platform.

Sequencing reads processing

DNA sequences (reads) and quality information were extracted independently of each MID's in fasta format from the Standard Flowgram Format file (SFF) using the sffinfo script from Roche's Newbler package (454 SFF Tools). We removed adapters, putative contaminant sequences (upon the UniVecdatabase and the *E. coli* genome sequence) and reads shorter than 50 bp were removed using the SeqClean (http://compbio.dfci.harvard.edu/tgi/software/) script. All reads with a mean quality score below 20 were trimmed, and the low-quality bases at the ends of the reads were also removed using PRINSEQ [20].

Sequencing reads post-processing

We determined whether the mitochondrial genes were present in sequencing reads by a BLAST analysis (v. 2.2.24) using available mitochondrial genome data (downloaded from NCBI) of parasitic flatworms (Table 1) as query. In particular we used the protein information

of *Taenia solium* [21], *Gyrodactylus derjavinoides* [22] and *Fasciola hepatica* [23] (Table B in S1 Tables file). For the analyses we applied the tBLASTn algorithm (e-value cut-off: 10^{-3}), using translation table 9 (echinoderm and flatworm mitochondrial code) to translate DNA information of the 454 reads in all six reading frames.

Mitochondrial genomes assembling, annotation, PCR amplification and re-sequencing

We first tried to assemble the DNA genome sequence using Newbler 2.6 (454 life Sciences, with settings:-urt-ml 40-mi 85-minlen 50), but with little success. Several short contigs, with a N50 length of about 400 nucleotides, were resolved. However, SeqMan software (DNASTAR, http://www.DNASTAR.com) resolved large nearly complete mtDNA sequences including all filtered 454. The assembled mitogenomes were annotated with Geneious Pro 6.1.7 [24]. Later, we validated the genome assemblies by further Sanger DNA sequencing. This experimental approach allowed us to determine the existence of, and thereby correct, some 454-induced sequence errors (e.g. frameshifts; [25]), to complete the molecules, and to confirm the gene order resulting from the assembled genomes. For such analysis, we designed 34 primers for PCR amplification in *C. alpina* and 20 primers for *Obama* sp. (Tables C and D in S1 Tables file) covering the whole length of the genomes. PCR reactions initially included: 1 μl of DNA, 5 μl of Promega 5X Buffer, 1 μl of dNTPs (10 mM), 0.5 μl of each primer (25 μM), 2 μl of $MgCl_2$ (25 mM), 0.15 μl of *Taq* polymerase (GoTaq Flexi DNA Polymerase, Promega). Double-distilled and autoclaved water was added to obtain a final 25 μl PCR volume for all molecules. In many cases PCR needed to be optimised by varying annealing temperatures or the concentrations of $MgCl_2$ or DNA. PCR products of low yield for direct sequencing were cloned using TOPO TA Cloning Kit of (Invitrogen) following manufacturers' instructions. For every PCR product cloned, five bacterial colonies on average were picked and sequenced in order to obtain representation of the different haplotypes. Cloned fragments were amplified using universal vector primers T3 and T7. All PCR amplicons were purified using the purification kit illustra (GFX PCR DNA and Gel Band of GE Healthcare) or by using a vacuum system ($MultiScreen_{HTS}$ Vacuum Manifold, Millipore). Sequencing reactions, using Big-Dye (3.1, Applied Biosystems) with the same primers used to amplify the fragment, were run on an automated sequencer ABI Prism 3730 (Unitat de Genòmica of Centres Científics i Tecnològics de la Universitat de Barcelona – CCiTUB) or at Macrogen Corporation (Amsterdam, the Netherlands). The chromatograms were visually checked. These additional DNA sequences were aligned and compared with the 454-based assemblies using the software Geneious 6.1.7, which was also used to obtain the final assemblies.

Prediction of protein-coding genes and rRNA genes

We determined the location of the protein-coding, rrnL and rrnS genes by using a combination of BLAST searches, ORF finder and the Glimmer plug-in in Geneious 6.1.7, MITOS online software [26], and using information from published Platyhelminthes sequences.

We used the online software GenDecoder v1.6 [27] in order to assign the genetic code of the triclads analyzed. As the expected code we used the Echinoderm and Flatworm Mitochondrial Code (translation table 9). We tested all different degrees of Shannon entropy available in the program and we let the removal of columns at 20% of gaps, as it is set as default. We compared our mitogenomes with the Metazoa reference data set, which also includes parasitic platihelmints.

Prediction of tRNAs

Putative tRNA genes were identified using a combination of the following software: ARWEN (http://130.235.46.10/ARWEN) [28], tRNAscan-SE 1.21 [29], MITOS [26] and DOGMA [30]. The tRNAs not found with these programs were found and annotated by eye with reference to known platyhelminth sequences. In addition to our mtDNA molecules, we included the published *D. japonica* mitochondrial genome [8] to double-check the annotation of the molecule.

Nucleotide composition bias analyses

Comparative analyses of nucleotide composition bias across species or among DNA regions is a powerful approach to determine the impact of mutational and selective pressures on genome evolution. In addition to the standard A+T (or G+C) content, we also estimated the putative nucleotide frequencies bias (NB statistic) from a single strand (the coding strand). Following Shields *et al.* (1988) [31], we defined the NB statistic as:

$$NB = \left[\sum_{i=1}^{4} (O_i - E_i)^2 / E_i\right] / n$$

Where O_i and E_i are the observed and the expected (under equifrequency) numbers of nucleotide variant i (i = 1, 2, 3, and 4 correspond to A, C, G, and T), and n is the total number of positions analyzed. We applied the NB statistic in different portions of the mitochondrial molecule: NBp, NB at the protein coding regions; NB2, NB at the second position of codons; NB3, NB at the third position of four-fold degenerate codons; NBr and NBt, NB at the ribosomal and tRNA genes, respectively.

We also estimated the particular AT and GC strand skews, using the Perna and Kocher (1995) [32] indices, where the AT skew (sAT) is computed as (A−T)/(A+T) and the GC skew (sGC) = (G−C)/(G+C); in both cases the nucleotide frequencies are those of the coding strand. These values range from −1 to +1, where a value of zero indicates that the frequency of A is equal to T (AT skew), or G equal to C (GC skew). We calculated these indices for each gene and for the whole mitochondrial genome of *C. alpina* and *Obama* sp., but also for other free-living flatworms with available mitochondrial genome sequence data, and for six selected parasitic species (Table 1). We also computed the sAT (and sGC index) in different functional regions of the mitochondrial molecule, being sATp, the sAT at the protein coding regions; sAT2, sAT at the second position of codons; sAT3, the sAT at the third position of four-fold degenerate codons; sATr and sATt, sAT at the ribosomal and tRNA genes, respectively.

Codon bias analyses

Analyses of codon bias offer an effective means of disentangling the effects of mutational and selective factors. We estimated the codon usage bias applying the scaled chi-squared (SC) [31], which is a measure based on the chi-square statistic normalized by the number of codons, and Effective Number of Codons statistics (ENC) [33]. For the SC calculation we conducted two types of analyses: for one we used as the expected values those values assuming codon equifrequency (the standard way to compute SC), for the other, we used the observed nucleotide frequencies to determine the expected codon frequency values. For the latter we conducted the analysis separately for each species, and using 4 different types of observed nucleotide frequencies: the SC statistic computed (SCp) using as the expected number of codons (at each codon class) those values based on the observed nucleotide frequency at the protein coding region (the average for all genes within a species); SC2, the SC using information of the observed nucleotide frequencies at the second position of codons; SC3, SC using information at the third

position of four-fold degenerate codons; and SCr and SCt, those SC values using the observed nucleotide frequencies at the ribosomal and tRNA genes, respectively.

Results

454 raw data processing, assembling and gene annotation

The summary statistics for the 454 sequencing are shown in Table E in S1 Tables file. The 454 reads of *C. alpina* and *Obama* sp. provided sufficient information to assemble the mitogenomes successfully (Fig. 2 and Table F in S1 Tables) while it was not possible for the other three species (see S1 File). The SeqMan assembly of *C. alpina* generated a single contig of 17,079 bp. The average coverage of the assembly was 29.1X. For *Obama* sp. we obtained a contig of 14,893 bp with an average coverage of 24.3X. In this case, the quality of the DNA sequence was poorer than that obtained for *C. alpina*, likely by an increased 454 error rate in *Obama* sp. caused by a higher frequency of homo-polymer sequences. Both assemblies included all mitochondrial genes but lacked a large portion of the non-coding regions.

We completed and checked the sequences of these preliminary assemblies by Sanger DNA sequencing. We carried out additional partial PCR amplifications on the basis of the first assembly, and identified missing and/or extra bases. For instance, in the first assembly of *C. alpina* there was a missing nucleotide (a 454 error) in *nad4* and *nad5*, leading to a putative (erroneous) frameshift. This situation also occurred in several genes of the *Obama* sp. assembly.

It was not possible to re-sequence by Sanger the complete mitogenome of *C. alpina* since the designed primers failed to PCR amplify a fragment containing a repetitive region of about 186 bp (consensus size) (Fig. 2A). Indeed, the 454 assembly of this region recovered only two copies of this repetitive sequence likely due to the limitation of 454 read lengths. However, when the 454 reads were mapped to the whole mitochondrial molecule this region showed much higher sequence coverage than the rest of. the molecule suggesting that there were more than two repeat units, likely around four. Hence we do not know the exact number of repeats present in this region, and thus the total length of the full mitogenome.

For *Obama* sp. we PCR amplified a band of around 2,000 bp from the 3' end of *rrnL* to the 5' end of *cob* gene. However, it was not possible to obtain clean Sanger sequences probably due to the presence of a repetitive region within this fragment (Fig. 2B), hence the complete mitogenome length is also unknown for this species.

The mitochondrial genome of *C. alpina* (estimated size >16,894 bp; GenBank ID: KP208776) and *Obama* sp. (estimated size ~16,600 bp; GenBank ID: KP208777) encode 12 protein-coding genes, 22 tRNA genes and 2 ribosomal genes (Fig. 2 and Tables G and H in S1 Tables file), all transcribed from the same strand. As with other platyhelminths *nad4l* gene was the single case of one protein coding gene overlapping another; in *Obama* sp. and *C. alpina* *nad4l* overlaps 32 bp with *nad4*. In *Obama* sp., there may be (i) an overlap of 17 bp with *cob*, or (ii) no overlap and an alternative stop codon for *cob* one nucleotide before the start of *nad4l* (a codon presenting two ambiguous positions: TWW).

GenDecoder results support the use of the Echinoderm and Flatworm Mitochondrial Code for *Obama* sp. and *Crenobia alpina*. We found differences between the expected and predicted translation for some codons; one or two for *Obama* sp. and one to five for *Crenobia alpina* depending on the degree of Shannon entropy. However, these alternative translations were weakly supported, considered as unreliable predictions, thus supporting our expected code.

Gene order

The protein coding gene (PCG) order is conserved across Tricladida, but it is radically different from the incomplete fragment available from another free-living flatworm, *Microstomum*, and

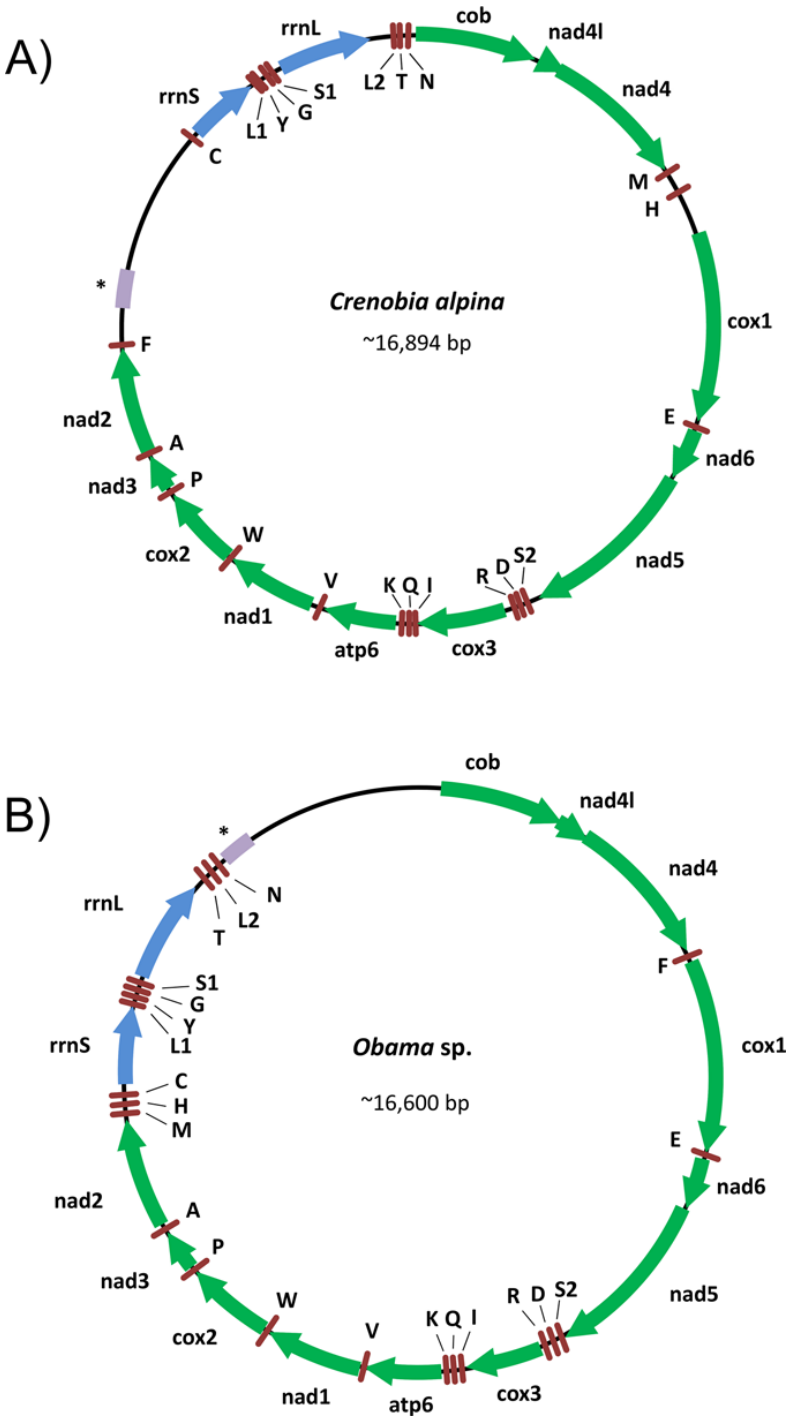


Fig 2. New freshwater flatworm mitogenomes obtained. Arrangement of the mitogenomes of *Crenobia alpina* (A) and *Obama* sp. (B). Green arrows correspond to the protein coding genes; blue arrows ribosomal genes; brown rods tRNAs; Purple bar indicates the putative repetitive region; * indicates that the relative length of the region may be different. Gene identifier: *rrnS*/*rrnL* = small and large subunit rRNA,

nad1–6,4L = NADH dehydrogenase subunits 1–6 + 4L, *cox1–3* = cytochrome c oxidase subunits 1–3, *atp6* = ATP synthase subunit 6, *cob* = cytochrome b. The tRNAs are shown according to the amino acid code letter.

doi:10.1371/journal.pone.0120081.g002

all the parasitic species (S1 Fig). Only three blocks of genes are conserved between parasites and triclads (S2 Fig). Our re-annotation of the *D. japonica* mitogenome entailed the change of three tRNAs to positions more similar, or identical, to those found in the other triclads: *trnC* is on the same strand as the rest of genes and *trnA* and *trnL1* are in the same relative position than in the other triclads (S3 Fig). In spite of these corrections all four triclad species (*C. alpina*, *Obama* sp., *S. mediterranea* and *D. japonica*) exhibit differences in the location of some tRNAs (S4 Fig).

The ribosomal genes are located close to the long non-coding region in the four Tricladida species, although in a different position. For *C. alpina* and *S. mediterranea* the long non-coding region is situated 5' upstream of the ribosomal genes while for *Obama* sp., and *D. japonica* it is situated at its 3' end. In contrast to other platyhelminth mitogenomes *rrnS* is situated upstream of *rrnL* amongst triclads (S1 Fig).

Start and terminal codons

We infer that four start codons are used in the two species analyzed. TTG and ATG are used at equivalent frequencies in *Obama* sp. while ATG is more frequent than TTG in *C. alpina*, TTA is also used in both species and GTG only in *Obama* sp. (Tables G and H in S1 Tables file). Stop codons are TAG and TAA. In *C. alpina*, *cox2* gene has a TAR stop codon, showing the presence of the two possible stop codons within the population (heterozygosity). Alternatively this could be a case of a truncated TA stop codon.

The length of the genes is very similar between the two species. However, in general the predictions for *Obama* sp. are slightly longer resulting in a more compact genome (shorter intergenic regions).

Transfer RNAs and ribosomal genes

Most tRNA genes in *C. alpina* and *Obama* sp. have the classical secondary structure (S5 and S6 Figs). The tRNAs *trnS2* and *trnT* lack the DHU arm in both species, while in *C. alpina* the *trnQ* could have two alternative structures: either lacking the TΨC arm or the DHU arm.

In *C. alpina*, four tRNAs overlap (*trnI*, *trnW*, *trnA*, *trnF*) with the last two bases of four genes (*cox3*, *nad1*, *nad3*, *nad2* respectively). Moreover, *trnL1* overlaps with *trnY*. In *Obama* sp., *trnF* and *trnV* overlap 1 nucleotide with genes *nad4* and *atp6* respectively. On the other hand, there are 3 cases of overlapping between tRNAs (*trnD* and *trnR*, 5 bp; *trnQ* and *trnK*, 8 bp; *trnY* and *trnG*, 4 bp). In the new annotation of *D. japonica* mitogenome the *trnA* and *trnL1* preserve the four arms while *trnC* lacks TΨC arm (S7 Fig).

Non-coding regions

C. alpina long non-coding region contains at least four repeats of 186 bp (consensus size) between two non-repetitive regions of 309–311 bp upstream and of 1,363 bp downstream. The total length of this large non-coding region is, at least, 2,028 bp. In the case of *Obama* sp. we only have the information of the length of the amplified fragment, around 2,000 bp, but we cannot establish the true number of repeat elements.

Nucleotide composition, strand skew and codon usage bias

Triclad mitogenomes have high A+T content values (>60%) (Fig. 3A). The per strand nucleotide frequency bias is also noticeably high, both in free-living and parasitic species (Fig. 3B; S8 Fig). We found such bias both at the whole molecule (NB statistic) and in different portions of the same (NBp, NB2, NB3, NBr and NBt), with bias at the third position of codons (NB3) being more pronounced. The A+T content at the third position of codons correlates with that frequency in the 1st, the 2nd, the rRNA and tRNA sites (Fig. 3C). These analyses separate the surveyed species into two clusters, parasitic and free-living species (with the exception of *C. alpina*).

In contrast to the A+T and NB values, free-living and parasitic species do not differentiate themselves from one another with respect to sAT or sGT values, either for the total data or for the values estimated at positions with different functional behavior (S9 and S10 Figs.). All sAT values are negative (in all genes and in all species), with the exception of the *rrnS* gene of *Obama* sp. and *T. sigani* where values are slightly positive (Fig. 4A and 4B). Thus, there is a clear prevalence of T over A in the coding strand. Moreover, the general sAT skew varies considerably among species (−0.187 to −0.4 Tricladida; −0.168 to −0.483 Neodermata), but it is consistent across genes; for instance *F. hepatica* has the highest overall sAT values, a feature exhibited in all of its genes (Fig. 4B). The sAT and A+T content, however, are uncoupled; for instance, *Obama* sp., the species with highest A+T content, exhibits nearly the lowest sAT values. The general sGC estimates also show important strand skews, ranging from 0.246 to 0.283 in triclads and 0.148 to 0.475 in parasites, which indicate a higher frequency of G than C. Although the sGC values also show some species-specific pattern it is much less consistent across genes. Overall, the analyses uncover a species-specific pattern that (i) is not correlated with the actual A+T content (S9 Fig), (ii) differs between sGC and sAT estimates, and (iii) does not cluster free-living or parasitic species separately.

The results of the codon usage analysis also show high levels of bias across the surveyed species (Fig. 5A and 5B), both using the SC or ENC estimators. Interestingly, and in agreement with the nucleotide frequency bias analyses, the free-living species again show the highest levels of codon bias (excepting *C. alpina*).

Discussion

Mitogenomes of Tricladida: general features

The mitogenomes of the newly characterized triclad species, *Crenobia alpina* and *Obama* sp., share the same gene composition with the majority of the Platyhelminthes sequenced so far, 12 PCGs while the *atp8* gene is absent. This gene is also absent in the mitogenomes of Chaetognatha, and Rotifera among lophotrochozoans as well as in some Bivalvia (Mollusca) and most Nematoda [6,34,35]. They also encode for the usual complement of 22 tRNAs, as found in almost all other platyhelminth genomes; two species of the digenean genus *Schistosoma* (*S. japonicum* and *S. mansoni*) have 23 due to a duplication of the *trnC* gene [6]. Also, all genes are transcribed from the same strand, a feature found in other Platyhelminthes, Cnidaria, Porifera, Tunicata and many other lophotrochozoan phyla [6,34].

The genetic code used by all triclad species is consistent with that used for the majority of Platyhelminthes, i.e., the EMBL-NCBI genetic code 9: Echinoderm and Flatworm. We found no evidence that codon TAA codes for Tyr (as proposed by Bessho *et al.* 1992 [36]); on the contrary TAA appears to be the stop codon for most of our predicted genes, and in some of *D. japonica* [8]. Hence the "alternative flatworm mitochondrial code", code 14 from EMBL-NCBI, proposed for some Platyhelminthes [36] and Nematoda is likely a feature exclusive to the latter.

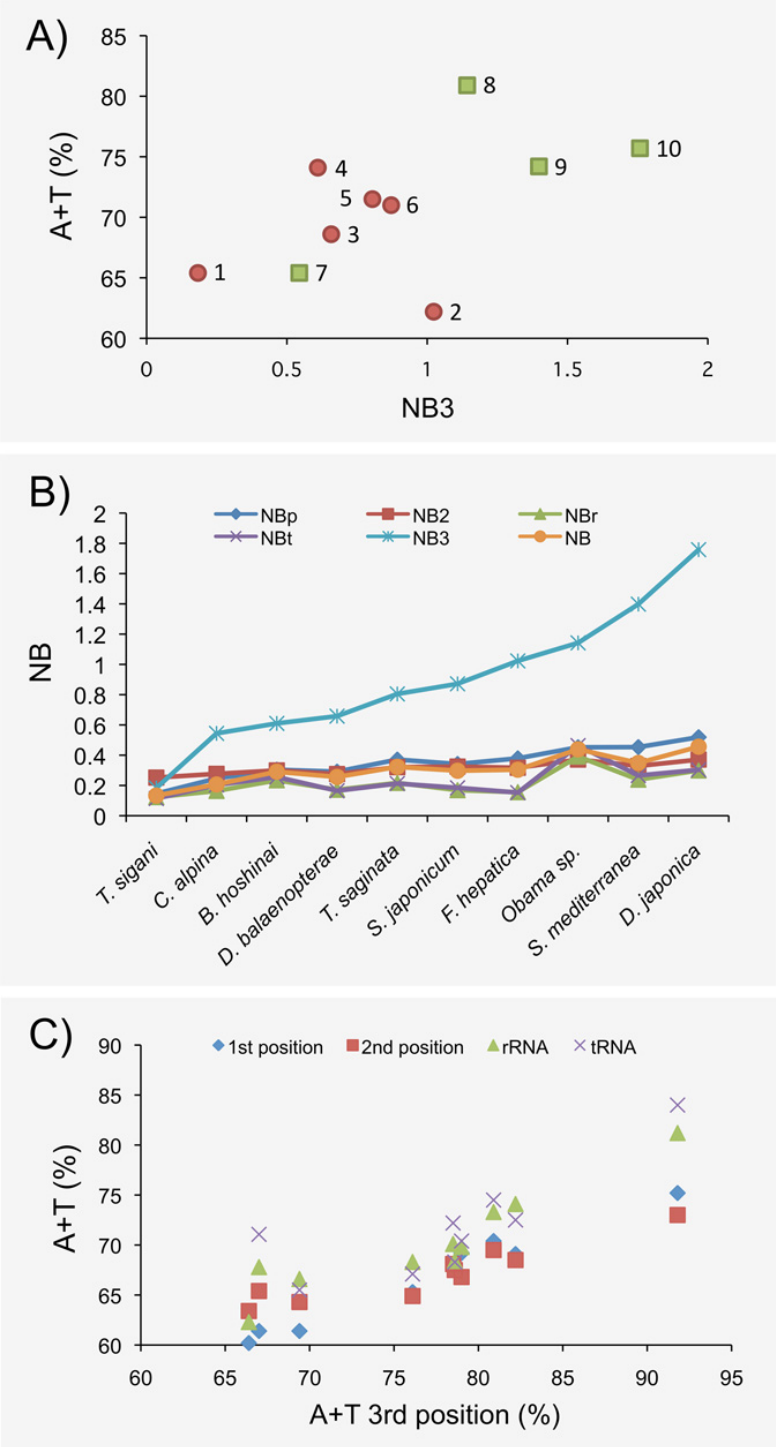


Fig 3. Nucleotide composition bias in the Platyhelminthes analyzed. A) Relationship between A+T content and NB3 (NB at the third position of four-fold degenerate codons) values. Green squares and red

circles indicate free-living and parasitic platyhelminths, respectively. The surveyed species are shown in numbers: 1, *T. sigani*; 2, *F. hepatica*; 3, *D. balaenopterae*; 4, *B. hoshinai*; 5, *T. saginata*; 6, *S. japonicum*; 7, *C. alpina*; 8, *Obama* sp.; 9, *S. mediterranea*; 10, *D. japonica*. B) Values of the different NB-based statistic across species. C) Relationship between A+T content for different genome portions and A+T content for the third positions.

doi:10.1371/journal.pone.0120081.g003

Gene order

The PCG order is identical in *C. alpina* and *Obama* sp. (Figs. 2 and S3), and also with the mitochondrial genomes of *D. japonica*, *D. ryukyuensis* and *S. mediterranea*. The only differences include the identity and arrangement of the tRNAs and the relative position of the long non-coding regions. The similarity in the situation of the non-coding region between *C. alpina* and *S. mediterranea* is surprising considering the closer phylogenetic relationships between *S. mediterranea* and *Dugesia* and *Obama*, all belonging to the superfamily Geoplanoidea, sister to the Planarioidea to which *Crenobia* belongs (Fig. 1B). On the other hand, the small number of changes in tRNAs order (S4 Fig) among all Tricladida is a notable feature given the very likely antiquity of the lineage.

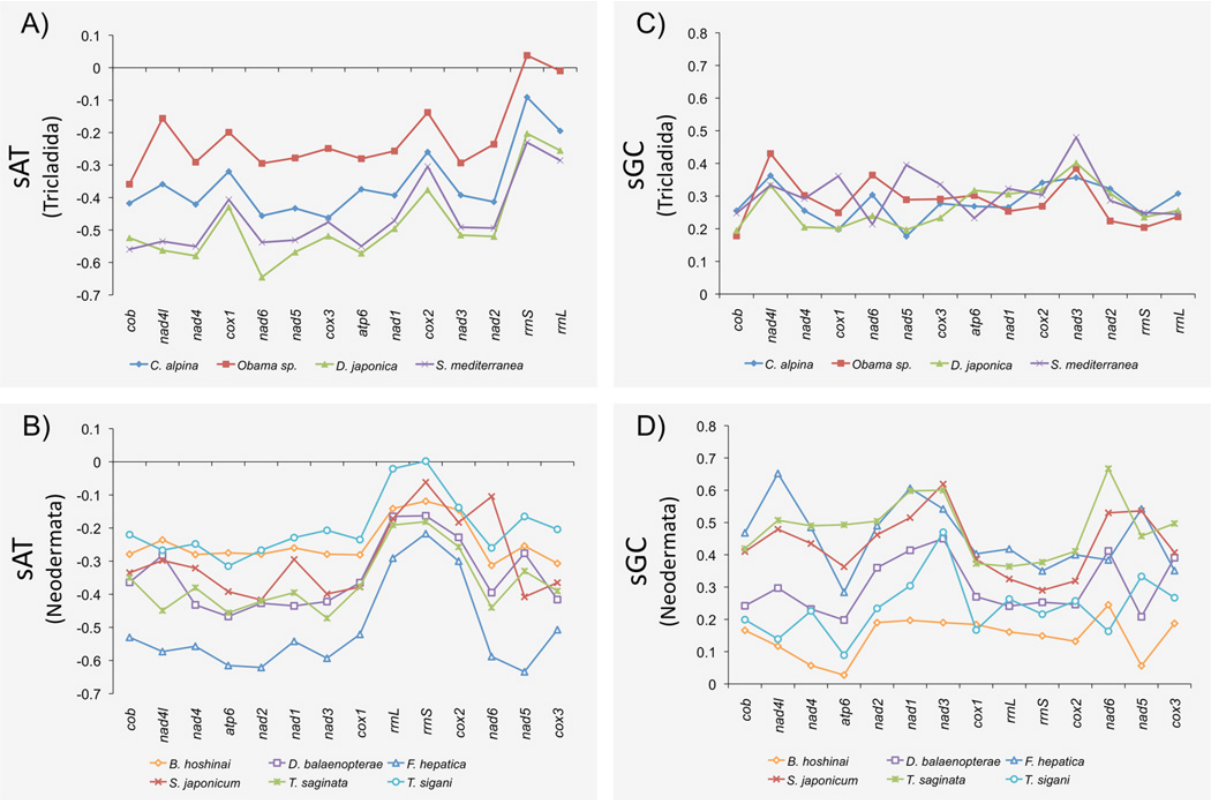


Fig 4. sAT and sGC values of the protein coding genes (PCG) along the mtDNA molecule. A) sAT of Tricladida; B) sAT of Neodermata; C) sGC of Tricladida; D) sGC of Neodermata.

doi:10.1371/journal.pone.0120081.g004

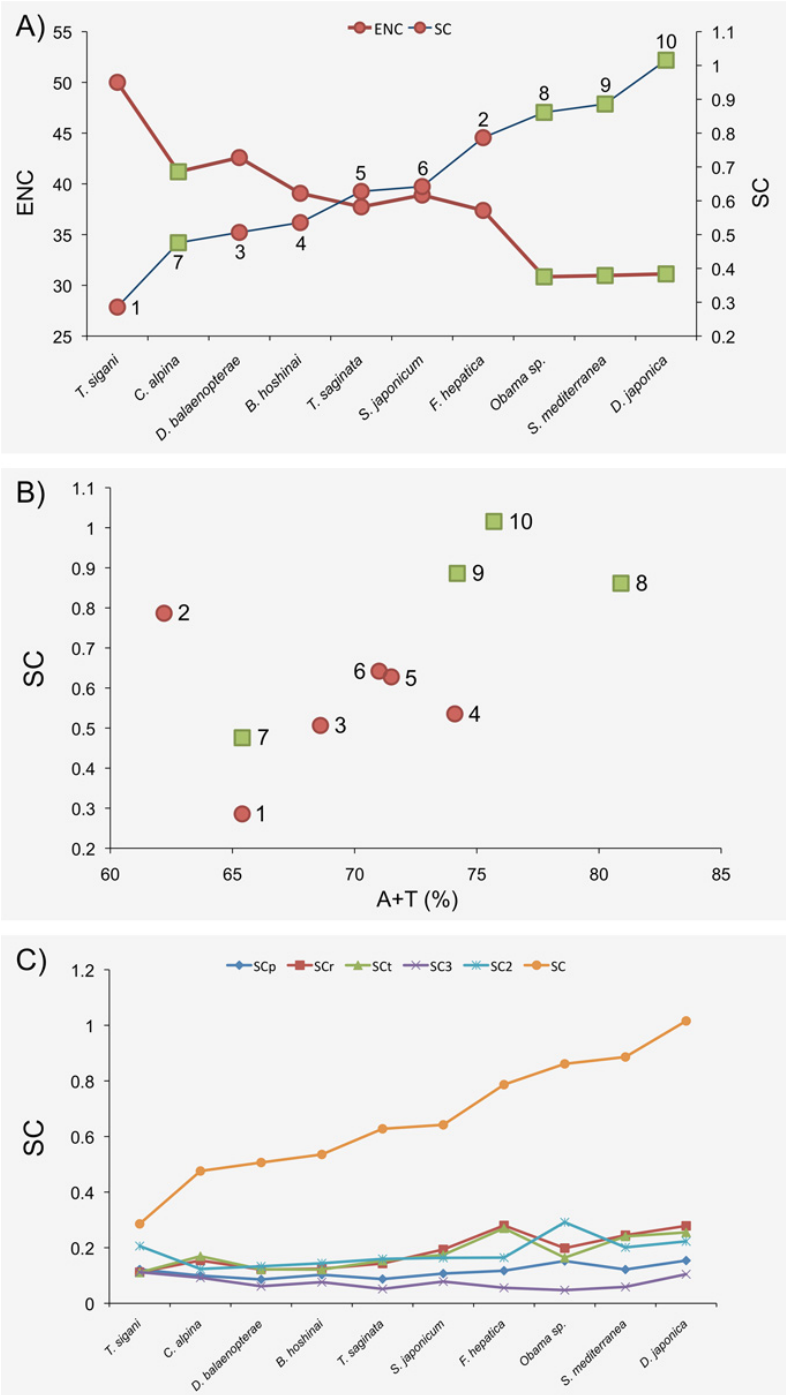


Fig 5. Relationship between different codon bias measures. A) Relationship between ENC and SC values. B) Relationship between SC and A+T% values. C) SC values across species (see Material and Methods text for acronym description). Green squares and red circles indicate free-living and parasitic Platyhelminthes, respectively. The surveyed species are shown in numbers: 1, *T. sigani*; 2, *F. hepatica*; 3, *D.*

balaenopterae; 4, *B. hoshinai*; 5, *T. saginata*; 6, *S. japonicum*; 7, *C. alpina*; 8, *Obama* sp.; 9, *S. mediterranea*; 10, *D. japonica*.

doi:10.1371/journal.pone.0120081.g005

The gene order among Tricladida differs considerably from that found in the parasitic platyhelminths and in *Microstomum*. One unique feature for Tricladida is the relative position of the two ribosomal genes; *rrnS* is located at 5' from *rrnL*, being the other way around in all the other platyhelminth mitogenomes characterized to date. Futhermore, in neodermatans *rrnL* and *rrnS* are flanked by *cox1* and *cox2*, whereas in triclads *rrnS* and *rrnL* are flanked by *nad2* and *cob*.

Start and terminal codon usage

While parasitic flatworms use only ATG and GTG as start codons, with the exception of a GTT used in *Hymenolepis diminuta* [6,37], Tricladida (Tables G, H and I in S1 Tables file; [8]) have much higher versatility. In addition to ATG and GTG, this group also appears to use TTG as start codon, and perhaps TTA and TAT. Moreover, the start codon for each gene is not conserved across Tricladida; in fact, only the start codon of *atp6* (TTG) is shared between all triclads. This diversity suggests independent origins of such codons across species. Although abbreviated stop codons (TA or T) are common in animal mitogenomes ([38] and references therein), we found that triclads have standard trinucleotide stop codons. In *Obama* sp., 10 out of the 12 PCG terminate in TAA, while *D. japonica* has the reverse situation 10 out of 12 PCG have TAG as stop codon. In *C. alpina* and *S. mediterranea* the usage of both stop codons is almost the same. The preference of the TAA stop codons in *Obama* sp. could be explained by the high frequency of A over G along its genome. The situation in the other three species with a similar proportion of A and G can explain the proportions of stop codons found in *S. mediterranea* and *C. alpina*, but not in *D. japonica*.

Although we used different methods to infer the start and stop codons for each gene, the lack of transcriptional information precludes any interpretation of boundaries with a high degree of confidence. Future studies involving transcriptomic analyses will help for a more accurate annotation of these species' genes.

A+T content and asymmetric strand bias

We have found that triclads have high A+T content values, a feature already detected in parasitic flatworms. Nevertheless, while some parasitic species have A+T content values around 70%, *Obama* sp. exhibits a much more extreme bias (over 80%), close to the highest described cases (Hymenoptera; [39]).

The surveyed triclad species exhibit negative sAT and positive sGC skew values in the coding strand, a typical feature also reported in other Platyhelminthes [6,40]. It has been proposed that this feature would be linked to the replication process [41–43]. That is, the longer strands are kept single during replication, the higher the likelihood of depurination mutations resulting in substitutions from A to G and from C to T (100 times more frequent). However, analysis of the sAT and sGC levels in the PCG as a function of their relative physical order does not show the predicted pattern; instead, there is a clear species-specific signature with contrasting values across species (Fig. 4). The fact that the A+T content (or the NB3 value) and skew values do not correlate across species (S9 and S10 Figs.) does not support the mutational input as a major source for the skew. The situation is the same when we consider the skews for only second or third sites within the coding regions (S9B and S9C Fig; S10B and S10C Fig). These results

suggest that the asymmetric nucleotide composition strand bias has some significance, a feature that could be related to the fact that all genes are located on the same strand (see [44]).

Effect of natural selection on free-living and parasitic species

It has been proposed that parasitic species might exhibit a relaxation of natural selection, as compared with free-living organisms, because of a putative reduction in their effective population sizes [45,46]. Changes in the selection regime may imprint a plethora of characteristic molecular hallmarks on DNA and protein sequences that eventually can be detected. For instance, the relaxation of the intensity of natural selection can cause an increase of the nucleotide and amino acid substitution rates, a decrease in the selective constraint levels (increased values of $\omega = d_N/d_S$ parameter), and an increase in the mutational bias. The effect of such relaxation on the codon usage bias, however, is likely to be more complex: a reduction of codon bias if the bias is actively maintained by natural selection, but an increase if mutation is the stronger force [47]. Here we have taken advantage of the availability of complete mtDNA data for a number of flatworm species to gain insights into this issue. Unfortunately, we cannot analyze either the putative different patterns left on the evolutionary rates (there is no reliable data of divergence times) or its impact of selective constraint levels because of the high saturation of d_S values.

The high A+T content value in all species analyzed, as expected, produces a substantial nucleotide frequency bias. Interestingly, the more pronounced bias corresponds to the NB3 statistic (Fig. 3B), where the highest biases are in species exhibiting the highest A+T content values (Fig. 3A). This result points to mutation, and not to natural selection, as the major evolutionary force responsible for the bias in the nucleotide frequencies. It can be argued that the high levels of A+T may be in fact driven by natural selection acting on the third positions of codons (to get a more efficient codon usage). Nevertheless, we can reject the selective hypothesis since the correlation of the A+T frequency with the frequencies at third codon positions is also observed at the 1st, the 2nd, the rRNA and tRNA sites (Fig. 3C). Remarkably, the free-living and parasitic species differ considerably in their nucleotide frequency bias, with free-living species having higher values (with the exception of *C. alpina*). Moreover, this pattern is consistent across the different NB measures (S8 Fig).

Interestingly, the pattern of codon usage bias reflects that shown by the nucleotide frequency analyses. The codon bias might be a by-product of the mutational input or might result from the action of natural selection for increased translational efficiency or accuracy [48–51]. To disentangle both effects we studied the level of codon bias adjusting for the observed mutation bias (Fig. 5C; S11 Fig). As expected if codon bias mainly results from some form of mutational bias, the SC values drop dramatically, and especially for SC3 values. However, we do not observe any clear pattern that differentiates free-living from parasitic species. Moreover, using different SC-mutational adjusting estimators yields different species-rank orders and, therefore, the separate clustering of free-living (except *C. alpina*) from parasites species on basis of their SC values disappears.

Our results on the impact of nucleotide and codon bias indicate that parasitic platyhelminth species do not exhibit a higher relaxation of natural selection than free-living species. On the contrary, three out of the four free-living species (Geoplanoidea representatives) exhibit patterns of A+T content and nucleotide frequency bias in clear agreement of mutation as the major evolutionary driver. Our results further reveal that the observed codon bias is primarily caused by mutation and not by natural selection mechanisms. Likewise, the high diversity of start codons uncovered in these free-living species and their usage of stop codons can also be explained by a putative relaxation of natural selection (see start and terminal codon usage section). Globally these results agree with that found for bacteria [47], although differ from some

studies of plants, in which mutation appears to have a higher impact than natural selection in parasitic relative to non-parasitic species [16]. In summary, although it has been proposed that life cycles of parasitic species render them more prone to suffering genetic bottlenecks that in turn may lead to putative reductions on the effective population sizes, we did not find the molecular hallmark of a relaxed selection force in the parasitic Platyhelminthes. On the contrary, free-living triclads appear to exhibit higher levels of relaxed selection. In fact their vagility and requirements for persistent habitats may render these species highly vulnerable, very susceptible to local extinctions and recolonizations, which in turn could explain these results. In any case, our conclusions suggest that the relaxed selection proposed for some parasites is not a general feature of parasitic organisms.

