Evolutionary Genomics of Panarthropoda: Study of the chemosensory gene families across phyla and the radiation of the spider genus *Dysdera* (Araneae, Dysderidae) in the Canary Islands

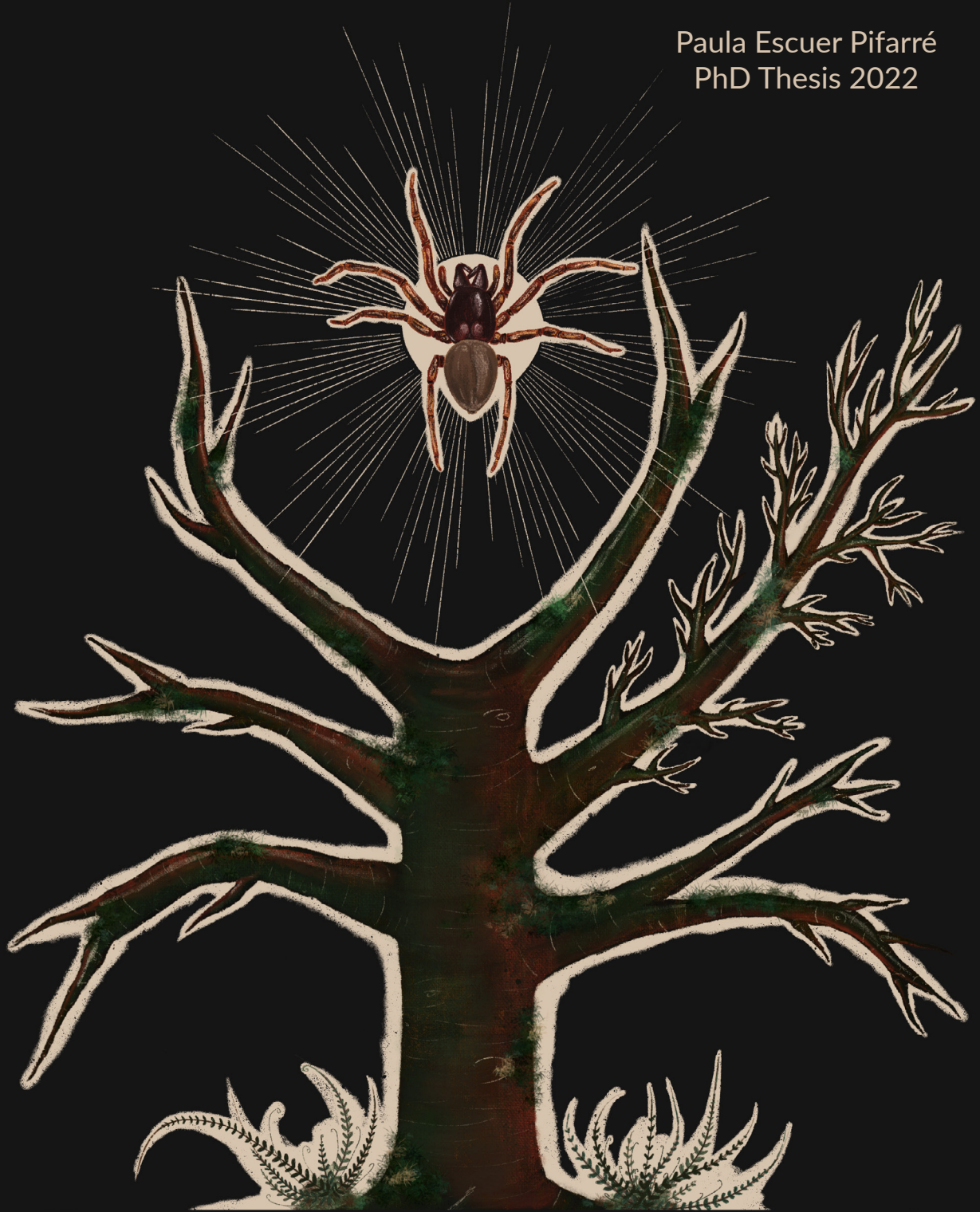
Paula Escuer Pifarré

Evolutionary Genomics of Panarthropoda:

Study of the chemosensory gene families across phyla and the radiation of the spider genus *Dysdera* (Araneae, Dysderidae) in the Canary Islands

Paula Escuer Pifarré

PhD Thesis 2022



UNIVERSITAT DE BARCELONA

Evolutionary Genomics of Panarthropoda: Study of the chemosensory gene families across phyla and the radiation of the spider genus *Dysdera* (Araneae, Dysderidae) in the Canary Islands

Memòria presentada per **Paula Escuer Pifarré**
per optar al Grau de Doctora per la Universitat de Barcelona.

Departament de Genètica, Microbiologia i Estadística

L'autora de la tesi
Paula Escuer Pifarré

El codirector i tutor de la tesi
Dr. Alejandro Sánchez-Gracia

Professor agregat
Departament de Genètica,
Microbiologia i Estadística Facultat
de Biologia
Universitat de Barcelona

El director de la tesi
Dr. Julio Rozas

Catedràtic de Genètica
Departament de Genètica,
Microbiologia i Estadística
Facultat de Biologia
Universitat de Barcelona

Barcelona, Setembre de 2022

"Qui no es mou, no sent les cadenes"
Rosa Luxemburg

Al Yayo i la Padrina

AGRAIMENTS

Als meus supervisors. Julio, gràcies per donar-me la oportunitat de poder fer la tesis en aquest grup, per ensenyar-me a ser detallista en la ciència (i en la estètica), per tots els dubtes que m'has respost pacientment i per ajudar-me a créixer com a investigadora. Alex, agraeixo que m'hagis ensenyat a ser una persona crítica i a apassionar-me per la genètica evolutiva. També vull fer una menció als meus primers supervisors del TFG, Eduardo, Ona i Marta, que em van ensenyar a fer els primers passos en la investigació i a les Postdocs, Sara i Sílvia, per fer-me de mentores i per la seva inestimable ajuda.

Als meus amics i companys del laboratori. Raquel, moltes gràcies per la nostra amistat, per haver recorregut aquest camí juntes i veure com n'hem sortit victorioses, has sigut un suport molt important per mi. A la Cris, la Eva, Juanma i Joel, gracies per les cerveses, pels dinars i els maldecaps compartits (Juanma, gràcies pel teu ajut amb les figures i la maquetació). Laia, Lisandra, Cristian, Dani, Edu, Joan, Ricard, Vadim, Gema, David i tots els que han passat per aquí, gràcies per fer aquest viatge molt més amè. Serà difícil trobar un altre lloc amb gent tan increïble com vosaltres.

Als meus camarades de Doctorands i OCR, per lluitar cada dia per un món més just.

Als companys de Sheffield que em van fer sentir com a casa.

Als meus amics, a la Raquel, la Carme, l'Oscar, la Mireia, el Josep, el Gepe, la Bàrbara, l'Adri, i en especial a la Eli, per aquesta portada tan bonica i a l'Alvaro, que em va animar a començar amb la bioinformàtica.

A tota la meva gran família, especialment als meus pares i germans, que sempre estan allí per recolzar-me. A la meva mare per donar sempre tant.

I finalment, a l'Arian, la persona que (probablement més m'ha aguantat durant aquests mesos d'estrès) fa que tot valgui la pena.

Abstract

During the last decade, advances in high-throughput sequencing (HTS) technologies have increased exponentially the generation of genomic data, especially of non-model species. In this thesis, we generated and used available HTS data to shed light on relevant evolutionary processes that shape genome structure and organization, such as the origin and evolution of multigene families, and the genomic determinants of adaptive radiations.