Supporting Information

S1 Fig. Linearized schemes of gene orders in Platyhelminthes. Those genes that are variable within each of the three parasitic groups (Cestoda, Monogenea and Trematoda) are in bold. Multiple genes in the same box indicate variable gene orders within the specific group. Gene identifier as in Fig. 2. The tRNAs are shown according to the amino acid code letter. Gene orders derived from the mt genomes of *Diphyllobothrium latum*, *D. nihonkaiense*, *Diplogonoporus balaenopterae*, *D. grandis*, *Echinococcus canadensis*, *E. equinus*, *E. granulosus*, *E. multilocularis*, *E. oligarthrus*, *E. ortleppi*, *E. shiquicus*, *E. vogeli*, *Hymenolepis diminuta*, *Spirometra erinaceieuropaei*, *Taenia asiatica*, *T. crassiceps*, *T. hydatigena*, *T. multiceps*, *T. pisiformis*, *T. saginata*, *T. solium*, *T. taeniaeformis* for CESTODA; *Benedenia hoshinai*, *B. seriolae*, *Tetrancistrum nebulosi* for MONOGENEA 1; *Gyrodactylus derjavinoides*, *G. salaris*, *G. thymalli* for MONOGENEA 2; *Microcotyle sebastis*, *Polylabris halichoeres*, *Pseudochauhanea macrorchis* for MONOGENEA 3; *Clonorchis sinensis*, *Fasciola hepatica*, *Opisthorchis felineus*, *Paragonimus westermani* for TREMATODA 1; *Schistosoma japonicum*, *Sc. mekongi*, *Trichobilharzia regent* TREMATODA 2; *Schistosoma haematobium*, *Sc. mansoni*, *Sc. spindale* for TREMATODA 3. Based on Wey-Fabrizius *et al*., 2013.
(PDF)

S2 Fig. Linearized scheme showing a comparison of the general protein coding and ribosomal genes between generalized mitogenomes of Neodermata (parasitic platyhelminths) and Continenticola (land planarians and freshwater triclads; all free-living).
(PDF)

S3 Fig. Comparison between the annotation proposed by Sakai and Sakaizumi (2012) for *Dugesia japonica* and the new annotation proposed in the present study.
(PDF)

S4 Fig. Comparison by pairs of the tRNA order of the different Tricladida species included in this work.
(PDF)

S5 Fig. Secondary structure of the 22 tRNA of *Crenobia alpina*. *trnQ in a box shows the alternative structure proposed for this tRNA.** The different tRNA parts are showed on *trnD*.
(PDF)

S6 Fig. Secondary structure of the 22 tRNA of *Obama* sp.
(PDF)

S7 Fig. Comparison between the Sakai and Sakaizumi (2012) *trnA*, *trnC* and *trnL1* secondary structure for *Dugesia japonica* based on their annotation and the secondary structure

based on our new proposed annotation.
(PDF)

S8 Fig. Values of the different NB-based statistic across species excluding the NB3 (NB at the third position of four-fold degenerate codons).
(PDF)

S9 Fig. Relationship between sAT, sGC values and NB3. sAT general skew; sAT2, sAT skew at the second positions; sAT3, sAT at the third positions. sGC, general skew; sGC2, sGC skew at the second positions; sGC3, sGC at the third positions. Green squares and red circles indicate free-living and parasitic platyhelminths, respectively. The surveyed species are shown in numbers: 1, *T. sigani*; 2, *F. hepatica*; 3, *D. balaenopterae*; 4, *B. hoshinai*; 5, *T. saginata*; 6, *S. japonicum*; 7, *C. alpina*; 8, *Obama* sp.; 9, *S. mediterranea*; 10, *D. japonica*.
(PDF)

S10 Fig. Relationship between sAT and sGC values and A+T content. sAT general skew; sAT2, sAT skew at the second positions; sAT3, sAT at the third positions. sGC, general skew; sGC2, sGC skew at the second positions; sGC3, sGC at the third positions. Green squares and red circles indicate free-living and parasitic Platyhelminthes, respectively. The surveyed species are shown in numbers: 1, *T. sigani*; 2, *F. hepatica*; 3, *D. balaenopterae*; 4, *B. hoshinai*; 5, *T. saginata*; 6, *S. japonicum*; 7, *C. alpina*; 8, *Obama* sp.; 9, *S. mediterranea*; 10, *D. japonica*.
(PDF)

S11 Fig. SC values across species adjusted for the observed mutation bias. Ordered ascending based on the Chi scales values for A) second positions of the PCG and B) for the third position of four-fold degenerate codons equifrequency.
(PDF)

S1 File. Supplementary information on negative results for *Polycelis felina* and *Dugesia subtentaculata*.
(DOCX)

S1 Tables. Table A. Locality and habitat information on species collected for this study. Table B. Data of mitochondrial proteins used to conduct the tBLASTx analyses in order to detect whether the mitochondrial genes were present in the 454 sequencing reads. Table C. Primers designed for the reamplification of *Crenobia alpina*. Table D. Primers designed for the reamplification of *Obama* sp. Table E. Summary statistics for the 454 sequencing. Table F. Summary of tBLASTn hits for raw reads against the mitochondrial proteins of the three parasitic flatworms. Table G. Annotation table for the mitochondrial genome of *C. alpina*. Table H. Annotation table for the mitochondrial genome of *Obama* sp. Table I. Annotation table for the mitochondrial genome of *S. mediterranea*.
(DOCX)

Acknowledgments

We want to thank Mrs. Jill McDonald who kindly contributed with samples, to M. Gorchs and L. Leria that helped in the collection of *C. alpina*, E. Mateos who helped us in collecting *P. felina* and to Jitka Aldhoun who gave support in some laboratory experiments.

Author Contributions

Conceived and designed the experiments: MR JR. Performed the experiments: ES MAP DTJL MR. Analyzed the data: ES MAP CFL JR MR. Wrote the paper: ES MAP CFL DTJL JR MR.

References

1. Ballard JWO, Pichaud N. Mitochondrial DNA: more than an evolutionary bystander. Funct Ecol. 2014; 28: 218–231.
2. Ballard JWO, Whitlock MC. The incomplete natural history of mitochondria. Mol Ecol. 2004; 13: 729–744. PMID: 15012752
3. Barr CM, Neiman M, Taylor DR. Inheritance and recombination of mitochondrial genomes in plants, fungi and animals. New Phytol. 2005; 168: 39–50. PMID: 16159319
4. Stöck M, Moritz C, Hickerson M, Frynta D, Dujsebayeva T, Eremchenko V, et al. Evolution of mitochondrial relationships and biogeography of Palearctic green toads (Bufo viridis subgroup) with insights in their genomic plasticity. Mol Phylogenet Evol. 2006; 41: 663–689. PMID: 16919484
5. Le TH, McManus DP, Blair D. Codon usage and bias in mitochondrial genomes of parasitic platyhelminthes. Korean J Parasitol. 2004; 42: 159–167. PMID: 15591833
6. Wey-Fabrizius AR, Podsiadlowski L, Herlyn H, Hankeln T. Platyzoan mitochondrial genomes. Mol Phylogenet Evol. 2013; 69: 365–375. doi: 10.1016/j.ympev.2012.12.015 PMID: 23274056
7. Ruiz-Trillo I, Riutort M, Fourcade HM, Baguñà J, Boore JL. Mitochondrial genome data support the basal position of Acoelomorpha and the polyphyly of the Platyhelminthes. Mol Phylogenet Evol. 2004; 33: 321–332. PMID: 15336667
8. Sakai M, Sakaizumi M. The complete mitochondrial genome of Dugesia japonica (Platyhelminthes; order Tricladida). Zoolog Sci. 2012; 29: 672–680. doi: 10.2108/zsj.29.672 PMID: 23030340
9. Riutort M, Álvarez-Presas M, Lázaro E, Solà E, Paps J. Evolutionary history of the Tricladida and the Platyhelminthes: an up-to-date phylogenetic and systematic account. Int J Dev. 2012; 56: 5–17.
10. Solà E, Sluys R, Gritzalis K, Riutort M. Fluvial basin history in the northeastern Mediterranean region underlies dispersal and speciation patterns in the genus *Dugesia* (Platyhelminthes, Tricladida, Dugesiidae). Mol Phylogenet Evol. 2013; 66: 877–888. doi: 10.1016/j.ympev.2012.11.010 PMID: 23182762
11. Álvarez-Presas M, Sánchez-Gracia A, Carbayo F, Rozas J, Riutort M. Insights into the origin and distribution of biodiversity in the Brazilian Atlantic forest hot spot: a statistical phylogeographic study using a low-dispersal organism. Heredity (Edinb). 2014; 112: 656–665. doi: 10.1038/hdy.2014.3 PMID: 24549112
12. Telford MJ, Herniou EA, Russell RB, Littlewood DT. Changes in mitochondrial genetic codes as phylogenetic characters: two examples from the flatworms. Proc Natl Acad Sci U S A. 2000; 97: 11359–11364. PMID: 11027335
13. Dowton M, Austin AD. Increased genetic diversity in mitochondrial genes is correlated with the evolution of parasitism in the Hymenoptera. J Mol Evol. 1995; 41: 958–965. PMID: 8587141
14. Page RD, Lee PL, Becher SA, Griffiths R, Clayton DH. A different tempo of mitochondrial DNA evolution in birds and their parasitic lice. Mol Phylogenet Evol. 1998; 9: 276–293. PMID: 9562986
15. Castro LR, Austin AD, Dowton M. Contrasting rates of mitochondrial molecular evolution in parasitic Diptera and Hymenoptera. Mol Biol Evol. 2002; 19: 1100–1113. PMID: 12082129
16. Bromham L, Cowman PF, Lanfear R. Parasitic plants have increased rates of molecular evolution across all three genomes. BMC Evol Biol. 2013; 13: 126. doi: 10.1186/1471-2148-13-126 PMID: 23782527
17. Álvarez-Presas M, Mateos E, Tudó À, Jones H, Riutort M. Diversity of introduced terrestrial flatworms in the Iberian Peninsula: a cautionary tale. PeerJ. 2014; 2: e430. doi: 10.7717/peerj.430 PMID: 24949245
18. Bessho Y, Tamura S, Hori H, Tanaka H, Ohama T, Osawa S. Planarian mitochondria sequence heterogeneity: relationships between the type of cytochrome c oxidase subunit I gene sequence, karyotype and genital organ. Mol Ecol. 1997; 6: 129–136. PMID: 9061940
19. Chomczynski P, Sacchi N. Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. Anal Biochem. 1987; 162: 156–159. PMID: 2440339
20. Schmieder R, Edwards R. Quality control and preprocessing of metagenomic datasets. Bioinformatics. 2011; 27: 863–864. doi: 10.1093/bioinformatics/btr026 PMID: 21278185
21. Nakao M, Sako Y, Ito A. The Mitochondrial Genome of the Tapeworm *Taenia solium*: A Finding of the Abbreviated Stop Codon U. J Parasitol. 2003; 89: 633–635. PMID: 12880275
22. Huyse T, Buchmann K, Littlewood DTJ. The mitochondrial genome of *Gyrodactylus derjavinoides* (Platyhelminthes: Monogenea)—a mitogenomic approach for *Gyrodactylus* species and strain identification. Gene. 2008; 417: 27–34. doi: 10.1016/j.gene.2008.03.008 PMID: 18448274
23. Le T, Blair D, Agatsuma T. Phylogenies inferred from mitochondrial gene orders—a cautionary tale from the parasitic flatworms. Mol Biol Evol. 2000; 17: 1123–1125. PMID: 10889225

24. Biomatters. 2014. Geneious. Biomatters, Inc. San Francisco, CA.
25. Huse SM, Huber JA, Morrison HG, Sogin ML, Welch DM. Accuracy and quality of massively parallel DNA pyrosequencing. Genome Biol. 2007; 8: R143. PMID: 17659080
26. Bernt M, Donath A, Jühling F, Externbrink F, Florentz C, Fritzsch G, et al. MITOS: improved de novo metazoan mitochondrial genome annotation. Mol Phylogenet Evol. 2013; 69: 313–319. doi: 10.1016/j.ympev.2012.08.023 PMID: 22982435
27. Abascal F, Zardoya R, Posada D. GenDecoder: genetic code prediction for metazoan mitochondria. Nucleic Acids Res. 2006; 34: W389–W393. PMID: 16845034
28. Laslett D, Canbäck B. ARWEN: a program to detect tRNA genes in metazoan mitochondrial nucleotide sequences. Bioinformatics. 2008; 24: 172–175. PMID: 18033792
29. Schattner P, Brooks AN, Lowe TM. The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. Nucleic Acids Res. 2005; 33: W686–9. PMID: 15980563
30. Wyman SK, Jansen RK, Boore JL. Automatic annotation of organellar genomes with DOGMA. Bioinformatics. 2004; 20: 3252–3255. PMID: 15180927
31. Shields DC, Sharp PM, Higgins DG, Wright F. "Silent" sites in *Drosophila* genes are not neutral: evidence of selection among synonymous codons. Mol Biol Evol. 1988; 5: 704–716. PMID: 3146682
32. Perna NT, Kocher TD. Patterns of nucleotide composition at fourfold degenerate sites of animal mitochondrial genomes. J Mol Evol. 1995; 41: 353–358. PMID: 7563121
33. Wright F. The "effective number of codons" used in a gene. Gene. 1990; 87: 23–29. PMID: 2110097
34. Gissi C, Iannelli F, Pesole G. Evolution of the mitochondrial genome of Metazoa as exemplified by comparison of congeneric species. Heredity (Edinb). 2008; 101: 301–320. doi: 10.1038/hdy.2008.62 PMID: 18612321
35. Breton S, Stewart DT, Hoeh WR. Characterization of a mitochondrial ORF from the gender-associated mtDNAs of *Mytilus* spp. (Bivalvia: Mytilidae): identification of the "missing" ATPase 8 gene. Mar Genomics. 2010; 3: 11–18. doi: 10.1016/j.margen.2010.01.001 PMID: 21798192
36. Bessho Y, Ohama T, Osawa S. Planarian mitochondria II. The unique genetic code as deduced from cytochrome c oxidase subunit I gene sequences. J Mol Evol. 1992; 34: 331–335. PMID: 1314909
37. Le T, Pearson M, Blair D, Dai N. Complete mitochondrial genomes confirm the distinctiveness of the horse-dog and sheep-dog strains of *Echinococcus granulosus*. Parasitology. 2002; 124: 97–112. PMID: 11811807
38. Boore J, Brown W. Complete sequence of the mitochondrial DNA of the annelid worm *Lumbricus terrestris*. Genetics. 1995; 141: 305–319. PMID: 8536978
39. Wei S, Shi M, He J, Sharkey M, Chen X. The complete mitochondrial genome of *Diadegma semiclausum* (hymenoptera: ichneumonidae) indicates extensive independent evolutionary events. Genome. 2009; 52: 308–319. doi: 10.1139/g09-008 PMID: 19370087
40. Weber M, Wey-Fabrizius AR, Podsiadlowski L, Witek A, Schill RO, Sugár L, et al. Phylogenetic analyses of endoparasitic Acanthocephala based on mitochondrial genomes suggest secondary loss of sensory organs. Mol Phylogenet Evol. 2013; 66: 182–189. doi: 10.1016/j.ympev.2012.09.017 PMID: 23044398
41. Tillier ERM, Collins RA. The contributions of replication orientation, gene direction, and signal sequences to base-composition asymmetries in bacterial genomes. J Mol Evol. 2000; 50: 249–257. PMID: 10754068
42. Necşulea A, Lobry JR. A new method for assessing the effect of replication on DNA base composition asymmetry. Mol Biol Evol. 2007; 24: 2169–2179. PMID: 17646257
43. Marín A, Xia X. GC skew in protein-coding genes between the leading and lagging strands in bacterial genomes: new substitution models incorporating strand bias. J Theor Biol. 2008; 253: 508–513. doi: 10.1016/j.jtbi.2008.04.004 PMID: 18486155
44. Francino MP, Chao L, Riley MA, Ochman H. Asymmetries generated by transcription-coupled repair in Enterobacterial genes. Science. 1996; 272: 107–109. PMID: 8600517
45. Huyse T, Poulin R, Théron A. Speciation in parasites: a population genetics approach. Trends Parasitol. 2005; 21: 469–475. PMID: 16112615
46. Woolfit M, Bromham L. Increased rates of sequence evolution in endosymbiotic bacteria and fungi with small effective population sizes. Mol Biol Evol. 2003; 20: 1545–1555. PMID: 12832648
47. Sharp PM, Emery LR, Zeng K. Forces that influence the evolution of codon bias. Philos Trans R Soc Lond B Biol Sci. 2010; 365: 1203–1212. doi: 10.1098/rstb.2009.0305 PMID: 20308095
48. Bernardi G, Bernardi G. Compositional constraints and genome evolution. J Mol Evol. 1986; 24: 1–11. PMID: 3104608

49. Poh Y-P, Ting C-T, Fu H-W, Langley CH, Begun DJ. Population genomic analysis of base composition evolution in *Drosophila melanogaster*. Genome Biol Evol. 2012; 4: 1245–1255. doi: 10.1093/gbe/evs097 PMID: 23160062

50. Lawrie DS, Messer PW, Hershberg R, Petrov DA. Strong purifying selection at synonymous sites in *D. melanogaster*. PLoS Genet. 2013; 9: e1003527. doi: 10.1371/journal.pgen.1003527 PMID: 23737754

51. Chen H, Sun S, Norenburg JL, Sundberg P. Mutation and Selection Cause Codon Usage and Bias in Mitochondrial Genomes of Ribbon Worms (Nemertea). PLoS One. 2014; 9: e85631. doi: 10.1371/journal.pone.0085631 PMID: 24454907

52. Sluys R, Solà E, Gritzalis K, Vila-Farré M, Mateos E, Riutort M. Integrative delineation of species of Mediterranean freshwater planarians (Platyhelminthes: Tricladida: Dugesiidae). Zool J Linn Soc. 2013; 169: 523–547.

53. Kang S, Kim J, Lee J, Kim S, Min G-S, Park J-K.The complete mitochondrial genome of an ectoparasitic monopisthocotylean fluke *Benedenia hoshinai* (Monogenea: Platyhelminthes). Mitochondrial DNA. 2012; 23: 176–178. doi: 10.3109/19401736.2012.668900 PMID: 22545965

54. Yamasaki H, Ohmae H, Kuramochi T. Complete mitochondrial genomes of *Diplogonoporus balaenopterae* and *Diplogonoporus grandis* (Cestoda: Diphyllobothriidae) and clarification of their taxonomic relationships. Parasitol Int. 2012; 61: 260–266. doi: 10.1016/j.parint.2011.10.007 PMID: 22079238

55. Jeon H-K, Kim K-H, Eom KS. Complete sequence of the mitochondrial genome of *Taenia saginata*: comparison with *T. solium* and *T. asiatica*. Parasitol Int. 2007; 56: 243–246. PMID: 17499016

56. Zhang J, Wu X, Li Y, Xie M, Li A. The complete mitochondrial genome of *Tetrancistrum nebulosi* (Monogenea: Ancyrocephalidae). Mitochondrial DNA. 2014; 1736: 1–2.

B

Evolution of Chemosensory Gene Families in Arthropods: Insight from the First Inclusive Comparative Transcriptome Analysis across Spider Appendages

Joel Vizueta, Cristina Frías-López, Nuria Macías-Hernández, MiquelA Arnedo, Alejandro Sánchez-Gracia, and Julio Rozas

2016, Genome Biology Evolution, 9(1):178–196.

GBE

Evolution of Chemosensory Gene Families in Arthropods: Insight from the First Inclusive Comparative Transcriptome Analysis across Spider Appendages

Joel Vizueta[1], Cristina Frías-López[1], Nuria Macías-Hernández[2], Miquel A. Arnedo[2], Alejandro Sánchez-Gracia[1,*], and Julio Rozas[1,*]

[1]Departament de Genètica, Microbiologia i Estadística and Institut de Recerca de la Biodiversitat (IRBio), Universitat de Barcelona, Spain

[2]Departament de Biologia Evolutiva, Ecologia i Ciències Ambientals and Institut de Recerca de la Biodiversitat (IRBio), Universitat de Barcelona, Spain

*Corresponding authors: E-mails: jrozas@ub.edu; elsanchez@ub.edu.

Accepted: December 16, 2016

Data deposition: This project has been deposited at the Sequence Read Archive (SRA) database under accession numbers SRX1612801, SRX1612802, SRX1612803 and SRX1612804 (Bioproject number: PRJNA313901).

Abstract

Unlike hexapods and vertebrates, in chelicerates, knowledge of the specific molecules involved in chemoreception comes exclusively from the comparative analysis of genome sequences. Indeed, the genomes of mites, ticks and spiders contain several genes encoding homologs of some insect membrane receptors and small soluble chemosensory proteins. Here, we conducted for the first time a comprehensive comparative RNA-Seq analysis across different body structures of a chelicerate: the nocturnal wandering hunter spider *Dysdera silvatica* Schmidt 1981. Specifically, we obtained the complete transcriptome of this species as well as the specific expression profile in the first pair of legs and the palps, which are thought to be the specific olfactory appendages in spiders, and in the remaining legs, which also have hairs that have been morphologically identified as chemosensory. We identified several ionotropic (*Ir*) and gustatory (*Gr*) receptor family members exclusively or differentially expressed across transcriptomes, some exhibiting a distinctive pattern in the putative olfactory appendages. Furthermore, these IRs were the only known olfactory receptors identified in such structures. These results, integrated with an extensive phylogenetic analysis across arthropods, uncover a specialization of the chemosensory gene repertoire across the body of *D. silvatica* and suggest that some IRs likely mediate olfactory signaling in chelicerates. Noticeably, we detected the expression of a gene family distantly related to insect odorant-binding proteins (OBPs), suggesting that this gene family is more ancient than previously believed, as well as the expression of an uncharacterized gene family encoding small globular secreted proteins, which appears to be a good chemosensory gene family candidate.

Key words: chemosensory gene families, specific RNA-Seq, *de novo* transcriptome assembly, functional annotation, chelicerates, arthropods.

Introduction

Chemoreception, the detection and processing of chemical signals in the environment, is a biological process that is critical for animal survival and reproduction. The essential role of smell and taste in the detection of food, hosts and predators and their participation in social communication make the molecular components of this system solid candidates for important adaptive changes associated with animal terrestrialization (Whiteman and Pierce 2008). In insects, chemical recognition occurs in specialized hair-like cuticular structures called sensilla, which can be found almost anywhere in the body (Joseph and Carlson 2015). In *Drosophila*, olfactory sensilla are concentrated on the antenna and the maxillary palps, while gustatory sensilla are spread across various body locations, such as the proboscis, the legs and the anterior margins of wings (Pelosi 1996; Shanbhag et al. 2001). The chemoreceptor proteins embedded within the membrane of sensory neurons (SN) innervating these sensilla are responsible for transducing the external chemical signal into an action potential. In the case of smell, olfactory SNs project the axons to

specific centers of the brain, where the signals are processed and engender a behavioral response to the specific external stimuli. The process can be facilitated by small soluble chemosensory proteins that are secreted in the lymph that bathes the dendrites of the SNs and are believed to solubilize and either transport the signaling molecules to membrane receptors or protect them from premature degradation (Vogt and Riddiford 1981; Pelosi et al. 2006). Although insect chemoreceptors and soluble chemosensory proteins are encoded by gene families exhibiting high gene turnover rates (see Sánchez-Gracia et al. 2011 for a comprehensive review), distant homologues of the members of these families have been identified in other arthropod lineages (Colbourne et al. 2011; Vieira and Rozas 2011; Chipman et al. 2014; Frías-López et al. 2015; Gulia-Nuss et al. 2016). Vertebrate functional counterparts of these gene families, however, are not evolutionarily related; indeed, the members of this subphylum use different molecules to perform the same general physiological function (Kaupp 2010).

Spiders comprise a highly diverse group of arthropods, including >45,000 described species (World Spider Catalog 2016), and are dominant predators in most terrestrial ecosystems. Given their potential as biological control agents as well as the engineering properties of silk and venom, these organisms are of great economic and medical relevance (Clarke et al. 2014). Because the Arachnida ancestors of these chelicerates colonized the land ~475 Ma, long after the split of the four major extant arthropod lineages (Rota-Stabelli et al. 2013), spiders are good models for comparative studies on the diverse strategies adopted by arthropod lineages during their independent adaptation to terrestrial environments. However, despite their biological and translational implications, there are relatively few genomic and transcriptomic studies conducted on these organisms compared with those conducted on insects, and studies on spiders almost exclusively focus on silk and venom research (Grbić et al. 2011; Clarke et al. 2014; Posnien et al. 2014; Sanggaard et al. 2014).

Spiders can detect volatile and nonvolatile compounds through specialized chemosensitive hairs distributed at the tips of various extremities and appendages, including legs and palps (Foelix 1970; Foelix and Chu-Wang 1973; Kronestedt 1979; Cerveira and Jackson 2012; Foelix et al. 2012). Nevertheless, the molecular nature of chelicerate chemoreceptors has remained elusive until recently. We and others have identified distant homologs of some insect gene families associated with chemosensation in the genomes of mites, ticks and spiders (Montagné et al. 2015; Gulia-Nuss et al. 2016), such as members of the gustatory (*Gr*) and ionotropic (*Ir*) receptor, and of the chemosensory protein (*Csp*), Niemann–Pick protein type C2 (*Npc2*) and sensory neuron membrane protein (*Snmp*) multigene families. In addition, chelicerates lack homologs of the typical insect olfactory receptor family *Ors*, which are thought to have originated later with the appearance of flying insects, and no *Obp* gene had been detected to date (Vieira and Rozas 2011; Chipman et al. 2014). Overall, available genomic studies suggest that the *Ir* gene family is responsible for smell not only in chelicerates but also in all nonneopteran arthropods (Croset et al. 2010; Colbourne et al. 2011; Chipman et al. 2014; Gulia-Nuss et al. 2016). Regarding taste, the presence of numerous copies of *Gr* and nonconserved *Ir* (a group of divergent IR proteins associated with gustatory function in insects, Croset et al. 2010) genes in chelicerate genomes clearly suggests that these families are responsible for contact chemoreception in this species.

Nevertheless, the simple comparative analysis of genomic sequences does not allow inferring which specific members of already known chemosensory families are involved in the different sensory modalities. Additionally, chelicerates could also use molecules completely different from those already known in insects during the water-to-land transition, which should also be different from those used by vertebrates (these molecules have also not been found in the available genome sequences); these uncharacterized genes (or annotated with incomplete gene models) would be not directly detectable only by comparative genomics. Instead, specific transcriptomic analyses of chemosensory tissues can provide useful insight into all these issues. Antennae-specific gene expression studies in lobsters and hermit crabs (Corey et al. 2013; Groh-Lunow et al. 2014), for example, have revealed the presence of several transcripts encoding IRs, supporting the active role in olfaction of this gene family in crustaceans. To gain insight into the specific proteins involved in chelicerate chemoreception, we recently performed a tissue-specific comparative transcriptomics study in the funnel-web spider *Macrothele calpeiana* (Frías-López et al. 2015). Unfortunately, we failed to detect the specific expression of *Ir* or *Gr* genes in the first pair of legs and in palps, the best candidate structures to hold olfactory hairs in chelicerates. This result might be caused by either the sedentary lifestyle of this mygalomorph spider, which may lead to a marginal role of chemical communication in this species, or the low sequencing coverage of this RNA-Seq study.

Here, in order to better characterize the chemosensory repertoire of a spider, we report a more comprehensive comparative transcriptomic analysis in an active nocturnal hunter spider, *Dysdera silvatica* Schmidt, 1981 (Araneae, Dysderidae) (fig. 1). This species, which is endemic to the Canary Islands, belongs to a genus characterized by long and protruding chelicerae used to capture and feed on woodlice (Crustacea: Isopoda: Oniscidea; fig. 1*B*). We have conducted a deep RNA-Seq experiment in four separated body parts, three of them likely containing chemosensitive hairs in spiders. Because the performance of the *de novo* assembly of short reads strongly depends on biological data (i.e., the complexity of the data is almost species specific), we first performed a comparative analysis among a set of commonly used software for transcriptome assembly. Based on the

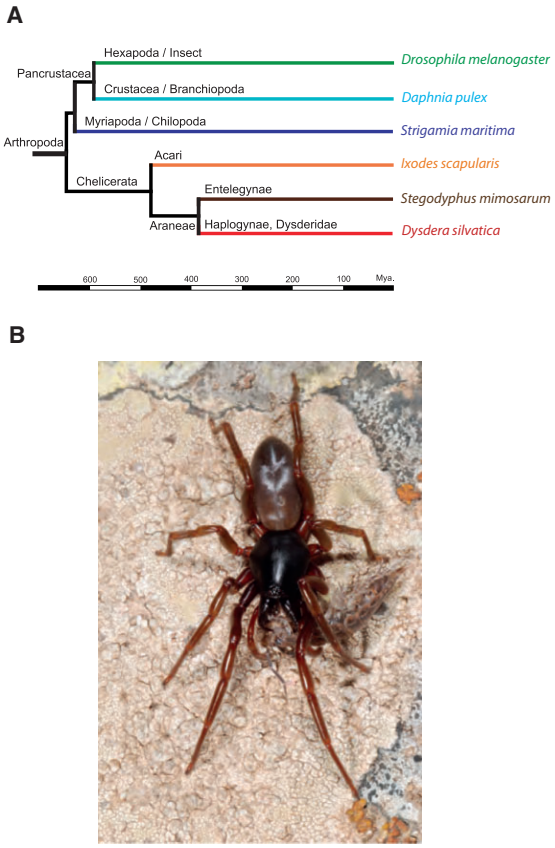


Fig. 1.—(*A*) Phylogenetic position of *Dysdera silvatica* within arthropods. Divergence times were obtained from TimeTree (Hedges et al. 2015). (*B*) *D. silvatica* feeding on a woodlouse.

best assembly and highly accurate functional annotations, we conducted a comparative analysis between the specific transcriptomes of the different body parts, emphasizing the detection of distinctive chemosensory profiles, especially in the palps and the first pair of legs, which has been reported to hold the peripheral olfactory structures in spiders. We then contextualized these results by applying a sound phylogenetic analysis including representative members of each arthropod chemosensory gene family.

We have identified several members of the *Ir* and *Gr* gene families specifically or differently expressed in some of the four surveyed transcriptomes (including a clear homolog of the co-receptor IR25a of *Drosophila melanogaster*) and some signs of chemosensory specialization across spider chemosensory structures. Moreover, we have also identified three genes distantly related to the insect *Obp* gene family and a new gene family encoding small secreted soluble proteins that might function as molecular carriers in the spider chemosensory system. We discuss these findings in the context of the origin and evolution of chemosensory gene families in arthropods and propose some candidate genes that may have an important chemoreceptor role in spiders.

Materials and Methods

Sample Collection, RNA Extraction and Library Preparation

We sequenced and analysed the transcriptome of four *D. silvatica* males (voucher specimens were deposited at the *Centre de Recursos de Biodiversitat Animal* of the Universitat de Barcelona under catalog numbers NMH2597-99 and NMH2601) collected from the Canary Islands, La Gomera and Las Tajoras (28.112736 N, 17.262511 W) in 2013. We used males because this sex has been shown to respond to sex-specific olfactory information (Nelson et al. 2012). We performed four separated RNA-Seq experiments, which included expressed sequences form the palps (*PALP*), the first pair of legs (*LEG#1*), all other pairs of legs (*LEG#234*)

and the remaining body structures (*REST*), henceforth referred to as experimental conditions. We dissected these body parts independently for each of the four males (after snap freezing in liquid nitrogen) and extracted the total RNA separately for each condition and sample using the RNeasy Mini kit (Qiagen, Venlo, The Netherlands) and TRIzol reagent (Invitrogen, Waltham, MA). We determined the amount and integrity of RNA using a Qubit Fluorometer (Life Technologies, Grand Island, NY) and Agilent 2100 Bioanalyzer (CCiTUB, Barcelona, Spain), respectively. We sequenced the transcriptome of each condition using the Illumina Genome Analyzer HiSeq 2000 (100 bp PE reads) according to the manufacturer's instructions (Illumina, San Diego, CA). Briefly, for each experimental condition, the mRNA was purified from 1 μg of total RNA using magnetic oligo(dT) beads and fragmented into small pieces. Double-stranded cDNA was synthesized with random hexamer (N6) primers (Illumina), and Illumina paired-end (PE) adapters were ligated to the ends of adenylated cDNA fragments. All library preparation steps and transcriptome sequencing were carried out in Macrogen Inc., Seoul, South Korea.

Raw Data Pre-Processing

Raw NGS data were pre-processed to eliminate all reads with a quality score $\leq$20 in at least the 30% of the read length and to remove reads with putative sequencing errors using NGSQCToolkit and SEECER v_0.1.3 (Patel and Jain 2012; Le et al. 2013). Before the assembly step, we performed an in silico normalization of filtered reads using Diginorm, an algorithm included in Trinity software (Haas et al. 2014). We set 50X as the targeted maximum coverage for the reads.

De Novo Transcriptome Assembly

First, to determine the best assembler for the *D. silvatica* RNA-Seq data, we compared the performance of five commonly used software programs in assembling the specific transcriptome of the experimental condition *REST*. We tested Trinity r2.1.1, Bridger r2014-12-01, SOAPdenovo-Trans release 1.03, Oases version_0.2.8, and ABySS version_1.3.7/trans-ABySS version1.4.8 (Birol et al. 2009; Schulz et al. 2012; Xie et al. 2014; Z. Chang et al. 2015). For this comparative analysis and depending on the specificities of the selected software (allowing single or multiple *k*-mer values), we applied several single *k*-mer lengths and *k*-mer ranges (see supplementary table S1, Supplementary Material online, for details).

After the assembly phase, we removed all contigs with evidence of contaminant sequences using the software Seqclean (ftp://occams.dfci.harvard.edu/pub/bio/tgi/software/; last accessed May 1, 2015) together with the sequences of the UniVec vector database and the genomes of *Escherichia coli, Pseudomonas aeruginosa, Staphylococcus aureus, Saccharomyces cerevisiae* and *Homo sapiens*. Clean contigs were then clustered into putative transcripts (analogous to the Trinity *components*). We determined the assembly performance of each software based on (1) the DETONATE score (Li et al. 2014), (2) the outcome of the assembled sequences in a set of sequence similarity and profile-based searches using different databases (see the "Results" section for more details), and (3) some commonly used descriptive statistics on assembly quality, namely the average sequence length, the N50, the maximum and minimum transcript lengths and the total bases in the assembly, calculated with the NGSQCToolkit software and some Perl scripts. All analyses were run in a 64-CPU machine with 750 Gb of RAM.

Protein Databases

We built two customized protein databases to assist the functional annotation of the *D. silvatica* transcriptome. The arthropodDB database contains the publicly available amino acid sequences of fully annotated proteins and protein models from a set of representative arthropod genomes and some appropriated external groups, along with their complete entry description, associated GO terms and InterPro identifiers (Ashburner et al. 2000; Mitchell et al. 2014). This database includes information for the following species: (1) the chelicerates *Ixodes scapularis* (Acari) (Gulia-Nuss et al. 2016), *Metaseiulus occidentalis* (Acari) (https://www.hgsc.bcm.edu/arthropods/western-orchard-predatory-mite-genome-project; last accessed May 1, 2015), *Tetranychus urticae* (Acari) (Grbić et al. 2011), *Mesobuthus martensii* (Scorpiones) (Cao et al. 2013), *Acanthoscurria geniculata* (Araneae, Theraphosidae) (Sanggaard et al. 2014), *Stegodyphus mimosarum* (Araneae, Eresidae) (Sanggaard et al. 2014), *Latrodectus hesperus* (Araneae) (https://www.hgsc.bcm.edu/arthropods/western-black-widow-spider-genome-project; last accessed May 1, 2015), *Loxosceles reclusa* (Araneae, Sicariidae) (https://www.hgsc.bcm.edu/arthropods/brown-recluse-spider-genome-project; last accessed May 1, 2015) and *Parasteatoda tepidariorum* (Araneae, Theridiidae) (https://www.hgsc.bcm.edu/arthropods/common-house-spider-genome-project; last accessed May 1, 2015); (2) the hexapods *D. melanogaster* (Diptera) (Adams et al. 2000), *Pediculus humanus* (Phthiraptera) (Kirkness et al. 2010) and *Bombyx mori* (Lepidoptera) (Mita et al. 2004); (3) the crustacean *Daphnia pulex* (Branchiopoda) (Colbourne et al. 2011); (4) the myriapod *Strigamia maritima* (Chilopoda, Geophilomorpha) (Chipman et al. 2014); (5) the tardigrade *Hypsibius dujardini* (http://badger.bio.ed.ac.uk/H_dujardini; last accessed May 1, 2015); and (6) the nematode *Caenorhabditis elegans*. In the cases where there was no functional description or associated GO term (e.g., the protein models from *A. geniculata*, *L. hesperus*, *L. reclusa*, *M. martensii*, *M. occidentalis* and *P. tepidariorum*), we approximated the functional annotation using InterProScan version 5.4.47 (Jones et al. 2014).

The chemDB database contains the amino acid sequences and the functional information of all well-annotated members

of the *Or*, *Gr Ir*, *Csp*, *Obp*, *Npc2* and *Snmp* gene families from a representative set of insect species, namely *D. melanogaster*, *Tribolium castaneum* (Coleoptera), *Apis mellifera* (Hymenoptera) and *Acyrthosiphon pisum* (Hemiptera), and from the noninsect species included in arthropodDB. Moreover, we also included in chemDB some vertebrate odorant binding proteins and olfactory and taste receptors identified by the InterPro signatures IPR002448, IPR000725 and IPR007960, respectively (see supplementary table S1*B* in Frías-López et al. 2015). Furthermore, we progressively updated chemDB by adding to this database all novel members of these chemosensory families (the conceptual translation of the identified transcripts) characterized in *D. silvatica*.

Functional Annotation of the *D. silvatica* Transcripts

We applied a similarity-based search approach to assist the annotation of the *D. silvatica* transcriptome. We first used *BLASTx* to search the translated transcripts against the SwissProt and arthropodDB databases (BLAST v2.2.29; Altschul et al. 1990; Altschul 1997). To search against NCBI-nr, we used GHOSTZ version1.0.0; this software is much faster than *BLAST*, especially for large databases without a substantial reduction of sensibility (Suzuki et al. 2014). We improved the functional annotation by searching for the specific protein-domain signatures in translated transcriptome sequences using InterProScan (Jones et al. 2014). We predicted signal peptides and transmembrane helices with SignalP and TMHMM, respectively (Krogh et al. 2001; Petersen et al. 2011). To carry out the profile-based searches, we created custom HMM models, one for each chemosensory family included in chemDB. These models are based on multiple sequence alignments (MSA) built with the program *hmmalign* (HMMER 3.1b1 package; Eddy 2011) using the specific core Pfam profile as a guide.

We conducted a GO-enrichment analysis with the BLAST2GO term suite using all functionally annotated transcripts with an associated GO term (Conesa et al. 2005). Moreover, we also searched these functionally annotated transcripts for KEGG enzymes and pathways (Kanehisa and Goto 2000), for CEG (Core Eukaryotic Genes) (Parra et al. 2007; Parra et al. 2009) and for the list of housekeeping (HK) genes used in supplementary table S1*A* in Frías-López et al. (2015).

To characterize the chemosensory gene repertory of *D. silvatica*, we first used the proteins in chemDB as query sequences to search for putative homologs among spider transcripts (using *tBLASTn* search; *E*-value cutoff of 10^{-3}). We only considered as positives those hits covering at least 2/3 of the query sequence length or the 80% of the total subject sequence. Then, we conducted some additional searches based on our custom HMM models and the conceptual translation of *D. silvatica* transcripts as subject sequences (using *hmmer* and an i-*E*-value of 10^{-3}). The integration of the results from these different analyses provided us a highly curated and trustworthy set of *D. silvatica* chemosensory-related transcripts.

Expression Profiling across Experimental Conditions

The pre-processed reads of each experimental condition (*LEG#1*, *LEG#234*, *PALP*, and *REST*) were back aligned to the final reference transcriptome using Bowtie version 1.0.0 (Langmead et al. 2009). We used RSEM 1.2.19 software to obtain read counts and TMM-normalized FPKMs (i.e., trimmed mean of M values-normalized fragments per kb of exon per million reads mapped) per transcript (Li and Dewey 2011). For the analysis, we consider that a gene is actually expressed when the FPKM values are >0.01, a reasonable cut-off given the low expression levels reported for other arthropod chemoreceptor proteins (Zhang et al. 2014). For the differential expression analysis, we considered that our data represent a single biological replicate (Robinson et al. 2010) and used EdgeR version 3.6.8 to calculate the negative binomial dispersion across conditions from the read counts of HK genes (Robinson et al. 2010). The *P* values from the differential expression analysis were adjusted for the false discovery rate (FDR; Benjamini and Hochberg 1995).

Phylogenetic Analyses

The quality of the MSA is critical to obtain a reliable phylogenetic reconstruction. This issue is very problematic in the face of highly divergent sequences, as in our case. To minimize this problem, we applied a profile-guided MSA approach based on highly curated Pfam core profiles, which generated MSAs with better TCS scores than other MSA approaches (Chang et al. 2014; J.-M. Chang et al. 2015). We used RAxML version 8.2.1 and the WAG protein substitution model with rate heterogeneity among sites to determine the phylogenetic relationships among the members of each chemosensory gene family in arthropods (Whelan and Goldman 2001; Stamatakis 2014). Node support was estimated from 500 bootstrap replicates. All phylogenetic tree images were created using the iTOL webserver (Letunic and Bork 2007). Trees were rooted according to available phylogenetic information; otherwise, we applied a midpoint rooting.

Results

Evaluation of the Best *De Novo* Assembly for *D. silvatica* Data

We obtained 441.8 million reads across the four experimental conditions, which dropped to 418.2 million (94.7%) after removing low-quality reads (table 1). We used the 98.4 million reads of the *REST* condition to evaluate the best *de novo* transcriptome assembler for our specific data. We found that among the assemblers using a single *k-mer* value of 25, SOAPdenovo-Trans and Trinity produced the largest number of contigs and the lowest N50 values (supplementary table S1,

Table 1
Summary of RNA-Seq Data Assembly and Annotation

	PALP	*LEG#1*	*LEG#234*	*REST*	Total	Total aligned
Total raw reads	114,986,182	118,017,386	104,967,256	103,865,040	441,835,864	441,835,864
GC (%)	41.41	41.38	41.39	41.55	41.43	41.43
Total qualified reads	108,490,938	112,102,210	99,231,056	98,380,850	418,205,054	418,205,054
Transcripts	130,908	144,442	147,737	149,796	236,283	214,969
Unigene transcripts (UT)	93,283	104,004	106,966	109,335	170,846	154,427
UT average length (in bp)	1,027	956	943	932	702	751
UT maximum length (in bp)	26,709	26,709	26,709	26,709	26,709	26,709
HK UT	1,134	1,134	1,131	1,133	1,136	1,136
CEG UT (CEG genes)	766 (456)	766 (457)	775 (457)	759 (457)	807 (457)	804 (457)
UT with GO annotation	20,481	21,799	22,332	23,471	29,879	28,157
UT with Interpro domain	21,436	22,735	23,293	24,435	30,886	29,168
UT with KEGG annotation	3,313	3,409	3,444	3,599	3,895	3,817
UT with functional annotation[a]	21,567	22,874	23,438	24,600	31,091	29,359
UT with genomic annotation[b]	27,043	28,922	29,645	31,236	41,046	38,317

[a]GO, Interpro or KEGG annotation.
[b]GO, Interpro, KEGG annotation or BLAST hit.

Supplementary Material online). The assembly based on Bridger provided the second best RSEM-EVAL score (after Trinity) but produced contigs with more positive BLAST hits against CEG and SwissProt proteins with a 100% alignment length filtering with an *E*-value of 10^{-3}. Increasing the *k*-mer size had a disparate effect on the number of contigs and on the N50, but the resulting assemblies were generally worse than those generated using *k-mer* 25 (based on RSEM-EVAL scores and positive BLAST hits). Only the assemblies obtained in Bridger and Trinity with a *k*-mer of 31 outperformed their respective assemblies with a *k*-mer of 25. However, the multiple *k-mer* strategies implemented in Trans-Abyss and Oases yielded very different assembly qualities. Trans-Abyss produced a highly fragmented transcriptome (i.e., with a large number of very short contigs) that was clearly outperformed by Oases using the clustered option. Nevertheless, Oases performed worse than Bridger and Trinity (*k*-mer = 31) in terms of RSEM-EVAL scores and positive BLAST hits. Hence, although the Trinity assembly provided a lower RSEM-EVAL score, Bridger produced a very similar value of this parameter while performing better based on all other calculated statistics. Consequently, we selected Bridger with a *k*-mer of 31 as the best strategy for the *de novo* assembly of *D. silvatica* data and used the transcriptome from this software for further analyses.

The initial assembly from Bridger (using the reads from the four conditions) was formed by 236,283 contigs (after removing contaminant sequences), which decreased to 170,846 putative nonredundant transcripts after the clustering of isoforms (table 1). We identified 807 transcripts with significant BLAST hits against 457 out of the 458 CEGs, 454 of them with alignment lengths longer than the 60% of CEG target gene (234 with 100% of this length; supplementary table S2, Supplementary Material online). These results clearly demonstrate the completeness of the assembled transcriptome.

Functional Annotation of the *D. silvatica* Transcriptome

As expected, arthropodDB received the most significant positive BLAST hits with an *E*-value of 10^{-3} when using *D. silvatica* transcripts as queries (supplementary table S3, Supplementary Material online). Of these hits, 85% corresponded to chelicerate subjects; the spiders *A. geniculata* and *S. mimosarum* and the scorpion *M. martensii* were the most represented species (supplementary fig. S1, Supplementary Material online).

The most frequent GO terms associated with the *D. silvatica* transcripts were "metabolic" and "cellular processes" (biological process), as well as "binding" and "catalytic activities" (molecular function) (supplementary fig. S2, Supplementary Material online). Moreover, we found that 3,895 (out of the 29,879 transcripts with an associated GO term) showed significant positive BLAST hits against 136 different entries of the KEGG database (supplementary table S4, Supplementary Material online), with Purine metabolism (2,030 transcripts), Thiamine metabolism (1,053 transcripts) and Biosynthesis of antibiotics (454 transcripts including, e.g., some spider glutamate synthases and dehydrogenases) being the most represented pathways.

Condition-Specific Gene Expression Analysis

Our comparative analysis identified 57,282 transcripts expressed in all four conditions (37.1%) (fig. 2). The number of condition-specific transcripts in *LEG#1*, *PALP* and *LEG#234* was rather similar (7,446, 6,000 and 8,605, respectively) and was much higher in *REST* (14,414), which is easily explained by the much larger number of tissues and physiological functions included in this condition. In the absence of separated biological replicates, we used the expression profile of HK genes to estimate the approximate dispersion of mean

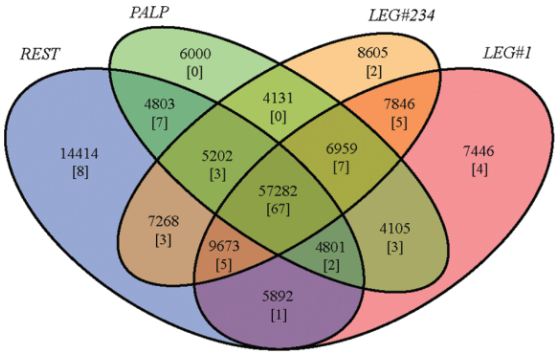


Fig. 2.—Venn diagram showing the total number of transcripts (154,427 transcripts) specifically expressed in each experimental condition and their intersections (red, orange, green and blue indicate *LEG#1*, *LEG#234*, *PALP* and *REST*, respectively). Numbers in brackets indicate putative chemosensory protein encoding transcripts (117 in total).

read counts across conditions to perform a rough differential expression analysis. The estimated dispersion across conditions of the 1,136 transcripts with significant positive BLAST hits to our set of HK genes (edgeR common dispersion value of 0.15) was used as the fold-change threshold for this analysis.

Our analyses show that *LEG#1* and *LEG#234* had rather similar transcriptomic profiles (supplementary fig. S3, Supplementary Material online). We found that only two transcripts were significantly overexpressed in *LEG#1* and the other two in *LEG#234*; taking these two conditions together, there were 27 overexpressed transcripts, none annotated as a chemosensory gene. These results contrast with those obtained in *PALP*, where 174 transcripts were significantly overexpressed. However, again, none of these transcripts encoded an annotated chemosensory function; they were enriched in signal peptide encoding sequences (Fisher's exact test, P value $= 2.63 \times 10^{-23}$), a feature characteristic of secreted proteins.

In addition, we found that the genes overexpressed in *PALP* were significantly enriched in the GO terms "metalloendopeptidase activity" (GO:0004222) and "proteolysis" (GO:0006508). In this specific tissue, these genes could be linked with the extra-oral digestion characteristic of these animals. However, we did not detect any GO term overrepresented in *LEG* or *REST*, and only 10 of the 27 genes significantly overexpressed in these structures had BLAST hits with an annotated sequence. Among these, we found genes encoding DNA-binding proteins, such as some transcription factors, hydrolases and proteins with transport activity.

Chemosensory Gene Families

To identify specific transcripts encoding chemosensory proteins in *D. silvatica*, we conducted additional exhaustive searches. We found many members of the *Gr*, *Ir*, *Npc2* and *Cd36-Snmp* families, as well as putative distant homologs of insect OBPs and one uncharacterized protein family that may be involved in chemosensory function in this spider. Nevertheless, we failed to find homologs of the *Csp* gene family, which is present in the genome of other chelicerates. As expected, the *D. silvatica* transcriptome did not encode insect OR proteins nor their vertebrate functional counterparts (supplementary table S5*A*, Supplementary Material online).

We identified 127 transcripts encoding IR/iGluR homologs (*Ir* transcripts), 57 exhibiting the specific domain signature of the ionotropic glutamate receptors (IPR001320). Some of these transcripts encoded some of the characteristic domains of the IR/iGluR proteins, such as the amino terminal (ATD-domain; PF01094), the ligand binding (LBD-domain; PF10613) and the ligand channel (LCD-domain; PF00060) (supplementary fig. S4, Supplementary Material online; see also Croset et al. 2010). Indeed, nine of them encoded all three domains, thus forming the typical complete iGluR structure, while 23 only had the two ligand-binding domains.

To understand the evolutionary diversification of the *Ir/iGluR* gene family in chelicerates, we carried out a protein domain-specific phylogenetic analysis. We used the information exclusively from the LCD domain because it is shared by all characterized arthropod IR/iGluR. For the analysis, we built an amino acid-based MSA including all *D. silvatica* transcript-coding LCD domains (70 transcripts) along with all reported sequences of this domain from *D. melanogaster*, *D. pulex*, *S. maritima*, *I. scapularis*, and *S. mimosarum* (i.e., in order to avoid large and unreadable trees, we included only one species per main arthropod lineage except for chelicerates, which were represented by a tick and a well annotated spider). We found that *D. silvatica* had representatives of all major IR/iGluR subfamilies, namely the AMPA, Kainate, NMDA (canonical iGluR subfamilies having all three Pfam domains), the two IR major subfamilies, the so called "conserved" IRs (encompassing the IR25a/IR8a members; having all three PFAM domains), and the remaining IR members (IR subfamily having only the LBD and LCD domains and that in *Drosophila* includes members with chemosensory function encompassing the so called "divergent" and the "antennal" IRs). In total, we identified 26 transcripts encoding canonical iGluR proteins plus another 44 encoding IRs (fig. 3 and supplementary fig. S5, Supplementary Material online), including a putative homolog of the highly conserved family of IR25a/IR8a proteins (transcript Dsil31989). Noticeably, this transcript is significantly overexpressed in *LEG#1* with respect to *REST* (~10 times more expression $-\log FC = 4$; $P < 0.01$ after FDR), although it also shows 2 and 4 times more FPKM values with respect to *PALP* and *LEG#234*, respectively (supplementary table S5*B*, Supplementary Material online).

Our phylogenetic analysis uncovered a set of *D. silvatica* transcripts phylogenetically related to some *D. melanogaster* antennal IRs, such as the IR21a (Dsil32714), the IR40a (Dsil150464) and the IR93a (Dsil55987, Dsil29850 and Dsil48134) proteins. These transcripts, however, did not show any clear differential expression pattern in *LEG#1* or

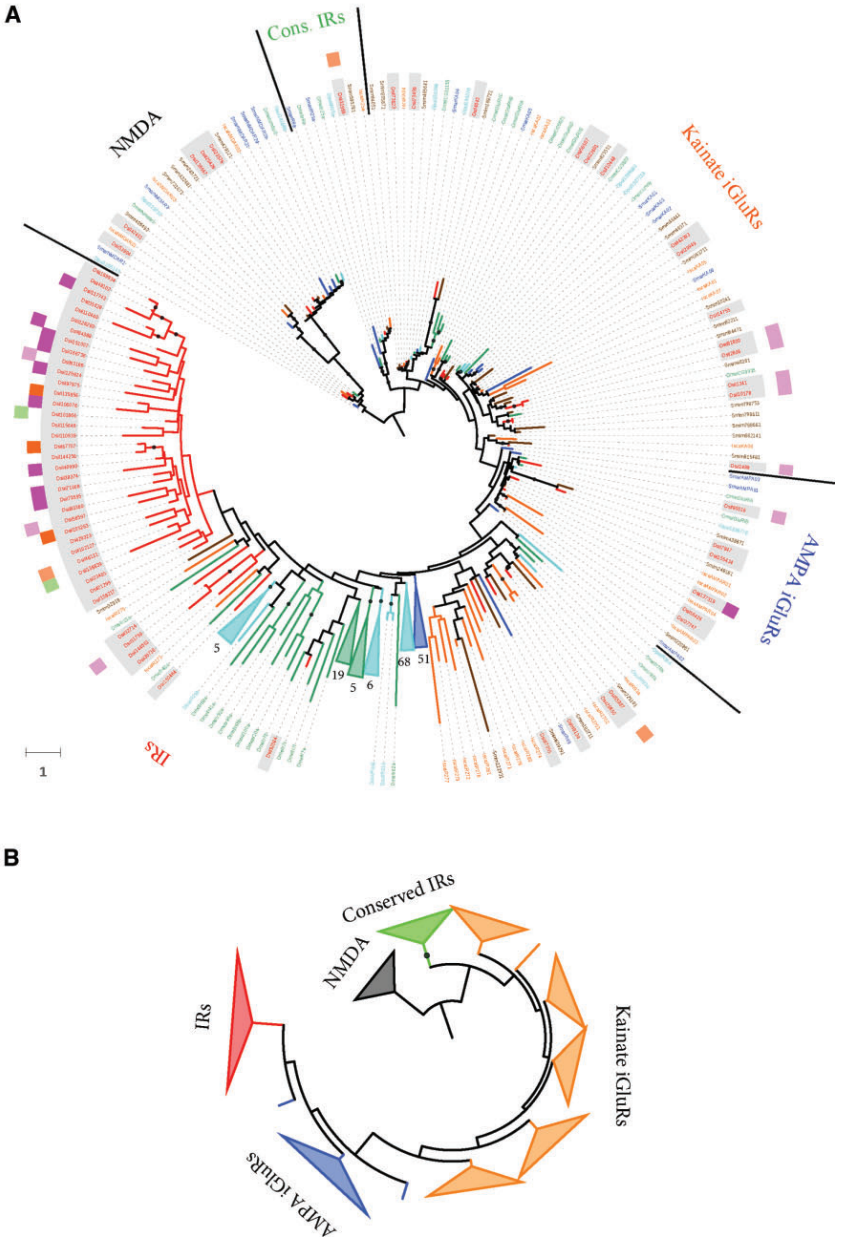


Fig. 3.—Maximum likelihood phylogenetic tree of the IR/iGluR proteins across arthropods. The tree is based on the MSA of the LCD domain (PF00060). (*A*) Sequences of *Drosophila melanogaster*, *Daphnia pulex*, *Strigamia maritima, Ixodes scapularis*, *Stegodyphus mimosarum* and *Dysdera silvatica* are depicted in green, light blue, dark blue, orange, brown and red, respectively. Additionally, the translation of the *D. silvatica* transcripts are shadowed in grey. Nodes with bootstrap support values >75% are shown as solid circles. Nodes with five or more sequences from the same species were collapsed; the actual number of collapsed branches is indicated in each case. The two surrounding circles provide information regarding the expression pattern of some *D. silvatica* genes. The most external circle indicates the genes specifically expressed in palps (PALP; in green), legs (both *LEG#1* and *LEG#234;* in pink) and palps and legs (*PALP*, *LEG#1* and *LEG#234;* in orange). The inner circle shows the genes overexpressed in these conditions using the same color codes but with two color intensities, one more intense color for overexpression levels >5× over *REST* and another lighter color for 2–5× overexpression values. The branch length scale is in numbers of amino acid substitutions per amino acid position. (*B*) Simplified phylogenetic tree highlighting the main *Ir* sub-families.

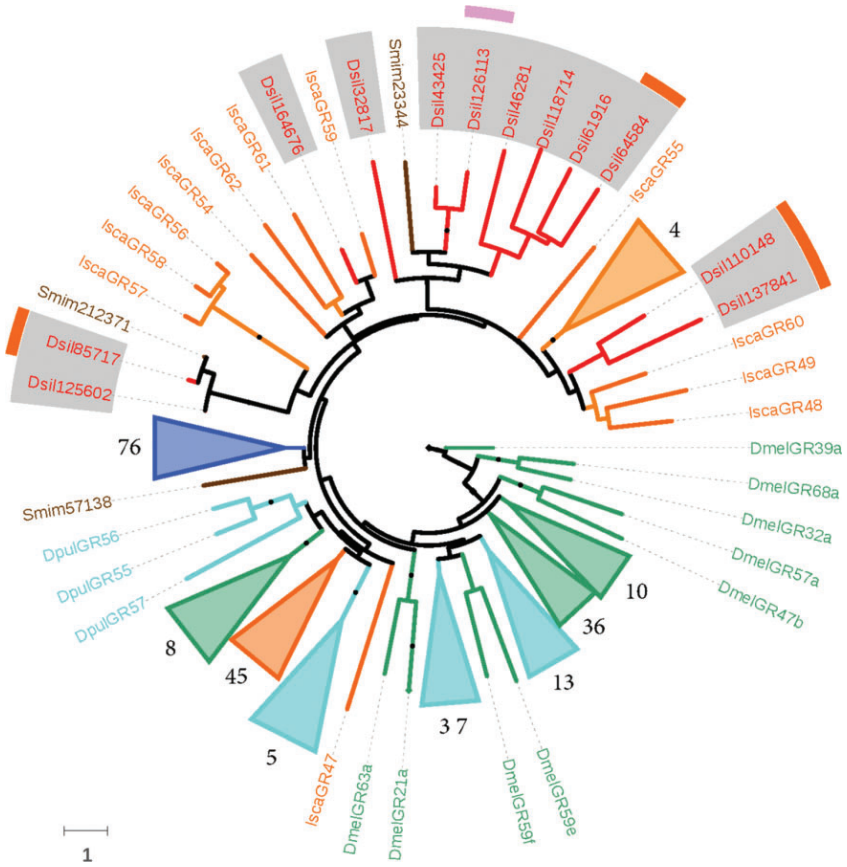


Fig. 4.—Maximum likelihood phylogenetic tree of the GR proteins across arthropods. Species names, node support features and surrounding circles are colored as in figure 3.

PALP, while two of them were clearly overexpressed in *REST*. Moreover, similarly to what occurs in other arthropods, many nonconserved IRs formed a species-specific monophyletic clade (33 transcripts). Interestingly, 11 of these receptors were condition specific, and 8 were overexpressed (or showed at least 2 times more FPKMs) in the examined appendages (i.e., *LEG#1*, *LEG#234* and *PALP* with respect to *REST*). Actually, *LEG#1* was the expression condition with the highest number of different nonconserved *Ir* transcripts; only 14 of the 43 nonconserved *Ir* members were not expressed in this appendage (supplementary table S5*B*, Supplementary Material online). Overall, the expression level of *Ir*s (including conserved *Ir*s) was lower than that of the *iGluR* transcripts.

We further identified 12 transcripts encoding GR proteins (*Gr* transcripts), although only four of them had one of the two specific InterPro signatures that characterize this family (7m_7, IPR013604 and Trehalose receptor, IPR009318). In addition, these 12 *Gr* transcripts were phylogenetically related to members of this family characterized in the spider *S. mimosarum* and in the deer tick *I. scapularis* (fig. 4 and supplementary fig. S6, Supplementary Material online). The expression levels of *D. silvatica Gr* genes were considerably low compared both with the overall expression levels and with the expression levels of other chemosensory families (supplementary table S5*C*, Supplementary Material online). Interestingly, only two *Gr* transcripts were condition specific (Dsil61916 and Dsil164676 in *REST*), and the other two were specifically expressed in both *LEG#1* and *PALP* (Dsil110148 and Dsil137841). The remaining *Gr* transcripts showed variable gene expression profiles across conditions, with some genes having a wide expression pattern and others being more restricted to particular conditions (supplementary table S5*C*, Supplementary Material online).

Our BLAST- and profile-based results revealed significant similarities between three spider transcripts and some insect

members of the *Obp* family (with *E*-values between 10^{-3} and 10^{-5}). The primary amino acid sequence and the cysteine pattern of the encoded proteins (hereafter designated OBP-like proteins) resembled those of OBPs and, one of them (Dsil553) showed a match to the PBP_GOBP InterPro domain (PBP_GOBP; IPR006170), uncovering a protein domain with folding features similar to those found in some insect OBPs. Using the three OBP-like sequences identified in the transcriptome of *D. silvatica* as a query in a BLASTp search against the NCBI-nr database (*E*-value of 10^{-3}), we detected six additional members of this novel family in the genomes of *S. mimosarum*, *I. scapularis* and *S. maritima* (two copies in each genome; fig. 5) but none in the annotated proteomes of crustaceans. The MSA of the nine copies identified in noninsect species and all characterized members of the *Obp* family in *D. melanogaster* and *Anopheles gambiae* would suggest that the *Obp*-like family is distantly related to the Minus-C *Obp* subfamily. Despite the particularly low sequence similarity and the large differences in protein length (not only between OBP-like and insect OBPs but also among OBP members), three different MSAs built with different alignment algorithms, i.e., MAFFT with the option L-INS-I (Katoh and Standley 2013), PROMAL3D (Pei et al. 2008) and PSI-coffee (Chang et al. 2012), yielded exactly the same pattern of cysteine homology in the region of the GOBP-PBP domain. Accordingly, with these MSAs, OBP-like proteins lacked the same two structurally relevant cysteines as insect Minus-C OBPs (except the *S. maritima* protein Smar010094 in the MAFFT alignment; supplementary fig. S7, Supplementary Material online). These results, however, must be taken with caution due to the fact that some OBP-like as well as several insect OBPs show large amino or carboxy terminal domains outside the conserved OBP domain, some of them including extra cysteines. If these cysteines are not correctly aligned in their true homologous positions, the interpretation of the cysteine pattern of OBP-like proteins could be erroneous.

We built a 3D protein model of both the conceptual translation of one of the *Obp*-like transcripts identified in *D. silvatica* (Dsil553) and of the *S. maritima* protein Smar010094 using the Phyre2 web portal (Kelley et al. 2015). As expected, the predicted models showed a globular structure very similar to that found in insect OBPs (fig. 6). In fact, the top 10 structural templates identified by the software and, therefore, the one selected for the final modeling (*A. gambiae* proteins OBP20 and OBP4 for Dsil553 and Smar010094, respectively) were insect OBPs. In addition, the models showed a high confidence in the region corresponding to the GOBP-PBP domain (56% and 59% of the query sequences were modeled with 89.2% and 81.6% confidence by the single highest scoring template, respectively). Remarkably, the amino acid alignment between Smar010094 and OBP4, used as a guide by Pyre2 for building the 3D model of this *S. maritima* OBP-like protein, coincided with the PROMAL3D and Psi-Coffee alignments but not with the MAFFT one (see above). Hence, we hypothesize that, given the wide expression of spider OBP-like across the four experimental conditions (supplementary table S5*D*, Supplementary Material online), these proteins, similar to those in insects, might be carriers of small soluble molecules acting in one or more physiological processes without ruling out a putative role in chemosensation.

We also identified 11 transcripts encoding putative NPC2 proteins, all of them having the characteristic IPR domain (MD-2-related lipid-recognition domain; IPR003172). The phylogenetic tree reconstructed from the MSA including these and other arthropod members of this family (including the members expressed in the antenna of *A. mellifera* and *Camponotus japonicus* (Ishida et al. 2014; Pelosi et al. 2014; fig. 7) uncovered a less dynamic gene family with neither large species-specific clades nor long branches. Nevertheless, internal node support was low and the precise phylogenetic relationships among arthropod NPC2s could not be determined with confidence. It is worth nothing, however, that this family underwent a moderate expansion in arthropods because it seems to be only one copy in both *C. elegans* and vertebrates. Only one putative *D. silvatica Npc2* transcript was *LEG#1* specific (Dsil113431), while two of them showed 11–4 times more FPKM in *PALP* (Dsil16636 and Dsil93094) and two others had 7 and 2 times more FPKM in *LEG#1* and *PALP* than in *REST* (Dsil56450 and Dsil793), respectively (supplementary table S5*E*, Supplementary Material online).

Finally, we identified 13 transcripts related to the *Cd36-Snmp* family, with 12 of them having the corresponding InterPro domain signature (CD36 antigen; IPR002159). Our phylogenetic analysis showed that *D. silvatica* had representatives of the three SNMP protein groups (Nichols and Vogt 2008; fig. 8), which would indicate that the origin of these subfamilies predated the diversification of the four major extant arthropod lineages. All four *D. silvatica Snmp* transcripts were similarly expressed in the four studied conditions, which would suggest either a nonchemosensory specific function of these proteins in spiders or a global general function within the chemosensory system (supplementary table S5*F*, Supplementary Material online).

A Novel Candidate Chemosensory Gene Family in Spiders

Furthermore, we conducted an exhaustive search on the 174 transcripts overexpressed in *LEG#1* and *PALP* to try to identify putative novel, previously uncharacterized spider olfactory chemosensory families. For this, we first searched for gene families (groups of 4 or more similar sequences) by performing a clustering analysis of the 174 transcripts with CD-HIT (Fu et al. 2012); then, we searched for the presence of a signal peptide or for signs of trans-membrane helices in the identified families. We found one family (with five copies) in which one member had the molecular hallmark of a signal peptide; the absence of such a mark in the other four members could be due to the failure to detect full-length transcripts in these

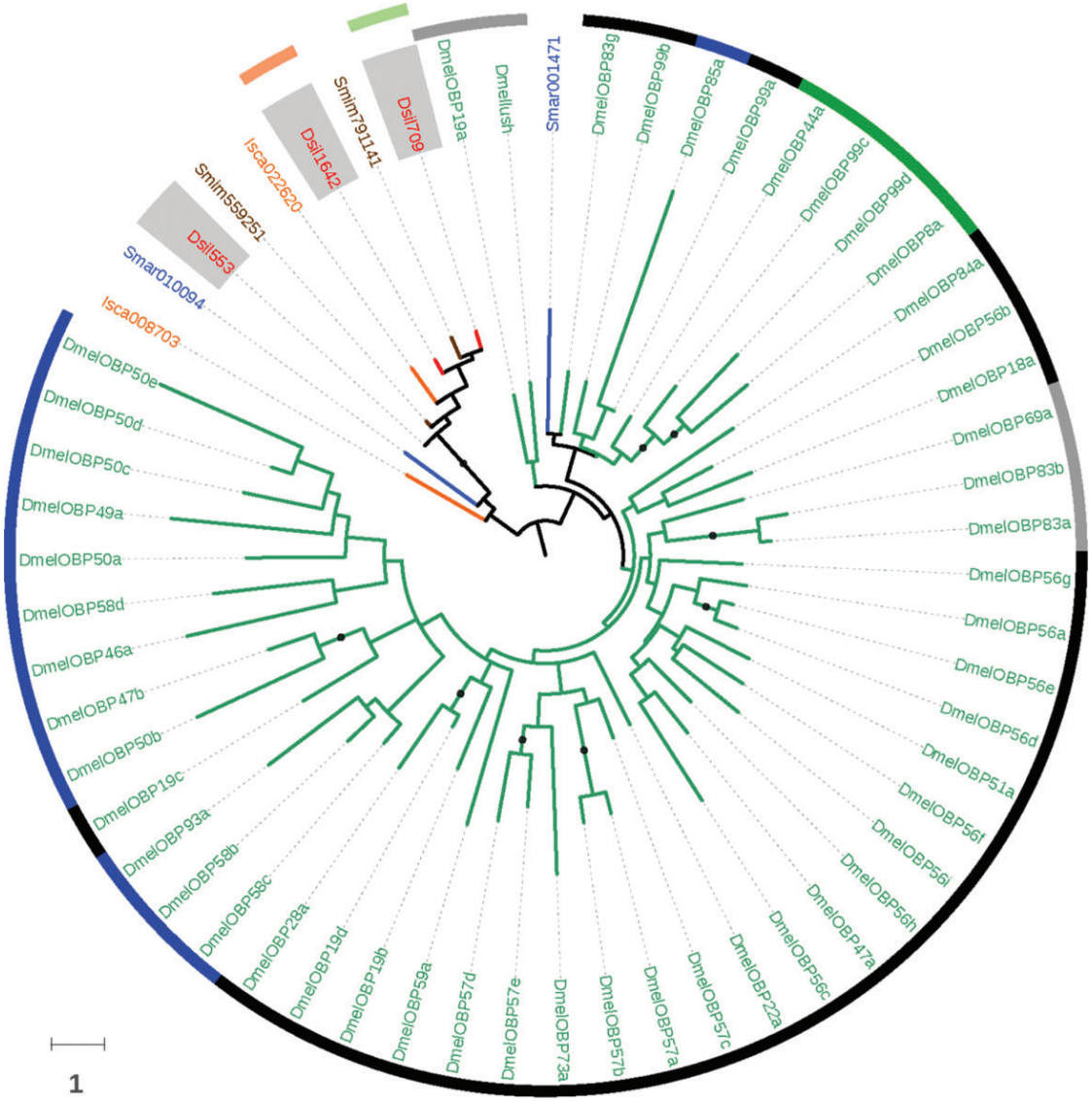


Fig. 5.—Maximum likelihood phylogenetic relationships of spider OBP-like and insect OBP proteins. Species names, node support features and surrounding circles are colored as in figure 3. The inner circle labels the previously defined OBP phylogenetic subfamilies (Classic, Minus-C, Plus-C and ABPII in black, green, blue and grey, respectively).

members (supplementary table S5*G*, Supplementary Material online). Using these five sequences as queries in a BLAST search against the complete *D. silvatica* transcriptome, we further detected seven more members of this family. New BLAST searches using all 12 proteins as queries identified homologous copies in other spiders but not in the genomes of either other chelicerate lineages or nonchelicerate species. A preliminary phylogenetic analysis including all new identified sequences indicated that this family (supplementary fig. S8, Supplementary Material online) was highly dynamic, with several species-specific clades of CCPs (one of them including all *D. silvatica* copies) and no clear orthologous relationships across spiders. All these spider sequences, however, were annotated as uncharacterized proteins in these genomes.

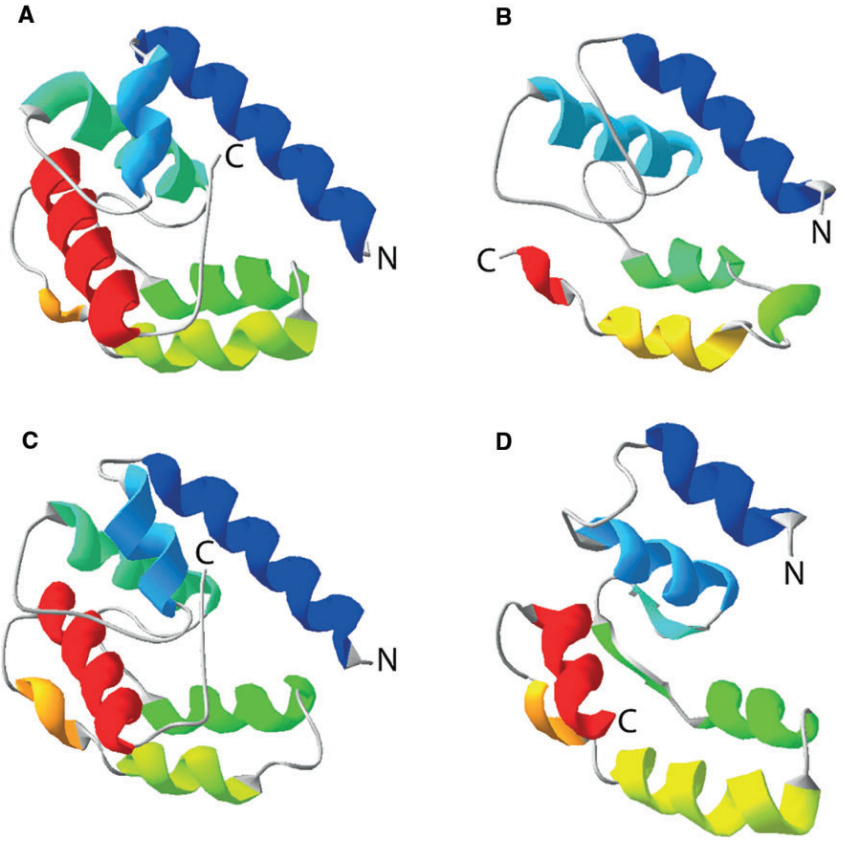


Fig. 6.—Predicted 3D structure of two OBP-like proteins. (*A*) Structure of *Anopheles gambiae* OBP20 (PDB 3V2L). (*B*) Structure of *A. gambiae* OBP4 (PDB 3Q8I). (*C*) 3D model of the protein encoded by the transcript Dsil553. (*D*) Predicted 3D model of the *Strigamia maritima* Smar010094 protein. PBD files were viewed and manipulated in Swiss-PdbViewer version 4.1 (Guex and Peitsch 1997).

The MSA of the members of this novel family revealed a conserved cysteine pattern similar to that observed in insect OBPs and CSPs. However, unlike the OBP-like proteins, we could not obtain a reliable 3D protein model of a member of this family in the Phyre2 webserver. The server was unable to identify reasonable templates with large alignment coverage for the modeling (all templates with confidences >15 had an alignment coverage <7%). We then used I-TASSER suite (Yang et al. 2015) to try to find template proteins of similar folds as our *D. silvatica* queries. Although two of the identified threading templates were OBPs, some artificially designed proteins were also included in the modeling, generating five highly heterogeneous folding models, most of them with unacceptable C-scores. Nevertheless, some of the estimated folding models showed a compact global structure that, along with the presence of a signal peptide and the gene expression data, would suggest that the members this novel gene family could also acts as carriers of small soluble molecules, as insect OBP do (hereinafter we will refer to this novel family as the *Ccp* gene family for candidate carrier protein family).

Discussion

A High-Quality *De Novo* Assembly of the *D. silvatica* Transcriptome

The key step to obtain a high-quality transcriptome is selecting the best *de novo* assembly strategy and software. Nevertheless, because most assemblers have been developed for specific NGS platforms or tested using reduced data sets with limited taxonomic coverage, it is very difficult to predict their performance with disparate datasets (Martin and Wang 2011). Obtaining a high-quality transcriptome depends on factors such as the organism (which determines DNA

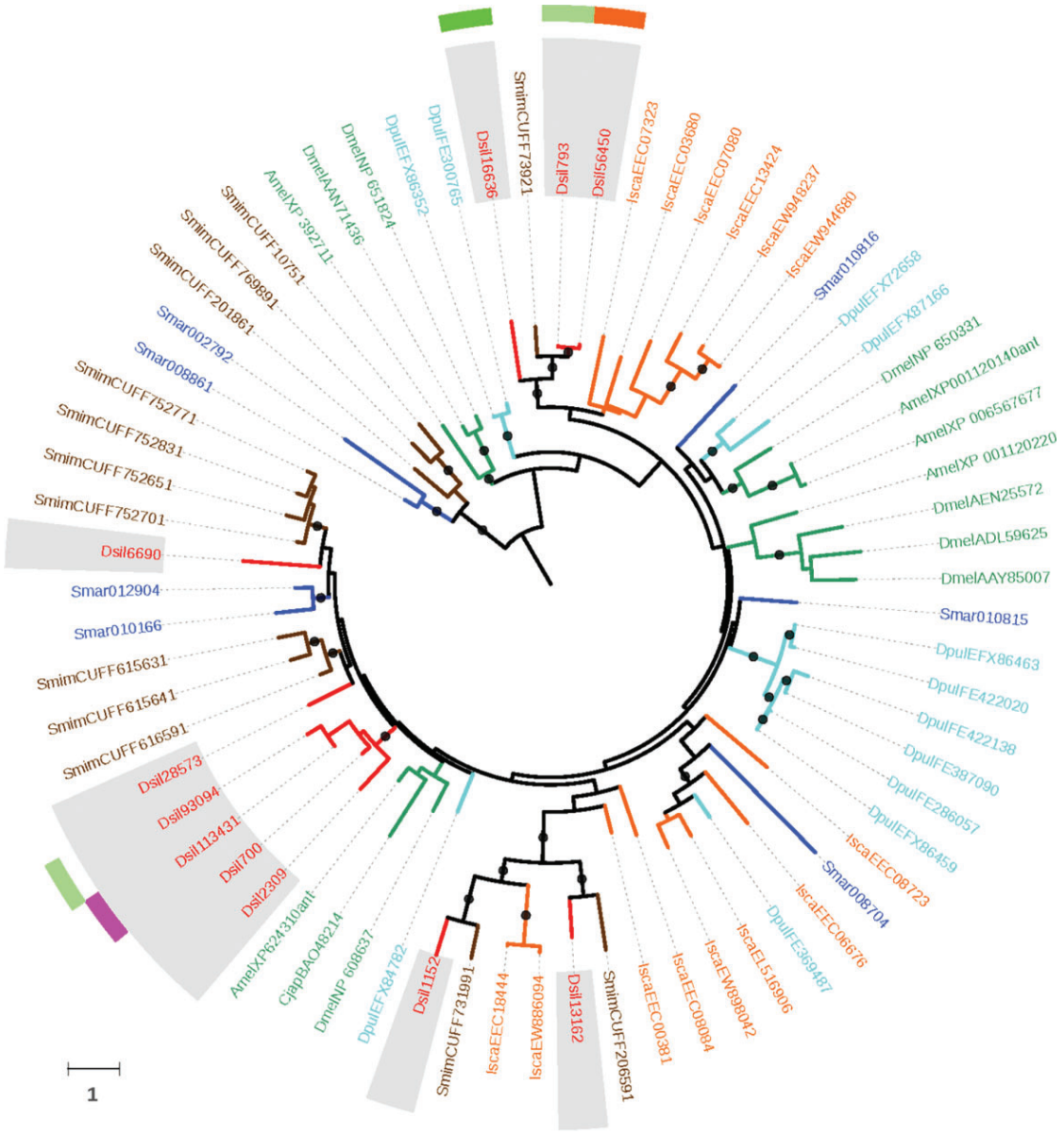


Fig. 7.—Maximum likelihood phylogenetic tree of the NPC2 proteins across arthropods. Species names, node support features and surrounding circles are colored as in figure 3. Sequences from *Apis mellifera* and *Camponotus japonicus* are colored in green.

complexity and heterozygosity levels), the read length and the sequencing depth. The best approach to determine the quality of different assemblies is to evaluate their accuracy (especially their completeness) in the context of a well-annotated, closely related reference genome (Marchant et al. 2015). Unfortunately, functionally annotated genomes of close relatives are usually not available for nonmodel organisms. In our case, the phylogenetically closest species with genome information, the spider *L. reclusa*, diverged from *D. silvatica* ~200 Ma (Binford et al. 2008), which prevented any reliable evaluation. To circumvent this limitation, we used a combination of two strategies to evaluate the performance of five competing assemblers, one based on information of the transcriptome completeness (using CEG and SwissProt databases as subjects)

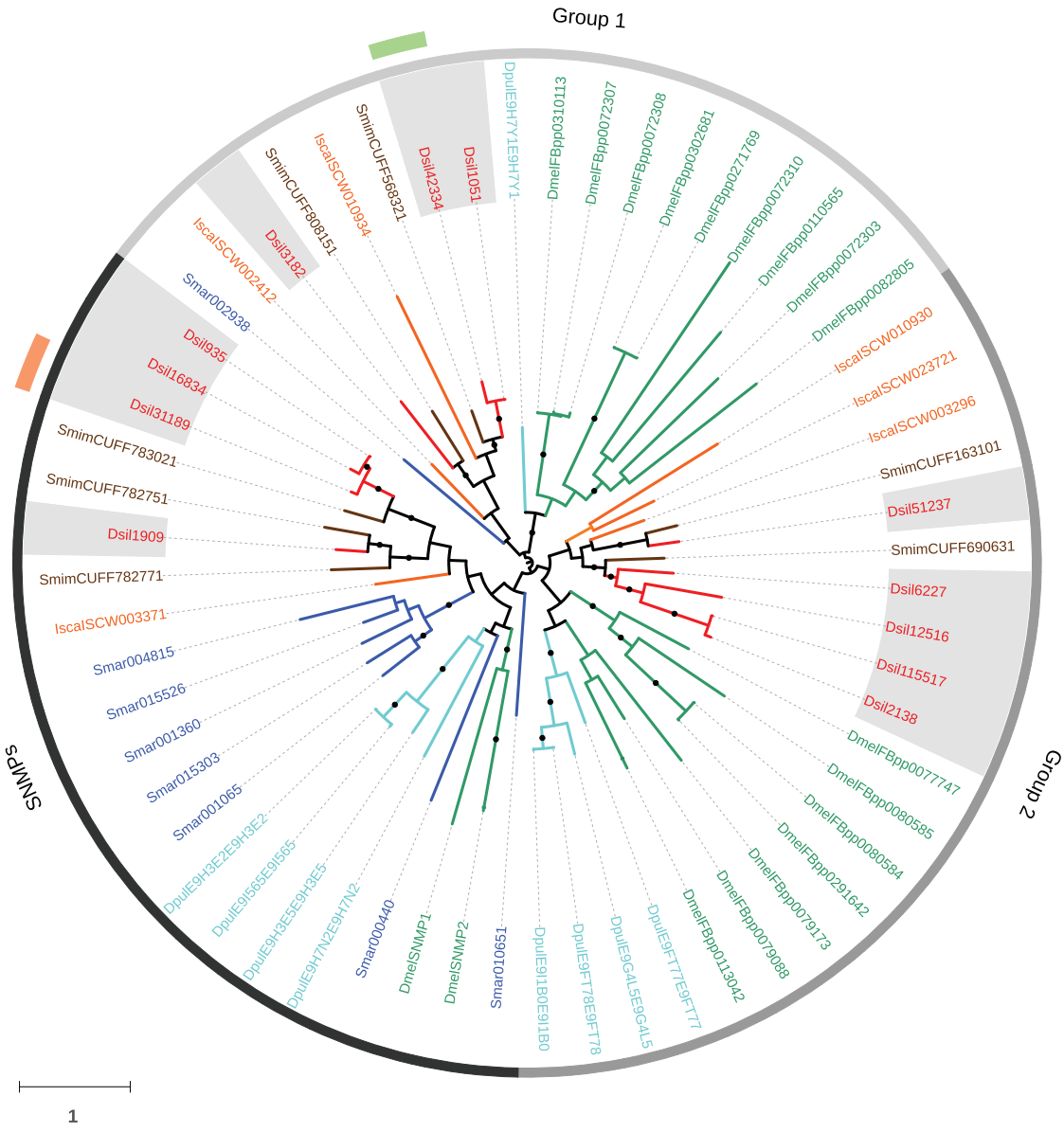


Fig. 8.—Maximum likelihood phylogenetic tree of CD36-SNMP proteins across arthropods. Species names, node support features and surrounding circles are colored as in figure 3. The inner circle shows the different subfamilies.

and the other based on some statistics measuring the assembly quality (Li et al. 2014). Using this combined strategy and after evaluating 11 assembly scenarios, we were able to obtain a high-quality assembly that probably covers most of the *D. silvatica* transcriptome and that has a large proportion of full-length transcripts.

A Comprehensive Annotated Transcriptome That Uncovers a Surprising Gene Loss in Chelicerata

The functional annotation of a *de novo* assembled transcriptome from a nonmodel organism is a daunting task, being usually slow and computationally intensive. The large number

of query sequences (transcripts) make similarity- and profile-based searches against general big databases, such as the NCBI-nr, very problematic, especially when using the free version of some software suites (e.g., BLAST2GO). Here, we used GHOSTZ instead of BLAST when searching against NCBI-nr, considerably reducing the computational time of the functional annotation step in >100 times, which is a relevant feature when testing assemblers in a comparative framework (i.e., a large number of independent annotations). Moreover, to increase the sensibility of the searches and reduce the computation time, we included only a representative set of phylogenetically close species to *D. silvatica* to build our specific databases (some annotated proteins are not yet available in NCBI-nr). Finally, we largely reduced the running time of the InterProScan searches (~10 times) by using only the Pfam database (Finn et al. 2014) as a query without a substantial loss in the number of positive hits.

Despite the exhaustive annotation process, a high number of *D. silvatica* transcripts (81.8%) could not be functionally annotated. These percentages, however, are commonly obtained in RNA-Seq studies and can be attributable to different causes. First, nonannotated transcripts are significantly shorter than annotated ones (P value = 2.2×10^{-16}), suggesting that many nonannotated transcripts are actually assembly errors or small fragments lacking any detectable protein domain signature (supplementary table S6, Supplementary Material online). Second, a fraction of these unannotated sequences could correspond to noncoding RNAs. Finally, the modest annotation of the genome of *L. reclusa*, the closest available relative to *D. silvatica*, could considerably reduce the success of our searches. In fact, an important number of *D. silvatica* transcripts without functional annotation (9,955 sequences) encoded proteins tagged as uncharacterized in the genome of *L. reclusa*.