First, we contribute to the genome assembly of the spider *Dysdera silvatica*, the first published assembly of a member of the Dysderidae family. For that we used a hybrid assembly strategy based on short (Illumina) and long (low coverage PacBio and Nanopore) reads to generate an assembly of moderate to low continuity (N50 of 38 kb). Then, we upgraded the assembly to a high quality, chromosome-level, using Chicago and Hi-C libraries, which raised the continuity of the previous draft considerably (N50 of 174.19 Mb). This newly generated genomic resource is not only a significant contribution to the field of spider genomics, but also the backbone for many of the analyses presented in this thesis. We identified 33,275 putative protein-coding genes in the genome of *D. silvatica* (87% of them were functionally annotated) and estimated that 52.87% of the genome are repetitive elements. We used this high-quality assembly to perform a comprehensive study of the chromosomal location and evolutionary divergence of chemoreceptor gene families. We identified 545 chemoreceptor genes, which are highly clustered in the genome of *D. silvatica* and show a clear correlation between physical and evolutionary distances.

Second, we used available genomics data, and newly generated transcriptomes, to perform a phylogenetically deep study of the evolution of the chemosensory gene families across Panarthropoda. We characterized the expression of chemoreceptor genes in the antenna of the velvet worm *Euperipatoides rowelli* and performed a comparative genomic analysis of these gene families across representatives of the phyla composing the Panarthropoda clade. Noticeably, we did not find the expression of the Ionotropic co-receptor IR25a in onychophorans, which questions the chemosensory role of this family in this lineage. Surprisingly, eutardigrades

lack genes encoding the DEG-ENaC and CD36/SNMPs gene families, a feature that could be related to their extraordinary resistance to desiccation. Finally, we propose the NPC2 gene family as the candidate chemosensory soluble protein in the Panarthropoda ancestor.

We also used the chromosome-level assembly of *D. silvatica* to study some aspects of the adaptive radiation of the spider genus *Dysdera* in the Canary Islands. We performed a population genomics study of a natural population of *D. silvatica* from La Gomera using a whole-genome re-sequencing strategy. We found unexpected high levels of nucleotide polymorphism and a lower X/autosomes polymorphism ratio than the theoretically anticipated from their corresponding effective number of chromosomes. Demographic inference based on the coalescence predicted that in the past this species likely underwent a long period characterized by a large effective population size (about 100 times greater than the estimate for the present). We hypothesize that the high levels of polymorphism currently observed in the population of *D. silvatica* is the molecular hallmark of a long period of ancestral population structure that dominated the history of this species for almost 200 Kya.

Globally, our results provide very valuable new knowledge about the genomic bases of adaptation at different evolutionary timescales, ranging from the deep evolutionary dynamics of chemosensory repertories across Panarthropoda phyla, to the very recent origin of new copies and the early steps of their evolution within a genome, including the genomic signals of positive selection and demographic history at the scale of an adaptive radiations. This thesis also demonstrates the great advantage of having highly continuous genome assemblies to obtain reliable data when working gene families, and other repetitive elements, or to perform comprehensive chromosome-level population genomics analyses.

Resum

Durant la darrera dècada, els avenços en les tecnologies de *High-Throughput Sequencing* (HTS) van augmentar exponencialment la generació de dades massives, proporcionant una gran quantitat d'informació genòmica d'espècies no models. L'objectiu principal d'aquesta tesi és investigar processos evolutius com l'origen i l'evolució de les famílies de gens i els mecanismes genòmics subjacents a l'adaptació en la diversificació d'espècies. Per això, en primer lloc, hem contribuït a la construcció de l'assemblatge genòmic de l'aràcnid *Dysdera silvatica*, el primer genoma de la família Dysderidae amb lectures curtes, amb una petita fracció de lectures llargues i un valor N50 de 38 kb. A continuació, el vam millorar a un genoma a nivell de cromosoma d'alta qualitat, utilitzant llibreries de Chicago i Hi-C, a més d'augmentar el valor de N50 a 174,19 Mb. Vam identificar 33.275 gens (el 87% amb anotacions funcionals), i determinar que el 52,87% del genoma estava representat per elements repetitius. Vàrem identificar 545 gens quimioreceptors, el 54% d'ells distribuïts en 83 clústers genòmics que es trobaven majoritàriament en els *scaffolds* més petits. D'altra banda, hem estudiat i caracteritzat les famílies multigèniques quimiosensorials a diverses espècies dels grup dels Panarthropoda, centrant-nos en Onychophora i Tardigrada. Hem caracteritzat les principals famílies quimiosensorials de l'onicòfor *Euperipatoides rowelli* i hem trobat que no tenen el gen que codifica el receptors ionotròpic IR25a, present en la majoria d'espècies de Protòstoms. A més, les famílies de gens DEG-ENaC i CD36/SNMPs són absents a les espècies de la classe Eutardigrada, mentre que la família de gens NPC2 seria l'única que codifica proteïnes solubles en l'ancestre Panartropoda. Finalment, hem estudiat la radiació adaptativa de l'aranya del gènere *Dysdera* a les Illes Canàries, mitjançant una anàlisi de genòmica de poblacions. Hem determinat que a pesar de ser una espècie endèmica i que viu a illes molt petites, presenta un alt nivell de polimorfisme nucleotídic. Les inferències demogràfiques suggereixen que aquesta espècie té una gran grandària efectiva poblacional, probablement causada per una estructuració poblacional ancestral. Els resultats d'aquesta tesi són un recurs valuós que ens ha permès caracteritzar detalladament les famílies multigèniques del sistema quimiosensorial, a més proporcionar nou coneixement sobre paper relatiu de la selecció positiva en les radiacions adaptatives.

Summary

Introduction

1 Evolutionary genomics

1.1 Molecular evolution

Biological evolution stands for the change across generations of heritable features, encoded by genes; this process, that acts within populations of a single species, is a last resort responsible for the biodiversity on Earth. Current molecular data supports that life is monophyletic, meaning that all life forms descend from a single ancestor (organism) that existed over 3.5 billion years ago; in the same way, we can assume that all extant genes derive from a single ancestral copy. The field of evolutionary genetics tries to understand how these genes have diverged across the tree of life and the driving forces underlying this process.

Even though many people have tried to explain the evolutionary process in different ways and means in the past, current scientific-based mechanisms of evolution were conceived very recently in human history. In 1859, Charles Darwin and Alfred R. Wallace not only reported many examples of the "descent with modification" concept, but also provided the idea on how favorable variants can prevail over others, as the key mechanism for evolution, since then known as natural selection. At the same time, Gregor Mendel, conceived its concept of inheritance from his experiments with pea plants in 1865; it was not until the 1900, however, that his knowledge transcended and triggered the growth of Genetics as a new science, founded on the study of the inheritance of qualitative traits, and that ultimately fostered the modern ideas about the evolutionary process[1]. During this period, it was established that genes are particular identities that can be transmitted over generations from parent to their offspring, and that their inheritance follows specific mathematical rules. Not much later, geneticists and statisticians such as Ronald Fisher, Jack Haldane and Sewall Wright contributed to the development of the basic theory of population genetics[2]. Based on this synthesis, they defined the five evolutionary mechanisms: mutation, genetic drift, migration, recombination and natural selection. Mutation is the ultimate source of variation that, jointly with recombination, increases genetic variation in populations. On the other

hand, natural selection is the only force that explains adaptation of species to their environment.

It is worth mentioning that all these advances took place without any knowledge of the molecular nature of the heritable material that is transmitted from the parents to the offspring: the DNA. This molecule was identified as the "transforming principle" by Avery, MacLeod and McCarty in 1943; and was some years later, in 1953, when Franklin and Wilkins provided the X-ray diffraction pattern that allowed Watson and Crick determining the first correct model of a DNA molecule. Finally, another remarkable discovery in the context of this thesis was in 1977 when Sanger, Nicklen and Coulson developed the most popular DNA sequencing technique that allowed the study of genetic variation for the first time[3].

1.1.1 The Neutral Theory and natural selection

For a long time, natural selection has been considered the (only) key force shaping genetic variation within populations. This theory, however, has some important caveats, such as the difficulties of explaining the unexpected high levels of intraspecific genetic variation (polymorphism) observed at the allozyme loci[1]. Then, in 1968-1969, Mooto Kimura completed one of the most important and revolutionary theories of molecular evolution, "The Neutral Theory of Molecular Evolution", or just "Neutral Theory"[4]. Kimura's ideas were beautifully simple but brilliant: the high levels of polymorphism observed within populations are the consequence of selectively neutral segregating variants, the fate of which is only shaped by genetic drift, the random fluctuation of allele frequencies across generations in finite populations (Figure 1). The Neutral Theory is the baseline of the standard neutral model (SNM), used as the null model to test for putative departures from neutrality of the observed patterns of variation in natural populations. This null hypothesis is very useful to understand the forces and mechanisms underlying molecular evolutionary change and thus, to determine the biological meaning of genetic variation.

Even though most of the observed variants in a genome can be considered as selectively neutral, the Neutral Theory states that most of the mutations occurring in DNA have deleterious effects and are eliminated from the population by purifying selection (they decrease the individual fitness). There is also room, in fact, for positively selected mutations in the Neutral Theory, but in this case, it is considered that they are a very small fraction compared to the total number of mutations. These variants confer an advantage to the carrier individual in particular environmental

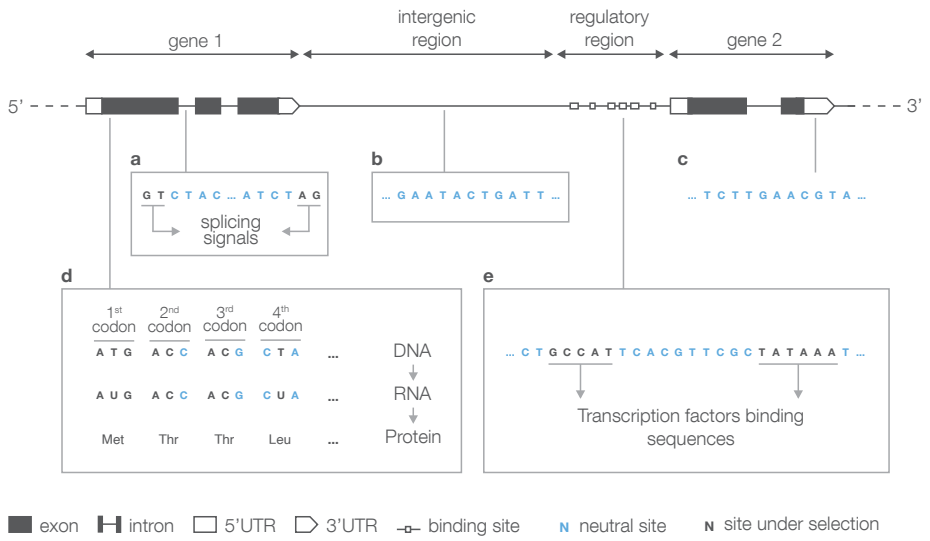


Figure 1. The effects of new mutations in different genomic regions. In blue, positions where mutations would not produce (or produce a little) a phenotypic change, i.e., silent sites that accumulate neutral mutations. These mutations are mostly observed in introns, intergenic regions, untranslated regions (UTRs) within genes, and synonymous sites in the coding regions (a, b, c and d, respectively). Positions in black are those in which mutations usually result in phenotypic changes (a, mutations in splicing sites; d, nonsynonymous changes in the coding sequence), or mutations in regulatory regions affecting gene expression. These positions are the target of natural selection, either positive or negative. Adapted from: Calvo Martín[5].

conditions increasing its fitness and driving them quickly to fixation. Overall, and in the absence of migration or recombination, the probability of fixation of a variant in a population will depend on the interaction of different factors, such as its initial frequency, the effective population size and the intensity of selection (in the case of deleterious and beneficial mutations). It is also possible that two beneficial variants under different conditions coexist in the same locus, producing what is known as balancing selection[1].

1.2 Evolutionary genomics approaches: transcriptomics, comparative and population genomics

Current genetic and genomic applications and tools are becoming indispensable to get insights into the evolutionary processes affecting biodiversity, which are required for many conservation aims. The technological improvements for the DNA sequencing since the apparition of high-throughput sequencing (HTS) technologies accelerated the production of large-scale biological data, leading to the

so called "omics" technologies[6]. Omics are a set of large-scale biological generating data and analyses, useful for understanding the structure, functions or interactions of the genome of organisms. In this thesis, we have used several approaches such as transcriptomics, comparative and population genomics to gain insights into structure of the genome or on evolutionary genomics questions.

Transcriptomics is a molecular biology technique that allows the acquisition of nucleotide sequence information of all transcribed genes from an organism in a specific developmental stage or condition. It has multiple applications, such as providing a whole set of different kinds of transcripts (mRNA, non-coding RNA and small RNAs); analyzing transcriptional structure of genes and differential splicing; quantifying levels of expression in a determined condition or tissue[7].

Genomics, and particularly comparative genomics, is much more useful to understand the evolutionary process. The sequencing and obtention of whole genome assemblies has apported a wider perspective of evolutionary processes than just analyzing patterns in a few loci. Comparative genomics are useful for multiple purposes, for instance, to understand the structure and function across individuals or species, characterize and study the origin and evolution of multigene families, or to obtain reference genomes even from non-model organisms (which can be also used for other "omics" inferences). Lastly, another remarkable application of genomics is population genomics. The concept itself was started to be considered in 2007, with the first large-scale polymorphism projects in Drosophila[8]. It is defined as the sequencing (re-sequencing) of samples of individuals from one or several natural populations; these studies allow estimating genome-wide polymorphism patterns, which are critical to obtain relevant evolutionary-based inferences[6].

1.3 Adaptive radiations

The accumulation of genetic changes across generations can end up with the generation of reproductive isolation barriers that could originate new species, usually a very slow process called speciation[9]. This process has been deeply studied in evolutionary biology; indeed, the understanding of the genetic bases is key for the comprehension of the origin and fate of current biodiversity. In some cases, speciation can also take place rapidly in a short period of time, which leads towards the apparition of multiple species from one or several ancestors due to environmental shifts or the arrival of a species to an unexploded ecological niche. This particular event is defined as adaptive radiation, and usually occurs in some isolated ecosystems, such as islands or lakes[10].

Oceanic Islands are a very interesting model to study adaptive radiations. They are considered natural laboratories because of their geographical isolation from the continent, avoiding gene flow. Consequently, islands harbor lots of endemisms, becoming hotspots of biodiversity, which makes them ideal to study evolutionary processes as speciation[11–13]. One of the most paradigmatic examples are Darwin's finches in Galapagos Islands[14] (which helped to develop the idea of natural selection as the key mechanism for evolution), in addition to African cichlid fishes or the Anolis lizards of the Caribbean[15]. In these cases, there are patterns of morphological or behavior differentiation between groups of phylogenetically close (and even sympatric) species[9]. The process starts from a small population that arrives to a new uncolonized niche, then, the availability of resources and lack of competence allows them to specialize in feeding from different sources, which results in multiple endemisms with low genetic differences and morphological adaptations towards their feeding behavior.

Even though adaptive radiation has been deeply studied, there is still a lot of unknowledge about the genetic bases of this process and the role of evolutive forces shaping genetic variability. For that, adaptive radiation is a unique and interesting kind of event appropriate to clarify some issues on the origin and diversification of species[16–18].