A relevant result of our functional annotation of the *D. silvatica* transcriptome is the absence of a transcript encoding a Trehalase (KOG0602), the only gene of the CEG database not identified in the *D. silvatica* transcriptome. This gene seems to also be absent in the genomes of other chelicerates because we failed to detect it even using powerful profile-based approaches. Intriguingly, this protein is essential for insects (Shukla et al. 2015) not only because of its function as hydrolase but also for its involvement in the development of the optic lobe (Chen et al. 2014). Given that this gene is certainly present in the genome of all other major arthropod lineages as well as in the tardigrade *H. dujardini* and the nematode *C. elegans*, the most likely explanation for its absence is specific gene loss in the ancestor of chelicerates. The apparent absence of this gene in this lineage is interesting and clearly demands further investigation. The study of this gene loss, jointly with that of the set of uncharacterized proteins found in the *D. silvatica* transcriptome, will provide new insight into some important biological processes specific to chelicerates.

The Chemosensory Transcriptome of *D. silvatica*

Unlike our previous survey in the mygalomorph species *M. calpeiana* (Frías-López et al. 2015), here we identified several transcripts encoding members of chemosensory gene families in the four studied body parts, albeit with low expression levels. The different levels of success of the two studies could be related to the much higher sequencing depth (i.e., >10 Gbp sequenced per condition) of the *D. silvatica* RNA-Seq experiment.

As expected from the genome annotations of some chelicerate species, the transcriptome of *D. silvatica* did not contain genes related to the vertebrate chemoreceptors or odorant-binding protein families, ruling out the possibility that these or other similar families play any role in spider chemosensation. Similarly, we failed to detect members of the insect *Or* gene family, adding further evidence of the complete absence of this family in all arthropod lineages other than winged insects (Missbach et al. 2015). Moreover, despite the presence of members of the *Csp* gene family in some chelicerates and myriapods (Chipman et al. 2014; Qu et al. 2015; Gulia-Nuss et al. 2016), we did not identify any transcript encoding a protein with significant similarity to this family in *D. silvatica*. Although this negative result might be explained by sequencing or assembly limitations, *Csp* genes are also absent in all other spider genomes available in public repositories. We postulate that this gene family could have been lost early in the diversification of arachnids.

Candidate Spider Chemoreceptor Gene Families

Here, we identified a maximum of 12 transcripts encoding GR proteins (i.e., some of them may form part of the same gene), a number that may seem surprisingly small in comparison with the large number of *Gr* genes identified in the tick *I. scapularis* (62), the myriapod *S. maritima* (77) and the water flea *D. pulex* (58) genomes, for example. Nevertheless, given the underrepresentation of the chemosensitive hairs with respect to the total amount of tissue examined in each specific transcriptome, the identification and comprehensive annotation of the complete set of *Gr* genes are quite challenging in standard RNA-Seq studies (Zhang et al. 2014). In addition, some *Gr* genes do not necessarily have to be expressed at the precise moment (i.e., developmental stage or environmental condition) of the experiment (this can also be applied to all other chemosensory families). Therefore, the *D. silvatica* genome likely encodes many more members of this family, and the 12 transcripts found in this study are only a first preliminary subset of the gustatory repertoire of this spider. These molecules seem to be expressed across different spider body parts and some show specific expression in particular appendages, with groups of copies broadly expressed, other groups that are never found in particular appendages and others that show an opposite pattern of specificity. This combinatorial manner of expression is similar to that the described for the

*Gr*s in *Drosophila*, which would suggest analogous gustatory coding mechanisms in these two arthropods (Depetris-Chauvin et al. 2015; Joseph and Carlson 2015). The two phylogenetically related *Gr* genes specifically expressed in *LEG#1* and *PALP* (Dsil110148 and Dsil137841) could be involved in the detection of some ecologically relevant signals, for example, partial pressure of CO_2, in a similar way as some insect *Gr* specifically expressed in *D. melanogaster* antenna, although the proteins encoded by spider and insect transcripts are phylogenetically unrelated. In fact, all *Gr* transcripts detected in the *D. silvatica* transcriptome (including *LEG#1* and *PALP* specific sequences) are members of a monophyletic group of chelicerate receptors for which we have no functional information. However, some *Gr* transcripts are also overexpressed or even exclusively expressed in the transcriptome of *REST*. The encoded proteins might participate in other, nonchemosensory physiological functions, as has also been observed in insects (Joseph and Carlson 2015). Even so, we cannot rule out that they actually act as chemoreceptors in other body structures, apart from palps and legs, such as in the mouthparts, which are included in *REST* transcriptome.

Unlike *Gr*s, we have detected in *D. silvatica* a substantial number of sequences (127) encoding putative *Ir* transcripts, including a putative homolog of the conserved *Ir* subfamily *Ir25a/Ir8a* (Dsil31989). The phylogenetic analysis of the members of this family in arthropods clearly reflects the effect of the long-term birth-and-death process acting on most members of this family. Remarkably, this effect is almost unnoticeable in iGluR and in conserved IRs proteins, ratifying the marked differences in gene turnover rates between subfamilies. This highly dynamic evolution of nonconserved IR jointly with that reported for other proteins associated with contact chemoreception has been suggested as a proof of the high adaptive potential of the molecular components of the gustatory system in arthropods (see Torres-Oliva et al. 2016, and references therein). Interestingly, some of the 10 nonconserved IRs not included in the *D. silvatica*-specific clade are phylogenetically related to some *D. melanogaster* antennal IRs, including one member that presumably plays an important role in thermosensation (IR21a). Nevertheless, the expression profiles of these five transcripts do not provide clues regarding their possible role in spider chemosensation (i.e., they do not show any specific gene expression pattern across conditions). Although the putative spider homolog of the *Ir25a/Ir8a* subfamily is also expressed in all four conditions, it is much more abundant in *PALP*, *LEG#1* and *LEG#234*, and even significantly overexpressed in *LEG#1* with respect to *REST*. The IR25a and IR28a proteins are widely expressed in *Drosophila* olfactory sensilla (and in olfactory organs of other arthropods; Croset et al. 2010) and have been involved in the trafficking to the membrane of the other IR and in a co-receptor function of food-derived chemicals and humidity and temperature preferences. Thus, our results indicate that the first pair of legs of spiders could be relevant for the detection of amines and/or aldehydes as well as for determining favorable ranges of certain environmental variables (Silbering et al. 2011; Min et al. 2013; Enjin et al. 2016). Finally, and similar to that observed in for *Gr* transcripts, some members of the nonconserved *Ir* subfamily are also detected in *REST*, further supporting their involvement in other nonchemosensory functions or, alternatively, the presence of chemosensory structures in body parts other than legs or palps.

Evolution of the IR Family in Arthropods

Since our phylogenetic analysis includes highly diverged sequences, we applied for first time domain-specific HMM profiles to guide the MSA of chemosensory families. This strategy has been especially useful for the *Ir/iGluR* families, exploiting the evolutionary information of the conserved ligand channel domain (LCD domain) clearly shared by all known members. The inferred tree mirrors the same focal phylogenetic groups obtained in previous works (Croset et al. 2010). Most tree reconstructions show that (1) the Kainate and AMPA proteins are closely related, and AMPA likely a derived linage, (2) the subfamily of the conserved IRs is the sister group of these Kainate/AMPA receptors, and (3) NMDA sequences represent the first offshoot. However, there are some important differences between the present study and findings regarding the putative origin of the nonconserved IRs. This group of IRs, which forms a supported monophyletic group in all tree reconstructions, is more closely related to non-NMDA receptors than to the remaining iGluRs in our tree, which could indicate that they originated from a Kainate- or AMPA-like receptor. Nevertheless, the poor support of some internal nodes, probably due to alignment artifacts caused by the diverse domain structure of *Ir/iGluR* families, precludes making definitive conclusions about the origin of these highly divergent receptors.

Novel Classes of Candidate Transport Proteins in Chelicerates

Pelosi et al. (Pelosi et al. 2014) proposed that some members of the *Npc2* family might be involved in the transport and solubilization of semiochemicals in noninsect arthropods, constituting an alternative to the insect OBP and CSP proteins involved in the peripheral events of olfaction. Here, we show that the spider *D. silvatica* has a similar repertoire of *Npc2* genes to that found in other surveyed arthropods, which seems to be expanded in arachnids. We identified one member of this family specifically expressed in *LEG#1* that may be a good candidate to participate in odor detection in spiders; this transcript, however, showed a relatively low expression level, in contrast to the very high expression levels observed in insect *Obp* and *Csp* genes. Although the remaining members of the *Npc2* family might also have other chemoreceptor functions in *Dysdera*, most of them probably perform other important physiological functions, such as

cholesterol lipid binding and transport, which is the known function of these proteins in vertebrates (Storch and Xu 2009).

One unexpected and remarkable result is the expression in *D. silvatica* of at least three genes encoding proteins with a secondary structure, conserved cysteine pattern (revealed in the MSAs that include insect OBPs and characteristic of the Minus-C subfamily) and predicted folding similar to that of insect OBPs. In fact, our searches using these newly identified OBP-like proteins as a query revealed that chelicerates and myriapods, but not crustacean or insects, have some copies of this family. In the absence of confirmation by functional experiments and structural data, these results suggest that the *Obp* superfamily was already present in the arthropod ancestor. We cannot confirm whether putative ancestors were actually members of the Minus-C subfamily because this group of proteins is polyphyletic in the OBP tree (Vieira and Rozas 2011). Nevertheless, the fact that chelicerate and myriapod genomes only carry Minus-C *Obp* genes supports them as the ancestral arthropod *Obp*. In *D. melanogaster*, the Minus-C *Obp*s are highly expressed in several tissues other than the head, including adult carcass, testis, male accessory glands, spermatheca and some larval tissues (data from FlyAtlas project; Chintapalli et al. 2007). The wide expression levels of OBP-like genes across all four experimental conditions, together with their low gene turnover rates in chelicerates, also indicate essential and multiple functional roles of these putative small soluble carriers, regardless of their possible function in the chemosensory system.

Lastly, the newly identified *Ccp* family encodes a protein with a clear signal peptide that shows similar folding characteristics to those of insect OBPs. Interestingly, half of their members are overexpressed in the proposed spider olfactory organs. In this case, however, we only detected homologous copies in the genomes of arachnids, where the products are annotated as uncharacterized proteins. Thus, both the NPC2 copy and the proteins encoded by the *Ccp* family are good candidate chelicerate counterparts of the insect OBP and the CSP proteins, and their specific function clearly deserves further exploration.

In this study, we report the first comprehensive comparative transcriptomic analysis across different body structures of a spider, including those that most likely carry the chemosensory hairs. Our results indicate that, as in other noninsect arthropods, gustatory and ionotropic receptor families are the best candidate peripheral chemoreceptors in chelicerates. Additionally, we found some noteworthy differences in the specific pattern of gene expression of the members of these chemosensory families across different body structures, some of them involving the putative olfactory system-containing organs, which can indicate some specialization of chemosensory structures across the body of *D. silvatica*. In addition, we identified a protein family in chelicerates that seems to be distantly related to the insect *Obp* family and have characterized a new gene family of small secreted soluble proteins analogous to the insect OBPs or CSPs that could act as molecular carriers in this species. Finally, we provide the first complete and functionally annotated transcriptome of a polyphagous predator species of the genus *Dysdera*, which will provide valuable information for further studies on this group, and a list of candidate genes suitable for further functional dissection. Our results will help better establish the specific role and sensory modality of each of these new identified genes and gene families in spiders while providing new insight into the origin and evolution of the molecular components of the chemosensory system in arthropods.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

This work was supported by the Ministerio de Economía y Competitividad of Spain (BFU2010-15484, CGL2012-36863 and CGL2013-45211) and the Comissió Interdepartamental de Recerca I Innovació Tecnològica of Catalonia, Spain (2009SGR-1287, 2014SGR-1055 and 2014SGR1604). J.V. was supported by an FPI grant (Ministerio de Economía y Competitividad of Spain, BES-2014-068437), C.F.-L. by an IRBio fellowship, and A.S.-G. by a Beatriu de Pinós grant (Generalitat de Catalunya, 2010-BP-B 00175), and J.R. and M.A.A. were partially supported by ICREA Academia (Generalitat de Catalunya). We acknowledge the Garajonay National Parks for granting collection permits and helping with lodging and logistics during fieldwork.

Literature Cited

Adams MD, et al. 2000. The genome sequence of *Drosophila melanogaster*. Science 287:2185–2195.

Altschul S. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25:3389–3402.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol. 215:403–410.

Ashburner M, et al. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 25:25–29.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc. 57:289–300.

Binford GJ, et al. 2008. Phylogenetic relationships of *Loxosceles* and *Sicarius* spiders are consistent with Western Gondwanan vicariance. Mol Phylogenet Evol. 49:538–553.

Birol I, et al. 2009. De novo transcriptome assembly with ABySS. Bioinformatics 25:2872–2877.

Cao Z, et al. 2013. The genome of *Mesobuthus martensii* reveals a unique adaptation model of arthropods. Nat Commun. 4:2602.

Cerveira AM, Jackson RR. 2012. Love is in the air: olfaction-based mate-odour identification by jumping spiders from the genus *Cyrba*. J Ethol. 31:29–34.

Chang J-M, Di Tommaso P, Lefort V, Gascuel O, Notredame C. 2015. TCS: a web server for multiple sequence alignment evaluation and phylogenetic reconstruction. Nucleic Acids Res. 43:W3–W6.
Chang J-M, Di Tommaso P, Notredame C. 2014. TCS: a new multiple sequence alignment reliability measure to estimate alignment accuracy and improve phylogenetic tree reconstruction. Mol Biol Evol. 31:1625–1637.
Chang JM, Di Tommaso P, Taly JF, Notredame C. 2012. Accurate multiple sequence alignment of transmembrane proteins with PSI-Coffee. BMC Bioinformatics 13(Suppl 4):S1.
Chang Z, et al. 2015. Bridger: a new framework for *de novo* transcriptome assembly using RNA-seq data. Genome Biol. 16:1–10.
Chen EA, et al. 2014. Effect of RNA integrity on uniquely mapped reads in RNA-Seq. BMC Res Notes 7:753.
Chintapalli VR, Wang J, Dow JAT. 2007. Using FlyAtlas to identify better *Drosophila melanogaste*r models of human disease. Nat Genet. 39:715–720.
Chipman AD, et al. 2014. The first myriapod genome sequence reveals conservative arthropod gene content and genome organisation in the centipede *Strigamia maritima*. PLoS Biol. 12:e1002005.
Clarke TH, et al. 2014. Multi-tissue transcriptomics of the black widow spider reveals expansions, co-options, and functional processes of the silk gland gene toolkit. BMC Genomics 15:365.
Colbourne JK, et al. 2011. The ecoresponsive genome of *Daphnia pulex*. Science 331:555–561.
Conesa A, et al. 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21:3674–3676.
Corey EA, Bobkov Y, Ukhanov K, Ache BW. 2013. Ionotropic crustacean olfactory receptors. PLoS One 8:e60551.
Croset V, et al. 2010. Ancient protostome origin of chemosensory ionotropic glutamate receptors and the evolution of insect taste and olfaction. PLoS Genet. 6:e1001064.
Depetris-Chauvin A, Galagovsky D, Grosjean Y. 2015. Chemicals and chemoreceptors: ecologically relevant signals driving behavior in *Drosophila*. Front Ecol Evol. 3:41.
Eddy SR. 2011. Accelerated profile HMM searches. PLoS Comput. Biol. 7:e1002195.
Enjin A, et al. 2016. Humidity sensing in *Drosophila*. Curr Biol. 26:1352–1358.
Finn RD, et al. 2014. Pfam: the protein families database. Nucleic Acids Res. 42:D222–D230.
Foelix RF, Chu-Wang IW. 1973. The morphology of spider sensilla. II. Chemoreceptors. Tissue Cell 5:461–478.
Foelix RF, Rast B, Peattie AM. 2012. Silk secretion from tarantula feet revisited: alleged spigots are probably chemoreceptors. J Exp Biol. 215:1084–1089.
Foelix RF. 1970. Chemosensitive hairs in spiders. J Morphol. 132:313–333.
Frías-López C, et al. 2015. Comparative analysis of tissue-specific transcriptomes in the funnel-web spider *Macrothele calpeiana* (Araneae, Hexathelidae). Peer J. 3:e1064.
Fu L, Niu B, Zhu Z, Wu S, Li W. 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 28:3150–3152.
Grbić M, et al. 2011. The genome of *Tetranychus urticae* reveals herbivorous pest adaptations. Nature 479:487–492.
Groh-Lunow KC, Getahun MN, Grosse-Wilde E, Hansson BS. 2014. Expression of ionotropic receptors in terrestrial hermit crab's olfactory sensory neurons. Front Cell Neurosci. 8:1–12.
Guex N, Peitsch MC. 1997. SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. Electrophoresis 18:2714–2723.
Gulia-Nuss M, et al. 2016. Genomic insights into the *Ixodes scapularis* tick vector of Lyme disease. Nat Commun. 7:10507.
Haas BJ, et al. 2014. De novo transcript sequence reconstruction from RNA-Seq: reference generation and analysis with Trinity. Nat Protoc. 8:1–43.
Hedges SB, Marin J, Suleski M, Paymer M, Kumar S. 2015. Tree of life reveals clock-like speciation and diversification. Mol Biol Evol. 32:835–845.
Ishida Y, et al. 2014. Niemann-Pick type C2 protein mediating chemical communication in the worker ant. Proc Natl Acad Sci U S A. 111:3847–3852.
Jones P, et al. 2014. InterProScan 5: genome-scale protein function classification. Bioinformatics 30:1236–1240.
Joseph RM, Carlson JR. 2015. Drosophila chemoreceptors: a molecular interface between the chemical world and the brain. Trends Genet. 31:683–695.
Kanehisa M, Goto S. 2000. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 28:27–30.
Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 30:772–780.
Kaupp UB. 2010. Olfactory signalling in vertebrates and insects: differences and commonalities. Nat Rev Neurosci. 11:188.
Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJ. 2015. The Phyre2 web portal for protein modeling, prediction and analysis. Nat Protoc. 10:845–858.
Kirkness EF, et al. 2010. Genome sequences of the human body louse and its primary endosymbiont provide insights into the permanent parasitic lifestyle. Proc Natl Acad Sci U S A. 107:12168–12173.
Krogh A, Larsson B, von Heijne G, Sonnhammer EL. 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol. 305:567–580.
Kronestedt T. 1979. Study on chemosensitive hairs in wolf spiders (Araneae, Lycosidae) by scanning electron microscopy. Zool Scr. 8:279–285.
Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 10:R25.
Le H-S, Schulz MH, McCauley BM, Hinman VF, Bar-Joseph Z. 2013. Probabilistic error correction for RNA sequencing. Nucleic Acids Res. 41:e109.
Letunic I, Bork P. 2007. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics 23:127–128.
Li B, Dewey CN. 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 12:323.
Li B, et al. 2014. Evaluation of *de novo* transcriptome assemblies from RNA-Seq data. Genome Biol. 15:553.
Marchant A, et al. 2015. Comparing *de novo* and reference-based transcriptome assembly strategies by applying them to the blood-sucking bug *Rhodnius prolixus*. Insect Biochem Mol Biol. 69:25–33.
Martin J. a, Wang Z. 2011. Next-generation transcriptome assembly. Nat Rev Genet. 12:671–682.
Min S, Ai M, Shin SA, Suh GSB. 2013. Dedicated olfactory neurons mediating attraction behavior to ammonia and amines in *Drosophila*. Proc Natl Acad Sci U S A. 110:E1321–E1329.
Missbach C, Vogel H, Hansson BS, Groβe-Wilde E. 2015. Identification of odorant binding proteins and chemosensory proteins in antennal transcriptomes of the jumping bristletail *Lepismachilis y-signata* and the firebrat *Thermobia domestica*: evidence for an independent OBP-OR origin. Chem Senses 40:615–626.
Mita K, et al. 2004. The genome sequence of silkworm, *Bombyx mori*. DNA Res. 11:27–35.
Mitchell A, et al. 2014. The InterPro protein families database: the classification resource after 15 years. Nucleic Acids Res. 43: D213–D221.
Montagné N, de Fouchier A, Newcomb RD, Jacquin-Joly E. 2015. Advances in the identification and characterization of olfactory receptors in insects. Prog Mol Biol Transl Sci. 130:55–80.

Nelson XJ, Warui CM, Jackson RR. 2012. Widespread reliance on olfactory sex and species identification by lyssomanine and spartaeine jumping spiders. Biol J Linn Soc. 107:664–677.
Nichols Z, Vogt RG. 2008. The SNMP/CD36 gene family in Diptera, Hymenoptera and Coleoptera: *Drosophila melanogaster*, *D. pseudoobscura*, *Anopheles gambiae*, *Aedes aegypti*, *Apis mellifera*, and *Tribolium castaneum*. Insect Biochem Mol Biol. 38:398–415.
Parra G, Bradnam K, Korf I. 2007. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics 23:1061–1067.
Parra G, Bradnam K, Ning Z, Keane T, Korf I. 2009. Assessing the gene space in draft genomes. Nucleic Acids Res. 37:289–297.
Patel RK, Jain M. 2012. NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. PLoS One 7:e30619.
Pei J, Kim BH, Grishin VN. 2008. PROMALS3D: a tool for multiple sequence and structure alignment. Nucleic Acids Res. 36:2295–2300.
Pelosi P, Iovinella I, Felicioli A, Dani FR. 2014. Soluble proteins of chemical communication: an overview across arthropods. Front Physiol. 5:320.
Pelosi P, Zhou J-J, Ban LP, Calvello M. 2006. Soluble proteins in insect chemical communication. Cell Mol Life Sci. 63:1658–1676.
Pelosi P. 1996. Perireceptor events in olfaction. J Neurobiol. 30:3–19.
Petersen TN, Brunak S, von Heijne G, Nielsen H. 2011. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods 8:785–786.
Posnien N, et al. 2014. A comprehensive reference transcriptome resource for the common house spider *Parasteatoda tepidariorum*. PLoS One 9:e104885.
Qu S-X, Ma L, Li H-P, Song J-D, Hong X-Y. 2015. Chemosensory proteins involved in host recognition in the stored food mite *Tyrophagus putrescentiae*. Pest Manag Sci. 72(8):1508–1516.
Robinson MD, McCarthy DJ, Smyth GK. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26:139–140.
Rota-Stabelli O, Daley AC, Pisani D. 2013. Molecular timetrees reveal a Cambrian colonization of land and a new scenario for ecdysozoan evolution. Curr Biol. 23:392–398.
Sánchez-Gracia A, Vieira FG, Almeida FC, Rozas J. 2011. Comparative genomics of the major chemosensory gene families in arthropods. Encycl Life Sci. 3:476–490.
Sanggaard KW, et al. 2014. Spider genomes provide insight into composition and evolution of venom and silk. Nat Commun. 5:3765.
Schulz MH, Zerbino DR, Vingron M, Birney E. 2012. Oases: robust *de novo* RNA-seq assembly across the dynamic range of expression levels. Bioinformatics 28:1086–1092.
Shanbhag SR, et al. 2001. Expression mosaic of odorant-binding proteins in *Drosophila* olfactory organs. Microsc Res Tech. 55:297–306.
Shukla E, Thorat LJ, Nath BB, Gaikwad SM. 2015. Insect trehalase: physiological significance and potential applications. Glycobiology 25:357–367.
Silbering AF, et al. 2011. Complementary function and integrated wiring of the evolutionarily distinct *Drosophila* olfactory subsystems. J Neurosci. 31:13357–13375.
Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313.
Storch J, Xu Z. 2009. Niemann-Pick C2 (NPC2) and intracellular cholesterol trafficking. Biochim Biophys Acta. 1791:671–678.
Suzuki S, Kakuta M, Ishida T, Akiyama Y. 2014. Faster sequence homology searches by clustering subsequences. Bioinformatics 31:1183–1190.
Torres-Oliva M, Almeida FC, Sánchez-Gracia A, Rozas J. 2016. Comparative genomics uncovers unique gene turnover and evolutionary rates in a gene family involved in the detection of insect cuticular pheromones. Genome Biol Evol. 8:1734–1747.
Vieira FG, Rozas J. 2011. Comparative genomics of the odorant-binding and chemosensory protein gene families across the arthropoda: origin and evolutionary history of the chemosensory system. Genome Biol Evol. 3:476–490.
Vogt RG, Riddiford LM. 1981. Pheromone binding and inactivation by moth antennae. Nature 293:161–163.
Whelan S, Goldman N. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol Biol Evol. 18:691–699.
Whiteman NK, Pierce NE. 2008. Delicious poison: genetics of *Drosophila* host plant preference. Trends Ecol Evol. 23:473–478.
World Spider Catalog. 2016. World Spider Catalog. Nat. Hist. Museum Bern:online at http://wsc.nmbe.ch; version 17.0.
Xie Y, et al. 2014. SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads. Bioinformatics 30:1660–1666.
Yang J, et al. 2015. The I-TASSER suite: protein structure and function prediction. Nat Methods 12:7–8.
Zhang Y, Zheng Y, Li D, Fan Y. 2014. Transcriptomics and identification of the chemoreceptor superfamily of the pupal parasitoid of the oriental fruit fly, *Spalangia endius* Walker (Hymenoptera: Pteromalidae). PLoS One 9:e87800.

Associate editor: Davide Pisani

C

Development of Anonymous Nuclear Markers for *Buthus* scorpions (Scorpiones: Buthidae) using massive parallel sequencing, with an overview of nuclear markers used in Scorpions phylogenetics

Pedro Sousa, Cristina Frías-López, D. James Harris, Julio Rozas, Miquel A. Arnedo

C

PAPER 2: Development of Anonymous Nuclear Markers for *Buthus* scorpions (Scorpiones: Buthidae) using massive parallel sequencing, with an overview of nuclear markers used in Scorpions phylogenetics

Pedro Sousa[1,2,3,‡], Cristina Frías-López [3,4,‡], D. James Harris[1,2], Julio Rozas[3,4], Miquel A. Arnedo[3,§]

1 CIBIO Research Centre in Biodiversity and Genetic Resources, InBIO, Universidade do Porto, Campus Agrário de Vairão, Vairão, Portugal
2 Departamento de Biologia, Faculdade de Ciências da Universidade do Porto, Porto, Portugal
3 Departament de Biologia Evolutiva, Ecologia i Ciències Ambientals, and Institut de Recerca de la Biodiversitat (IRBio), Universitat de Barcelona, Barcelona, Spain
4 Departament de Genètica, Microbiologia i Estadística, and Institut de Recerca de la Biodiversitat (IRBio), Universitat de Barcelona, Barcelona, Spain

‡ - The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors; §Corresponding author

Abstract:

Multilocus datasets are routinely used to reconstruct phylogenetic relationships among groups of organisms, and to uncover phylogeographical patterns underlying species genetic diversity. However, comparatively few markers have been used to infer evolutionary histories in scorpions, of which we offer an overview, and many of the nuclear markers used are too conserved to be useful at or below the species level. Here we used a reduced representation library (RRL) combined with massive parallel sequencing, to develop five new Anonymous Nuclear Markers (ANM) that amplify in the scorpion genus *Buthus* Leach, 1815. Nucleotide diversity of the ANMs ranged from 2.2% to 5.6% for the five Iberian *Buthus* mtDNA lineages, and average uncorrected sequence divergence between lineages ranged from 0.23% to 5.28%. These results demonstrate the potential utility of these ANMs to infer the phylogeographical patterns of the Iberian *Buthus*. Furthermore, we demonstrated that two of the developed ANMs and two other nuclear markers that have been used in *Mesobuthus* Vachon, 1950, cross-amplify in the Buthidae, at least within the *Buthus* group of genera, and therefore have the potential to help reconstructing the phylogeny of the Buthidae Family, which contains almost half of all known scorpion species.

Keywords:

Reduced representation library, N.G.S., nuclear loci, multilocus, Phylogeny, Phylogeography, non-model organisms

Introduction

The PCR revolution made multilocus DNA sequencing data ever more present, and the number of Loci is increasing fast (e.g. Garrick et al. 2015), especially with the maturation of Next Generation Sequencing (NGS) and Genomics, the implications of which are far reaching (Lemmon and Lemmon, 2013; McCormack et al., 2013; Morey et al., 2013).

Multilocus studies, based on unlinked markers, have many advantages over single-locus studies (Sánchez-Gracia and Castresana, 2012). These derive from the augmented resolving power they provide, and a combination of both mtDNA and nuDNA has been shown to be particularly desirable (Sánchez-Gracia and Castresana, 2012). Multilocus datasets have a wide range of applications, not limited to phylogenetic reconstruction (Yang and Rannala, 2012), but also species delimitation (Yang and Rannala, 2010), conservation biology (Fennessy et al., 2016), etc. They are also essential for inferring species-trees, which can be different from individual gene-trees (Degnan and Rosenberg, 2009) and have wide impacts on the determination of speciation times (Nichols, 2001).

The development of new nuclear markers in non-model organism can be achieved with different methodologies. These in turn will result in different types of markers that are informative at different levels of the phylogenetic reconstruction (reviewed in Lemmon and Lemmon, 2013; Thomson et al., 2010). The methodologies to obtain new DNA sequences include expressed sequence tag libraries (EST) (e.g. Gantenbein and Keightley 2004), genomic libraries (e.g. Amaral et al. 2009, Bidegaray-Batista et al. 2011) and increasingly, NGS based approaches (Ferreira et al., 2014; Lemmon and Lemmon, 2012). The genomic library preparation can itself be obtain in several ways (reviewed in Lemmon and Lemmon, 2013; McCormack et al., 2013). According to Thomson et al. (2010) these markers can be grouped into three categories: 1) Nuclear Protein Coding Loci (NPCL); 2) Exon-primed Intron-crossing (EPIC); 3) Anonymous Nuclear Markers (ANM). ANM are attractive because they require the least amount of previous knowledge and because they have a strong probability of being highly variable and thus useful at lower levels of phylogenetic reconstruction including species' phylogeny and phylogeography. This characteristic is intrinsic to their development as they are constructed from random portions of the genome, and as most of the genome in Eukaryota is non-coding, amplification of regions of high mutation rate is expected (Thomson et al., 2010).

In Scorpiones, molecular phylogenetic studies have not been numerous. In fact the first Cladistic study in Scorpiones was, according to Soleglad and Fet (2003), that of

Stockwell (1989, unpublished), based on morphological characters. The first molecular phylogenetic studies studying the phylogeny of a genus used allozymes (Gantenbein et al., 2000b, 1998a, 1998b), allozyme and mtDNA (Gantenbein et al., 2000a, 1999; Scherabon et al., 2000) and nuDNA (Ben Ali et al., 2000), many of which already in a multilocus approach. Nevertheless the use of DNA sequences in multilocus datasets has been scarce even in the present, although it is growing, which makes Gantenbein and Keightley (2004) even more noteworthy. These authors developed eight new ANM to reconstruct the evolutionary history of *Mesobuthus gibbosus* (Brullé, 1832) and *M. cyprius* (Gantenbein and Kropf, 2000). Several studies have relied upon the usage of conserved and slow evolving regions of the nuclear rDNA (5.8S, 18S and 28S), but also using the faster evolving Internal transcribed spacers (ITS1 and ITS2) that can be amplified concomitantly (Schlötterer et al., 1994).

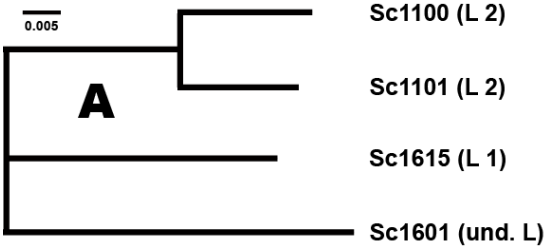
The scorpion genus *Buthus* Leach, 1815 (Buthidae C. L. Koch, 1837) currently comprises 52 species that occur in south-western Europe, North Africa and the Middle East. The phylogeography of the Western Mediterranean range of the genus was first evaluated by Gantenbein and Largiadèr (2003) using mtDNA and nuDNA. These authors found three main lineages, namely a European, a Moroccan and a Tunisian lineage. The European populations were further studied by Sousa et al. (2010) using only mtDNA. Their broader geographic sampling uncovered two previous unknown mtDNA lineages, revealing that the evolutionary history of the genus *Buthus* was more complex than previously reported in Iberia. At the same time the taxonomy of the genus in the Iberian Peninsula also changed. For a long time only *B. occitanus* (Amoreux, 1789) was accepted, but *B. ibericus* Lourenço & Vachon, 2004, and *B. montanus* Lourenço & Vachon, 2004 were described and a forth species, *B. elongatus* Rossi, 2012, was also added to the Iberian fauna. Sousa et al. (2012) confirmed that the four Iberian *Buthus* species, together with samples from Northerner Morocco, form one of the four main *Buthus* mtDNA lineages in the Western Mediterranean.

Our objective was to develop new ANMs using a reduced representation library (RRL) combined with NGS (Lemmon and Lemmon, 2012 approach) to improve our knowledge of the phylogeography of the *Buthus* lineages/species found in the Iberian Peninsula (Sousa et al., 2010). As no systematic overview of the nuclear markers used on lower rank phylogenies of Scorpiones has been published we present such overview to promote the usage of comparable datasets in the future.

Material and methods

Target species and cross-amplification

We surveyed four *Buthus* individuals from three distinct Iberian mtDNA lineages (*sensu* Sousa et al. 2010). These consisted of three samples from two different mtDNA lineages of *B. ibericus* (Sc1110 and Sc1101 from Alcalá de los Gazules, Spain – lineage 2, and Sc1615 from São Brás de Alportel, Portugal – lineage 1) and one sample from *B. montanus* (Sc1601 from Refugio Poqueira, Capileira, Spain – lineage 4) (Fig. 1).



B

	Sc1100	Sc1101	Sc1615
Sc1100			
Sc1101	3.3%		
Sc1615	7.4%	6.8%	
Sc1601	7.6%	8.5%	7.8%

Figure III-1. A – Phylogenetic neighbour-joining tree of the *cox1* mtDNA marker of the 4 *Buthus* individuals used to construct the genomic reduced representation libraries (RRL). **B** – Uncorrected sequence p-distances of the same *cox1* data. All Iberian lineages are part of the *occitanus* group of *Buthus* species (*sensu* Sousa et al Paper 1. Lineage numbering according to Sousa et al (2010).

For testing the variability of the newly design primers, we chose 10 *Buthus* individuals, two from each of the five Iberian mtDNA lineages (*sensu* Sousa et al. 2010). To further test the utility of the primers in a broader taxonomic sample, we cross-amplified them on individuals belonging to Moroccan *Buthus* lineages (*occitanus* group Sc2409 or Sc2533 and *mardochei* group, Sc 1568; *sensu* Sousa et al. Paper 1, other Buthidae genera (*Androctonus mauritanicus*, Sc2408, *Compsobuthus* sp., Sc2591; *Mesobuthus* sp., Sc2520), and two additional families: Scorpionidae Latreille, 1802 (*Scorpio* sp., Sc2405) and Iuridae Thorell, 1876 (*Calchas* sp., Sc2523).

General lab procedures

Whole genomic DNA was extracted from freshly preserved (96% ethanol) muscle tissue from the whole animal, excluding only the digestive system organs and the exoskeleton, using either the SPEEDTOOLS Tissue DNA Extraction Kit (BIOTOOLS) or phenol/chloroform extraction (two samples).

Polymerase chain reactions (PCR) were performed in a final volume of 25 µL using Sigma's REDTaq DNA polymerase with the REDTaq ReadyMix PCR Reaction Mix with MgCl2 (Sigma-Aldrich). General PCR conditions are given in Table 1. Amplified DNA templates were sequenced in both directions using one of the PCR primers and sequenced in an ABI 3700 automated sequencer at the Centres Científics i Tecnològics

de la Universitat de Barcelona (CCiTUB). DNA sequences were edited and assembled using Geneious software v.6.1.8 (Kearse et al., 2012).

In addition to the novel markers, we amplified and sequenced mitochondrial and nuclear markers that have been used in scorpion research. A partial fragment of the *cox1* mitochondrial gene was amplified using the Folmer et al. (1994) primers LCO1490 (GGT CAA CAA ATC ATA AAG ATA TTG G) and HC02198 (TAA ACT TCA GGG TGA CCA AAA AAT CA). We further tested eight nuclear fragments (Table 1). The ribosomal internal transcribed spacer (ITS2) using primers ITS4F (White et al., 1990) and ITS-5.8Sv2 (Agnarsson, 2010). The Histone 3 (H3) using primers H3aF and H3aR (Colgan et al., 1998). The 28S rDNA large subunit domain D3 (28S) using primers 28SO (Hedin and Maddison, 2001) and 28SBv2 (M. Arnedo NEW). The Protein kinase (PK) using primers 03B09F and 03B09R (Gantenbein et al., 2003). And three gene fragments from Gantenbein and Keightley (2004), Methyl transferase (MetT), Defensin 4kD (D4kD) and Lysozyme precursor C (Lys-C). All additional primer sequences can be found in the Supplemental Table 1.

Table III-1. Primer sequences for seven anonymous nuclear loci developed from a NGS approach of Iberian *Buthus*. Names indicate the loci, forward and reverse primers, PCR annealing temperature (AT), Extension time (ExT), size of amplicon (bp). Performance taxa tested: IP, Iberian Peninsula *Buthus* lineages; Moroccan *Buthus* mtDNA groups (*sensu* Sousa et al. Paper 1: occ, *occitanus* group; mar, *mardochei* group; A, *Androctonus*; M, *Mesobuthus*; C, *Compsobuthus*; S, Scorpionidae; I, Iuridae. Performance coded as follow: ?, not tested; strikethrough, unsuccessful PCR or Sequencing; or otherwise successful. n.a. unnamed in the publication; H. & M. – Hedin and Maddison, 2001.

Loci	Primer Name	Primer sequence (5'-3')	AT (ºC)	ExT (s)	Amplicon size	Performance	Source
c0037	0037F	TGTTTAGCAGATTTCGTCGGA	60º	30s	248	**IP, occ, mar, ?**	NEW
	0037R	AGCTGACTGTTTAATTCTCGCTG			to 261		
c0061	0061F	ATCAACTCGGATGTAACATCAC	53º	45s	248	**IP, occ, mar, ~~A, M, C, S, I~~**	NEW
	0061R	AGCATCAGAAACGTTAGACAAGAG					
c0118	0118F	TCTGCGAGTCACACCTTCAC	60º	30s	366	**IP, occ, ?**	NEW
	0118R	CCCTAGAACTGCTGTCTGCC					
c0717	0717F	CGGATTCTCTCGCTGAACCG	50º	45s	493	**IP, ?**	NEW
	0717R	AGGTGTACCTCAAGGCTCTG					
c0791	0791F	CGCTGCCAATGTAGCTCCAG	53º	45s	293	**IP,?**	NEW
	0791R	GTTCGATTCCCGGCGTGG			to 301		
c0971	0971F	CACGGTTAATGGAAGAAAGAGC	53º	45s	467	**IP, occ, mar, A, M, ~~S, I~~**	NEW
	0971R	AAGTTCGCATCAGTAAACAGCG			to 500		
c5070	5070F	CGACACTTTGCCAACTTCAAC	64º	30s	780	**IP, occ, mar, ~~A, M, C, S, I~~**	NEW
	5070R	GCATTGGTCTGTGGCGAATC			to 845		
28S	28SO,	GAAACTGCTCAAAGGTAAACGG	52º	45s	≈727	**IP, occ, mar, A, ?**	H. & M.
	28SBv2	TCGGAAGGAACGAGCTAC					NEW
PK	03B09F,	TCTGATGTATGGCAGATGGCAATG	45º	30s	362	**IP, occ, mar, A, M, C, ~~S, I~~**	SupT 1
	03B09R	CGAACTCAAGATCCACTCCTGTACTCG					
MetT	n.a.	TGGGTTCCAGCTCGCAGCGGTAACG	60º	30s	456	**IP, occ, mar, M, C, ~~S, I~~**	SupT 1
	n.a.	AACTTCGTAGTCGGAATACGAATGTTCTC			to 466		

Preparation of the genomic reduced representation libraries (RRL)

We obtained a reduced representation genome fragment by digesting the genome DNA with the rare-cutting restriction endonuclease *NotI* (recognition sequence: 5' GC/GGCCGC 3') (New England Biolabs), which generates large genomic fragments (Lambert et al., 2008). We subsequently selected fragments ranging from 2.5 to 3 kb by excising the corresponding bands from the agarose gel (1% concentration). DNA was purified with a QIAquick Gel Extraction Kit (Qiagen). Because of the reduced amount of DNA recovered, we conducted a round of genome re-amplification, using the illustra GenomiPhi V2 DNA Amplification Kit (GE Healthcare), following the manufacturer's isothermal reaction specifications. This method conducts a global amplification via multiple strand displacements (Paez et al., 2004), using the Φ29 DNA polymerase (Blanco et al., 1989) and random hexamer primers. The DNA was purified again with the QIAquick PCR Purification Kit (Qiagen). We constructed four separated libraries (one per individual) that were individually tagged with MID's (multiplexed identifier) barcodes. The DNA sequencing was performed in a 1/2 454 plate of the GS-FLX titanium platform.

Pre-processed and assembling of NGS data

We processed the 454 reads independently, according to their MID tag. First, adapters and putative contaminants were discarded using the SeqClean (http://compbio.dfci.harvard.edu/tgi/software/) script. We then removed exact duplicate reads (forward and reverse complement) and those with low complexity using the dust algorithm with the PRINSEQ (Schmieder and Edwards, 2011) software; this step reduces the computation time and the number of false nucleotide variants. Moreover, we also removed read fragments with low-quality bases at the ends of the sequences, and all reads shorter than 100 bp with a mean quality score below 20 sequences using NGS QC Toolkit v.2.1 (Patel and Jain, 2012).

The pre-processed reads were used to conduct the *de novo* assembling (independently per each species) using two iterative rounds of CAP3. Then, we mapped the reads to the assembled contigs belonging to the same individual using the algorithm BWA-MEM (bwa-0.6.1) to determine the individual depth and removed contigs and reads related with multiple alignments (generating 4 BAM files, one per species). This excluded for the next SNP discovery step the reads that align to multiple locations and the contigs involved. Then, we performed a second alignment round, mapping all the filtered reads onto the four filtered assemblies (generating 12 BAM files, three per species) in order to identify the variant positions against the individuals used for generating the RRLs. Then we applied several filters to discard putative false nucleotide variants using a combination of SAMtools 'view' (Li et al., 2009) and a number of Perl script developed

ad hoc. In particular, we 1) removed the alignments with a CIGAR string with a 10% of hard clipping larger than the length of the aligned sequence*;* 2) realigned mismatches positions around indels; 3) removed reads that map to multiple locations (using SAMtools view option '-q 1') and removed sequences with the XA:Z flag; and 4) mark duplicated reads and add read groups using picard [http://picard.sourceforge.net.]. Afterwards, we used the filtered files and SAMtools 'mpileup' to obtain the coordinates of the variable positions of every individual sequence against each other and we exclude the nucleotide variant positions with a depth bigger than the two-fold of average coverage (obtained in the first mapping step).

Finally, we conducted the SNP calling step using the above filtered pileup files and *in house* Perl scripts to translate these pileup files to a matrix (using a value of a '0' for the non-variant positions and '1' for the polymorphic positions) to identify contigs mapped for reads belonging to at least two individuals (or the individuals of interest), with a variable region larger than 300 bp, a percentage of variability between 1-10%, and flanked by conserved regions with a length of 30-50 bp. The filtered contigs were visually explored using Geneious software v.6.1.8 (Kearse et al., 2012) to identify some erroneous mapped sequences, for instance, with homopolymers, or contigs with an excess of heterozygotic positions, that might imply a bad assembly (or mapping).

Primer pairs for the ANM (Table 1) were designed with the Primer3 software (Rozen and Skaletsky, 1999) as implemented in the Geneious software v.6.1.8 (Kearse et al., 2012). Several primers were tweaked to guaranty that their 3'-end was a C or G to promote binding, while also retaining annealing temperature, G-C content and other primer design features requirements.

All new sequences obtained in this study are available in GenBank.

Data analyses

The haplotype phases were resolved using a two-step procedure. First, for sequences that were heterozygous for insertions or deletions, we used Champuru software online v1.0 (Flot, 2007), which implements the method described by Flot et al. (2006). Second, nucleotide polymorphisms were resolved using the Bayesian algorithm implemented in PHASE (Stephens et al., 2001). Phase was run five times per dataset.

The protein coding genes *cox1* and PK were aligned with Muscle (Edgar, 2004) and no indel were found. They were translated to amino acids and show no stop codons. The remaining genes were aligned with the MAFFT (v7.017) method G-INS-i (Katoh et al., 2002; Katoh and Standley, 2013) in Geneious v.6.1.8 (Kearse et al., 2012).

Uncorrected genetic p-distances between mtDNA *cox1* lineages were estimated with MEGA v6.06 (Tamura et al., 2013). Standard deviation was assessed by conducting

1000 bootstraps. Genetic diversity indices were estimated using DnaSP v.5.10.01 (Librado and Rozas, 2009). We calculated the number of segregation sites (S), the number of segregating sites per 100 bp (S_{100}), the nucleotide diversity (π), and the haplotype number (H) and diversity (Hd). Non-neutral evolution was evaluated with Tajima's D test (D) (Tajima, 1989). Recombination was investigated using the minimum number of recombination events (R_M) (Hudson and Kaplan, 1985) and the linkage disequilibrium statistic (ZZ), which can also detect intragenic recombination (Rozas et al., 2001). The significance of the results was assessed using coalescent simulations with the algorithm implemented in DnaSP.

We made a bibliographic search for all relevant literature published until December 2016 that presented a molecular phylogeny or phylogeography of the Order Scorpiones, bellow the family rank, in order to a review all nuclear markers used. Studies focusing on venom nuclear markers were not considered.

Results

We obtained a total of 487,357 raw reads across all four samples, which represent about 0.7% of the genome, assuming a random distribution of restriction sites, or 7.7 Mbp assuming an average genome size of about 1.1 Gbp for scorpions [from 0.90 Gbp in *Centruroides vittatus* (Hanrahan and Johnston, 2011) to 1.35 Gbp in *Mesobuthus martensii* (Cao et al., 2013)]. We removed low quality reads and, given the properties of the fragments obtained, searched and discarded the sequences with repetitive motifs, low complexity, and with high levels of entropy, removing 51% of the total reads. After the pre-processing step, we used two iterative rounds of CAP3 (-o 150 –p 90, -o 100 –p 90), and assembled a total of 9,183 contigs (Suppl. Table 2.2) with a N50 of 758. We also reduced the number of duplicate/paralogous sequences performing a previous individual mapping step using the assembled contigs as reference sequences and the reads assembled of the same individual. We therefore, repeated the mapping step, aligning the filtered reads of the four individuals using also the four filtered assemblies as reference sequences (the reads and contigs filtered in the previous step related with multimapping flags). The BAM files obtained were filtered to obtain only the reads with a single/UNIQ alignment performed and without hard clipped bases which reduced the percentage mapping around ~30% (Suppl. Table 2.2). To identify polymorphic positions, we used SAMtools to generate the 'mpileup' file with the alignment information of every individual, through mapping the reads of the other three individuals onto every assembly.

We also removed the alignments with a higher coverage than the two-fold of its average depth. Filtered pileup files were analysed using *in house* Perl script's to identify sequences mapped for the other individuals with a variable region of at least 300 bp, a variability range between 1-10%, and flanked by conserved fragments of 30-50 bp.

We identified 67 contigs that fulfilled the defined rules. These were then individually analysed and 18 were selected (16 different markers and two length variants) for which we design primers and tested for PCR amplification and variability. Only seven markers (ANM) could be amplified and sequenced (Table 1), although only five of them were consistently recovered (Table 1, 2).

Table III-2. Summary diversity statistics for 12 nuclear sequence markers for five Iberian *Buthus* lineages plus one Moroccan *Buthus*. Nº, number of specimens, IP lin., number of Iberian lineages; Mor, Moroccan mtDNA groups represented (*sensu* Sousa et al. Paper 1, for *occitanus* group: *occ*1 - Sc2409, *occ*2 - Sc2533; The length in bp for each locus (L) after sequences end-trimming, excluding sites with gaps. The summed lengths of indels in bp (Indels). The number of segregating sites (S), the number of segregating sites per 100 bp (S100), haplotype number (H), haplotype diversity (Hd), nucleotide diversity (π), minimum number of recombination events (RM) of Hudson (1985), linkage disequilibrium statistic (ZZ) of Rozas et al. (2001), Tajima's D test (D) of Tajima (1989). Not significant (ns) and significant (*) values at P < 0.05 of statistics after 10.000 coalescence simulations. *cox1* is presented twice, with and without Moroccan samples for more appropriate comparison with the different Loci results. 1 – Different specimens from the same lineages were used due to difficulties during sequencing. 2- Includes a previous unidentified Iberian lineage that we used due to limited available results.

Locus	Nº	IP	Mor	L	Indels	S	S_{100}	π	H	Hd	Rm	ZZ	D
cox1	11	5	*mar.*	641	0	136	21.2%	0.088	11	1.000	45 ns	0.006 ns	0.019 ns
c0037	**7**	**3**	*mar.*	264	21	36	14.8%	0.056	4	0.810	2 ns	-0.002 ns	-0.6 ns
c0061	11	5	*occ*1	226	6	27	12.3%	0.037	10	0.992	1 ns	0.002 ns	0.256 ns
c5070	11	5	*mar.*	778	31	76	10.2%	0.022	13	0.944	**2 ***	0.139 ns	-0.875 ns
c0971	11	5	*mar.*	451	27	40	9.4%	0.022	12	0.926	4 ns	0.045 ns	-0.689 ns
c0118	**7**[1]	5	*occ*2	366	0	33	9.0%	0.029	10	0.923	6 ns	0.161 ns	-0.093 ns
MetT	11	5	*mar.*	385	5	27	7.1%	0.024	13	0.944	4 ns	0.136 ns	0.76 ns
PK	11	5	*mar.*	362	0	13	3.6%	0.008	12	0.887	2 ns	-0.014 ns	-0.631 ns
28S	11[1]	5	*mar.*	727	0	7	1.0%	0.004	6	0.801	1 ns	0.015 ns	1.222 ns
ITS2	**4**	**2**	*mar.*	475	16	26	5.7%	0.032	3	0.833	0 ns	0.042 ns	-0.159 ns
H3	**2**	**1**	*mar.*	328	0	2	0.6%	0.006	2	1.000	0 ns	0.000	n.a.
cox1	10	5	none	641	0	118	18.4%	0.080	10	1.000	38 ns	0.031 ns	0.381 ns
c0717	**4**	**3**[2]	none	493	0	24	4.9%	0.019	7	0.964	2 ns	0.009 ns	-0.019 ns
c0791	**2**	**2**[2]	none	302	10	8	2.7%	0.027	2	1.000	0 ns	0.000	n.a.

We were able to amplify, albeit with different success rates, six of the former nuclear markers available for scorpion research, (Suppl. Table 2.1). We only amplified two specimens for the **H3,** although due to the extremely low S_{100} (0.61%) (Table 2) we did not investigate this marker any further. Although, we initially tried to sequence the fragment spawning the **18S** plus **ITS1** region used by Gantenbein and Largiadèr (2003), this proved to be difficult and we used the **ITS2** instead. Although **ITS2** had potential (S_{100} = 5.7%) we did not pursue it due to the lack of *Buthus* sequences available for comparison in Genbank. The **D4kD** and **Lys-C** could not be amplified in any *Buthus*

samples. The remaining three nuclear markers, Met T, PK and 28S were successfully amplified and sequenced (Table 2).

The success of cross-amplification varied considerably between the loci tested. All amplified the five Iberian lineages tested and the Moroccan representative of the *occitanus* mtDNA group and all that were tested also most amplified the distant Moroccan *mardochei* mtDNA group (Table 1). The loci c0971, 28S, PK and MetT, amplified all the Buthidae genera in which they were tested, but we were unsuccessful in amplifying either the Scorpionidae or Iuridae samples (Table 1).

Genetic divergences were calculated for the *cox1* and the c0037, c0061, c0118, c0971, c5070, MetT, PK and 28S nuDNA (Suppl. Table 2.3). For the Iberian *Buthus* lineages, estimates ranged from 0.27% (c0971) to 4.07% (c0061) between lineages 1 and 2, 0.34% (28S) to 5.28% (c0037) between lineages 1 and 3, 0.38% (28S) to 4.33% (MetT), between lineages 1 and 4, 0.34% (28S) to 4.84% (c0037) between lineages 1 and 5, 0.23% (c0971) to 2.23% (c5070) between lineages 2 and 3, 0.58% (28S) to 4.19% (c0061) between lineages 2 and 4, 0.28% (28S) to 4.75% (c0061) between lineages 2 and 5, 0.58% (28S) to 4.19% (c0061) between lineages 3 and 4, 0.28% (28S) to 4.76% (c0061) between lineages 3 and 5, 0.28% (28S) to 4.66% (MetT) between lineages 4 and 5 (Suppl. Table 2.3). The *cox1* mtDNA fragment was found to be twice as variable as the ANM c0037 and c0061, ten times more variable than PK and more than twenty times more variable than 28S. Remarkably the ANM c0037 was found to be a little bit more variable (x1.1) than the *cox1* locus when comparing Iberian and Moroccan samples (Suppl. Table 2.3).

No intragenic recombination was detected (Rozas' et al. ZZ), with the data conforming to the expected linkage disequilibrium, although we found that the minimum number of recombination events deviated from what was expected for the Loci c5070 (Hudson and Kaplan's Rm).

Table III-3. Nuclear Loci used in 30 molecular phylogenetic or phylogeographic studies of Scorpiones, ordered chronologically. The list does not include works that have relied upon venom markers, including venom gland transcriptomes or cytogenetics, and only include works that have focused below the family rank in Scorpiones. Notes: 1 – internal region sequenced; 2 – only amplifies in *Centruroides vittatus* according to the authors. For primer sequences and references see Supplemental Table 1. a – the authors also sequenced a small portion of the end of 18S and beginning of 28S; b – only the 18S + ITS1 region was sequenced.

Loci or Marker type	Works
Allozymes	Gantenbein et al., 1998a, 1998b, 2000, 2001; Gantenbein 2004
ITS1+ 5.8S + ITS2	Ben Ali et al., 2000; Bryson et al., 2014[a]
18S + ITS1 + 5.8S	Gantenbein & Largiadèr, 2003[b]; Salomone et al., 2007
18S rRNA	Soleglad & Fet, 2003; Li et al. 2009; González-Santillán & Prendini 2014; Santibáñez-López et al., 2014 Ceccarelli et al., 2016a?
28S rRNA	Prendini et al., 2003; Bryson et al., 2013a; Bryson et al., 2013b; González-Santillán & Prendini 2014; Santibáñez-López et al., 2014; Talal et al., 2015; Ceccarelli et al., 2016a; Ojanguren-Affilastro et al., 2016; Luna-Ramirez et al., 2017
Protein kinase	Gantenbein et al., 2003; Gantenbein & Keightley, 2004; Shi et al., 2013
Chaperonin 10, Defensin, Lysozyme precursor C, Methyl transferase, Unknown protein, Thioredoxin1	Gantenbein & Keightley, 2004
Serinproteinase inhibitor, Serin-type endopepdidase	Gantenbein & Keightley, 2004; Shi et al., 2013
non-LTR retrotransposons	Glushkov et al., 2006
Microsatellites	Ji et al. 2008
RAPD	Abdel-Rahman et al., 2009
ITS2	Bryson et al., 2013a; Bryson et al., 2013b; Graham et al., 2013
ANM (Locus 1075)[2]	Yamashita & Rhoads, 2013
Genomics	Sharma et al., 2015
Actin 5C	Ceccarelli et al., 2016b
wingless	Ceccarelli et al., 2016b
SNP	Bryson et al., 2016

In Table 3 we present a bibliographic overview of all nuclear markers that have been used in molecular phylogenenetic studies on scorpions in Table 3. We have found 30 published articles up to December 2016, the majority of which (21) have relied on Sanger sequencing of nuclear DNA. Most of the molecular systematics and phylogeographic studies found used a limited number of nuclear Loci using Sanger sequencing (1.86 average Loci per study; 1.50 removing Gantenbein and Keightley 2004 work) (Table 3). For comparison, earlier allozyme studies had an average of 16.8 loci analysed per study.

Discussion

We were able to successfully design new nuclear markers that were informative at the species and population level in *Buthus* scorpions, using the approach described in Frías-López et al. (2016, supplementary material), based on the combination of restricted representation libraries and massive parallel sequencing (Lemmon and Lemmon, 2012). The novel markers designed in the present study remain anonymous (ANM) because no significant matches were recovered in BLAST searches. Although the assembled complete genome of *M. martensii* has been made available by Cao et al. (2013), it is of limited use due to the lack of accurate annotations.

We found wide disparities in the results when comparing the average uncorrected inter-lineage sequence divergence (Suppl. Table 2.3), as expect if we were successful at amplifying portions of the nuclear genome that are evolving at different rates. Nevertheless, were surprise to find similar but higher divergences using the locus c0037 and not the *cox1* mtDNA loci when comparing Iberian and Moroccan lineages. When comparing only Iberian *Buthus* lineages the *cox1* mtDNA performed as expected, whit at least double the amount of inter-lineage sequence divergence. This might reflect different evolutionary rates in different branches of *Buthus* phylogeny.

Most of the published works found in the bibliographic search have used a limited number of nuclear Loci (Table 3). Most works relied on two nuclear genes, 18S and 28S, probably due to easiness of amplification (Hillis and Dixon 1991). As predicted, these markers were found to be highly conserved and their use for shallow relationship was very limited (Bryson Jr et al., 2013; Talal et al., 2015). We obtained a similar result while testing them in *Buthus* (Suppl. Table 2.3). Comparing NPCL markers, these were found to yield similar results to ours in Ceccarelli et al. (2016) study. However, the PK marker that we tested was found to be much more variable in *Mesobuthus* species (S_{100} = 16.5% vs 3.6% in *Buthus*) (we combined three works in this analysis: Gantenbein et al., 2003;

Gantenbein and Keightley, 2004; Shi et al., 2013). As expected, the Internal transcribed spacer (ITS1, ITS2 or combined with portions of rDNA), used in three studies (Bryson Jr et al., 2014, 2013; Salomone et al., 2007) (Table 3), was more variable. These markers were found to be at least as variable (S_{100} ranging from 10.0% to 20.3%) as the most variable ANM developed in our work (*e.g.* c0037, 14.8%; c5070, 10.2%). Comparison of interspecific sequence divergence yield similar results. The amount of divergence between two pairs of species calculated with the most variable nuclear markers (ITS) (Salomone et al., 2007) was similar to what we found in *Buthus* with the two most variable ANM (c0037, 4.29%; c0061, 3.62%; average p-distances). The results of interspecific sequence divergence (p-distance) using the PK alignment described above (10 species pairs of *Mesobuthus,* number sequences = 97) was on average 2.04%. This was more than double what we found for *Buthus* (0.90%), but it was very similar to the divergence found between *M. gibbosus* and *M. cyprius* (0.96%), both from the Aegean region. Both marker variability and sequence divergence suggests either an older divergence time for the *Mesobuthus* species studied or an accelerated rate of mutation in the PK marker.

In this study we were able to demonstrate that two new ANM (c0037 and c0971) can be sequenced in three Buthidae genera. This is the most specious scorpion family, comprising almost half of all known scorpion species (Rein, 2016). We also demonstrated that two other markers, PK (Gantenbein et al., 2003), a NPCL marker, and MetT (Gantenbein and Keightley, 2004), an EPIC marker, can both be sequenced beyond the *Mesobuthus* genus. These four markers were applied only in the *Buthus* group of Buthidae genera (*sensu* Fet et al., 2005), but if successfully applied in the broader Buthidae, they can provide a framework for a coherent molecular systematic study of this diverse and venomous scorpion family (Chippaux and Goyffon, 2008), which remains largely unexplored (Fet et al., 2003; Sharma et al., 2015).

The methodological approach we followed proved successful to develop five new ANM that seem promising to investigate evolutionary relationships at least within the genus *Buthus*. Moreover, other massive parallel sequencing techniques that provide greater coverage should facilitate the assembly steps of the genomic RRL pipeline. This approach is also very flexible because the NGS data acquired can be used for other objectives, for example creating microsatellites markers to study recent population-level events in the Iberian *Buthus* species.

Acknowledgments

Pedro Sousa is most grateful to Sara Guirao-Rico, for her teaching on how to prepare the RRLs, and for her invaluable “long distance” troubleshooting of our attempts at it. Sónia Ferreira offered comments on an early version of the manuscript. We are also thankful to Enric Planas and our late colleague Margarita Metallinou for providing us with important cross-amplification taxa, namely all the Eastern Mediterranean samples. And to all our colleagues and friends from Cibio-UP who participated during fieldwork and gave advices on the work developed here, in particular Arie van der Meijden. We would also express our thanks to our colleagues and friends from the Departament de Biologia Animal de la Universitat de Barcelona that helped in the lab, in particular Gema Blasco, Vera Opatova, Paola Mazzuca, Elisa Mora and Leticia Bidegaray-Batista. And to Sara and Francisca that patiently helped us “navigate” through theirs’ Departament de Genètica lab. We also thanks the Centres Científics i Tecnològics de la Universitat de Barcelona (CCiTUB) for the NGS library sequencing.

This work was partially supported by FEDER through the COMPETE program, Portuguese national funds through the Portuguese Fundação para a Ciência e Tecnologia (FCT), financed by the Programa Operacional Potencial Humano (POPH) – Quadro de Referência Estratégico Nacional (QREN) from the European Social Fund and Portuguese Ministério da Educação e Ciência: PS, PhD grant SFRH/BD/74934/2010; DJH, IF-contract IF/01627/2014. And by the Ministerio de Educacion y Ciencia of Spain, Nº BFU2010-15484 and Nº CGL2013-45211 to JR, and Nº CGL2012-36863 to MAA.

References

Agnarsson, I., 2010. The utility of ITS2 in spider phylogenetics: notes on prior work and an example from Anelosimus. J. Arachnol. 38, 377–382.

Amaral, A.R., Silva, M.C., Möller, L.M., Beheregaray, L.B., Coelho, M.M., 2009. Anonymous nuclear markers for cetacean species. Conserv. Genet. 11, 1143–1146. doi:10.1007/s10592-009-9903-3

Avise, J.C., 2000. Phylogeography: the history and formation of species. Harvard University Press, Cambridge & London.

Ben Ali, Z., Boursot, P., Said, K., Lagnel, J., Chatti, N., Navajas, M., 2000. Comparison of ribosomal ITS regions among Androctonus spp. scorpions (Scorpionida: Buthidae) from Tunisia. J. Med. Entomol. 37, 787–790. doi:10.1603/0022-2585-37.6.787

Bidegaray-Batista, L., Gillespie, R.G., Arnedo, M.A., 2011. Bringing spiders to the multilocus era: novel anonymous nuclear markers for Harpactocrates ground-dwelling spiders (Araneae: Dysderidae) with application to related genera. J. Arachnol. 39.

Blanco, L., Bernad, A., Lázaro, J.M., Martín, G., Garmendia, C., Salas, M., Bernads, A., Lharo, J.M., Martins, G., 1989. Highly Efficient DNA Synthesis by the Phage Φ29 DNA Polymerase. J. Biol. Chem. 264, 8935–8940.

Bryson Jr, R.W., Prendini, L., Savary, W.E., Pearman, P.B., 2014. Caves as microrefugia: Pleistocene phylogeography of the troglophilic North American scorpion Pseudouroctonus reddelli. BMC Evol. Biol. doi:10.1186/1471-2148-14-9

Bryson Jr, R.W., Savary, W.E., Prendini, L., 2013. Biogeography of scorpions in the Pseudouroctonus minimus complex (Vaejovidae) from south-western North America: Implications of ecological specialization for pre-Quaternary diversification. J. Biogeogr. 40, 1850–1860. doi:10.1111/jbi.12134

Cao, Z., Yu, Y., Wu, Y., Hao, P., Di, Z., He, Y., Chen, Z., Yang, W., Shen, Z., He, X., Sheng, J., Xu, X., Pan, B., Feng, J., Yang, X., Hong, W., Zhao, W., Li, Z., Huang, K., Li, T., Kong, Y., Liu, H., Jiang, D., Zhang, B., Hu, J., Hu, Y., Wang, B., Dai, J., Yuan, B., Feng, Y., Huang, W., Xing, X., Zhao, G., Li, X., Li, Y., Li, W., 2013. The genome of Mesobuthus martensii reveals a unique adaptation model of arthropods. Nat. Commun. 4, 2602. doi:10.1038/ncomms3602

Ceccarelli, F.S., Pizarro-Araya, J., Ojanguren-Affilastro, A.A., 2016. Phylogeography and population structure of two Brachistosternus species (Scorpiones: Bothriuridae) from the Chilean coastal desert - the perils of coastal living. Biol. J. Linn. Soc. doi:10.1111/bij.12877

Chippaux, J.-P., Goyffon, M., 2008. Epidemiology of scorpionism: a global appraisal. Acta Trop. 107, 71–79. doi:10.1016/j.actatropica.2008.05.021

Colgan, D.J., McLauchlan, A., Wilson, G.D.F., Livingston, S.P., Edgecombe, G.D., Macaranas, J., Cassis, G., Gray, M.R., 1998. Histone H3 and U2 snRNA DNA sequences and arthropod molecular evolution. Aust. J. Zool. 46, 419–437. doi:10.1071/ZO98048

Degnan, J.H., Rosenberg, N.A., 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends Ecol. Evol. 24, 332–40. doi:10.1016/j.tree.2009.01.009

Edgar, R.C., 2004. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32, 1792–1797. doi:10.1093/nar/gkh340

Felsenstein, J., 2004. Inferring phylogenies. Sinauer Associates, Inc., Sunderland.

Fennessy, J., Bidon, T., Reuss, F., Vamberger, M., Fritz, U., Janke Correspondence, A., Kumar, V., Elkan, P., Nilsson, M.A., Janke, A., 2016. Multilocus Analyses Reveal Four Giraffe Species Instead of One. Curr. Biol. 26, 1–7. doi:10.1016/j.cub.2016.07.036

Ferreira, S., Lorenzo-Carballa, M.O., Torres-Cambas, Y., Cordero-Rivera, A., Thompson, D.J., Watts, P.C., 2014. New EPIC nuclear DNA sequence markers to improve the resolution of phylogeographic studies of coenagrionids and other odonates. Int. J. Odonatol. 17, 135–147. doi:10.1080/13887890.2014.950698

Fet, V., Gantenbein, B., Gromov, A. V., Lowe, G., Lourenço, W.R., 2003. The first molecular phylogeny of Buthidae (Scorpiones). Euscorpius 4, 1–10.

Fet, V., Soleglad, M.E., Lowe, G., 2005. A new Trichobothrial character for the High-Level Systematics of Buthoidea (Scorpiones: Buthida). Euscorpius 23, 1–40.

Flot, J.-F., 2007. Champuru 1.0: a computer software for unraveling mixtures of two DNA sequences of unequal lengths. Mol. Ecol. Notes 7, 974–977. doi:10.1111/j.1471-

8286.2007.01857.x

Flot, J.-F., Tillier, A., Samadi, S., Tillier, S., 2006. Phase determination from direct sequencing of length-variable DNA regions. Mol. Ecol. Notes 6, 627–630. doi:10.1111/j.1471-8286.2006.01355.x

Folmer, O., Black, M., Hoeh, W., Lutz, R., Vrijenhoek, R., 1994. DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. Mol. Mar. Biol. Biotechnol. 3, 294–299.

Fore, J., Wiechers, I.R., Cook-Deegan, R., 2006. The effects of business practices, licensing, and intellectual property on development and dissemination of the polymerase chain reaction: case study. J. Biomed. Discov. Collab. 1. doi:10.1186/1747-5333-1-7

Frías-López, C., Sánchez-Herrero, J.F., Guirao-Rico, S., Mora, E., Arnedo, M.A., Sánchez-Gracia, A., Rozas, J., 2016. DOMINO: development of informative molecular markers for phylogenetic and genome-wide population genetic studies in non-model organisms. Bioinformatics btw534. doi:10.1093/bioinformatics/btw534

Gantenbein, B., Braunwalder, M.E., Scholl, A., 1998a. Allozyme studies on scorpions from the Aegean Region and from Morocco. Abstr. XIV Int. Congr. Arachnol. \ 22 Annu. Meet. Am. Arachnol. Soc. Chicago 51.

Gantenbein, B., Büchi, L., Braunwalder, M.E., Scholl, A., 1998b. The genetic population structure of Euscorpius germanus (C.L. Koch) (Scorpiones: Chactidae) in Switzerland., in: Selden, P.A. (Ed.), Proceedings of the 17th European Colloquium of Arachnology, Edinburg, 1997. British Arachnological Society, Burnham Beeches, Bucks., pp. 33–40.

Gantenbein, B., Fet, V., Barker, M.D., Scholl, A., 2000a. Nuclear and mitochondrial markers reveal the existence of two parapatric scorpion species in the Alps: Euscorpius germanus (CL Koch, 1837) and E. alpha Caporiacco, 1950, stat. nov. (Euscorpiidae). Rev. suisse Zool. 107, 843–869.

Gantenbein, B., Fet, V., Gromov, A. V., 2003. The first DNA phylogeny of four species of Mesobuthus (Scorpiones, Buthidae) from Eurasia. J. Arachnol. 31, 412–420. doi:10.1636/H01-23

Gantenbein, B., Fet, V., Largiader, C.R., Scholl, A., 1999. First DNA pylogeny of Euscorpius Thorell, 1876 (Scorpiones, Euscorpiidae) and its bearing on taxonomy and biogeography of the genus. Biogeographica 75, 49–65.

Gantenbein, B., Keightley, P.D., 2004. Rates of molecular evolution in nuclear genes of east Mediterranean scorpions. Evolution (N. Y). 58, 2486–97.

Gantenbein, B., Kropf, C., Largiadèr, C.R., Scholl, A., 2000b. Molecular and morphological evidence for the presence of a new Buthid taxon (Scorpiones: Buthidae) on the Island of Cyprus. Rev. Suisse Zool. 107, 213–232.

Gantenbein, B., Largiadèr, C.R., 2003. The phylogeographic importance of the Strait of Gibraltar as a gene flow barrier in terrestrial arthropods: a case study with the scorpion *Buthus occitanus* as model organism. Mol. Phylogenet. Evol. 28, 119–130. doi:10.1016/S1055-7903(03)00031-9

Garrick, R.C., Bonatelli, I. a S., Hyseni, C., Morales, A., Pelletier, T.A., Perez, M.F., Rice, E., Satler, J.D., Symula, R.E., Thome, T.C., Carstens, B.C., 2015. The evolution of phylogeographic data sets. Mol. Ecol. 24, 1164–1171. doi:10.1111/mec.13108

Hanrahan, S.J., Johnston, J.S., 2011. New genome size estimates of 134 species of arthropods. Chromosom. Res. 19, 809–823. doi:10.1007/s10577-011-9231-6

Hedin, M.C., Maddison, W.P., 2001. A combined molecular approach to phylogeny of the jumping spider subfamily Dendryphantinae (Araneae: Salticidae). Mol. Phylogenet. Evol. 18, 386–403. doi:10.1006/mpev.2000.0883

Hennig, W., 1966. Phylogenetic Systematics. University of Illinois Press, Urbana.

Hillis, D.M., Dixon, M.T.., 1991. Ribosomal DNA: molecular evolution and phylogenetic inference. Q. Rev. Biol. 66, 411–453.

Hudson, R.R., Kaplan, N.L., 1985. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics 111, 147–164.

Katoh, K., Misawa, K., Kuma, K., Miyata, T., 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 30, 3059–3066. doi:10.1093/nar/gkf436

Katoh, K., Standley, D.M., 2013. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Mol. Biol. Evol. 30, 772–780. doi:10.1093/molbev/mst010

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., Thierer, T., Ashton, B., Meintjes, P., Drummond, A., 2012. Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28, 1647–1649. doi:10.1093/bioinformatics/bts199

Lambert, A.R., Sussman, D., Shen, B., Maunus, R., Nix, J., Samuelson, J., Xu, S.-Y., Stoddard, B.L., 2008. Structures of the Rare-Cutting Restriction Endonuclease NotI Reveal a Unique Metal Binding Fold Involved in DNA Binding. Structure 16, 558–569. doi:10.1016/j.str.2008.01.017

Lemmon, A.R., Lemmon, E.M., 2012. High-throughput identification of informative nuclear loci for shallow-scale phylogenetics and phylogeography. Syst. Biol. 61, 745–761. doi:10.1093/sysbio/sys051

Lemmon, E.M., Lemmon, A.R., 2013. High-Throughput Genomic Data in Systematics and Phylogenetics. Annu. Rev. Ecol. Evol. Syst. 44, 99–121. doi:10.1146/annurev-ecolsys-110512-135822

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 1000 Genome Project Data Processing Subgroup, 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079. doi:10.1093/bioinformatics/btp352

Librado, P., Rozas, J., 2009. DnaSP v5: A software for comprehensive analysis of DNA polymorphism data. Bioinformatics 25, 1451–1452. doi:10.1093/bioinformatics/btp187

McCormack, J.E., Hird, S.M., Zellmer, A.J., Carstens, B.C., Brumfield, R.T., 2013. Applications of next-generation sequencing to phylogeography and phylogenetics. Mol. Phylogenet. Evol. 66, 526–38. doi:10.1016/j.ympev.2011.12.007

Morey, M., Fernández-Marmiesse, A., Castiñeiras, D., Fraga, J.M., Couce, M.L., Cocho, J. a., 2013. A glimpse into past, present, and future DNA sequencing. Mol. Genet. Metab. 110, 3–24. doi:10.1016/j.ymgme.2013.04.024

Nichols, R., 2001. Gene trees and species trees are not the same. Trends Ecol. Evol. 16, 358–364. doi:10.1016/S0169-5347(01)02203-0

Paez, J.G., Lin, M., Beroukhim, R., Lee, J.C., Zhao, X., Richter, D.J., Gabriel, S., Herman, P., Sasaki, H., Altshuler, D., Li, C., Meyerson, M., Sellers, W.R., 2004. Genome coverage and sequence fidelity of Φ29 polymerase-based multiple strand displacement whole genome amplification. Nucleic Acids Res. 32, e71. doi:10.1093/nar/gnh069

Patel, R.K., Jain, M., 2012. NGS QC toolkit: A toolkit for quality control of next generation sequencing data. PLoS One 7. doi:10.1371/journal.pone.0030619

Rozas, J., Gullaud, M., Blandin, G., Aguadé, M., 2001. DNA variation at the rp49 gene region of Drosophila simulans: Evolutionary inferences from an unusual haplotype structure. Genetics 158, 1147–1155.

Rozen, S., Skaletsky, H.J., 1999. Primer3 on the WWW for general users and for biologist programmers. pp 365-386, in: Misener, S., Krawetz, S. (Eds.), Bioinformatics Methods and Protocols: Methods in Molecular Biology. Humana Press, New Jersey, p. 500. doi:10.1385/1592591922

Saiki, R.K., Gelfand, D.H., Stoffel, S., Scharf, S.J., Higuchi, R., Horn, G.T., Mullis, K.B., Erlich, H. a, 1988. Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. Sci. 239, 487–491. doi:10.1126/science.2448875

Salomone, N., Vignoli, V., Frati, F., Bernini, F., 2007. Species boundaries and phylogeography of the “Euscorpius carpathicus complex” (Scorpiones: Euscorpiidae) in Italy. Mol. Phylogenet. Evol. 43, 502–14. doi:10.1016/j.ympev.2006.08.023

Sánchez-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. doi:10.1371/journal.pone.0030239

Scherabon, B., Gantenbein, B., Fet, V., 2000. A new species of scorpion from Austria, Italy, Slovenia, and Croatia: Euscorpius gamma Caporiacco, 1950, stat. nov.(Scorpiones: Euscorpiidae). Ekológia (Bratislava) 19, 253–262.

Schlötterer, C., Hauser, M.T., Von Haeseler, A., Tautz, D., 1994. Comparative evolutionary analysis of rDNA ITS regions in Drosophila. Mol. Biol. Evol. 11, 513–22.

Schmieder, R., Edwards, R., 2011. Quality control and preprocessing of metagenomic datasets. Bioinformatics 27, 863–864. doi:10.1093/bioinformatics/btr026

Sharma, P.P., Fernández, R., Esposito, L.A., González-Santillán, E., Monod, L., 2015. Phylogenomic resolution of scorpions reveals multilevel discordance with morphological phylogenetic signal. Proc. R. Soc. B Biol. Sci. 282B. doi:10.5061/dryad.n0qr5

Shi, C.-M., Ji, Y.-J., Liu, L., Wang, L., Zhang, D.-X., 2013. Impact of climate changes from Middle Miocene onwards on evolutionary diversification in Eurasia: insights from the mesobuthid scorpions. Mol. Ecol. 22, 1700–1716. doi:10.1111/mec.12205

Simon, C., Buckley, T.R., Frati, F., Stewart, J.B., Beckenbach, A.T., 2006. Incorporating Molecular Evolution into Phylogenetic Analysis, and a New Compilation of Conserved Polymerase Chain Reaction Primers for Animal Mitochondrial DNA. Annu. Rev. Ecol. Evol. Syst. 37, 545–579. doi:10.1146/annurev.ecolsys.37.091305.110018

Soleglad, M.E., Fet, V., 2003. High-level systematics and phylogeny of the extant scorpions (Scorpiones: Orthosterni). Euscorpius 11, 1–175.

Sousa, P., Froufe, E., Alves, P.C., Harris, D.J., 2010. Genetic diversity within scorpions

of the genus *Buthus* from the Iberian Peninsula: mitochondrial DNA sequence data indicate additional distinct cryptic lineages. J. Arachnol. 38, 206–211. doi:10.1636/H08-98.1

Sousa, P., Harris, D.J., Froufe, E., van der Meijden, A., 2012. Phylogeographic patterns of *Buthus* scorpions (Scorpiones: Buthidae) in the Maghreb and South-Western Europe based on CO1 mtDNA sequences. J. Zool. 288, 66–75. doi:10.1111/j.1469-7998.2012.00925.x

Stephens, M., Smith, N.J., Donnelly, P., 2001. A new statistical method for haplotype reconstruction from population data. Am. J. Hum. Genet. 68, 978–989. doi:10.1086/319501

Stockwell, S.A., 1989. Revision of the phylogeny and higher classification of the scorpions (Chelicerata). University of California, Berkeley.

Tajima, F., 1989. The effect of change in population size on DNA polymorphism. Genetics 123, 597–601.

Talal, S., Tesler, I., Sivan, J., Ben-Shlomo, R., Muhammad Tahir, H., Prendini, L., Snir, S., Gefen, E., 2015. Scorpion speciation in the Holy Land: Multilocus phylogeography corroborates diagnostic differences in morphology and burrowing behavior among Scorpio subspecies and justifies recognition as phylogenetic, ecological and biological species. Mol. Phylogenet. Evol. 91, 226–237. doi:10.1016/j.ympev.2015.04.028

Tamura, K., Stecher, G., Peterson, D., Filipski, A., Kumar, S., 2013. MEGA6: Molecular evolutionary genetics analysis version 6.0. Mol. Biol. Evol. 30, 2725–2729. doi:10.1093/molbev/mst197

Thomson, R.C., Wang, I.J., Johnson, J.R., 2010. Genome-enabled development of DNA markers for ecology, evolution and conservation. Mol. Ecol. 19, 2184–2195. doi:10.1111/j.1365-294X.2010.04650.x

White, T.J., Bruns, S., Lee, S., Taylor, J., 1990. Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. Pp. 315-322, in: Innis, M.A., Gelfand, D.H., Sninsky, J.J., White, T.J. (Eds.), PCR Protocols: A Guide to Methods and Applications. Academic Press, Inc., New York, p. 482. doi:dx.doi.org/10.1016/B978-0-12-372180-8.50042-1

Yang, Z., Rannala, B., 2012. Molecular phylogenetics: principles and practice. Nat. Rev. Genet. 13, 303–314. doi:10.1038/nrg3186

Yang, Z., Rannala, B., 2010. Bayesian species delimitation using multilocus sequence data. PNAS 107, 9264–9. doi:10.1073/pnas.0913022107

D

A bacterial GH6 cellobiohydrolase with a novel modular structure

Liliana Cerda-Mejía, Susana Valeria Valenzuela, Cristina Frías, Pilar Diaz, F. I. Javier Pastor

2017, Appl Microbiol Biotechnol, 101:2943–2952

Appl Microbiol Biotechnol (2017) 101:2943–2952
DOI 10.1007/s00253-017-8129-4

BIOTECHNOLOGICALLY RELEVANT ENZYMES AND PROTEINS

A bacterial GH6 cellobiohydrolase with a novel modular structure

Liliana Cerda-Mejía[1] · Susana Valeria Valenzuela[1] · Cristina Frías[1] · Pilar Diaz[1] · F. I. Javier Pastor[1]

Received: 9 November 2016 / Revised: 19 December 2016 / Accepted: 28 December 2016 / Published online: 25 January 2017

Abstract Cel6D from *Paenibacillus barcinonensis* is a modular cellobiohydrolase with a novel molecular architecture among glycosyl hydrolases of family 6. It contains an N-terminal catalytic domain (family 6 of glycosyl hydrolases (GH6)), followed by a fibronectin III-like domain repeat ($Fn3_{1,2}$) and a C-terminal family 3b cellulose-binding domain (CBM3b). The enzyme has been identified and purified showing catalytic activity on cellulosic substrates and cellodextrins, with a marked preference for phosphoric acid swollen cellulose (PASC). Analysis of mode of action of Cel6D shows that it releases cellobiose as the only hydrolysis product from cellulose. Kinetic parameters were determined on PASC showing a K_m of 68.73 mg/ml and a V_{max} of 1.73 U/mg. A series of truncated derivatives of Cel6D have been constructed and characterized. Deletion of CBM3b caused a notable reduction in hydrolytic activity, while deletion of the Fn3 domain abolished activity, as the isolated GH6 domain was not active on any of the substrates tested. Mutant enzymes Cel6D-D146A and Cel6D-D97A were constructed in the residues corresponding to the putative acid catalyst and to the network for the nucleophilic attack. The lack of activity of the mutant enzymes indicates the important role of these residues in catalysis. Analysis of cooperative activity of Cel6D with cellulases from the same producing *P. barcinonensis* strain reveals high synergistic activity with processive endoglucanase Cel9B on hydrolysis of crystalline substrates. The characterized cellobiohydrolase can be a good contribution for depolymerization of cellulosic substrates and for the deconstruction of native cellulose.

✉ F. I. Javier Pastor
fpastor@ub.edu

[1] Department of Genetics, Microbiology and Statistics, Faculty of Biology, University of Barcelona, Av. Diagonal 643, 08028 Barcelona, Spain

Keywords Exoglucanase · Modular GH6 · *Paenibacillus*

Introduction

Cellulose, the most abundant component of biomass, is a focus of intense research for the production of biofuels and new biomaterials (Tuck et al. 2012). The production of high-added-value products from this polysaccharide requires its depolymerization or modification by physical, chemical, or biological technologies, among which catalytic breakdown by enzymes can be a key tool for cellulose transformation to novel and sustainable industrial products (Chandel et al. 2012; Delidovich et al. 2014; Hubbe et al. 2015).

The chemical composition of cellulose consists in linear chains of β-1,4-linked glucose molecules connected by extensive hydrogen bonding, which pack a highly crystalline structure recalcitrant to microbial degradation. Depolymerization of cellulose depends on a battery of hydrolytic enzymes, cellulases, which act synergistically to solubilize the polymer. They are classified in endoglucanases, which randomly attack the molecule internally producing new chain ends, cellobiohydrolases, which hydrolyze the chain ends into cellobiose, and β-glucosidases that cleave cellobiose and cellodextrins (Lynd et al. 2002; Bayer et al. 2006). Cellobiohydrolases are exo-type enzymes with preference for the crystalline regions of cellulose, which processively release cellobiose from one end (reducing or nonreducing) of the glucose chains (Teeri 1997). According to their amino acid sequence and hydrophobic cluster analysis, they are classified in several glycosyl hydrolase (GH) families (CAZy, Lombard et al. 2014). Family 6 of glycosyl hydrolases (GH6) comprises several nonreducing end cellobiohydrolases (EC 3.2.1.91) from bacterial and fungal origin. Many of these enzymes display a modular structure that, besides the catalytic

module GH6, comprises a carbohydrate-binding module (CBM) which significantly enhances the activity against crystalline cellulose (Zhang et al. 1995; Tomme et al. 1998). CBMs can widely differ in their binding kinetics and specificity (Carrard et al. 2000; Boraston et al. 2004). They target enzymes to their specific substrates, enhancing carbohydrate degradation as a result of the increased local concentration of each enzyme around its substrate (Hervé et al. 2010; Gilbert et al. 2013). Besides CBMs, glycosyl hydrolases often show a variety of other ancillary modules, such as Fn3, whose role in cellulose hydrolysis by microbial enzymes remains to be understood (Mingardon et al. 2011).