2 Genomic sequencing and Next Generation Sequencing (NGS) technologies

2.1 Brief history of genomic sequencing, and the raise of NGS

2.1.1 The beginning of the DNA sequencing technologies: the first-generation DNA sequencing

The beginning of the genomic era is very contemporary compared to other science disciplines. Back to 1869, the first successful DNA isolation was obtained by Friedrich Mietscher, but it took almost 80 years, in 1953, the determination of the three-dimensional structure of DNA thanks to crystallographic data produced by Rosalind Franklin and Maurice Wilkins[19,20]. In those days, the technologies to sequence protein chains were available but, unfortunately, DNA was more difficult since molecules are longer and hard to separate properly, leading to many failed attempts[21].

It was not until 1977 that Sanger & Coulson developed their technique based on radiolabeled partially digested fragments called "chain termination method"[3,22] . Using this approach, they were able to sequence hundreds of nucleotides from DNA in a few hours. Parallelly, Maxam & Gilbert developed another technique based on the chemical modification of DNA chains and posterior fragmentation, which they called "chemical cleavage technique"[23]. Contrary to the Sanger method, the approach of Maxam & Gilbert does not rely on a DNA polymerase. Even though both techniques are rooted in a similar concept, the first one was more successful due to its precision, robustness and simplicity. This technique, also called nowadays "first generation DNA sequencing", was ruling over the DNA sequencing field for more than 30 years. At that time, the development of the recombinant DNA technologies[24] and later the polymerase chain reaction (PCR)[25], empowered new methodologies to generate the quantity and quality of DNA necessary for the sequencing process[21]. Several improvements were achieved from that moment, such as, for instance, the first automated machines for Sanger sequencing, developed in 1987 by Leroy Hood and Applied Biosystems. These machines were capable of

sequencing contiguous fragments of up to 1 kb of length using dye terminators, allowing the parallel sequencing of hundreds of samples[21,26].

2.1.2 The second-generation DNA sequencing revolution

The Human Genome Project (HGP) accelerated an advance in DNA sequencing methodologies during the decades of the 1980s and 90's. In 1996, the pyrosequencing technique, developed by Pål Nyrén and collaborators, appeared as a revolutionary DNA sequencing method, being considered the start of the second generation of sequencing technologies (also known as HTS or NGS -next generation sequencing). This novel technique included important advances such as the use of non-modified nucleotides (instead of the modified ones in chain termination methods), which can be detected directly by the luminescence generated during the incorporation of the nucleotide by the DNA polymerase (sequencing-by-synthesis technology). For that, pyrosequencing can be already considered as HTS[27–29]. In fact, the patent of this technology was later purchased by 454 Life Sciences, where it evolved into the first major successful commercial NGS technology. The first platforms based on this approach appeared in 2005 (as Roche 454) but they suffered important advances very quickly. New and quick improvements made that just a few years later this platform fell into disuse, being the Illumina company, currently one of the most successful companies, facilitating the access of the HTS technologies to many labs in the world due its effectiveness and affordable cost.

2.1.3 Third-generation DNA sequencing era

The beginning of the "third-generation sequencing" is still a matter of debate, indeed, it is sometimes difficult to establish the limit with the previous generation. It is mainly assumed that only "single molecule real-time sequencing" (SMRT) and "nanopore sequencing" are third-generation sequencing[30,31]. The principal representatives of these technologies are the SMRT of Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT). The first one is based on the parallel synthesis of single, very long reads (more than 10 kb of length), sequenced in a short matter of time (a few hours). On the other hand, ONT was the first company to create in 2014 some small devices (the nanopore sequencers known as GridION or MinION), that can sequence large DNA fragments much faster and cheaper than any other sequencer up to date[32–35]. These sequencers use the changes in electrical conductivity produced by DNA strands when passing through biological nanopores to identify the nucleotide that is in the pore.

These impressive developments have reduced the cost of DNA sequencing several orders of magnitude during the last decades, while the speed and quality of the final output have been successfully increased. As an example, it took almost 15 years to complete the Human Genome Project, finished in 2003 with an estimated cost of over $3000 million (about 1$ per base pair). Currently, a complete human genome can be sequenced in less than one day for about $300[30], that is 10^7 base pairs per 1$, or a reduction of seven orders of magnitude. It is very exciting to look forward to the new technologies that are yet to come and that will constitute the fourth-generation of DNA sequencing, for instance the so-called telomere-to-telomere technologies[36,37]. The future technologies will sequence (may be reading) the continuous and complete information of single chromosomes, with few error sequencing rates, and without the need to carry out the assembly step.

2.2 Chromosome conformation capture technologies

The spatial organization of the genome in the nucleus is nonrandom, and affects many genome functions, including transcription, replication, and DNA repair[38]. There are methods that have taken advantage of such features to develop improved genome scaffolding tools. For instance, methods based on chromosome conformation capture[39] can be used to determine chromosome structure and organization by using the physical proximity of pairs of regions with molecular biology techniques (Figure 2). These techniques preserve chromatin interactions by cross-linking before the fragmentation, ligation and sequencing of DNA fragments, allowing the further recognition of the interacting sequences and the reconstruction of genomic 3D domains[38].

This methodology has multiple applications such as, for instance, quantifying the interaction between two regions (3C), between one region and the genome (4C) and the interaction of all regions and the genome[39,40]. In this last application, also called Hi-C ("HTS genome-wide chromosome conformation capture"; Figure 2), all genomic fragments are marked before the ligation process, labeling ligation junctions. Hi-C methodology includes several steps. First, formaldehyde is added to the sample to fix the cross-links in the chromatin between the sites that are physically interacting. After the digestion of DNA by a restriction enzyme, or through sonication, proximity ligation is performed between DNA ends captured on the same complex. Then, the detection of the ligation junctions is made and these labeled junctions can be sequenced directly with Illumina reads, representing a great advance for the analyses of genome organization at kb resolution[41]. Hi-C data can be applied to obtain gene expression profiles, genome-wide maps of

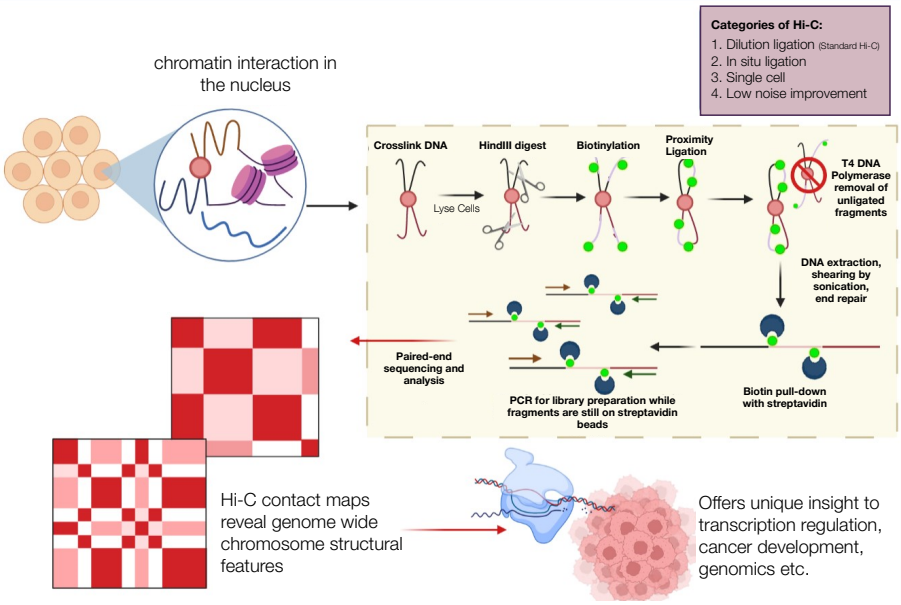


Figure 2. Schematic representation of the workflow in Standard Hi-C technology, and their possible applications. The main steps include: i) DNA crosslink in the chromatin between interacting sites, ii) DNA digestion, iii) proximity ligation between DNA ends, iv) removal of unligated fragments, v) detection and labeling of the ligation junctions and, vi) pair-end sequencing of labeled junctions and downstream analyses. Adapted and modified from Prakrutiuday, distributed under a CC-BY-SA-4.0 license.

chromatin modifications and to facilitate the scaffolding process (e.g., using the Hi-Rise software[42]) to generate chromosome-level genomes[39].

2.3 HTS methods in population genomics

Classic population genetics inferences relied on molecular biology techniques such as allozyme electrophoresis, RFLPs, microsatellite data, and DNA sequencing of mitochondrial DNA or a few nuclear loci using Sanger approach[43]. However, these techniques provide information from a very limited number of genetic markers across individuals. Recent advances in HTS technologies have boosted the proliferation of new and powerful population genomics studies, using a large amount of fine-scale genetic data distributed across the genome, from multiple individuals and even in non-model organisms[44]. Importantly, these techniques, including RNAseq, RADseq and low-coverage whole-genome re-sequencing (lcWGR) approach (Figure 3), do not require any a priori knowledge of the genomic sequences of the surveyed or closely related species[6]; nevertheless, having a reference genome is always desirable.

lcWGR has emerged as a powerful and cost-effective approach for population genomic studies in non-model species. However, since read depths are too low (low genome coverage; usually lower than 15x-20x), it requires the use of specialized analytical tools that explicitly (statistically) account for the genotype uncertainty; that is to confidently call individual genotypes (in diploid organisms). A growing number of such tools have become available in recent years, but it is still difficult to establish the appropriate balance between sequencing costs and inference accuracy and versatility[45]. Methods based on the maximum likelihood calculation of the site-frequency spectrum considering genotype uncertainty (genotype likelihood), rather than directly called genotypes, appear powerful for analyses in which the retention of rare alleles are critical[6]. Currently, the most widely used program for lcWGR analysis is ANGSD[46], a comprehensive package that implements the larger number of analysis options and utilities; due to its robustness and versatility, it is the most popular tool for population genomic inference based on low-coverage data[45], and this is the main tool used in the last chapter of this thesis.

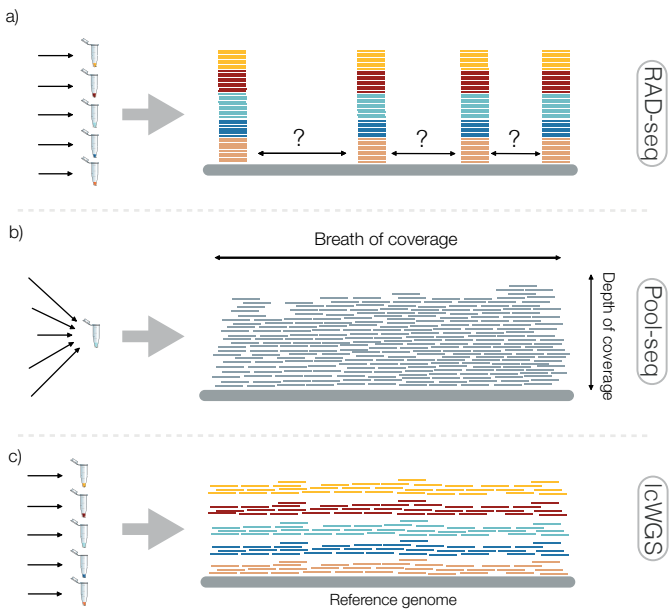


Figure 3. Comparison among different genome sequencing strategies used in population genomics. Images represent sequencing depth and breadth of coverage of short sequencing reads mapped to a reference genome. (a) RAD-seq, (b) Pool-seq and (c) lcWGS. In RAD-seq and lcWGS, samples from different individuals are sequenced separately, while in the Pool-seq approach, samples are combined and sequenced in the same experiment. Adapted from Lou et al.[45]

3 The chemosensory system in Arthropoda

Chemoreception refers to the detection and processing of chemicals from the environment, a crucial system for the survival and reproduction of organisms. The chemosensory system (CS) plays an important role in the fitness of organisms: they are necessary to detect preys, predators or hosts, to communicate with other conspecific and elicit feeding and reproductive responses. In addition, its role on mate recognition and choice may push to reproductive isolation and speciation processes, making chemoreception a particularly interesting system to study the impact of natural selection in molecular adaptation.

The perception of chemical signals in all animals is a complex process mediated by both small soluble carriers and large transmembrane receptor proteins. The process of detecting and processing chemicals begins with the entry of volatile molecules in specialized organs, such as the hair-like sensilla in insects (Figure 4a). In the lymph of these structures are high concentrations of soluble carriers, which bind, solubilize and transport small hydrophobic molecules to the surrounding of chemoreceptors located in the dendritic membrane of olfactory neurons. The chemoreceptor translates the chemical signal into an electrical one, which is processed in the mushroom body's (in insects) causing a particular behavioral response. The chemical signals can be volatile molecules, peptides or gas like oxygen or carbon dioxide, which are usually detected at very low concentration (picomolar or millimolar). Depending on their biological function, they are designated as just odorants, or pheromones if they affect the behavior of the receiving individual[47].

Despite the same function, and general structure of the system, the molecular identity of receptors and carriers in vertebrates and insects are fundamentally different; Arthropoda chemoreceptors are not homologous to their vertebrate and Nematoda counterparts (see below). Comparably, small insect carrier proteins are not homologous to those used in vertebrates, differing in protein size and structural folding[48]. The similarities between the performance of the CS in vertebrates and arthropods is one of the most interesting examples of evolutionary convergence.

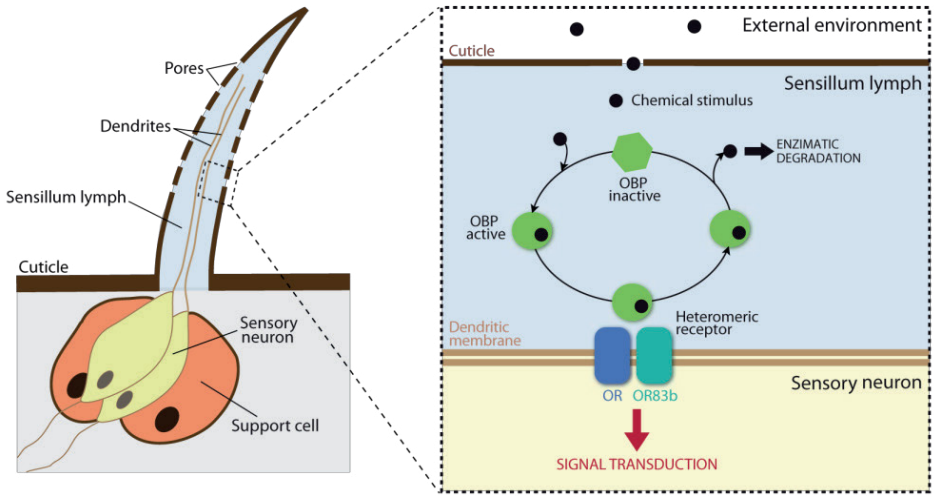


Figure 4. Schematic representation of an insect's olfactory hair. a) Gustative sensilla has a similar structure except for the presence of a unique porous at the apical part. b) Simplified scheme showing a hypothetical model of the perireceptor events occurring in an insect olfactory sensilla. The signaling molecule enters through the cuticle pores and activates the odorant binding protein (OBP) when they enter in contact. The active OBP interacts with either an odorant receptor (OR) or some other membrane components to deliver the ligand. The OBP is subsequently degraded to avoid the continuous stimulation of receptors. Modified from Sánchez-Gracia et al.[53]

3.1 The chemosensory system in Arthropoda: multigenic families

In arthropods, the proteins involved in chemoreception belong to multigene families composed of many genes[48], most of them arranged in small clusters. Their genomic evolution is consistent with the birth-and-death model[49–52] of gene turnover evolution, that could satisfactorily explain the gene member repertoire. The CS gene families are divided in three different types: First, gene families that encode small soluble carriers, such as the odorant binding (OBPs), the chemosensory proteins (CSPs) and the Niemann-Pick type C2 (NPC2). Second, the membrane chemoreceptors include the olfactory (ORs) and gustatory (GRs) receptor families, and the family of ionotropic chemosensory receptors (IRs) and finally, the sensory neuron membrane proteins (SNMPs), which belong to the CD36 family of receptors, involved in performing an adequate response of chemosensory receptors.