Efficient degradation of crystalline cellulosic substrates requires also a recently described type of enzymes that cleave cellulose by an oxidizing mechanism (Horn et al. 2012). These enzymes, named lytic polysaccharide monooxygenases (LPMOs), have been recently classified in families 9 and 10 of auxiliary activities for carbohydrate depolymerization (AA9 and AA10) (Lombard et al. 2014). They boost the enzymatic degradation of insoluble polysaccharides and are proposed to play an important role in enzyme cocktails to depolymerize native cellulose for biomass valorization (Forsberg et al. 2014).

The enzyme characterized in our work belongs to the multiple enzyme β-glycanase system of *Paenibacillus barcinonensis* BP-23, which shows high potential to degrade polysaccharides (Sánchez et al. 2005). Up to now, several of the cellulases of the strain have been identified, including endoglucanases and exoglucanases (Blanco et al. 1998; Sánchez et al. 2003; Chiriac et al. 2010). Among them, processive endoglucanase Cel9B has been evaluated in biotechnological upgrading of lignocellulosic pulp, showing a biorefining effect, which can produce important energy savings in pulp and paper industries (Cadena et al. 2010). We have characterized in the present work a new enzyme, Cel6D, that shows unique multidomain structure and biochemical properties among cellobiohydrolases. The economy and social interest for the valorization of agricultural, industrial, and urban wastes stimulate the search for new types of cellulose-depolymerizing and/or cellulose-modifying enzymes, among which cellulases with novel properties can make an important contribution. The properties of the enzyme identified make it a good candidate for the design of novel enzymatic cocktails to deconstruct and upgrade plant biomass.

Materials and methods

Bacterial strains and plasmids

The DNA fragment encoding the complete cellobiohydrolase gene, *cel6D* (KY050765), from *P. barcinonensis* BP-23 genomic DNA (CECT 7022; DSM 15478) was PCR amplified (Phusion High-Fidelity DNA Polymerase, Thermo Scientific™, Rockford, IL, USA) with oligonucleotide primers FwCel6D (5′-CTTT*CCATGG*TGCGTAAAGGTGTA-3′) and BwCel6D (5′-GGCCGC*AAGCTT*TGGCTCAATGCCCCA-3′) (restriction sites in italics). The amplified product was inserted between the *Nco*I and *Hin*dIII sites of pET28a (Novagen, Madison, WI, USA), to generate the plasmid pET28aCel6D that was introduced into *Escherichia coli* BL21 star (DE3) (Invitrogen, Carlsbad, CA, USA) to express the recombinant enzyme linked to a C-terminal His tag. The same strategy was used to clone and express the truncated forms of the enzyme. Oligonucleotide primers FwCel6D and BwFn3 (5′-GGCCGC*AAGCTT*ACTTGTCGGTACAAC-3′) were used to give rise to the truncated enzyme GH6-Fn3, while FwCel6D and BwGH6 (5′-GGCCGC*AAGCTT*CACCAGCTGTACAAA-3′) were used to express the GH6 isolated catalytic module. For the expression of CBM3, primers FwCBM3 (5′-AAGT*CCATGG*ACTTGGTACTGCAA-3′) and BwCel6D were used.

Mutants Cel6D-D97A and Cel6D-D146A were constructed by QuikChange® (Agilent, Santa Clara, CA, USA) from plasmid pET28aCel6D. The oligonucleotide primers used were FwD97 (5′- TTGCCGGGCCGGGcTTGTCATGCCCTCG CATCTAACGGGGAGCTT-3′) and BwD97 (5′-GAGG GCATGACAAgCCCGGCCCGGCAAATTATAAATAACA AA-3′); and FwD146 (5′-ATTATTGAACCGGcCAGTCTGC CGAATCTGGTAACGAACCTTAGT-3′) and BwD146 (5′-ATTCGGCAGACTGgCCGGTTCAATAATGGCAATG ATCCGAATGTC-3′), respectively. The recombinant plasmids obtained were cloned in *E. coli* BL21 star (DE3). FastDigest restriction nucleases and T4 DNA ligase (Thermo Scientific™, Rockford, IL, USA) were used according to the manufacturer specifications. All DNA constructs were verified by sequencing. Sequence homology was analyzed by BLAST (https://blast.ncbi.nlm.nih.gov/Blast.cgi).

Expression and purification of recombinant proteins in *E. coli*

Cel6D, GH6-Fn3, GH6, CBM3, Cel6D-D146A, and Cel6D-D97A were purified from cell extracts of the corresponding *E. coli* BL21 star (DE3) recombinant clones by immobilized metal affinity chromatography (IMAC) on a fast protein liquid chromatography system (ÄKTA FPLC, GE Healthcare, Uppsala, Sweden) as described (Valenzuela et al. 2016). Buffer exchange and protein concentration were performed in Centricon centrifugal filter units of 3-kDa molecular mass cutoff (Millipore, Darmstadt, Germany). Protein concentration was determined by Bradford (1976) using bovine serum albumin as the standard. Additionally, it was quantified measuring absorbance at 280 nm by NanoDrop® ND-1000 (NanoDrop Technologies, Thermo Fisher Scientific, Waltham, MA, USA). Proteins were analyzed by sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) (Laemmli 1970).

Enzyme assays

Cellulase activity was assayed by measuring the amount of reducing sugar released from cellulose using the method of Nelson and Somogyi (Spiro 1966). The assay mixture contained 1.5% phosphoric acid swollen cellulose (PASC), obtained from Avicel PH-101 (Fluka, St. Louis, USA) treated with 70% PO_4H_3 (Wood 1988), in a final volume of 0.1 ml of 50 mM acetate buffer (pH 4.0). The mixture was incubated at 47 °C for 30 min. Color development was measured at 520 nm. To test for substrate specificity, PASC was replaced by Avicel PH-101 (Fluka), carboxymethyl cellulose (CMC) or *p*-nitrophenyl-cellobioside *p*NPG2 (Sigma-Aldrich, St. Louis, USA), regenerated amorphous cellulose (RAC), and filter paper (Whatman no. 1, Chicago, IL, USA). RAC was prepared as described by Zhang et al. (2006). One unit of enzymatic activity was defined as the amount of enzyme that releases 1 μmol of reducing sugar equivalent per minute under the assay conditions described.

The effect of temperature and pH on cellulase activity was evaluated by response surface methodology (RSM) as described (Padilha et al. 2014). The Briton-Robinson buffer in a pH range between 2.6 and 5.4 and temperatures ranging from 36 to 64 °C were used in the study. The effect of pH and temperature on cellulase stability was evaluated by incubating Cel6D in the Briton-Robinson buffer (Britton 1952) at different pH values ranging from 2.0 to 10.0 and temperatures ranging from 25 to 55 °C for intervals of up to 3 h. The residual activity was then determined by the standard assay. All the activity and stability values shown are the means of at least three replicates of two independent experiments. The influence of metal ions and chemicals on cellulase activity was determined as described by Padilha et al. (2014).

The kinetics of Cel6D was characterized in terms of Michaelis–Menten kinetic constants (K_m and V_{max}) by assaying the enzyme activity on PASC concentrations ranging from 1.25 to 40.0 mg/ml in 50 mM phosphate buffer (pH 3.9) at 47 °C for 30 min. The study of enzyme kinetics was done using the Graph Pad software version 4.0.

Cellulose-binding properties were evaluated by mixing the purified enzyme or CBM3 with Avicel PH-101 at 5% final concentration in 0.5 ml of 50 mM acetate buffer (pH 4.0) and keeping the mixture at 4 °C with gentle shaking for 1 h. Samples were then centrifuged and supernatants collected. Pellets were washed three times in buffer before being resuspended in 0.5 ml of 10% SDS and boiled for 10 min to release bound protein. Samples were then analyzed by SDS-PAGE.

Thin-layer chromatography

Of purified Cel6D, 0.6 μg were incubated with 1.5% PASC or 0.6 mg/ml of cellobiose, cellotriose, cellotetraose, or cellopentaose in 50 mM acetate buffer (pH 4.0) at 47 °C; the samples were taken at regular intervals, centrifuged and supernatants kept frozen until spotted on silica gel plates (Kieselgel 20 F254 20 × 20 cm, Merck, Darmstadt, Germany). The solvent used was chloroform/acetic acid/water (3:6:1, *v*/*v*/*v*). Oligosaccharides were detected by spraying the plates with an ethanol/concentrated sulfuric acid mixture (95:5, *v*/*v*) and heating at 120 °C.

Analysis of synergism

To evaluate the synergistic effect of Cel6D with *P. barcinonensis* cellulases, endoglucanases Cel5A and Cel9B were previously produced. Cel5A samples were cell extracts from the recombinant clone *E. coli* 5K/pC11 overproducing the enzyme (Blanco et al. 1998), while purified Cel9B was obtained from *E. coli* BLR(DE3)/pET28aCel9B cell extracts by binding and elution from Avicel, as previously described (Chiriac et al. 2010). To determine the enzyme dosage to be used in the experiments, preliminary tests were performed and kinetic curves were generated to verify the minimum enzyme concentration required to obtain measurable reducing sugars after 30 min of reaction.

For the synergism assays, 0.2 μmol of Cel6D and Cel9B were mixed in a final volume of 0.4 ml of 50 mM acetate buffer containing 1.25% PASC or 1.125% Avicel, CMC, or filter paper. The mixtures were incubated at 50 °C for 1 h, and the amount of sugar released was determined. Synergistic activity between Cel6D and Cel5A contained 11 μg of each enzyme preparation.

Results

Cloning and biochemical characterization of cellobiohydrolase Cel6D

Genomic analysis of *P. barcinonensis* BP-23 identified an open reading frame showing high identity to glycosyl hydrolases of family 6 (Fig. 1). To evaluate its functionality, it was PCR amplified and cloned under the control of the high expression T7 promoter from plasmid pET28a in *E. coli* BL21 star (DE3). Extracts from the recombinant clone showed hydrolytic activity on phosphoric acid swollen cellulose (PASC), indicating that the identified open reading frame coded for an

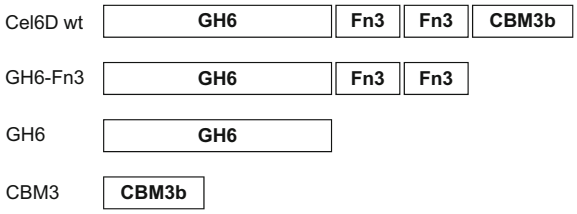


Fig. 1 Modular structure of Cel6D and derived truncated enzymes

active enzyme, which was named Cel6D. The recombinant enzyme contained a fused C-terminal His-tail that facilitated its purification by affinity chromatography to His-trap columns. Electrophoretic analysis showed that the enzyme exhibited an apparent molecular mass of 84 kDa, in accordance with theoretical molecular weight of the mature protein deduced from sequence (83,837.9 Da) (Fig. 2).

Substrate specificity of Cel6D was determined by evaluating the release of reducing sugars from different substrates such as CMC, PASC, RAC, and filter paper, as well as from barley glucan, lichenan, xylans, starch, pectin, and *p*NPG2. Hydrolytic activity of the enzyme was only found on cellulosic substrates. Among these substrates, Cel6D showed the highest activity on PASC (0.66 U/mg) while on Avicel and RAC showed much lower activity (0.01 U/mg). On the contrary, Cel6D did not hydrolyze CMC or filter paper, nor it showed activity on the other substrates tested (Table 1).

Hydrolysis products from PASC were analyzed by thin-layer chromatography, which showed that cellobiose was the only product released from this substrate (Fig. 3). Cellobiose was also found as only hydrolysis product from Avicel and RAC (data not shown). To evaluate the activity of the enzyme on cellooligosaccharides, products released from these oligomers were also analyzed in thin-layer chromatograms. Cel6D was not active on cellobiose, while cellotriose, cellotetraose, and cellopentaose were cleaved to cellobiose as the main product (Fig. 3). This mode of action of the enzyme, which liberates cellobiose as only main reaction product, together with the preference for insoluble cellulose, indicates that Cel6D is a cellobiohydrolase.

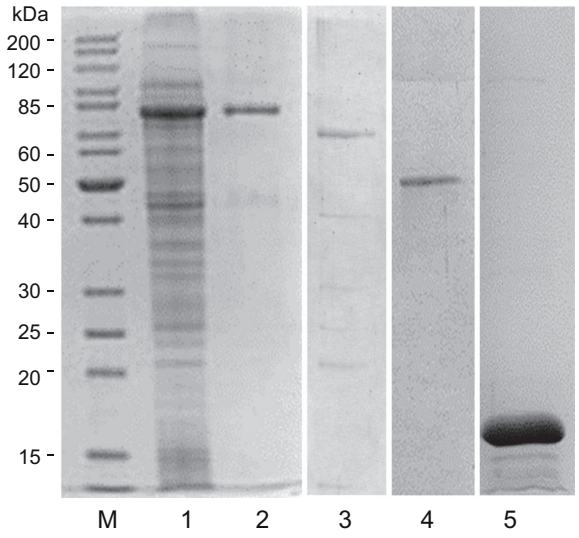


Fig. 2 SDS-PAGE analysis of Cel6D and derived truncated enzymes. *Lanes*: *1*, cell extracts of recombinant *E. coli* strain expressing Cel6D; *2*, purified Cel6D; *3*, purified GH6-Fn3; *4*, purified GH6; *5*, purified CBM3; *M*, molecular mass standard proteins

Table 1 Substrate specificity of Cel6D and truncated derivatives

Substrate	Specific activity (U/mg)		
	Cel6D wt	GH6-Fn3	GH6
PASC	0.66	0.07	ND
Avicel	0.01	ND	ND
RAC	0.01	ND	ND
CMC	ND	ND	ND
Filter paper	ND	ND	ND
*p*NPG2	ND	ND	ND

ND no activity detected

The influence of temperature and pH on activity of purified Cel6D was determined by RSM. The enzyme showed maximum activity at 47 °C and pH 3.9. Analysis of stability of the enzyme at different conditions showed that it retained more than 60% of activity after 3-h incubation at 47 °C in the pH range from 4 to 6, while at 55 °C, it was completely inactivated after 3-h incubation at pH 4. Kinetic constants on PASC were determined at optimal conditions of activity. The enzyme showed a K_m of 68.73 mg/ml and a V_{max} of 1.73 U/mg. The influence of metal ions and chemical agents on activity of the enzyme was also determined. Cel6D was completely inhibited by 1 mM Hg^{+2}, 0.5% Tween 80, and Triton X-100, while 1 mM Ca^{+2} and Li^{+1} produced a small stimulating effect.

Sequence analysis and protein engineering

Analysis of amino acid deduced sequence of Cel6D showed it is a multidomain enzyme containing a catalytic module of GH6, followed by a duplicated fibronectin-like domain (Fn3) and a carbohydrate-binding module of family 3b (CBM3) at the C-terminal portion of the enzyme. This modular structure is novel among GH6 cellobiohydrolases, with only another example recently described in Cel6A from *Paenibacillus curdlanolyticus* B-6 (Baramee et al. 2016). Cel6D shows an N-terminal sequence of 30 amino acids with the features of a signal peptide which would direct the enzyme to the extracellular compartment of the producing bacteria, *P. barcinonensis*. A comparison of the catalytic module of Cel6D with sequences contained in databases showed that it had a maximum identity (72%) to the catalytic module of *P. curdlanolyticus* Cel6A. Catalytic modules of *Cellulomonas fimi* Cel6B (Meinke et al. 1994), *Thermobifida fusca* YX Cel6B (Zhang et al. 1995), *Streptomyces* sp. M23 CBHII (Park et al. 2005), and *Trichoderma reesei* QM9414 Cel6A (Fägerstam and Pettersson 1980) showed 58, 48, 47, and 34% identity to the catalytic module of Cel6D, respectively. Analysis of amino acid sequence also showed that CBM of Cel6D showed

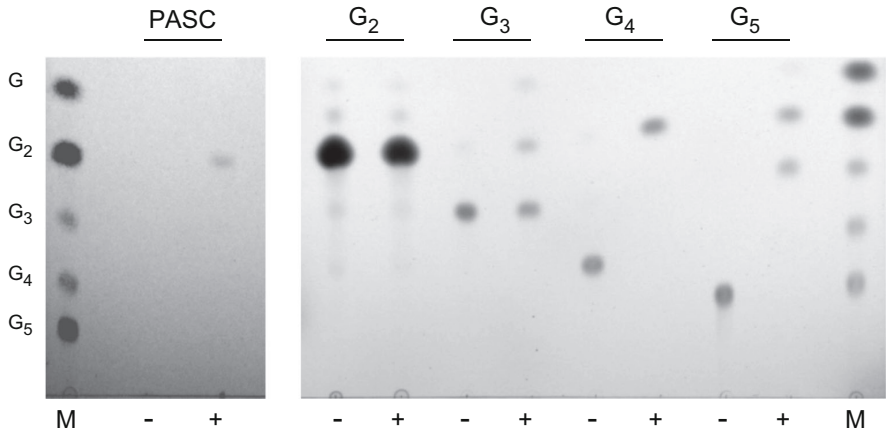


Fig. 3 Thin-layer chromatograms of hydrolysis products from PASC and cellooligosaccharides. Purified Cel6D (0.6 μg) was mixed with 1.5% PASC or 0.6 mg/ml of cellobiose, cellotriose, cellotetraose, or cellopentaose in 150 mM phosphate buffer (pH 4.0) and incubated at 47 °C for 0 (−) or 30 min (+). *Lanes M* contain size markers of glucose (G1), cellobiose (G2), cellotriose (G3), cellotetraose (G4), and cellopentaose (G5)

maximum identity of 55% to CBM3 of *P. curdlanolyticus* Cel6A (Baramee et al. 2016) and 42% identity to CBM3 of *Clostridium thermocellum* (Yaniv et al. 2012).

Amino acid sequence of Cel6D was aligned with cellobiohydrolases of family GH6, including the deeply characterized enzymes Cel6A from *T. reesei* and Cel6B from *T. fusca*. Cel6D showed several of the conserved residues in these enzymes, including the two aspartic acid residues directly involved in hydrolysis as acid and base catalysts and a third aspartic acid residue proposed to participate in the proton-transferring network for the nucleophile attack (Vuong and Wilson 2009b; Sandgren et al. 2013). Among these residues, D146, corresponding to the putative acid catalyst of Cel6D, and D97, putative residue contributing to proton transfer, were mutated to alanine residues to confirm their role in catalytic activity. The mutant enzymes constructed, Cel6D-D146A and Cel6D-D97A, were purified from soluble extracts and analyzed by SDS-PAGE, confirming their correct expression (data not shown). However, none of them showed activity on PASC or other celluloses, clearly indicating the implication of these two Asp residues in catalytic activity.

To check the contribution of the different modules of Cel6D to enzyme activity, three truncated derivatives of the enzyme were constructed. They were composed of the enzyme devoid of its CBM (GH6-Fn3), the isolated catalytic domain (GH6), or the isolated CBM (CBM3) (Fig. 1). The truncated forms of the enzyme were purified and analyzed by SDS-PAGE showing apparent molecular sizes of 69, 50, and 15 kDa, respectively, in accordance to their theoretical molecular weight deduced from sequence (Fig. 2). Truncated cellulase GH6-Fn3 showed a very diminished activity on PASC when compared to wild-type enzyme, while activity of Avicel or RAC was undetectable, indicating the important contribution of the CBM3 to enzyme activity. On its side, the isolated catalytic domain GH6 did not show activity on any of the substrates tested (Table 1). The activity of truncated enzyme GH6-Fn3 on soluble cellooligosaccharides was evaluated by TLC analysis of hydrolysis products. The truncated enzyme was still active on these substrates, giving a similar product pattern to that of wild-type enzyme (data not shown).

The ability of intact Cel6D and of the isolated CBM3 to bind crystalline cellulose was examined. The purified enzymes were mixed with Avicel and incubated in binding conditions, and unbound and bound fractions were collected and analyzed by SDS-PAGE (Fig. 4). Wild-type Cel6D appeared as a prominent protein band in the fraction bound to Avicel, while in the unbound fraction, it was almost undetectable. The isolated CBM3 show similar binding properties, although binding was not as complete as in the wild-type enzyme, as a small amount of CBM3 was found in the unbound fraction (Fig. 4).

Evaluation of synergism with endoglucanases

The cooperation on cellulose depolymerization between Cel6D and endoglucanases belonging to the cellulolytic system of *P. barcinonensis* was evaluated. The enzymes tested in the synergism analysis were Cel5A, a single domain enzyme, and Cel9B, a processive endoglucanase of modular structure, previously characterized (Blanco et al. 1998; Chiriac et al. 2010).

As first approach, we tested the simultaneous action on PASC of the combination of Cel6D and processive endoglucanase Cel9B. The treatment was performed at pH 4.0 and 50 °C, conditions in which both enzymes show more than 80% of maximum activity. Simultaneous activity of the enzymes released a higher amount of reducing sugars from PASC than the sum of the amount of sugars released by each enzyme separately (Fig. 5). The effect was notable up to the first hour of treatment (1.9-fold), but at longer incubation times, no increment in the sugar release was detected, probably due to enzyme inhibition by reaction products. The results indicate a synergistic effect of Cel6D with endoglucanase Cel9B on PASC degradation.

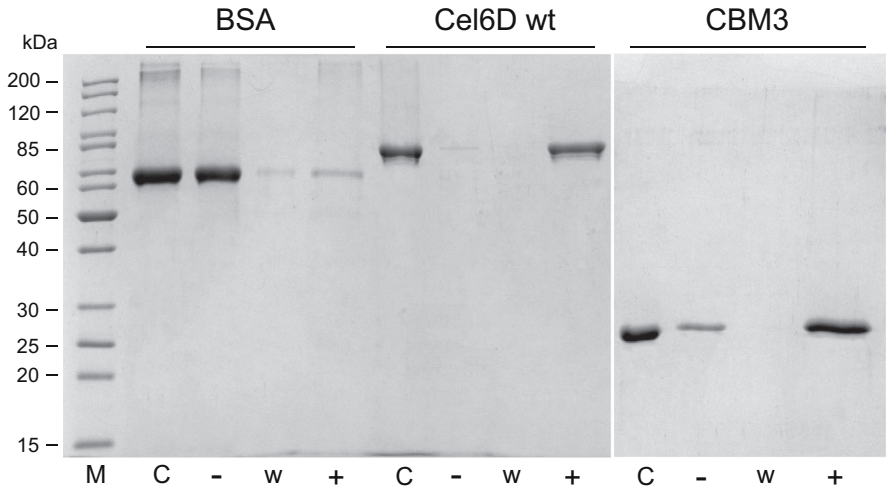


Fig. 4 SDS-PAGE analysis of binding to Avicel. Cel6D or CBM3 were mixed with Avicel for 1 h, and bound and unbound fractions were separated by centrifugation and analyzed by SDS-PAGE. *Lanes*: (−), unbound fraction; (*w*), wash; (+), bound fraction; *C*, control protein. Bovine serum albumin (BSA) was used as a binding control. The positions of molecular mass standard proteins are indicated

To explore the cooperativity on other cellulosic substrates, the combined effect of the enzymes was also tested on Avicel, filter paper, and CMC. On filter paper and Avicel, substrates on which the enzymes were not able to release sugars individually or released an almost undetectable amount (as the case of Cel6D on Avicel), the combined action of the enzymes released an important amount of sugars. The results indicate the important synergism between these two cellulases on crystalline celluloses. On the contrary, the simultaneous action of the two enzymes on CMC, an amorphous cellulose substrate not cleaved by Cel6D, released lower amount of sugars than those released by the individual action of Cel9B. This suggests an inhibition of endoglucanase Cel9B activity on CMC by cellobiohydrolase Cel6D. Evaluation of the synergistic effect of the two enzymes at pH 5.0, closer to optimum pH of Cel9B, gave similar results to those found at pH 4.0. To evaluate the cooperative effect of Cel6D and Cel9B on pretreated biomass, the individual action of the cellulases on *Eucalyptus* bleached pulps was tested and compared with the combined action of the enzymes. Cel9B released high amount of reducing sugars from pulp, while Cel6D released a minor amount. However, the combined action of the enzymes released 1.5-fold the amount of sugar released by the enzymes separately, indicating a synergistic effect on pulp, in agreement with the results found on insoluble cellulases.

The cooperation of Cel6D with Cel5A, a single domain endoglucanase, on PASC hydrolysis was also evaluated as described before, although at pH 4.0 and 40 °C, optimal conditions for activity of endoglucanase Cel5A and in which Cel6D has more than 80% of maximum activity. The simultaneous application of the enzymes resulted in similar amount of sugars released from PASC than those liberated by the enzymes separately, clearly showing the lack of synergism of these two enzymes on cellulose degradation.

Discussion

We have identified and characterized Cel6D from *P. barcinonensis* BP-23, a new cellobiohydrolase belonging to GH6. The enzyme is a multidomain cellulase with the modular structure GH6-Fn3$_{1,2}$-CBM3, which is novel among

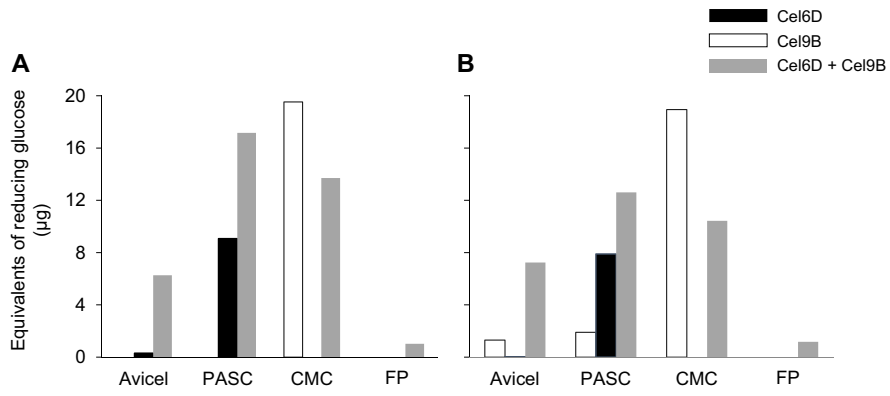


Fig. 5 Evaluation of synergistic activity between Cel6D and processive endoglucanase Cel9B. Reducing sugars released from Avicel, PASC, CMC, or filter paper (FP) by Cel6D, Cel9B, or the combined action of Cel6D and Cel9B. The substrates were mixed with 0.2 μmol of each of the enzymes and incubated for 1 h at 50 °C in 50 mM acetate buffer at pH 4.0 (**a**) or 5.0 (**b**)

cellobiohydrolases of family GH6. Only recently, an enzyme showing similar structure has been characterized in *P. curdlanolyticus* (Baramee et al. 2016). The enzyme cellobiohydrolase Cel6A shows high identity to *P. barcinonensis* Cel6D (66% identity). Cel6D also shows high identity with several putative cellulases of *Paenibacillus* and *Bacillus* contained in databases, although they have not been characterized up to date. Most of characterized GH6 cellobiohydrolases belong to bacteria phylogenetically distant from *Paenibacillus* or to fungi. Among them, two representative and deeply studied enzymes are Cel6B from *T. fusca* (*Tf*Cel6B) and Cel6A from *T. reesei* (*Tr*Cel6A). Similarly to Cel6D, they are multidomain cellobiohydrolases, although showing a different domain structure and composition. They show a CBM2 (*Tf*Cel6B) or a CBM1 (*Tr*Cel6A) at the N terminal side of the enzymes, while Cel6D has a CBM of a different family, CBM3, located at the C terminus of the enzyme, and besides, it has a duplicated Fn3 at a central position of the modular structure. These enzymes differ also from Cel6D in substrate specificity. *Tr*Cel6A shows high activity on Avicel and PASC and is also active on CMC (Tomme et al. 1988; Koivula et al. 1998; Poidevin et al. 2013), while *Tf*Cel6B shows maximum activity on PASC, lower activity on CMC and is active on filter paper (Zhang et al. 1995; Watson et al. 2002). On its side, Cel6D shows a narrower substrate range with a marked preference for PASC and is not active on CMC or filter paper. Substrate specificity of Cel6D resembles that of *P. curdlanolyticus* Cel6A, which shows also high activity on PASC and does not hydrolyze CMC. However, the enzymes differ in activity on cellooligosaccharides. While Cel6D hydrolyzes cellotriose and cellotetraose, *P. curdlanolyticus* Cel6A does not hydrolyze cellotriose and is not able to cleave cellotetraose completely (Baramee et al. 2016). The different activity of the enzymes on cellooligomers is probably related to differences in their catalytic domain, which show 72% identity. The two *Paenibacillus* enzymes are the only examples of GH6 cellobiohydrolases with a CBM of family 3. CBMs of this family, together with the CBMs of families 1 and 2, found in *Tr*Cel6A and *Tf*Cel6B belong to the type A of CBMs, which promote surface binding to highly crystalline celluloses (Boraston et al. 2004). However, subtle differences in structure and specificity have been reported among them (Tormo et al. 1996; Tomme et al. 1998). In fact, even in family 3 of CBMs, several differences are found, which have prompted the classification of its members in three subfamilies. One of them, CBM3c, is considered as an auxiliary domain without binding properties (Tormo et al. 1996; Jindou et al. 2006). The CBM3 of Cel6D belongs to a different subfamily, CBM3b, which has been shown to bind crystalline and amorphous cellulose. We have shown its binding to crystalline cellulose and that it plays an important role in catalysis as its deletion from the enzyme causes a remarkable reduction of activity (90% of Cel6D wt). On the contrary, the effect of CBM deletion on *Tr*Cel6A and *Tf*Cel6B activity is less pronounced as the truncated enzymes retained more than 50% of activity on most of their substrates (Tomme et al. 1988; Zhang et al. 1995). There are several examples of endoglucanases with a CBM3b, such as Cel9B of *P. barcinonensis* (Chiriac et al. 2010) and CelI from *C. thermocellum* (Gilad et al. 2003), where their deletion causes important decrease of activity on crystalline substrates. Regarding Fn3 modules, there are no published data on their influence on enzyme activity, although they have been reported to modify cellulose surface (Kataeva et al. 2002), suggesting that they can facilitate cellulose depolymerization. Our results showing that Fn3 deletion from the truncated enzyme GH6-Fn3 abolishes the activity of the isolated GH6 catalytic module indicate that Fn3 modules, so frequent in bacterial glycosyl hydrolases (Ficko-Blean et al. 2009), contribute to enzymatic hydrolysis of cellulose. The results obtained indicate the important contribution of CBM3 and Fn3 domains to enzymatic activity of Cel6D. However, the lack of reported data on activity of engineered derivatives of *P. curdlanolyticus* Cel6A precludes to ascertain a general role of these domains in GH6 cellobiohydrolases.

Sequence alignment of the catalytic module of Cel6D to those of *Tf*Cel6B and *Tr*Cel6A identified several residues in the enzyme that are conserved among GH6 cellobiohydrolases, including *P. curdlanolyticus* Cel6A (Vuong and Wilson 2009b; Baramee et al. 2016). To check their role in catalysis, mutant enzymes Cel6D-D146A and Cel6D-D97A were constructed, in which the corresponding aspartic acid residues were changed to alanine residues. These amino acids correspond to the putative acid catalyst and to a residue participating in the nucleophilic attack. None of the mutant enzymes showed catalytic activity. The results obtained with the mutant Cel6D-D146A indicate that aspartic acid 146 is the acid catalyst of Cel6D, in agreement with the results on active site characterization of *Tf*Cel6B and *Tr*Cel6A (Koivula et al. 2002; Vuong and Wilson 2009b). However, the results with the second mutant, Cel6D-D97A, are not in accordance with those reported for *Tf*Cel6B and *Tr*Cel6A mutants. While mutation in this residue causes a total loss of activity in Cel6D, the corresponding mutants of *Tr*Cel6A and *Tf*Cel6B retain activity although notably diminished (Vuong and Wilson 2009b; Sandgren et al. 2013). In fact, this residue was initially postulated as the catalytic base of these enzymes (Wohlfahrt et al. 2003), although it was later proposed to participate in the nucleophilic attack by two water molecules by a novel hydrolysis mechanism that seems to require a proton-transferring network involving several conserved residues, which carry out the catalytic base function (Vuong and Wilson 2009b). All these conserved residues, including a serine close to the aspartic acid residue mentioned, are present in Cel6D. The difference in activity found could be related to the low identity of the catalytic domain of Cel6B to those of *Tf*Cel6B and *Tr*Cel6A (48 and 34%, respectively) or to some unidentified trait of the enzyme.

Analysis of the cooperative activity of Cel6D with cellulases from the *P. barcinonensis* producing strain revealed that the enzyme shows synergism with endoglucanase Cel9B, a processive endoglucanase of modular structure, on depolymerization of PASC, Avicel, and filter paper. On the contrary, no cooperativity effect was found with endoglucanase Cel5A. Synergism found with Cel9B resembles previous results showing the cooperative action of this *P. barcinonensis* endoglucanase with cellobiohydrolase *Tf*Cel6B (Sánchez et al. 2004). Our results are also in accordance with the studies of synergism between GH6 and GH9 cellulases from *T. fusca*, which show synergistic activity between *Tf*Cel6B and processive endoglucanase *Tf*Cel9A (Watson et al. 2002). The synergistic effect of *P. curdlanolyticus* Cel6A with endoglucanase Cel9R from *C. thermocellum* on PASC and Avicel degradation has been reported (Baramee et al. 2016), although showing lower synergistic ratio (1.2–1.5) than that we have found for the combined action of *P. barcinonensis* Cel6D and Cel9B. In fact, Avicel, which was hardly hydrolyzed by the individual enzymes, was efficiently hydrolyzed by the joint action of Cel6D and Cel9B. Additionally, substrates not degraded by these individual enzymes, as filter paper, were hydrolyzed by their cooperative activity, clearly indicating the remarkable synergistic activity of Cel6D and Cel9B. Synergism between cellobiohydrolases and endoglucanases is a common feature of enzymatic depolymerization of cellulose, where endoglucanases create new ends for cellobiohydrolase activity, fostering substrate degradation. However, several other factors, besides exo or endo mode action of the enzymes, can affect synergism. Among them, processivity can differently influence synergistic activity between enzymes, depending on their particular traits (Vuong and Wilson 2009a). The lack of synergism found with Cel5A could be related to the nonprocessive mode of action of the endoglucanase. The enzyme that we have identified in our work is one of the few examples of characterized bacterial cellobiohydrolases of family GH6. Cel6D shows, together with the recently described Cel6A from *P. curdlanolyticus*, a unique modular structure which differs widely from that of cellobiohydrolases from this family. However, Cel6D shows distinctive traits and catalytic properties. The synergistic effect found on cellulose and pulp depolymerization makes Cel6D a good candidate for biomass transformation, where new cellulases and auxiliary activity enzymes are needed to formulate enzyme cocktails for bioethanol production and to upgrade lignocellulosic materials into biotechnological products.

Acknowledgements This work was partially supported by the Spanish Ministry of Economy and Competitivity, grant no. CTQ2013-48995-C2-2-R. Liliana Cerda-Mejía held a grant SENESCYT (Ecuador). The experiments described in this article have been performed complying with the Spanish current laws.

Compliance with ethical standards

Conflict of interest The authors declare that they have no competing interests.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

Baramee S, Teeravivattanakit T, Phitsuwan P, Waeonukul R, Pason P, Tachaapaikoon C, Kosugi A, Sakka K, Ratanakhanokchai K (2016) A novel GH6 cellobiohydrolase from *Paenibacillus curdlanolyticus* B-6 and its synergistic action on cellulose degradation. Appl Microbiol Biotechnol 1–14. doi: 10.1007/s00253-016-7895-8

Bayer EA, Shoham Y, Lamed R (2006) Cellulose-decomposing bacteria and their enzyme systems. In: Dworkin M, Falkow S, Rosenberg E, Schleifer K-H, Stackebrandt E (eds) The prokaryotes. Vol. 2: ecophysiology and biochemistry. Springer, New York, pp 578–617

Blanco A, Díaz P, Martínez J, Vidal T, Torres AL, Pastor FI (1998) Cloning of a new endoglucanase gene from *Bacillus* sp. BP-23 and characterisation of the enzyme. Performance in paper manufacture from cereal straw. Appl Microbiol Biotechnol 50:48–54

Boraston AB, Bolam DN, Gilbert HJ, Davies GJ (2004) Carbohydrate-binding modules: fine-tuning polysaccharide recognition. Biochem J 382:769–781. doi:10.1042/BJ20040892

Bradford MM (1976) A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. Anal Biochem 72:248–254

Britton HTS (1952) Hydrogen ions, 2nd edn. Chapman and Hall, London

Cadena EM, Chriac AI, Pastor FIJ, Diaz P, Vidal T, Torres AL (2010) Use of cellulases and recombinant cellulose binding domains for refining TCF kraft pulp. Biotechnol Prog 26:960–967. doi:10.1002/btpr.411

Carrard G, Koivula A, Söderlund H, Béguin P (2000) Cellulose-binding domains promote hydrolysis of different sites on crystalline cellulose. Proc Natl Acad Sci U S A 97:10342–10347. doi:10.1073/pnas.160216697

Chandel AK, Chandrasekhar G, Silva MB, Silvério da Silva S (2012) The realm of cellulases in biorefinery development. Crit Rev Biotechnol 32:187–202. doi:10.3109/07388551.2011.595385

Chiriac AI, Cadena EM, Vidal T, Torres AL, Diaz P, Javier Pastor FI, Pastor FIJ (2010) Engineering a family 9 processive endoglucanase from *Paenibacillus barcinonensis* displaying a novel architecture. Appl Microbiol Biotechnol 86:1125–1134. doi:10.1007/s00253-009-2350-8

Delidovich I, Leonhard K, Palkovits R (2014) Cellulose and hemicellulose valorisation: an integrated challenge of catalysis and reaction engineering. Energy Environ Sci 7:2803–2830. doi:10.1039/C4EE01067A

Fägerstam LG, Pettersson LG (1980) The 1,4-β-glucan cellobiohydrolases of *Trichoderma reesei* QM 9414. FEBS Lett 119:97–100. doi:10.1016/0014-5793(80)81006-4

Ficko-Blean E, Gregg KJ, Adams JJ, Hehemann J-H, Czjzek M, Smith SP, Boraston AB (2009) Portrait of an enzyme, a complete structural analysis of a multimodular β-N-acetylglucosaminidase from *Clostridium perfringens*. J Biol Chem 284:9876–9884. doi:10.1074/jbc.M808954200

Forsberg Z, Mackenzie AK, Sorlie M, Rohr AK, Helland R, Arvai AS, Vaaje-Kolstad G, Eijsink VGH (2014) Structural and functional characterization of a conserved pair of bacterial cellulose-oxidizing

lytic polysaccharide monooxygenases. Proc Natl Acad Sci 111: 8446–8451. doi:10.1073/pnas.1402771111

Gilad R, Rabinovich L, Yaron S, Bayer EA, Lamed R, Gilbert HJ, Shoham Y (2003) Cell, a noncellulosomal family 9 enzyme from *Clostridium thermocellum*, is a processive endoglucanase that degrades crystalline cellulose. J Bacteriol 185:391–398. doi:10.1128/JB.185.2.391-398.2003

Gilbert HJ, Knox JP, Boraston AB (2013) Advances in understanding the molecular basis of plant cell wall polysaccharide recognition by carbohydrate-binding modules. Curr Opin Struct Biol 23:669–677. doi:10.1016/j.sbi.2013.05.005

Hervé C, Rogowski A, Blake AW, Marcus SE, Gilbert HJ, Knox JP (2010) Carbohydrate-binding modules promote the enzymatic deconstruction of intact plant cell walls by targeting and proximity effects. Proc Natl Acad Sci U S A 107:15293–15298. doi:10.1073/pnas.1005732107

Horn S, Vaaje-Kolstad G, Westereng B, Eijsink VG (2012) Novel enzymes for the degradation of cellulose. Biotechnol Biofuels 5:45. doi:10.1186/1754-6834-5-45

Hubbe MA, Rojas OJ, Lucia LA (2015) Green modification of surface characteristics of cellulosic materials at the molecular or nano scale: a review. Bioresources 10:6095–6206

Jindou S, Xu Q, Kenig R, Shulman M, Shoham Y, Bayer EA, Lamed R (2006) Novel architecture of family-9 glycoside hydrolases identified in cellulosomal enzymes of *Acetivibrio cellulolyticus* and *Clostridium thermocellum*. FEMS Microbiol Lett 254:308–316. doi:10.1111/j.1574-6968.2005.00040.x

Kataeva IA, Seidel RD, Shah A, West LT, Li X-L, Ljungdahl LG (2002) The fibronectin type 3-like repeat from the *Clostridium thermocellum* cellobiohydrolase CbhA promotes hydrolysis of cellulose by modifying its surface. Appl Environ Microbiol 68:4292–4300. doi:10.1128/AEM.68.9.4292-4300.2002

Koivula A, Kinnari T, Harjunpää V, Ruohonen L, Teleman A, Drakenberg T, Rouvinen J, Jones TA, Teeri TT (1998) Tryptophan 272: an essential determinant of crystalline cellulose degradation by *Trichoderma reesei* cellobiohydrolase Cel6A. FEBS Lett 429:341–346

Koivula A, Ruohonen L, Wohlfahrt G, Reinikainen T, Teeri TT, Piens K, Claeyssens M, Weber M, Vasella A, Becker D, Sinnott ML, Zou JY, Kleywegt GJ, Szardenings M, Ståhlberg J, Jones TA (2002) The active site of cellobiohydrolase Cel6A from *Trichoderma reesei*: the roles of aspartic acids D221 and D175. J Am Chem Soc 124: 10015–10024. doi:10.1021/ja012659q

Laemmli UK (1970) Cleavage of structural proteins during the assembly of the head of bacteriophage T4. Nature 227:680–685