3.1.1 Evolution of multigene families

In the CS gene families, the new genes are born through a gene duplication (via unequal crossing-over), where some of them could be maintained, for a relatively long time, or could be lost (via deletion; or transitory via pseudogenization). Initially, these genes perform the same or related function, that could diverge over time. This process generates gene tandems, arrays or clusters of genes. The CS gene families of insects can include a large number of members, especially the chemoreceptors, which can contain up to 400 in some arthropod species[54]. Moreover, the number of genes of these families are extremely variable, even between phylogenetically close species[55,56].

After the gene duplication event, there will be two identical copies, defined as paralogues, initially performing the same function (Figure 5a). This functional redundancy could drive towards different fates. If these copies accumulate differentially mutations, could have different possible outcomes (Figure 5b): first, a copy could develop a new function (neofunctionalization process), that eventually will become fixed in the population by positive selection. Otherwise, both copies

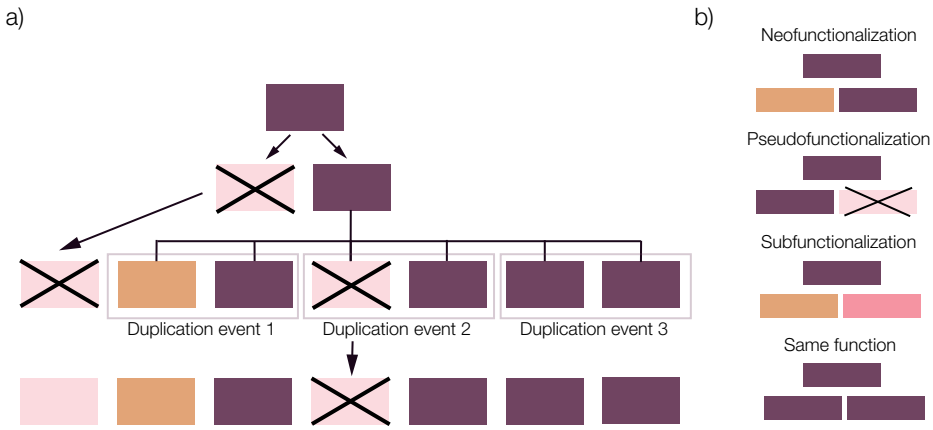


Figure 5. Origin and evolution of gene families. a) Schematic representation of the birth and death process, with a possible result of gene gains and losses. b) Possible outcomes after a duplication event: i) one copy acquires a new function (neofunctionalization process); ii) one of the copies becomes a pseudogene (gene without function) or is lost. iii) Both copies accumulate complementary mutations (subfunctionalization event; it may be that none of the copies performs the ancestral function by itself); iv) both copies maintain the (same) original function. Adapted from Conrad and Antonarakis[58].

could also diverge but perform complementary functions (subfunctionalization event); in this case, purifying selection might have a relevant role. Moreover, the accumulation of deleterious mutations in one copy could make them lose its function (pseudogenization event), which eventually could also be eliminated from the genome. Finally, both copies would maintain the same original function and eventually generate an increased number of transcripts[57].

3.1.2 Extracellular ligand-binding proteins

Binding proteins involved in CS response are soluble proteins that transport odorant molecules mainly located at insect sensilla lymph or other tissues (Figure 4a, b). In insects the main binding protein families are the OBP, CSP and NPC2. OBPs and CSPs are small globular proteins (10-30 kDa) secreted by non-neuronal support cells into the sensilla lymph[59]. These proteins are characterized by a specific domain of six α helices, joined by two (CSPs, some OBPs) or three (OBPs) disulfide bonds[60,61] that play essential roles in transporting semichemicals to the proximity of membrane receptors (Figure 4b), which are usually hydrophobic[62,63]. Even though they perform a similar physiological role, OBPs from vertebrates are not homologous to the insect ones and also differ in structure and size[64]. Several phylogenetic subfamilies of OBP exist in insects, classified by differences in structure and function: PBP/GOBP, Classic, Minus-C, Plus-C, Dimer, ABPI and ABPII, CRLBP, and D7 subfamilies[48]. These subfamilies are distributed heterogeneously across arthropods, even being absent in some species. The CSP gene family, on the contrary, is highly conserved in insects. OBPs and CSPs can also be expressed in non-chemosensory organs in insects, being some of these CSPs involved in embrionary development and regeneration, which points towards a possible original function related to these biological processes[65,66]. It has been suggested by several authors[67,68] that the OBP and CSP gene families may have shared a most recent common ancestor (MRCA) near the origin of the arthropods (Figure 6). On the other hand, Vieira and Rozas[69] proposed that OBP could have originated from the CSPs (already present in the Arthropoda ancestor). Nevertheless, it is possible that the role of this protein was not related to the CS since in arthropods it has been identified a very few number of CSPs (except in a few insects). Further research must be provided to resolve this question. Remarkably, a variant of OBPs called OBP-like have been found across arthropods[56,70]. They present up to 1-4 copies for species and are present in all chelicerates. However, the function of this family might be not related to the CS, as they were found expressed in all tissues in the transcriptome of the spider *Dysdera silvatica*[71].

NPC2 is a water-soluble protein composed of a flexible beta-structure and binds several types of potential hydrophobic chemicals. These proteins are involved in lipid and