Lombard V, Golaconda Ramulu H, Drula E, Coutinho PM, Henrissat B (2014) The carbohydrate-active enzymes database (CAZy) in 2013. Nucleic Acids Res 42:D490–D495. doi:10.1093/nar/gkt1178

Lynd LR, Weimer PJ, van Zyl WH, Pretorius IS (2002) Microbial cellulose utilization: fundamentals and biotechnology. Microbiol Mol Biol Rev 66:506–577. doi:10.1128/MMBR.66.3.506-577.2002

Meinke A, Gilkes NR, Kwan E, Kilburn DG, Warren RA, Miller RC (1994) Cellobiohydrolase A (CbhA) from the cellulolytic bacterium *Cellulomonas fimi* is a β-1,4-exocellobiohydrolase analogous to *Trichoderma reesei* CBH II. Mol Microbiol 12:413–422

Mingardon F, Bagert JD, Maisonnier C, Trudeau DL, Arnold FH (2011) Comparison of family 9 cellulases from mesophilic and thermophilic bacteria. Appl Environ Microbiol 77:1436–1442. doi:10.1128/AEM.01802-10

Padilha IQM, Valenzuela SV, Grisi TCSL, Diaz P, de Araújo DAM, Pastor FIJ (2014) A glucuronoxylan-specific xylanase from a new *Paenibacillus favisporus* strain isolated from tropical soil of Brazil. Int Microbiol 17:175–184. doi:10.2436/IM.V17I3.136532

Park C, Kawaguchi T, Sumitani J-I, Takada G, Izumori K, Arai M (2005) Cloning and sequencing of an exoglucanase gene from *Streptomyces* sp. M 23, and its expression in *Streptomyces lividans* TK-24. J Biosci Bioeng 99:434–436. doi:10.1263/jbb.99.434

Poidevin L, Feliu J, Doan A, Berrin J-G, Bey M, Coutinho PM, Henrissat B, Record E, Heiss-Blanquet S (2013) Insights into exo- and endoglucanase activities of family 6 glycoside hydrolases from *Podospora anserina*. Appl Environ Microbiol 79:4220–4229. doi:10.1128/AEM.00327-13

Sánchez MM, Pastor FIJ, Diaz P (2003) Exo-mode of action of cellobiohydrolase Cel48C from *Paenibacillus* sp. BP-23. A unique type of cellulase among Bacillales. Eur J Biochem 270:2913–2919. doi:10.1046/j.1432-1033.2003.03673.x

Sánchez MM, Irwin DC, Pastor FIJ, Wilson DB, Diaz P (2004) Synergistic activity of *Paenibacillus* sp. BP-23 cellobiohydrolase Cel48C in association with the contiguous endoglucanase Cel9B and with endo- or exo-acting glucanases from *Thermobifida fusca*. Biotechnol Bioeng 87:161–169. doi:10.1002/bit.20099

Sánchez MM, Fritze D, Blanco A, Spröer C, Tindall BJ, Schumann P, Kroppenstedt RM, Diaz P, Pastor FIJ (2005) *Paenibacillus barcinonensis* sp. nov., a xylanase-producing bacterium isolated from a rice field in the Ebro River delta. Int J Syst Evol Microbiol 55:935–939. doi:10.1099/ijs.0.63383-0

Sandgren M, Wu M, Karkehabadi S, Mitchinson C, Kelemen BR, Larenas EA, Ståhlberg J, Hansson H (2013) The structure of a bacterial cellobiohydrolase: the catalytic core of the *Thermobifida fusca* family GH6 cellobiohydrolase Cel6B. J Mol Biol 425:622–635. doi:10.1016/j.jmb.2012.11.039

Spiro RG (1966) Analysis of sugars found in glycoproteins. Methods Enzymol 8:3–26

Teeri TT (1997) Crystalline cellulose degradation: new insight into the function of cellobiohydrolases. Trends Biotechnol 15:160–167. doi:10.1016/S0167-7799(97)01032-9

Tomme P, Van Tilbeurgh H, Pettersson G, Van Damme J, Vandekerckhove J, Knowles J, Teeri T, Claeyssens M (1988) Studies of the cellulolytic system of *Trichoderma reesei* QM 9414. Analysis of domain function in two cellobiohydrolases by limited proteolysis. Eur J Biochem 170:575–581

Tomme P, Boraston A, McLean B, Kormos J, Creagh AL, Sturch K, Gilkes NR, Haynes CA, Warren RA, Kilburn DG (1998) Characterization and affinity applications of cellulose-binding domains. J Chromatogr B Biomed Sci Appl 715:283–296

Tormo J, Lamed R, Chirino AJ, Morag E, Bayer EA, Shoham Y, Steitz TA (1996) Crystal structure of a bacterial family-III cellulose-binding domain: a general mechanism for attachment to cellulose. EMBO J 15:5739–5751

Tuck CO, Pérez E, Horváth IT, Sheldon RA, Poliakoff M (2012) Valorization of biomass: deriving more value from waste. Science 337:695–699. doi:10.1126/science.1218930

Valenzuela SV, Lopez S, Biely P, Sanz-Aparicio J, Pastor FIJ (2016) The glycoside hydrolase family 8 reducing-end xylose-releasing exo-oligoxylanase Rex8A from *Paenibacillus barcinonensis* BP-23 is active on branched xylooligosaccharides. Appl Environ Microbiol 82:5116–5124. doi:10.1128/AEM.01329-16

Vuong TV, Wilson DB (2009b) Processivity, synergism, and substrate specificity of *Thermobifida fusca* Cel6B. Appl Environ Microbiol 75:6655–6661. doi:10.1128/AEM.01260-09

Vuong TV, Wilson DB (2009a) The absence of an identifiable single catalytic base residue in *Thermobifida fusca* exocellulase Cel6B. FEBS J 276:3837–3845. doi:10.1111/j.1742-4658.2009.07097.x

Watson DL, Wilson DB, Walker LP (2002) Synergism in binary mixtures of *Thermobifida fusca* cellulases Cel6B, Cel9A, and Cel5A on BMCC and Avicel. Appl Biochem Biotechnol 101:97–111. doi:10.1385/ABAB:101:2:097

Wohlfahrt G, Pellikka T, Boer H, Teeri TT, Koivula A (2003) Probing pH-dependent functional elements in proteins: modification of carboxylic acid pairs in *Trichoderma reesei* cellobiohydrolase Cel6A. Biochemistry 42:10095–10103. doi:10.1021/BI034954O

Wood TM (1988) Preparation of crystalline, amorphous, and dyed cellulase substrates. In: Meth Enzymol. pp 19–25

Yaniv O, Petkun S, Shimon LJW, Bayer EA, Lamed R, Frolow F (2012) A single mutation reforms the binding activity of an adhesion-deficient family 3 carbohydrate-binding module. Acta Crystallogr D Biol Crystallogr 68:819–828. doi:10.1107/S0907444912013133

Zhang S, Lao G, Wilson DB (1995) Characterization of a *Thermomonospora fusca* exocellulase. Biochemistry 34:3386–3395

Zhang Y-HP, Cui J, Lynd LR, Kuang LR (2006) A transition from cellulose swelling to cellulose dissolution by *o*-phosphoric acid: evidence from enzymatic hydrolysis and supramolecular structure. Biomacromolecules 7:644–648. doi:10.1021/bm050799c

E

A DNA barcode-assisted annotated checklist of the spider (Arachnida, Araneae) communities associated to white oak woodlands in Spanish National Parks

Luís C Crespo, Marc Domènech, Alba Enguídanos, Jagoba Malumbres-Olarte, Pedro Cardoso, Jordi Moya-Laraño, Cristina Frías-López, Nuria Macías-Hernández, Eva De Mas, Paola Mazzuca, Elisa Mora, Vera Opatova, Enric Planas, Carles Ribera, Marcos Roca-Cusachs, Dolores Ruiz, Pedro Sousa, Vanina Tonzo, Miquel A. Arnedo

2018, Biodiversity Data Journal 6: e29443

Biodiversity Data Journal 6: e29443
doi: 10.3897/BDJ.6.e29443

Taxonomic Paper

A DNA barcode-assisted annotated checklist of the spider (Arachnida, Araneae) communities associated to white oak woodlands in Spanish National Parks

Luís C Crespo[‡,§], Marc Domènech[‡], Alba Enguídanos[‡], Jagoba Malumbres-Olarte[‡,§,|], Pedro Cardoso[§], Jordi Moya-Laraño[¶], Cristina Frías-López[#], Nuria Macías-Hernández[§,¤], Eva De Mas[¶], Paola Mazzuca[‡], Elisa Mora[‡], Vera Opatova[‡,«], Enric Planas[‡], Carles Ribera[‡], Marcos Roca-Cusachs[»], Dolores Ruiz[¶], Pedro Sousa[^], Vanina Tonzo[‡], Miquel A. Arnedo[‡]

‡ Department of Evolutionary Biology, Ecology and Environmental Sciences & Biodiversity Research Institute (IRBio), Universitat de Barcelona, Av. Diagonal 643, E-08028, Barcelona, Spain
§ Laboratory for Integrative Biodiversity Research, Finnish Museum of Natural History, University of Helsinki; PO Box 17, 00014, Helsinki, Finland
| cE3c - Centre for Ecology, Evolution and Environmental Changes, University of the Azores; Rua Capitão João d´Ávila, Pico da Urze, 9700-042 , Angra do Heroísmo, Terceira, Azores, Portugal
¶ Department of Functional and Evolutionary Ecology, Estación Experimenta de Zonas Áridas (EEZA, CSIC); Carretera de Sacramento, s/n. La Cañada de San Urbano 04120, Almeria, Spain
# Department of Genetics, Microbiology and Statistics, & Biodiversity Research Institute (IRBio), Universitat de Barcelona, Av. Diagonal 643, E-08028, Barcelona, Spain
¤ Island Ecology and Evolution Research Group, Instituto de Productos Naturales y Agrobiologíia, C/Astrofísico Francisco Sánchez 3, La Laguna, Tenerife, Canary Islands, Spain
« Department of Entomology and Nematology, University of California, Davis, CA 95616, Davis, United States of America
» Laboratory of Systematic Entomology in the Department of Applied Biology of Chungnam National University, Daejeon, Korea, South
^ CIBIO, Centro de Investigação em Biodiversidade e Recursos Genéticos, Universidade do Porto, Vila do Conde, Portugal

Corresponding author: Miquel A. Arnedo (marnedo@gmail.com)

Academic editor: Gergin Blagoev

Received: 31 Aug 2018 | Accepted: 08 Nov 2018 | Published: 29 Nov 2018

Citation: Crespo L, Domènech M, Enguídanos A, Malumbres-Olarte J, Cardoso P, Moya-Laraño J, Frías-López C, Macías-Hernández N, De Mas E, Mazzuca P, Mora E, Opatova V, Planas E, Ribera C, Roca-Cusachs M, Ruiz D, Sousa P, Tonzo V, Arnedo M (2018) A DNA barcode-assisted annotated checklist of the spider (Arachnida, Araneae) communities associated to white oak woodlands in Spanish National Parks. Biodiversity Data Journal 6: e29443. https://doi.org/10.3897/BDJ.6.e29443

Abstract

Background

A large scale semi-quantitative biodiversity assessment was conducted in white oak woodlands in areas included in the Spanish Network of National Parks, as part of a project aimed at revealing biogeographic patterns and identify biodiversity drivers. The semi-quantitative COBRA sampling protocol was conducted in sixteen 1-ha plots across six national parks using a nested design. All adult specimens were identified to species level based on morphology. Uncertain delimitations and identifications due to either limited information of diagnostic characters or conflicting taxonomy were further investigated using DNA barcode information.

New information

We identified 376 species belonging to 190 genera in 39 families, from the 8,521 adults found amongst the 20,539 collected specimens. Faunistic results include the discovery of 7 new species to the Iberian Peninsula, 3 new species to Spain and 11 putative new species to science. As largely expected by environmental features, the southern parks showed a higher proportion of Iberian and Mediterranean species than the northern parks, where the Palearctic elements were largely dominant. The analysis of approximately 3,200 DNA barcodes generated in the present study, corroborated and provided finer resolution to the morphologically based delimitation and identification of specimens in some taxonomically challenging families. Specifically, molecular data confirmed putative new species with diagnosable morphology, identified overlooked lineages that may constitute new species, confirmed assignment of specimens of unknown sexes to species and identified cases of misidentifications and phenotypic polymorphisms.

Keywords

DNA barcoding, faunistics, COBRA protocol, Mediterranean region, Iberian Peninsula, Dictynidae, Gnaphosidae, Linyphiidae, *Philodromidae*

Introduction

The Iberian Peninsula is one of the most diverse regions in the Mediterranean Basin because of its location at the crossroads between Europe and Africa and its complex orography and variable climate, ranging from a central and southern Mediterranean climate to a northern Eurosiberian one. The high level of species richness in the Iberian Peninsula is particularly evident in spiders (Carvalho et al. 2012, Carvalho et al. 2011, Jiménez-

Valverde et al. 2010), where approximately 1,400 species have been catalogued to date (Morano et al. 2014). The Iberian biota is also highly endemic and threatened, with most of the south of the peninsula being identified as one of the most important biodiversity hotspots in the Mediterranean region (Medail and Quezel 1999). Amongst the order Araneae, 18% of the species in the Iberian Peninsula are Iberian endemics, a value that rises above 50% in families such as Dysderidae C. L. Koch, 1837, Zodariidae Thorell, 1881 or Nemesiidae Simon, 1889 (Cardoso and Morano 2010, Morano et al. 2014).

Despite the high number of spiders recorded in the Iberian Peninsula, the species-richness is lower than in neighbouring countries of similar size, yet less complex or younger geological history, such as France (1587 species) (Nentwig et al. 2017) or Italy (1632 species) (Pantini and Isaia 2017). The relatively shorter tradition in natural history of Iberian countries leads us to suspect that the Iberian arachnofauna is fewer because it is far from being fully catalogued, which is one of the main impediments for invertebrate conservation in the region (Cardoso et al. 2011). Gradually, new faunistic records are helping to build up our knowledge on both the richness and distribution of Iberian species (Barrientos and Fernández 2015, Barrientos et al. 2014, Barrientos et al. 2015a, Barrientos et al. 2015b, Barrientos et al. 2016, Barriga et al. 2007, Cardenas and Barrientos 2011, Jimenez-Segura et al. 2017, Melic et al. 2016, Pérez 2016, Pérez and Castro 2016, Pérez et al. 2015). Unfortunately, many specimens acquired in local ecological assessments frequently remain unidentified in collections due to either a lack of expertise or informative taxonomic literature. Diverse taxa such as Nemesiidae, Dysderidae, Gnaphosidae or Oonopidae continue to demand revisionary taxonomic work, which, given the current downward trend in funding for basic taxonomic research and the time needed to complete these thorough works, can only be afforded by a decreasing number of taxonomists.

The use of DNA barcoding – standardised, short fragments of DNA, as a species identifier (Hebert et al. 2003) – has become very popular amongst spider taxonomists (Astrin et al. 2016, Barrett and Hebert 2005, Blagoev et al. 2013, Blagoev et al. 2015, Candek and Kuntner 2015, Castalanelli et al. 2014, Robinson et al. 2009). Although the use of DNA barcoding is not yet fully incorporated into standard diversity assessments, when available, this tool provides many advantages to the taxonomists working on a medium- or large-sized collection of spiders. DNA barcodes can facilitate and accelerate taxonomic research by increasing the ability of matching individuals regardless of sex, stage or body parts, identifying specimens with morphological diagnostic characters either subtle, difficult to visualise or absent or reassessing intraspecific polymorphisms.

Here we present the checklist of spider species identified from the adult specimens collected as part of a large-scale biodiversity assessment of the spider communities in white-oak (*Quercus* L.) woodlands across the Spanish Natural Parks Network (hereafter referred to as the IBERCODING project). Specimens were collected using the COBRA protocol (Cardoso 2009), a semi-quantitative sampling protocol initially designed to assess biodiversity patterns in Mediterranean spider communities and then adapted to other habitats (Malumbres-Olarte et al. 2017) and potentially extendable to other taxa. The identification of the collected specimens is the first necessary step towards calculating α-

and β-diversity values across broad geographic and climatic ranges and ultimately inferring the drivers responsible for those patterns.

We chose to focus on white-oak forests because they represent common forests in the focal national parks and their high levels of endemicity (Franco 1990), relevance for conservation (García and Mejías 2009, Marañón and Pérez-Ramos 2009) and relatively well-characterised evolutionary history in the Iberian Peninsula (Olalde et al. 2002, Petit et al. 2002).

As part of the IBERCODING project, we generated DNA barcodes for more than 3,200 specimens with the aim of revealing fine scale geographic patterns in genetic diversity, retrieving phylogenetic information for assessing phylogenetic diversity of communities and facilitating sorting and identification of the specimens.

The present publication focuses on the identification of the individuals collected, with comments on their distribution and spatial location, as well as new records to the region and the discovery of putative new species. The availability of DNA barcodes helped identification and delimitation in some taxonomically challenging groups, such as the families Dictynidae, Gnaphosidae, Linyphiidae or *Philodromidae* We characterised the biogeographic patterns of the different plots and parks based on the species distribution information available in the literature and complemented it with our own data.

Materials and methods

Study area

Spider communities were sampled in white oak and related oak forests from six Spanish national parks (Fig. 1), namely Picos de Europa (P), Ordesa y Monte Perdido (O), Aigüestortes i Estany de Sant Maurici (A) (hereafter referred to as the northern parks) Fig. 2), Monfragüe (M), Cabañeros (C) and Sierra Nevada (S) (hereafter referred to as the southern parks) (Fig. 3). The chosen parks fulfilled three conditions: (1) they had representative white oak forests, (2) they represented the main biogeographic areas within the Iberian Peninsula (Atlantic, Alpine and Mediterranean) (European Environment Agency 2012) and (3) they covered a broad latitudinal and elevational gradient within the Iberian Peninsula. The selected parks spanned distances ranging from 80 km apart (A to O) to 720 km (S to A) and elevations from 320 m (Monfragüe) to 1786 m (Sierra Nevada). Sampling was conducted between May and June, when the richness and abundance of adult spiders in Mediterranean habitats are at its maximum (Cardoso et al. 2007), in two consecutive years, 2013 for the northern parks and 2014 for the southern parks. Two replicates (plots) were set up in each park, except in Picos de Europa and Cabañeros, where two different types of oak forest were available and hence two replicates were set up per forest type, resulting in a total of 16 plots (northern parks P=4, O=2, A=2; southern parks M=2, C=4 S=2, respectively). Additional details of the sampling plots are available in Table 1.

Table 1.

Information on the sampling sites. Site codes are derived from abbreviated park names. Geographical coordinates are given in the format of decimal degrees (DD).

Site code	Region	Province	Locality	Coordinates (Lat. / Lon.)	Altitude (m)	Collection dates	Habitat
P1	Castilla y Leon	Leon	Monte Robledo	43.14450 / -4.92675	1071.6	7.VI.2013–21.VI.2013	*Quercus petraea*
P2	Castilla y Leon	Leon	Joyoguelas	43.17771 / -4.90579	764.0	7.VI.2013–22.VI.2013	*Quercus faginea*
P3	Castilla y Leon	Leon	Las Arroyas	43.14351 / -4.94878	1097.1	8.VI.2013–23.VI.2013	*Quercus petraea*
P4	Castilla y Leon	Leon	El Canto	43.17227 / -4.90857	943.5	9.VI.2013–24.VI.2013	*Quercus faginea*
O1	Aragon	Huesca	O Furno	42.60677 / 0.13135	1396.7	12.VI.2013–26.VI.2013	*Quercus subpyrenaica*
O2	Aragon	Huesca	Rebilla	42.59427 / 0.15290	1158.1	13.VI.2013–27.VI.2013	*Quercus subpyrenaica*
A1	Catalonia	Lleida	Sola de Boi	42.54958 / -0.87254	1759.8	15.VI.2013–29.VI.2013	*Quercus pubescens*
A2	Catalonia	Lleida	Sola de Boi	42.54913 / 0.87137	1738.7	16.VI.2013–30.VI.2013	*Quercus pubescens*
M1	Extremadura	Cáceres	Peña Falcón	39.83296 / -6.06410	320.6	23.V.2014–6.VI.2014	*Quercus faginea*
M2	Extremadura	Cáceres	Fuente del Frances	39.82800 / -6.03249	320.7	24.V.2014–7.VI.2014	*Quercus faginea*
C1	Castilla-La Mancha	Ciudad Real	Valle Brezoso	39.35663 / -4.35912	756.6	27.V.2014–9.VI.2014	*Quercus pyrenaica*
C2	Castilla-La Mancha	Ciudad Real	Valle Brezoso	39.35159 / -4.35890	739.3	28.V.2014–10.VI.2014	*Quercus pyrenaica*
C3	Castilla-La Mancha	Ciudad Real	La Quesera	39.36177 / -4.41733	767.6	29.V.2014–11.VI.2014	*Quercus faginea*
C4	Castilla-La Mancha	Ciudad Real	La Quesera	39.36337 / -4.41704	772.3	30.V.2014–12.VI.2014	*Quercus faginea*
S1	Andalucia	Granada	Soportujar	36.96151 / -3.41881	1786.6	31.V.2014–14.VI.2014	*Quercus pyrenaica*
S2	Andalucia	Granada	Camarate	37.18377 / -3.26282	1714.0	1.VI.2014–15.VI.2014	*Quercus pyrenaica*

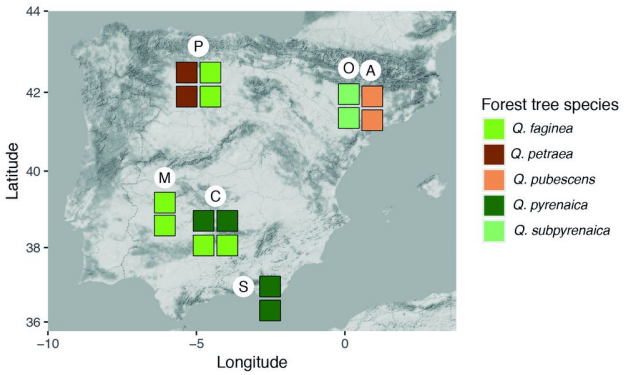


Figure 1.

Map of the Iberian Peninsula with the location of the national parks and the plots where the sampling protocol COBRA was applied. For each park, squares denote the number of plots and the oak forest type (colour code labels in the inset). Northern parks are Picos de Europa (P), Ordesa (O), Aigüestortes (A). Southern parks Monfragüe (M), Cabañeros (C), Sierra Nevada (S). See Table 1 for additional information on plots and parks.

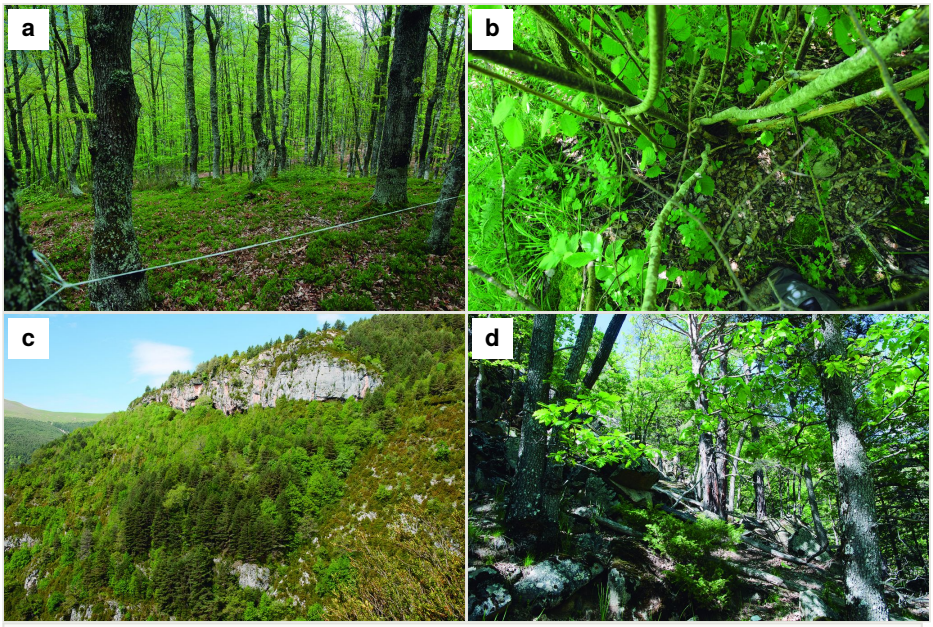


Figure 2.

Pictures of collection localities

a: P1 plot Monte Robledo, *Quercus petraea* forest
b: P4 plot El Canto, *Q. faginea* forest (detail leaflitter)
c: O1 plot O Furno, *Q. subpyrenaica* forest
d: A2 plot Sola de Boi, *Q. pubescens* forest

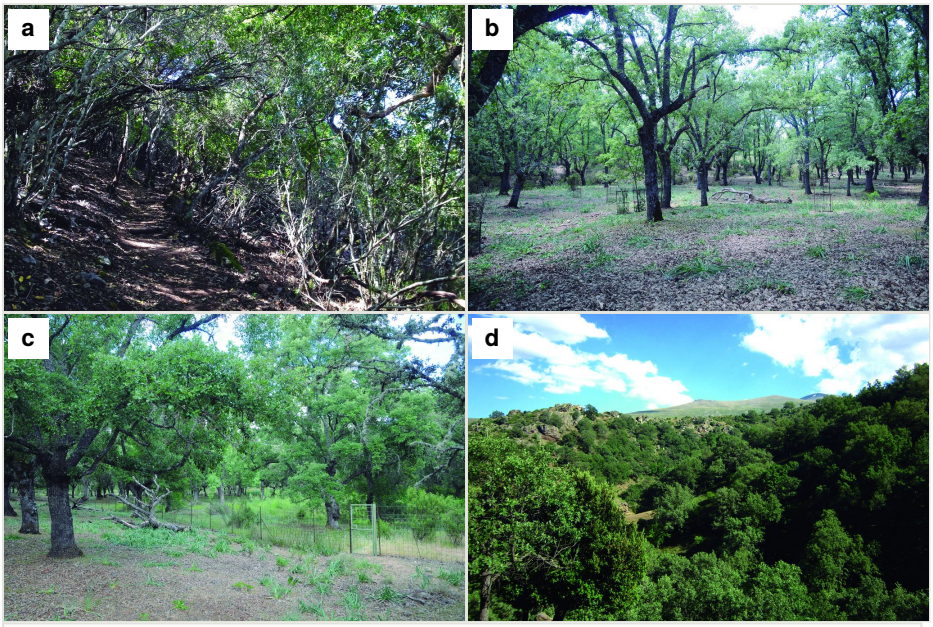


Figure 3.

Pictures of collection localities:

a: M2 plot Fuente del Frances, *Quercus faginea* forest
b: C1 plot Valle Brezoso, *Q. pirenaica*
c: C3 plot La Quesera, *Q. faginea* forest
d: S2 plot Camarate, *Q. pyrenaica*

Sample collection

In each plot, a COBRA 50 sampling protocol was conducted, which is specifically designed to collect 50% of the spider diversity in the sampling area in an optimised manner (Cardoso 2009). This protocol consists of using different sampling methods to obtain the maximum possible number of species. Direct sampling (methods that require the presence of the collector and her/ his active participation in the specimen capture) were foliage beating, vegetation sweeping and aerial hand collection. For foliage beating, a 1 m^2 beating tray and a wooden pole were used to beat tree branches as high as possible. For vegetation sweeping, a round sweep net with a diameter of 58 cm was used to sweep tall plants and bushes, below the collector's waist. Aerial hand collection was done through visual inspection and hand-capture (aided by forceps, pooter or brush, if needed) on the vegetation above knee-level. The maximum possible number of spiders was caught and transferred to a vial with ethanol. Each sampling consisted of one hour of continuous collecting by one collector. In two plots (P1, A1), we conducted two additional ground hand collecting samples but focused on specimens present below knee-level.

Indirect sampling (techniques that do not involve the presence of the collector), consisted in the use of pitfall traps, i.e. vessels 7.5 cm in diameter buried in the ground with the rim at

the ground level and filled with propylene glycol, which preserved spiders for both morphologic and genetic analyses. A few detergent drops were added to the liquid to break the surface tension and to allow spiders to sink to the bottom of the vessel. Pitfalls were covered with labelled plastic caps, held about 1 cm above the ground by four short wires anchored to the ground, in order to prevent the fall of debris into the trap and propylene glycol dilution or overflow caused by rainwater.

Direct sampling in each plot consisted of 2 hours of diurnal and 2 hours of nocturnal foliage beating, 2 hours of diurnal and 2 hours of nocturnal vegetation sweeping and 4 hours of nocturnal aerial hand collecting, which totals 12 hours of sampling, equating to 12 man-hours of sampling. Indirect samples were uniformly distributed within each plot in groups of 4 pitfalls, set in squares with 5 m sides. The traps were left active during two weeks. For subsequent analyses, each group of 4 contiguous pitfall traps were combined and considered as a single sample, which totals 12 indirect samples (Carvalho et al. 2011). All in all, the study included 388 samples (24 samples per plot x 16 plots + 2 extra ground samples x 2 plots, P1 and A1, respectively).

Identification of specimens

All adults were identified, when possible, to species level. Amongt a wide spectrum of taxonomic literature, the “Araneae: Spiders of Europe” database was used to identify most of the known species found in the samples (Nentwig et al. 2017). Identifications were made mainly with the use of a ZEISS Stemi 2000 stereomicroscope. Images were taken with a Leica DFC 450 camera attached to a Leica MZ 16A stereomicroscope, using the software Leica Application Suite v4.4. After collection, specimens were stored in 95% ethanol and kept at -20ºC in Falcon vials until these were sequentially sorted and identified (materials from the northern parks were collected and sorted before the materials from the southern parks), from which they were moved to smaller vials of 2 ml and returned to -20ºC, for subsequent genetic analyses.

Annotated checklist

For each species, we provided the number of male and female specimens identified by plot (see abbreviations in Table 1) and collecting technique, namely foliage beating (beating), vegetation sweeping (sweeping), aerial hand collection and pitfall trapping. Unidentified morphs and putative new species were provisionally labelled using the genus name and a sequential numeration (e.g. *Brigittea* sp04). Distributions were based on information available in public databases (Nentwig et al. 2017, World Spider Catalog 2018).

Molecular procedures

DNA barcodes were obtained from all sampled species – five individuals were analysed per morpho-species and per plot when possible, as many species collected without taxonomic targeting are usually found in singletons or doubletons. Legs were used for DNA extraction and the rest of the individual was kept as a voucher, although for small species, the entire

specimen was used. In these cases, the extractions were non-destructive (i.e. specimens were not ground up) and specimens were recovered as vouchers after the lyses of soft internal tissues. Total genomic DNA was extracted using the REDExtract-N-Amp™ Tissue PCR Kit Protocol from Sigma-Aldrich, following the manufacturer's protocol and performed in 96 well-plates. The primers used for amplification are listed in Table 2. The LCOI1490/HCO2198 was the preferred combination, while Nancy was used as a replacement for HCOI2198 and Ron as replacement for LCOI1490 (in that order). For problematic amplifications, we used the internal primers mlCOIintF/jgHCOI2198. The polymerase chain reaction (PCR) was performed in 96-well plates using 8 µl REDExtract-N-Amp™ PCR ReadyMix from Sigma-Aldrich, primers forward and reverse, 4 µl of diluted DNA and ultrapure, distilled water up to a total reaction volume of 20 µl. PCR conditions were as follows: initial denaturing step at 95°C for 5 min, 35 amplification cycles (94°C for 30 s, 45°C for 35 s, 72°C for 45 s) and a final step at 72°C for 5 min. In some cases a Touchdown protocol was used, consisting of 16 cycles of annealing temperature starting at 62°C and decreasing 1°C each cycle and 25 additional cycles of annealing at 46°C. PCR products were cycle-sequenced in both directions at Macrogen Inc. (Seoul, South Korea).

Table 2.

Primers used for amplification.

Location	Name	Sequence	Reference
C1-J-1490	LCO1490	GGTCAACAAATCATAAAGATATTGG	Folmer et al. 199
C1-N-2198	HCO2198	TAAACTTCAGGGTGACCAAAAAATCA	Folmer et al. 199
C1-N-2191	Nancy	ACCCGGTAAAATTAAAATATAAACTTC	Simon et al. 199
C1-J-1751	Ron	GGATCACCTGATATAGCATTCCC	Simon et al. 199
C1-J-1834	mlCOIintF	GCWACWGGWTGAACWGTWTAYCCYCC	Leray et al. 2013
C1-N-2198	jgHCO2198	TAIACYTCIGGRTGICCRAARAAYCA	Leray et al. 2013

Raw chromatograms were assembled, edited and further manipulated using the software Geneious v7.1.9 (Kearse et al. 2012).

DNA barcode analysis

Although DNA barcodes were obtained for all species, we decided to investigate a step further with several families that presented us with cases of incongruence between morphology-based identification and genetic-based identification, namely the Dictynidae, Gnaphosidae, Linyphiidae and *Philodromidae*. Alignments were obtained by combining all DNA barcodes of focal species (see Results) for each family. We inferred the maximum-likelihood tree for each alignment by finding the best partition scheme first (Kalyaanamoorthy et al. 2017), followed by tree inference using the edge-linked partition model in the IQ-TREE software v1.6.1 (Chernomor et al. 2016, Nguyen et al. 2015). The best tree was then used to delimit putative species using the mPTP algorithm (Kapli et al. 2017). The mPTP method allows identifying species boundaries based on branch lengths

obtained from a single-locus, without the need for an ultrametric tree and has been shown to generate more stable outputs than alternative approaches (Blair and Bryson 2017). Genetic distances within and between the clusters identified by mPTP were estimated using the Kimura 2 parameter model (Kimura 1980) in MEGA v.6 (Tamura et al. 2013). One DNA barcode per genetic cluster was further used for automatic identification using BOLD (Ratnasingham and Hebert 2007).

Biogeographic composition

The delimited and identified species were subsequently grouped in four groups, namely "Cosmopolitan", "Palearctic", "Mediterranean" and "Iberian", based on the distribution information available at the World Spider Catalog 2018, further refined with our own species presence data (see Suppl. material 1. We note that species found only in Iberia were considered Iberian and not Mediterranean and the species recorded only in countries of the Mediterranean basin (including north-African countries) were considered Mediterranean and not Palearctic. Percentage of each of the four groups per plot were estimated and visualised using R (R Development Core Team 2017).

F

The draft genome sequence of the spider *Dysdera silvatica* (Araneae, Dysderidae): A valuable resource for functional and evolutionary genomic studies in chelicerates

Jose Francisco Sánchez-Herrero, Cristina Frías-López, Paula Escuer, Silvia Hinojosa-Alvarez, Miquel A. Arnedo, Alejandro Sánchez-Gracia and Julio Rozas

2019, GigaScience, 8, 1–9

GigaScience, 8, 2019, 1–9

doi: 10.1093/gigascience/giz099
Data Note

DATA NOTE

The draft genome sequence of the spider *Dysdera silvatica* (Araneae, Dysderidae): A valuable resource for functional and evolutionary genomic studies in chelicerates

Jose Francisco Sánchez-Herrero [1], Cristina Frías-López [1], Paula Escuer [1], Silvia Hinojosa-Alvarez [1,2], Miquel A. Arnedo [3], Alejandro Sánchez-Gracia [1,*] and Julio Rozas [1,*]

[1]Departament de Genètica, Microbiologia i Estadística, Universitat de Barcelona (UB) and Institut de Recerca de la Biodiversitat (IRBio), Diagonal 643, 08028 Barcelona, Spain ; [2] Jardín Botánico, Instituto de Biología, Universidad Nacional Autónoma de México, Tercer Circuito Exterior S/N, Ciudad Universitaria Coyoacán, 04510 México DF, México and [3]Departament de Biologia Evolutiva, Ecologia i Ciències Ambientals, Universitat de Barcelona (UB) and Institut de Recerca de la Biodiversitat (IRBio), Diagonal 643, 08028 Barcelona, Spain

*Correspondence address. Julio Rozas, Departament de Genètica, Microbiologia i Estadística, Universitat de Barcelona (UB) and Institut de Recerca de la Biodiversitat (IRBio), Diagonal 643, 08028 Barcelona, Spain. E-mail: jrozas@ub.edu http://orcid.org/0000-0003-4543-4577; Alejandro Sánchez-Gracia, Departament de Genètica, Microbiologia i Estadística, Universitat de Barcelona (UB) and Institut de Recerca de la Biodiversitat (IRBio), Diagonal 643, 08028 Barcelona, Spain; . E-mail: elsanchez@ub.edu http://orcid.org/0000-0002-6839-9148

Abstract

Background: We present the draft genome sequence of *Dysdera silvatica*, a nocturnal ground-dwelling spider from a genus that has undergone a remarkable adaptive radiation in the Canary Islands. **Results:** The draft assembly was obtained using short (Illumina) and long (PaciBio and Nanopore) sequencing reads. Our *de novo* assembly (1.36 Gb), which represents 80% of the genome size estimated by flow cytometry (1.7 Gb), is constituted by a high fraction of interspersed repetitive elements (53.8%). The assembly completeness, using BUSCO and core eukaryotic genes, ranges from 90% to 96%. Functional annotations based on both *ab initio* and evidence-based information (including *D. silvatica* RNA sequencing) yielded a total of 48,619 protein-coding sequences, of which 36,398 (74.9%) have the molecular hallmark of known protein domains, or sequence similarity with Swiss-Prot sequences. The *D. silvatica* assembly is the first representative of the superfamily Dysderoidea, and just the second available genome of Synspermiata, one of the major evolutionary lineages of the "true spiders" (Araneomorphae). **Conclusions:** Dysderoids, which are known for their numerous instances of adaptation to underground environments, include some of the few examples of trophic specialization within spiders and are excellent models for the study of cryptic female choice. This resource will be therefore useful as a starting point to study fundamental evolutionary and functional questions, including the molecular bases of the adaptation to extreme environments and ecological shifts, as well of the origin and evolution of relevant spider traits, such as the venom and silk.

Received: 6 May 2019; **Revised:** 27 June 2019; **Accepted:** 30 July 2019

Keywords: Araneomorphae; hybrid genome assembly; genome annotation; Canary Islands

Figure 1 Male of *Dysdera silvatica* from Teselinde (La Gomera, Canary Islands). Photo credit: Miquel Arnedo.

Data Description

Spiders are a highly diverse and abundant group of predatory arthropods, found in virtually all terrestrial ecosystems. Approximately 45,000 spider species have been recorded to date [1]. The nocturnal ground family Dysderidae ranks 17th out of 118 currently accepted spider families in number of species. The type genus of the family, *Dysdera* Latreille, 1804, includes half of the family diversity (282 species). This genus is remarkable in several aspects. First, it represents one of the few cases of stenophagy, i.e., prey specialization, across spiders [2]. Many species in the genus have evolved special morphological, behavioral, and physiological adaptations to feed on woodlice, including modifications of mouthparts, unique hunting strategies, and effective restriction to assimilation of metals into its tissues [3–7]. Because of their chemical defenses and ability to accumulate heavy metals from the soil, woodlice are usually avoided as prey by most spiders, including generalist *Dysdera* [2,4, 5,7]. Although mostly circumscribed to the Mediterranean region, *Dysdera* has colonized all the Macaronesian archipelagoes and has undergone a remarkable species diversification in the Canary Islands [8]. As many as 55 species have been recorded across the 7 main islands and islets of this archipelago, being most of them single-island endemics [9]. Although multiple colonization events may account for the initial origin of species diversity the bulk of this diversity is the result of *in situ* diversification [8]. *Dysdera* spiders have adapted to a broad range of terrestrial habitats within the Canary Islands [9]. Interestingly, many co-occurring species significantly differ in mouthpart sizes and shapes, presumably owing to adaptations to a specialized diet [6,7], suggesting that stenophagy has evolved multiple times independently in these islands [10]. Although behavioral and physiological experiments have revealed a close correlation between morphological traits and prey preference in *Dysdera*, little is known about the molecular basis of trophic adaptations in this genus.

Here we present the draft assembly and functional annotation of the genome of the Canary Island endemic spider *Dysdera silvatica* Schmidt, 1981 (NCBI:txid477319; Fig. 1). This study is the first genomic initiative within its family and just the second within the Synspermiata [11], a clade that includes most of the families formerly included in Haplogynae, which was recently shown to be paraphyletic [12,13] (Fig. 2). Remarkably, a recent review on arachnid genomics identified the superfamily Dysderoidea (namely, Dysderidae, Orsolobidae, Oonopidae, and Segestriidae) as one of the priority candidates for genome sequencing [14]. The new genome, intended to be a reference genome for genomic studies on trophic specialization, will also be a valuable source for the ongoing studies on the molecular components of the chemosensory system in chelicerates [15]. Besides, because of the numerous instances of independent adaptation to caves [16], the peculiar holocentric chromosomes [17], and the evidence for cryptic female choice mechanisms [18,19] within the family, the new genome will be a useful reference for the study of the molecular basis of adaptation to extreme environments, karyotype evolution, and sexual selection. Additionally, a new fully annotated spider genome will greatly improve our understanding of key features, such as the venom and silk. The availability of new genomic information in a sparsely sampled section of the tree of life of spiders [14] will further provide valuable knowledge about relevant scientific questions, such as gene content evolution across main arthropod groups, including the consequences of whole-genome duplications, or the phylogenetic relationships with Araneae.

Sampling and DNA extraction

We sampled adult individuals of *D. silvatica* in different localities of La Gomera (Canary Islands) in March 2012 and June 2013 (Supplementary Table S1-1). The species was confirmed in the laboratory, and samples were stored at −80°C until its use. For Illumina and PaciBio libraries (see below), we extracted genomic DNA using Qiagen DNeasy Blood & Tissue Kit (Qiagen, Hilden, Germany, 74104)) according to the manufacturer's protocol. For the Oxford Nanopore libraries, we used a modified version of the Blood & Cell Culture DNA Mini Kit (Qiagen). Due to the high amount of chitin present in spiders we incubated fresh original samples 48 h at 32°C, avoiding a centrifugation step prior to sample loading to Qiagen Genomic tips, permitting the solution to precipitate by gravity. We also added an extra wash with 70% ethanol and centrifuged the solution at >5,000*g* for 10 min at 4°C. We quantified the genomic DNA in a Qubit fluorometer (Life Technologies, Thermo Fisher Scientific Inc., USA) using the dsDNA BR (double stranded DNA Broad Range) Assay Kit and checked its purity in a NanoDrop 2000 spectrophotometer (Thermo Fisher Scientific Inc.).

DNA sequencing

We sequenced the genome of *D. silvatica* using 4 different sequencing platforms (Table 1; Supplementary Table S1-2). First, we used the Illumina HiSeq2000 to obtain the genome sequence of a single male (100 bp, paired-end [PE] reads, 100 PE; TruSeq library). The flow-cell lane generated ∼51 Gb of sequence, representing a genome coverage of 30× (assuming a genome size of ∼1.7 Gb; see below). The genome of a female was sequenced using a mate pair (MP) approach; for that we used Nextera 5 kb-insert 100 PE libraries and the HiSeq2000 to generate ∼40 Gb of sequence (∼23× of coverage). A third individual (male) was used for single-molecule real-time (SMRT) sequencing (PacBio long reads). We used 8 SMRT libraries (20 kb SMRT bell templates), which were sequenced using the P6-C4 chemistry in a PacBio RSII platform. We obtained a yield of ∼9.6 Gb (raw coverage of

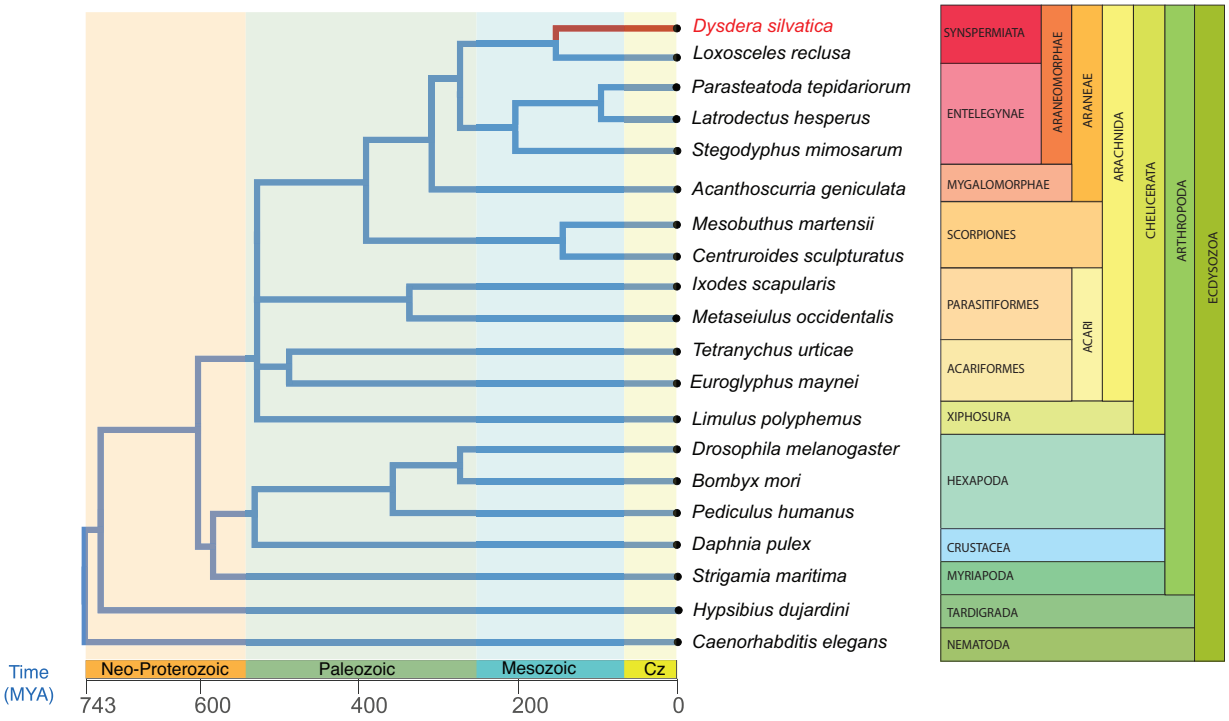


Figure 2 Phylogenetic relationships of the species used for the *D. silvatica* genome annotation (see Supplementary Table S1-11 for further details) and completeness analysis. Because the chelicerata phylogeny is controversial (e.g., [20], [21]), we set the most conflictive clades as polytomies. Divergence times were obtained from Carlson et al. (2017) [22] and the TimeTree web server (http://www.timetree.org/). Cz, cretaceous period.

Table 1. Sequencing data and library information

Run ID	Library	Insert size	Read lengths	Lanes	Total bases	Raw read pairs	Coverage (×)[a]
PE	Illumina HiSeq200 - Truseq	370 bp	100×100 PE	1	51,202,445,102	506,954,902	30
MP	Illumina HiSeq200 - Nextera	5 kb	100×100 PE	1	39,609,522,995	392,173,495	23
Nanopore	Nanopore 1D Libraries	-	Nanopore	5	23,193,357,481	20,534,058	14
PacBio	PacBio RSII 20 Kb SMRTbell	-	SMRT	8	9,652,844,880	1,455,288	6

[a]Based on the genome size estimated by flow cytometry ~1.7 Gb.

~6×). Finally, 2 additional females were used for the 5 runs of Nanopore sequencing (Nanopore 1D libraries). We got a yield of ~23.2 Gb (~14× coverage) (Table 1; Supplementary Table S1-2).

D. silvatica chromosome and genome size

D. silvatica has a diploid chromosome set of 6 pairs of autosomes and 2 (females are XX; 2n = 14) or 1 (males are X0) sex chromosomes (M. A. Arnedo, unpublished results). Using flow cytometry and the genome of the German cockroach *Blattella germanica* (1C = 2.025 Gb, J. S. Johnston, personal communication; see also [23]) as reference, we determined that the haploid genome size of *D. silvatica* is ~1.7 Gb. For the analysis, we adapted the Hare and Johnston [24] protocol for spiders species, without using male palps and chelicers to avoid analyzing haploid or endoreplicated cells, respectively [25,26]. Shortly, we isolated cells from the head of the male cockroach, and legs and palps from female spiders. We incubated the cells in LB0.1 with 2% of tween [27], propidium iodide (50 μg/mL), and RNAse (40 μg/mL). After 10 minutes, the processed tissue was filtered using a nylon mesh of 20 μm. We determined the DNA content of the diploid cells through the relative G0/G1 peak positions of the stained nuclei using a Gallios flow cytometer (Beckman Coulter, Inc, Fullerton, CA); the results were based on the average of 3 spider replicates, counting a minimum of 5,000 cells per individual.

In addition, we also estimated the *D. silvatica* genome size from the distribution of *k*-mers (from short reads) with Jellyfish v.2.2.3 (Jellyfish, RRID:SCR_005491) [28]. The distribution of *k*-mers of size 17, 21, and 41 (GenomeScope (GenomeScope, RRID:SCR_017014) [29]) resulted in a haploid genome size of ~1.23 Gb (Supplementary Fig. S1). The discrepancy between *k*-mer– and cytometry-based estimates may be caused by the presence of repetitive elements [30], which can affect *k*-mer estimates.

Read preprocessing

To avoid including contaminants in the assembly step, we searched the raw reads for mitochondrial, bacterial, archaeal, and virus sequences. We downloaded all genomes of all these kinds available in the GenBank database (Supplementary Table S1-3) and used BLASTN v2.4.0 (BLASTN, RRID:SCR_001598) [31] to detect and filter all contaminant reads (E-value $<10^{-5}$;

>90% alignment length; >90% identity). We preprocessed raw reads using PRINSEQ v.0.20.3 (PRINSEQ, RRID:SCR_005454) [32]. We estimated some descriptive statistics, such as read length and *k*-mer representation, and calculated the amount of adapter sequences and exact duplicates.

Quality-based trimming and filtering was performed according to the chemistry, technology, and library used (Supplementary Table S1-4). For the short-insert 100 PE library, we used Trimmomatic v0.36 (Trimmomatic, RRID:SCR_011848) [33] with specific lists of adapters of the TruSeq v3 libraries to filter all reads shorter than 36 bp or with minimum quality scores < 30 along 4-bp sliding windows. We also filtered trailing and leading bases with a quality score < 10. Long-insert MP libraries were preprocessed using NxTrim v0.4.1 [34] with default parameters (Supplementary Table S1-4a and b). We preprocessed the raw PacBio reads using the SMRT Analysis Software (SMRT Analysis Software, RRID:SCR_002942) [35], by generating circularized consensus sequence to further perform a polishing analysis with Pilon v1.22 (Pilon, RRID:SCR_014731) [36] based on short reads (Supplementary Table S1-4c).

De novo genome assembly

We used MaSuRCA v3.2.9 (MaSuRCA, RRID:SCR_010691) [37] for a hybrid *de novo* assembly of the *D. silvatica* genome (Supplementary Fig. S2). Additionally, we performed a scaffolding phase using AGOUTI (minimum number of joining reads pairs support, *k* = 3) [38], and the raw reads from a *D. silvatica* RNA sequencing (RNAseq) experiment [39] (Supplementary Table S1-5 and S1-6). During the assembly phase, we chose for each software the parameter values that generated the best assembly (Supplementary Table S1-7) in terms of (i) continuity and contig size statistics, such as the N50, L50, and the total number of sequences and bases assembled; and (ii) completeness measures, obtained as the fraction (and length) of a series of highly conserved proteins present in the draft genome. Particularly, we used 5 datasets, BUSCO v3 (BUSCO, RRID:SCR_015008) with genome option [40] using (i) the Arthropoda or (ii) the Metazoa dataset, (iii) the 457 core eukaryotic genes (CEGs) of *Drosophila melanogaster* [41], (iv) the 58,966 transcripts in the *D. silvatica* transcriptome [39], and (v) the 9,473 1:1 orthologs across 5 *Dysdera* species, *D. silvatica*; *D. gomerensis* Strand, 1911; *D. verneaui* Simon, 1883; *D. tilosensis* Wunderlich, 1992; and *D. bandamae* Schmidt, 1973 obtained from the comparative transcriptomics analysis of these species [42]. Finally, we performed an additional search to identify and remove possible contaminants in the generated scaffolds (Supplementary Table S1-7). We discarded 16 contaminant sequences > 5 kb. The final assembly size of the *D. silvatica* genome (Dsil v1.2) was ~1.36 Gb, with an N50 of ~38 kb (Table 2).

We determined the average genome coverage for each sequencing library with SAMtools v1.3.1 (SAMtools, RRID:SCR_002105) [43], by mapping short reads (using bowtie2 v2.2.9 [bowtie2, RRID:SCR_005476] [44]) or long reads (using minimap2 [45]) to the final draft assembly (Table 1; Supplementary Table S1-8; Supplementary Fig. S3).

Repetitive DNA sequences

We analyzed the distribution of repetitive sequences in the genome of *D. silvatica*, using either a *de novo* with RepeatModeler v1.0.11 (RepeatModeler, RRID:SCR_015027) [46], or a database-guided search strategy with RepeatMasker v.4.0.7 (RepeatMasker, RRID:SCR_012954) [47]. We used 3 different databases

Table 2. *Dysdera silvatica* nuclear genome assembly and annotation statistics

Genome assembly[a]	Value
Assembly size (bp)	1,359,336,805
% AT/CG/N	64.91%/34.83%/0.26%
Number of scaffolds	65,205
Longest scaffold	340,047
N50	38,017
L50	10,436
Repeat statistics[b]	
Number of elements	3,284,969
Length (bp) [% Genome]	731,540,381 [53.81%]
Genome annotation[a]	
Protein-coding genes	48,619
Functionally annotated	36,398 (74.86%)
Without functional annotation	12,221 (25.14%)
tRNA genes	33,934

[a]See also Supplementary S1-7.
[b]Summary of the RepeatMasker analysis (See also Supplementary Table S1-9).

of repetitive sequences, (i) *D. silvatica*–specific repetitive elements generated with RepeatModeler v1.0.11 [46], (ii) the Dfam_Consensus [48] (version 20170127), and (ii) the RepBase (version 20170127) [49,50]. We identified 2,604 families of repetitive elements, where 1,629 of them (62.6%) were completely unknown. Repetitive sequences accounted for ~732 Mb, which represent 53.8% of the total assembly size (Table 2; Supplementary Table S1-9a). Remarkably, most abundant repeats are from unknown families, 22.6% of the assembled genome. The repetitive fraction of the genome also include DNA elements (16.8%), LINEs (10.7%), and SINEs (1.85%), and a small fraction of other elements, including LTR elements, satellites, simple repeats, and low-complexity sequences. We found that the 10 most abundant repeat families among the 2,604 identified in *D. silvatica* account for ~7% of the genome and encode 5 unknown, 3 SINEs, and 2 LINEs, with an average length of ~193, ~161, and ~1,040 bp, respectively (Supplementary Table S1-9b).

We also studied the distribution of the high-covered genome regions to describe the spacing pattern among repetitive sequences. In particular, we searched for genomic regions that have a higher than average sequencing coverage above a particular threshold. Because repetitive regions are more prone to form chimeric contigs in the assembly step, we only used MaSuRCA super reads, and longer than 10 kb and free of Ns (34,937 contigs; 1.12 Gb). We estimated the coverage after mapping the short reads (from the 100PE library) to those contigs. We defined as high-coverage regions (HCRs) those with a coverage ≥2.5× or 5× the genome-wide average (~30×), in a region of ≥150, ≥500, ≥1,000, or ≥5,000 bp (Supplementary Fig. S4a; Supplementary Table S2). We found a large number of contigs encompassing ≥1 HCR. For instance, 21,614 contigs (~61.9%) include ≥1 HCR of 150 bp with >2.5× coverage (an average of 2.48 HCRs per contig; 77.7 HCR per Mb) (Supplementary Table S2-2a). For HCRs of >5× coverage, the results are also remarkable (10,604 contigs have ≥1 HCR of 150 bp, corresponding to 25.6 HCR per Mb). As expected, the longer the HCR the smaller the fraction in the genome; indeed, we found that the genome is encompassing ~5 HCR per Mb (HCR, longer than 1 kb at 2.5×). The distances between consecutive HCRs doenot show clear differences between the 2.5× and 5× thresholds (Supplementary Fig. 4b and S5; Supplementary Table S2-2b).

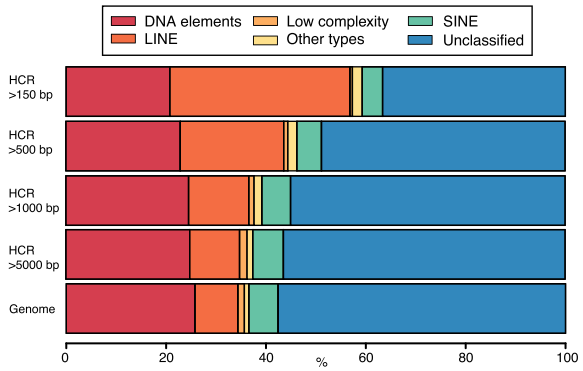


Figure 3 Bar plot of the annotation of the repetitive elements within the HCRs (2.5× threshold) at different intra-HCR length cutoffs (150, 500, 1,000, and 5,000 bp) (Supplementary Table S2-2a). Colors represent the type of repeat element identified by RepeatMasker. "Other types" class includes the LTR elements, small RNA, and satellite information that represent a small fraction.

We found a strong relationship between the length of the HCR and the type of the included repetitive elements (Fig. 3; Supplementary Table S2-3). For instance, while LINEs represent 8.62% of the repetitive elements in the whole genome, they are clearly enriched in the HCRs (36.12% in HCRs longer than 150 bp; 12.08% in HCRs longer than 5,000 bp) (Fig. 3; Supplementary Table S2-3a); the same was found for the small RNA fraction (ribosomal RNA). In contrast, the fraction of low-complexity repetitive sequences is much less represented in small HCRs than in the whole genome (~1.3%). We also found that the coverage threshold has little effect on the results (Supplementary Table S2-3; Supplementary Fig. S6), either for the main families or across subfamilies (Supplementary Table S2-4 and S2-5).

Given that the HCR analysis covers an important fraction of the assembled bases (~82%), the present results can likely be extrapolated to the whole genome. Therefore, the relatively low N50 of the *D. silvatica* genome draft is very likely to be caused by abundant interspersed repeats preventing genome continuity. Despite the low N50 we estimated that the draft presented here is mostly complete in terms of functional regions (see below).

Transcriptome assembly and genome annotation

We used the newly generated genome sequence to obtain a reference-guided assembly of the *D. silvatica* transcriptome with the RNAseq data from Vizueta et al. [39]. We used HISAT2 v2.1.0 (HISAT2, RRID:SCR_015530) [51] to map the RNAseq reads to the reference and Trinity v2.4.0. (Trinity, RRID:SCR_013048) [52] (genome-guided bam, max intron = 50 kb, min coverage = 3) to assemble the transcriptome (named "Dsil-RefGuided transcriptome"; Supplementary Table S1-10). We used the MAKER2 v2.31.9 (MAKER2, RRID:SCR_005309) [53] genome annotation pipeline for the structural annotation of *D. silvatica* genes (Supplementary Fig. S2), using both *ab initio* gene predictions and annotation evidences from *D. silvatica* and other sources. For the *ab initio* gene predictions we initially trained Augustus v3.1.0 (Augustus, RRID:SCR_008417) [54] and SNAP (SNAP, RRID:SCR_002127) [55] softwares using scaffolds longer than 20 kb, and BUSCO gene models generated from completeness searches. Then we iteratively included a reliable set of proteins for a further training. This dataset was composed of the 9,473 orthologs 1:1 identified in 5 *Dysdera* species and the 1:1 orthologs among spiders available at OrthoDB v10 (OrthoDB, RRID:SCR_011980) [56] (8,792). After several iterative training rounds, we applied MAKER2, Augustus, and SNAP, adding other sources of evidence: (i) transcript evidence (Dsil-RefGuided transcriptome), (ii) RNAseq reads exon junctions generated with HISAT2 [51] and regtools [57], and (iii) proteins annotated in other arthropods, especially chelicerates (Fig. 2; Supplementary Table S1-11). The annotation process resulted in 48,619 protein-coding and 33,934 transfer RNA (tRNA) genes. The mean annotation edit distance (AED) upon protein-coding genes was 0.32 (Supplementary Fig. S6), which is typical of a well-annotated genome [58, 59]. After each training and iterative annotation round, we checked the improvement of the annotation by means of the cumulative fraction of AED (Supplementary Table S1-12a; Supplementary Fig. S7).

We searched for the presence of protein domain signatures in annotated protein-coding genes using InterProScan v5.15-54 (InterProScan, RRID:SCR_005829) [60,61], which includes information from public databases (see additional details in Supplementary Table S1-7). Additionally, we used NCBI BLASTP v.2.4.0 (BLASTP, RRID:SCR_001010) [31] (E-value cutoff $<10^{-5}$; >75% alignment length) against the Swiss-Prot database to annotate *D. silvatica* genes. We found that 74.9% (36,398 genes) of the predicted protein-coding genes have hits with records of either InterPro (32,322 genes) (InterPro, RRID:SCR_006695) or Swiss-Prot (17,225 cases) (Table 2; Supplementary Table S1-7).

Table 3. Completeness analysis[a]

BLAST analysis[b]	Number Identified (%)
Parasteatoda genes (n = 30,041)	19,580 (65.2)
Single-copy *Dysdera* (n = 9,473)	8,420 (88.9)
Single-copy spiders (n = 2,198)	2,141 (97.4)
CEG (n = 457)	438 (95.8)
BUSCO analysis[c]	
Metazoa (n = 978)	
Identified BUSCO	882 (90.2)
Complete (C)	689 (70.5)
Single copy (S)	662 (67.7)
Duplicated (D)	27 (2.8)
Fragmented (F)	193 (19.7)
Missing (M)	96 (9.8)
Artrhopoda (**n = 1,066**)	
Identified BUSCO	959 (89.9)
Complete (C)	736 (69.1)
Single copy (S)	702 (65.9)
Duplicated (D)	34 (3.2)
Fragmented (F)	223 (20.9)
Missing (M)	107 (10.0)

[a]Completeness analysis of the 36,398 functional annotated proteins of *D. silvatica*.
[b]BLASTP searches against different datasets. E-value cutoff $< 10^{-3}$, alignment length cutoff > 30%, and identity cutoff > 30%.
[c]BUSCO analysis using default parameters against different datasets (BUSCO, RRID:SCR_015008).

Completeness

We determined the completeness of the *D. silvatica* genome assembly (Table 3) using BLASTP (E-value cutoff $<10^{-3}$; >30% of alignment length and identity > 50%). We searched for homologs of the functionally annotated peptides (36,398) (i) among CEG genes of *Drosophila melanogaster* [41]; (ii) among the pre-

dicted peptides of *Parasteatoda tepidariorum*, a spider with a well-annotated genome [62]; (iii) among the 9,473 1:1 orthologs across 5 *Dysdera* species; and (iv) among the 2,198 single-copy genes identified in all spiders and available in OrthoDB v10 [56]. We found in *D. silvatica* a high fraction of putative homologs (95.8% of CEG genes, and 97.4% spider-specific single-copy genes; Table 3). Furthermore, the analysis based on the putative homologs of the single-copy genes included in the BUSCO dataset (BUSCO, RRID:SCR_015008) [40], applying the default parameters for the genome and protein mode, also demonstrated the high completeness of the genome draft. Indeed the analysis recovered the ~90% of Metazoa or Arthropoda genes (v9), and nearly 70% of them are complete in *D. silvatica*.

We extended the search for *D. silvatica* homologs to a broader taxonomic range (Fig. 2; Supplementary Table S1-11) by including other metazoan lineages and performing a series of local BLASTP searches (E-value cutoff $< 10^{-3}$; >30% alignment length). We found that a great majority of *D. silvatica* genes are shared among arthropods (57.9%), 11,995 of them (32.95%) also being present in Ecdysozoa (Fig. 4a). Remarkably, 9,560 genes appears to be spider-specific, 4,077 of them being specific (unique) of *D. silvatica*. Despite almost all these species-specific genes having interproscan signatures, the annotation metrics are poor compared with genes having homologs in other species (Supplementary Table S1-12b; Supplementary Figs S7 and S9); indeed, they have an average number of exons (2.8) and gene length (~168aa), which may reflect their partial nature. They could be part of very large genes interspersed by repeats or complex sequences difficult to assemble. The analysis using OrthoDB (v10) [56] across 5 chelicerates (including *D. silvatica*) identified 1,798 genes, with 1:1 orthologous relationships (Fig. 4b), while 12,101 *D. silvatica* genes showed other more complex orthologous/homologous relationships (Fig. 4b, Supplementary Table S1-12c and S3-1). The analysis across the genome annotations of some representative arthropods identified 950 genes with 1:1 orthologous relationships (Supplementary Fig. S8, Supplementary Table S1-12c and S3-2).

Mitochondrial genome assembly and annotation

We assembled the mitochondrial genome of *D. silvatica* (mtDsil) from 126,758 reads identified in the 100PE library by the software NOVOPlasty [63]. Our *de novo* assembly yielded a unique contig of 14,440 bp (coverage of 878×) (Supplementary Table S1-13). CGVIEW (CGVIEW, RRID:SCR_011779) [64] was used to generate a genome visualization of the annotated mtDsil genome (Supplementary Fig. S10). We identified 2 ribosomal RNAs, 13 protein-coding genes, and 15 tRNAs (out of the putative 22 tRNAs). Based on the contig length and the inability of standard automatic annotation algorithms to identify tRNA with missing arms, as reported for spiders [65], the complete set of tRNAs is most likely present for this species.

Conclusion

We have reported the assembly and annotation of the nuclear and mitochondrial genomes of the first representative of the spider superfamily Dysderoidea and the second genome of a Synspermiata, one of the main evolutionary lineages within the "true spiders" (Araneomorphae) and still sparsely sampled at the genomic level [14]. Despite the high coverage and the hybrid assembly strategy, the repetitive nature of the *D. silvatica* genome precluded obtaining a high-continuity draft. The characteristic holocentric chromosomes of Dysderidae [17] may also explain the observed genome fragmentation; indeed, it has been recently shown that genome-wide centromere-specific repeat arrays are interspersed among euchromatin in holocentric plants (Rhynchospora, Cyperceae) [66].

Nevertheless, the completeness and the extensive annotations achieved for this genome, as well as the new reference-guided transcriptome, make this draft an excellent source tool for further functional and evolutionary analyses in this and other related species, including the origin and evolution of relevant spider traits, such as venom and silk. Moreover, the availability of new genomic information in a lineage with remarkable evolutionary features such as recurrent colonizations of the underground environment or complex reproductive anatomies indicative of cryptic female choice, to cite 2 examples, will further provide valuable knowledge about relevant scientific questions, such as the molecular basis of adaptation to extreme habitats or the genetic drivers of sexual selection, along with more general aspects related to gene content across main arthropod groups, the consequences of whole-genome duplications, or phylogenetic relationships with the Araneae. Additionally, because this genus experienced a spectacular adaptive radiation in the Canary Islands, the present genome draft could be useful to further studies investigating the genomic basis of island radiations.

Availability of supporting data and materials

The whole-genome shotgun project has been deposited at DDBJ/ENA/GenBank under accession number QLNU00000000 and project ID PRJNA475203. The version described in this article is version QLNU01000000. This project repository includes raw data, sequencing libraries information, and assemblies of the mitochondrial and nuclear genomes. Other relevant datasets such as annotation, reference-guide assembled transcripts, repeat, and HCR data, as well as other data relevant for the reproducibility of results, are available in the GigaDB dataset [67].

Additional file

File S1. Supplemental Material Summary
SanchezHerrero_Dsilvatica_SupMaterial_Summary.pdf

Availability of supporting source code and requirements

The scripts employed and developed in this project are available under the github repository:
Project name: Genome assembly of *Dysdera silvatica*
Project home page: https://github.com/molevol-ub/Dysdera_silvatica_genome
Operating system(s): Platform independent
Programming language: Bash, Perl, Python, R
License: MIT

Abbreviations

AED: annotation edit distance; AGOUTI: Annotated Genome Optimization Using Transcriptome Information; BLAST: Basic Local Alignment Tool; bp: base pair; BUSCO: Benchmarking Universal Single Copy Orthologs; CEG: core eukaryotic gene; Cz: Cretaceous period; Dsil: *Dysdera silvatica*; Gb: gigabase pairs; GC: guanine cytosine; GO: Gene Ontology; HCR: high-coverage re-

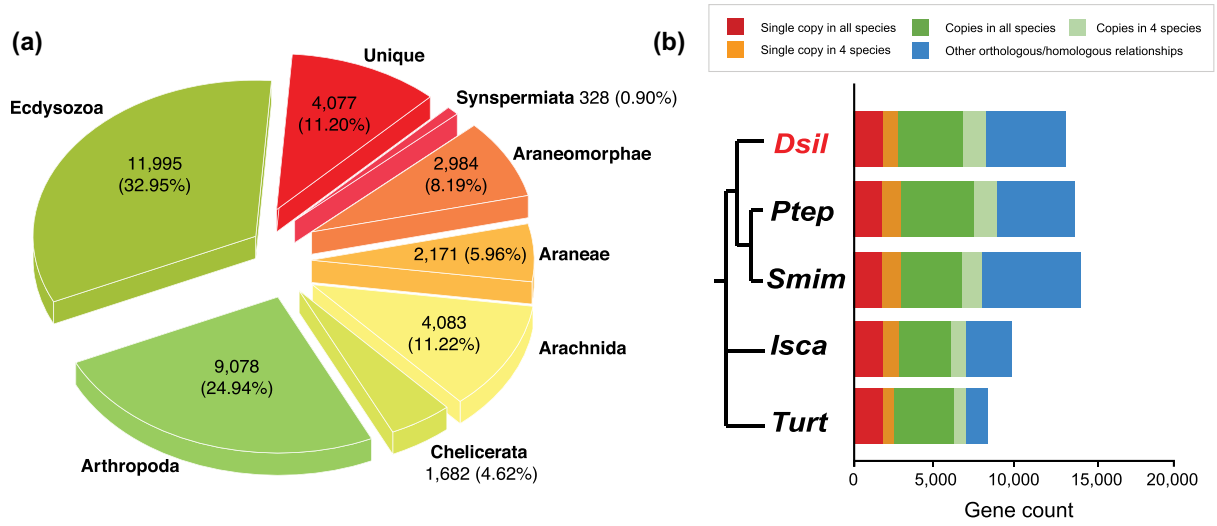


Figure 4 (a) Pie chart illustrating the taxonomic distribution of positive BLAST hits of the *D. silvatica* protein-coding genes against the sequence data of species included in Fig. 2. **(b)** Homology relationships among *D. silvatica* (Dsil) and different chelicerates genomes available in OrthoDB v10 [56], *Parasteatoda tepidariorum* (Ptep), *Stegodyphus mimosarum* (Smim), *Ixodes scapularis* (Isca), and *Tetranychus urticae* (Turt). Red and orange bars indicate the fraction of single-copy genes (1:1 orthologs) identified in all species, and in all but 1 (e.g., missing in 1 species), respectively. The dark and light green bar indicate the fraction of orthologs present in all species and in all but 1, respectively, that are not included in previous categories. The blue bar (other orthology/homology) shows other more complex homologous relationships. The results were generated by uploading *D. silvatica* proteins to the OrthoDB web server.

gions; Isca: *Ixodes scapularis*; kb: kilobase pairs; LINE: long interspersed nuclear element; LTR: long terminal repeats; MaSuRCA: Maryland Super-Read Celera Assembler; Mb: megabase pairs; MP: mate pair; Mya: million years ago; NCBI: National Center for Biotechnology Information; PacBio: Pacific Biosciences; PE: paired-end; PRINSEQ: PReprocessing and INformation of SEQuence data; Ptep: *Parasteatoda tepidariorum*; RNAseq: RNA sequencing; SINE: short interspersed nuclear element; Smim: *Stegodyphus mimosarum*; SMRT: Single-Molecule Real Time; tRNA: transfer RNA; Turt: *Tetranychus urticae*.

Competing interests

The authors declare that they have no competing interests.

Funding

This study was supported by the Ministerio de Economía y Competitividad of Spain (CGL2012-36863, CGL2013-45211, and CGL2016-75255), and by the Comissió Interdepartamental de Recerca I Innovació Tecnològica of Catalonia, Spain (2014SGR-1055 and 2014SGR1604). J.F.S.-H. was supported by a Formación del Profesor Universitario (FPU) grant (Ministerio de Educación of Spain, FPU13/0206); C.F.-L. by an IRBio PhD grant; S.H-A by Becas Postdoctorales en el Extranjero CONACyT; A.S.-G. by a Beatriu de Pinós grant (Generalitat de Catalunya, 2010-BP-B 00175); and J.R. and M.A.A. were partially supported by ICREA Academia (Generalitat de Catalunya).

Authors' contributions

J.R., A.S.-G., and M.A.A designed the study. C.F.-L., J.F.S.-H., P.E., and S.H-A. processed the samples and extracted DNA. J.F.S.-H. performed the bioinformatics analysis and drafted the manuscript. J.F.S.-H., A.S.-G., and J.R. interpreted the data. All authors revised and approved the final manuscript.

Acknowledgments

We acknowledge the Garajonay National Parks for granting collection permits and helping with lodging and logistics during fieldwork. We also thank CNAG (Centro Nacional de Análisis Genómico) for the Nanopore sequencing facilities.

References

1. World Spider Catalog (2018). 2018. http://wsc.nmbe.ch. Accessed on April 2019.
2. Pekár S, Toft S. Trophic specialisation in a predatory group: the case of prey-specialised spiders (Araneae). Biol Rev 2015;**90**(3):744–61.
3. Hopkin SP, Martin MH. Assimilation of zinc, cadmium, lead, copper, and iron by the spider *Dysdera crocata*, a predator of woodlice. Bull Environ Contam Toxicol 1985;**34**:183–87.
4. Pekár S, Líznarová E, Řezác M. Suitability of woodlice prey for generalist and specialist spider predators: a comparative study. Ecol Entomol 2016;**41**(2):123–30.
5. Toft S, Macías-Hernández N. Metabolic adaptations for isopod specialization in three species of *Dysdera* spiders from the Canary Islands. Physiol Entomol 2017;**42**(2):191–98.
6. Řezáč M, Pekár S. Evidence for woodlice-specialization in *Dysdera* spiders: behavioural versus developmental approaches. Physiol Entomol 2007;**32**(4):367–71.
7. Řezác M, Pekár S, Lubin Y. How oniscophagous spiders overcome woodlouse armour. J Zool 2008;**275**(1):64–71.
8. Arnedo MA, Oromí P, Ribera C. Radiation of the spider genus *Dysdera* (Araneae, Dysderidae) in the Canary Islands: cladistic assessment based on multiple data sets. Cladistics 2001;**17**:313–353.
9. Macias-Hernandez N, de la Cruz López S, Roca-Cusachs M, et al. A geographical distribution database of the genus *Dysdera* in the Canary Islands (Araneae, Dysderidae). Zookeys 2016;**625**(625):11–23.
10. Arnedo MA, Oromí P, Múrria C, et al. The dark side of an is-

land radiation: systematics and evolution of troglobitic spiders of the genus *Dysdera* Latreille (Araneae: Dysderidae) in the Canary Islands. Invertebr Syst 2007;**21**(6):623.
11. Michalik P, Ramírez MJ. Evolutionary morphology of the male reproductive system, spermatozoa and seminal fluid of spiders (Araneae, Arachnida) - current knowledge and future directions. Arthropod Struct Dev 2014;**43**(4):291–322.
12. Wheeler WC, Coddington JA, Crowley LM, et al. The spider tree of life: phylogeny of Araneae based on target-gene analyses from an extensive taxon sampling. Cladistics 2017;**33**(6):574–616.
13. Fernández R, Kallal RJ, Dimitrov D, et al. Phylogenomics, diversification dynamics, and comparative transcriptomics across the spider tree of life. Curr Biol 2018;**28**(9):1489–97.
14. Garb JE, Sharma PP, Ayoub NA. Recent progress and prospects for advancing arachnid genomics. Curr Opin Insect Sci 2018;**25**:51–7.
15. Vizueta J, Rozas J, Sánchez-Gracia A. Comparative genomics reveals thousands of novel chemosensory genes and massive changes in chemoreceptor repertories across chelicerates. Genome Biol Evol 2018;**10**(5):1221–36.
16. Deeleman-Reinhold CL. The genus *Rhode* and the harpacteine genera *Stalagtia, Folkia, Minotauria*, and *Kaemis* (Araneae, Dysderidae) of Yugoslavia and Crete, with remarks on the genus *Harpactea*. Rev Arachnol 1993;**10**(6):105–35.
17. Diaz MO, Maynard R, Brum-Zorrilla N. Diffuse centromere and chromosome polymorphism in haplogyne spiders of the families dysderidae and segestriidae. Cytogenet Genome Res 2010;**128**(1-3):131–8.
18. Uhl G. Two distinctly different sperm storage organs in female *Dysdera erythrina* (Araneae: Dysderidae). Arthropod Struct Dev 2000;**29**(2):163–9.
19. Burger M, Kropf C. Genital morphology of the haplogyne spider *Harpactea lepida* (Arachnida, Araneae, Dysderidae). Zoomorphology 2007;**126**(1):45–52.
20. Ballesteros JA, Sharma PP. A critical appraisal of the placement of Xiphosura (chelicerata) with account of known sources of phylogenetic error. Syst Biol 2019, doi:10.1093/sysbio/syz011.
21. Lozano-Fernandez J, Tanner AR, Giacomelli M, et al. Increasing species sampling in chelicerate genomic-scale datasets provides support for monophyly of Acari and Arachnida. Nat Commun 2019;**10**:2295.
22. Carlson DE, Hedin M. Comparative transcriptomics of Entelegyne spiders (Araneae, Entelegynae), with emphasis on molecular evolution of orphan genes. PLoS One 2017;**12**(4):e0174102.
23. Gregory TR. Animal Genome Size Database. 2018. http://www.genomesize.com.
24. Hare EE, Johnston JS. Genome size determination using flow cytometry of propidium iodide-stained nuclei. Methods Mol Biol 2011;**772**:3–12.
25. Rasch EM, Connelly BA. Genome size and endonuclear DNA replication in spiders. J Morphol 2005;**265**(2):209–14.
26. Gregory TR, Shorthouse DP. Genome sizes of spiders. J Hered 2003;**94**(4):285–90.
27. Dpooležel J, Binarová P, Lcretti S. Analysis of nuclear DNA content in plant cells by flow cytometry. Biol Plant 1989;**31**(2):113–20.
28. Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics 2011;**27**(6):764–70.
29. Vurture GW, Sedlazeck FJ, Nattestad M, et al. GenomeScope: fast reference-free genome profiling from short reads. Bioinformatics 2017;**33**(14):2202–4.
30. Austin CM, Tan MH, Harrisson KA, et al. De novo genome assembly and annotation of Australia's largest freshwater fish, the Murray cod (*Maccullochella peelii*), from Illumina and Nanopore sequencing read. GigaScience 2017;**6**(8):1–6.
31. Altschul SF, Gish W, Miller W, et al. Basic Local Alignment Search Tool. J Mol Biol 1990;**215**(3):403–10.
32. Schmieder R, Edwards R. Fast identification and removal of sequence contamination from genomic and metagenomic datasets. PLoS One 2011;**6**(3):e17288.
33. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 2014;**30**(15):2114–20.
34. O'Connell J, Schulz-Trieglaff O, Carlson E, et al. NxTrim: optimized trimming of Illumina mate pair reads. Bioinformatics 2015;**31**(12):2035–7.
35. PacBio. Single Molecule Real Time (SMRT). https://www.pacb.com/products-and-services/analytical-software/smrt-analysis/.
36. Walker BJ, Abeel T, Shea T, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One 2014;**9**(11):e112963.
37. Zimin AV, Marçais G, Puiu D, et al. The MaSuRCA genome assembler. Bioinformatics 2013;**29**(21):2669–77.
38. Zhang SV, Zhuo L, Hahn MW. AGOUTI: improving genome assembly and annotation using transcriptome data. GigaScience 2016;**5**(1):31.
39. Vizueta J, Frias-López C, Macías-Hernández N, et al. Evolution of chemosensory gene families in arthropods: insight from the first inclusive comparative transcriptome analysis across spider appendages. Genome Biol Evol 2017;**9**(1):178–96.
40. Simão FA, Waterhouse RM, Ioannidis P, et al. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 2015;**31**(19):3210–2.
41. Parra G, Bradnam K, Ning Z, et al. Assessing the gene space in draft genomes. Nucleic Acids Res 2009;**37**(1):289–97.
42. Vizueta, J., Macías-Hernández, N., Arnedo, MA., Rozas, J.and Sánchez-Gracia, A.(2019) Chance and predictability in evolution: the genomic basis of convergent dietary specializations in an adaptive radiation. Mol. Ecol. doi:10.1111/mec.1519931359512
43. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 2009;**25**(16):2078–9.
44. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods 2012;**9**(4):357–9.
45. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 2018;**34**(18):3094–100.
46. Smit AF, Hubley R. RepeatModeler Open-1.0. 2008–2015. http://www.repeatmasker.org.
47. Smit AF, Hubley R, Green P. RepeatMasker Open-3.0. 1996–2010. http://www.repeatmasker.org.
48. Wheeler TJ, Clements J, Eddy SR, et al. Dfam: a database of repetitive DNA based on profile hidden Markov models. Nucleic Acids Res 2012;**41**(D1):D70–D82.
49. Bao W, Kojima KK, Kohany O. Repbase Update, a database of repetitive elements in eukaryotic genomes. Mobile DNA 2015;**6**(1):11.
50. Jurka J, Kapitonov VV, Pavlicek A, et al. Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res 2005;**110**(1-4):462–7.
51. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nat Methods

2015;**12**(4):357–60.
52. Haas BJ, Papanicolaou A, Yassour M, et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc 2013;**8**(8):1494–512.
53. Holt C, Yandell M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics 2011;**12**(1):491.
54. Stanke M, Steinkamp R, Waack S, et al. AUGUSTUS: a web server for gene finding in eukaryotes. Nucleic Acids Res 2004;**32**(Web Server issue):W309–12.
55. Korf I. Gene finding in novel genomes. BMC Bioinformatics 2004;**5**(1):59.
56. Kriventseva EV, Kuznetsov D, Tegenfeldt F, et al. OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. Nucleic Acids Res 2019;**47**(D1):D807–D811.
57. Feng YY, Ramu A, Cotto KC, et al. RegTools: integrated analysis of genomic and transcriptomic data for discovery of splicing variants in cancer. bioRxiv 2018, doi:10.1101/436634.
58. Eilbeck K, Moore B, Holt C, et al. Quantitative measures for the management and comparison of annotated genomes. BMC Bioinformatics 2009;**10**(1):67.
59. Yandell M, Ence D. A beginner's guide to eukaryotic genome annotation. Nature Rev Genet 2012;**13**(5):329–42.
60. Mitchell AL, , Attwood TK, Babbitt PC, et al. InterPro in 2019: improving coverage, classification and access to protein sequence annotations. Nucleic Acids Res 2019;**47**(D1):D351–60.
61. Jones P, Binns D, Chang HY, et al. InterProScan 5: genome-scale protein function classification. Bioinformatics 2014;**30**(9):1236–40.
62. Schwager EE, Sharma PP, Clarke T, et al. The house spider genome reveals an ancient whole-genome duplication during arachnid evolution. BMC Biol 2017;**15**(1):62.
63. Dierckxsens N, Mardulyn P, Smits G. NOVOPlasty: de novo assembly of organelle genomes from whole genome data. Nucleic Acids Res 2016;**45**(4):gkw955.
64. Stothard P, Wishart DS. Circular genome visualization and exploration using CGView. Bioinformatics 2005;**21**(4):537–9.
65. Masta SE, Boore JL. The complete mitochondrial genome sequence of the spider *Habronattus oregonensis* reveals rearranged and extremely truncated tRNAs. Molec Biol Evol 2004;**21**(5):893–902.
66. Marques A, Ribeiro T, Neumann P, et al. Holocentromeres in *Rhynchospora* are associated with genome-wide centromere-specific repeat arrays interspersed among euchromatin. Proc Natl Acad Sci U S A 2015;**112**(44):13633–8.
67. Sánchez-Herrero JF, Frías-López C, Escuer P, et al. Supporting data for "The draft genome sequence of the spider *Dysdera silvatica* (Araneae, Dysderidae): a valuable resource for functional and evolutionary genomic studies in chelicerates." GigaScience Database 2019; http://dx.doi.org/10.5524/100628.