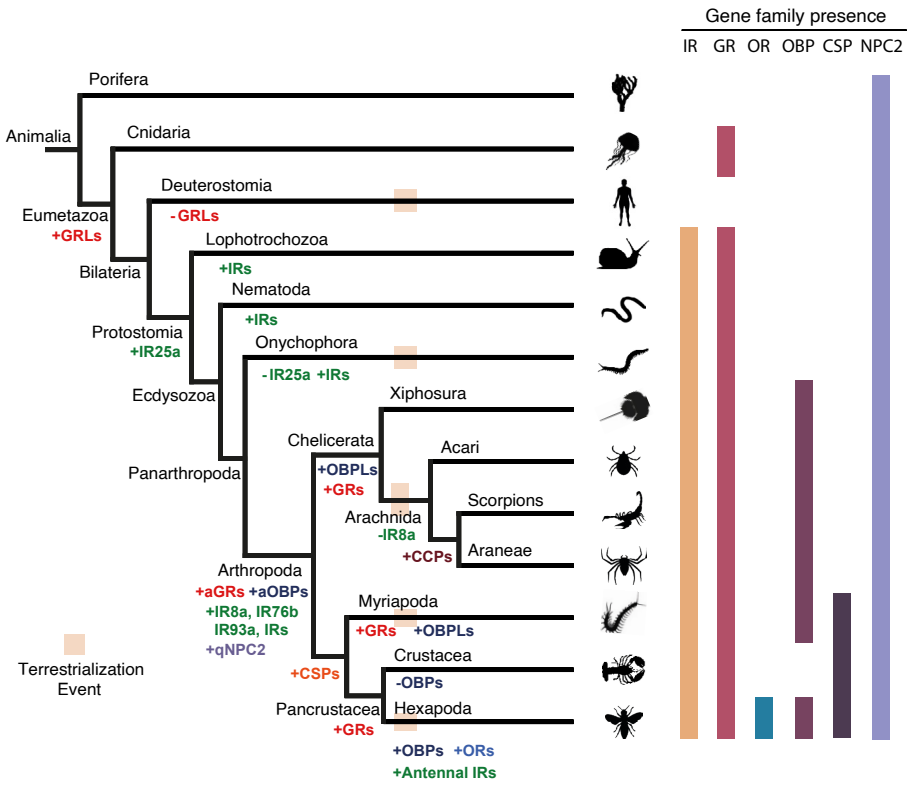


Figure 6. Origin and evolution of main CS multigenic families in Arthropoda. Plus and minus signs indicate the gains and losses of gene families. The light brown squares within the lineages show the origin of the independent terrestralization events. Adapted from Vizueta[74].

cholesterol metabolism[72] being present in all animals, specially conserved in vertebrates where there is only one copy per species[73]. In tardigrades, onychophorans and arthropods, this family has expanded its repertoire (between 2 and 16 copies), however, its function remains unknown. Remarkably, some research suggests that NPC2 are involved in the CS, as they are expressed in the antenna of some insects[72]. Further research must be made to infer its role in the CS in these lineages.

3.1.3 Chemoreceptors

Chemoreceptors are proteins anchored at the membrane of neurons located in the chemosensory organs, as the sensilla in insects (Figure 4b). In Arthropoda, three main chemoreceptor families have been described: ORs, GRs and IRs.

ORs and GRs are two phylogenetically related families involved mainly in the reception of olfaction and taste, that belong to an insect chemoreceptor superfamily with 400 aminoacids of length and seven transmembrane domains (7TM; Figure 7a). Even though insect and vertebrate receptors share some structural similarities, the membrane topology is reversed (insects have the C-terminus at the extracellular space[49]) and they evolved independently[75]. Moreover, several analyses have indicated that insect ORs function as odor-gated ion channels, on the contrary to the metabotropic vertebrate receptors[76,77]. In *Drosophila melanogaster*, ORs are expressed in the antenna and maxillary palp, while the GRs mainly in the gustatory organs (proboscis, legs and wings) but also in olfactory structures[78]. These receptors are differentially expressed in gustatory neurons, allowing the organism to recognize and respond to soluble chemicals such as sugars, amino acids or carbon dioxide and pheromones[79,80]. Some ORs form a complex with a co-receptor to perform their function, such as the co-receptor ORCO in *Drosophila*, one of the most conserved genes in insects[81,82]. The number of genes displayed in these families is very heterogeneous in arthropods, specially of GRs, where some species can have more than 300 or 400 copies[83]. It has been shown that ORs and GRs have a common origin, being the ORs originated from GRs as an adaptation to the terrestrialization[84,85], where the GRs have been identified in some basal species of Metazoa[84,86,87] (Figure 6). This GRs from species outside the Arthropoda phylum are named GR-like (GRL), due to the fact that they have a structure and sequence similar to the GRs but their functions could not be related to the CS; for instance, in Cnidaria are involved in embrionary development[87].

The third class of chemosensory receptors are the IRs, which are ligand-gated ion channels that use odor molecules as ligands[88]. This family is closely related to Ionotropic chemosensory glutamate receptors (iGluRs), a conserved family present in eukaryotes and prokaryotes that display an important role for synaptic transmission and plasticity[55]. All members of iGluRs family are composed of two protein subunits: the extracellular amino-terminal domain (ATD), and the ligand-binding domain (LBD). Between the two half-domains S1 and S2 of the LBD localizes the ion channel pore[89] (Figure 7a). Four canonical gene families have been described: α-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid (AMPA), Kainate and N-methyl-D-aspartate (NMDA) and delta[90]. The IRs family is extremely divergent, with sequence identity values up to 70%. Even though they still present the conserved ligand binding domains (ligand channel domain (LCD) and LBD) but lack the extracellular domain N-terminal (ATD; Figure 7b). The most divergent part from the iGluRs is the LBD, lacking several residues that contact with the ligand and the conserved part is the ion water pore, suggesting that IRs still

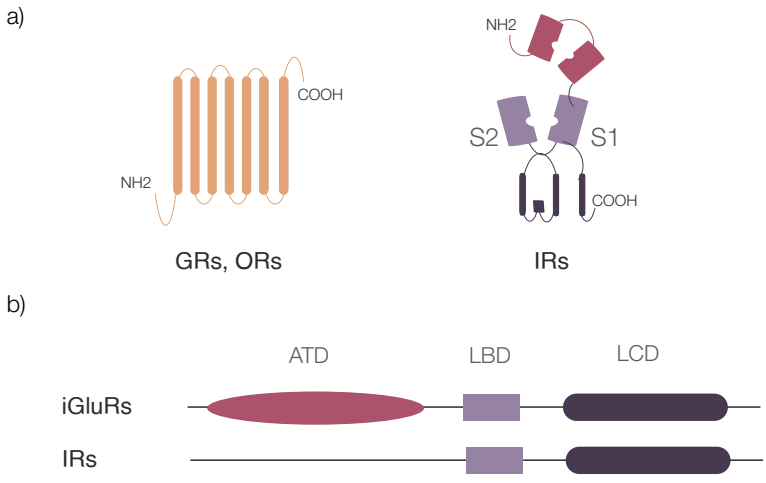


Figure 7. Schematic presentation of the transmembrane and domain structure of Arthropod chemoreceptors. a) GRs and ORs are seven-transmembrane receptors with an inverted membrane topology compared to vertebrate analogs. IRs are ionotropic transmembrane receptors derived from non-chemosensory iGluRs. b) Simple representation of protein domain organization in IRs/iGluRs. The LBD and LCD domains are conserved in all the two subfamilies. ATD represents the amino terminal domain only present in iGluRs (the coreceptor IR25a shows a highly diverged version of this domain).

have ion-conducting properties[90]. IRs have a very ancient origin, as they are present across all protostomes (molluscs, nematodes and arthropods, among others; Figure 6). One of the most conserved genes in this family is the co-receptor IR25a, which is suggested to have appeared at the common ancestor of protostomes from iGluRs and diverged to be able to bind other substrates besides glutamate[55,84].

Besides, other receptors, also encoded by multigene families, have been related to the CS, such as the DeG/ENaC and the SNMPs families. The first one (also known as pickpocket proteins, PPKs), have some members reported as gustative receptors in *Drosophila*[91,92]. The SNMPs, which are transmembrane proteins homologous to the CD36, are very conserved across animals[93,94] except for tardigrades where some species have lost this important family[95]. This family encodes a taste receptor that in insects are involved in important metabolic processes such as cholesterol transport, and fatty acid recognition[94] as well as in pheromone detection and olfactory response[96–99]. Nevertheless, there are few experimental data outside insects, which preclude making inferences across arthropods.

4 Research models

4.1 The Tardigrada and Onychophora phyla

Tardigrades, or also known as 'water bears', are tiny but complex aquatic animals that present a characteristic bilateral symmetrical body[100]. Adults measure about half millimeter, are long and transparent, and they have four body segments, each one with a ganglion, musculature and a pair of 'lobopods' legs, which end in claws or sucking disks[101]. It can be found tardigrades fossils from the Cambrian age, indicating that it is a very ancient phylum[102]. They occupy a huge diversity of niches in freshwater, marine and limno-terrestrial environments throughout the world, and more than ~1,300 species have been described up to date[103]. Even though they need a film of water to grow and reproduce, some species from the Eutardigrada class (Figure 8) have developed very interesting adaptations to resist unfavorable environmental conditions by entering a latent state called cryptobiosis. When they suffer from abiotic stress, they dehydrate themselves completely, a process that is reversible[104]. In this condition, they can survive in extreme temperatures (from −273 °C to almost 100 °C)[105–107], radiation[108,109], high pressures (up to 7.5 GPa)[110], and even being exposed to open space[111]. Despite its popularity, tardigrades are one of the less known groups of protostomes and, surprisingly, the molecular mechanisms underlying this fascinating resistance response to extreme habitats are still unknown. Currently there is available information from only a few transcriptomes[112,113] (SRX973445; SRX973444), and from five genome assemblies, all of them from the Eutardigrada class[104,114–117]. Some of the genomes belong to the popular *Hypsibius exemplaris* (Doyère, 1840; formerly referred to as "*H. dujardini*"), and another to the species *Ramazzottius varieornatus*[104,116,118], both used in this thesis.

Polemical research on the genome of *H. exemplaris* concluded that in this species a high number of genes came from horizontal gene transfer (HGT) from bacteria[117]. This work was further rebated demonstrating that contaminated sequences were

Figure 8. Lobopodia hypothesis[119]. Phylogenetic relationships among major Panarthropoda groups. Classical higher-level tardigrade relationships are shadowed in light pink. The position of Mesotardigrada is considered *nomen dubium* due to lack of information about this group.

responsible for the wrong interpretation[118]. The above-mentioned discrepancies also explain the huge differences in the estimates of *H. exemplaris* genome size between the two studies (~212 and about 100 Mbp, respectively).

Onychophorans are small, soft elongated bodied invertebrates, with 180 described species that inhabit tropical and temperate forests of the southern hemisphere and around the equator[120]. They are found in soil, leaf-litter and decomposing logs. These animals, also called velvet worms, are very interesting animals for multiple reasons. One of them is their particular mechanism of prey-capturing or defense that consists in ejecting a glue substance from oral slime papillae[121]. Recent results show that their distribution was caused by the ancient Pangaea regionalization and Gondwanan vicariance[122]. They are said to be kind of a "living fossil" because of their anatomic conservation respect to lobopodians[123] (they present 13-43 pairs of lobopod legs), which are considered their Cambrian ancestors[124–128]. Even though they are such charismatic organisms, there is only a single genome draft assembly available in public databases, from the species *Euperipatoides rowelli* (PRJNA203089); unfortunately, the poor continuity and completeness of this low-quality assembly make it of little utility to comparative genomics studies.

Unlike insects, tardigrades and onychophorans don't have specialized chemosensory organs; instead, they have some cuticular ciliary receptors present in some body regions, mostly in the anterior end in tardigrades[129] and in the tips of antennae in onychophorans. Nevertheless, the lip papillae surrounding the mouth of onychophorans also have sensory cells responding to chemical stimuli[130,131].

From the evolutionary point of view, tardigrades and onychophorans are especially relevant since they are the closest living relatives of arthropods; forming the

Panarthropoda clade[132,133]. The phylogenetic relationships between arthropods, onychophorans, and tardigrades, and even the Panarthropoda clade itself, are still under debate[120], However, onychophorans have been consistently recovered as the sister group of arthropods in most molecular phylogenetic analyses[134–137.] Resolving the relationships among these groups is fundamental for understanding key questions of invertebrate evolution. In this thesis, we have used available genomic and transcriptomic sequences for these animals and also generated new HTS data, to provide new knowledge about the origin and evolution of chemosensory gene families in Panarthropoda.

4.2 The phylum Arthropoda

One of the most evolutionary successful and diversified phylums on Earth is Arthropoda, with over 1.5 million described species and found in almost all ecosystems[138]. The phylum includes four extant subphyla: Chelicerata, Hexapoda, Crustacea and Myriapoda, plus the extinct subphylum of Trilobita[135]. The Chelicerata subphylum is a large and diverse group that contains 10% of Arthropoda species, the majority of which belong to Arachnida class (with ~130,000 species)[139–141]. Arachnids exhibit tremendous species richness and show adaptations that are of relevant biomedical, industrial and agricultural importance[140], such as the production of the silk fibers and venom toxins useful for the pharmaceutical industry or their capacity to participate in pest control. Surprisingly, despite this importance in many fields, not enough attention has been paid to spiders, with very few genomic resources available[54,140,142]. To date, only six chromosome-level Arachnid genomes have been published, all of them representing very recent studies[143–148]. Recently, the monophyly of Arachnida has been questioned[149,150], advocating for morphological convergence resulting from adaptations to life in terrestrial habitats as the main driver of the historical artificial perception of the monophyly of arachnids, paralleling the history of numerous other invertebrate terrestrial groups. Spiders (order Araneae) are the largest group within Arachnida, with more than ~50,000 species known[151].

4.3 The spider genus *Dysdera*

The genus *Dysdera* Latreille 1804 contains nearly half of the diversity of the family Dysderidae (i.e., 296 species[151]). *Dysdera* spiders are active nocturnal hunters and ground-dwelling organisms that rest in cocoons under stones, dead logs or leaves during the day, some species having adapted to live even in caves[152]. Some

members of this genus colonized the Macaronesian archipelagos, experiencing an extraordinary adaptive radiation in the Canary Islands and Madeira, with ~60 and 12 endemic *Dysdera* species, respectively; a single endemic species per archipelago has been found in Azores, Islas Salvajes and Cabo Verde[152–158]. Based on phylogenetic data, three independent colonization events have been described in Canary Islands[152,153], the last one corresponding to *Dysdera lancerotensis*, a species closely related to the Morocco members of this genus that colonized eastern Canary Islands very recently[159]. The species from the western and the eastern Canary Islands belong to distinct clades but the phylogenetic relationships among them remain mostly unsolved.

Dysdera spiders are known for being one of the few arthropods that are oniscophagous, meaning that they developed a prey specialization towards terrestrial woodlouse (Crustacea: Isopoda). Some species are able to follow a generalist diet that include these isopods but others are highly specialized in feeding nearly exclusively on them, with some intermediate cases that show different degrees of specialization[160]. Woodlice are not a common prey; in fact, they are rejected from most predators[161]. The main reason is that they are protected with a hard exoskeleton, accumulate heavy metals, and secrete toxic components, which gives them a very unpleasant flavor and makes them difficult to digest and extract the essential nutrients[162]. In addition, they also developed solid defense against predation strategies, such as pestilent secretions, nocturnal habits and the characteristic ball-form defense strategy. Interestingly, *Dysdera* spiders have several morphological, behavioral and metabolic adaptations that circumvent or mitigate the effect of these defenses, allowing them to feed on this unpleasant prey.

Oniscophagy has evolved independently multiple times in this genus in different geographic areas, being considered one of the major drivers of the adaptive radiation in the Canary Islands[163]. The species with a more specialized diet, adapted the chelicerae with shapes that match with the different capture strategies to prey woodlouse. They also developed key metabolic adaptations to extract and assimilate nutrients from these isopods more efficiently than from other. In fact, recent studies have shown that the degree of chelicerae modifications correlates with the levels of specialization in *Dysdera* species (Figure 9). That is, species with extremely modified chelicerae exclusively capture woodlice, while unmodified chelicerae appear in species that feed more frequently on soft-bodied arthropods[163]. Despite all this research on the morphological and metabolic adaptations in these spiders, their genomic bases are still unknown.

At the time of starting this thesis, the research group had generated several

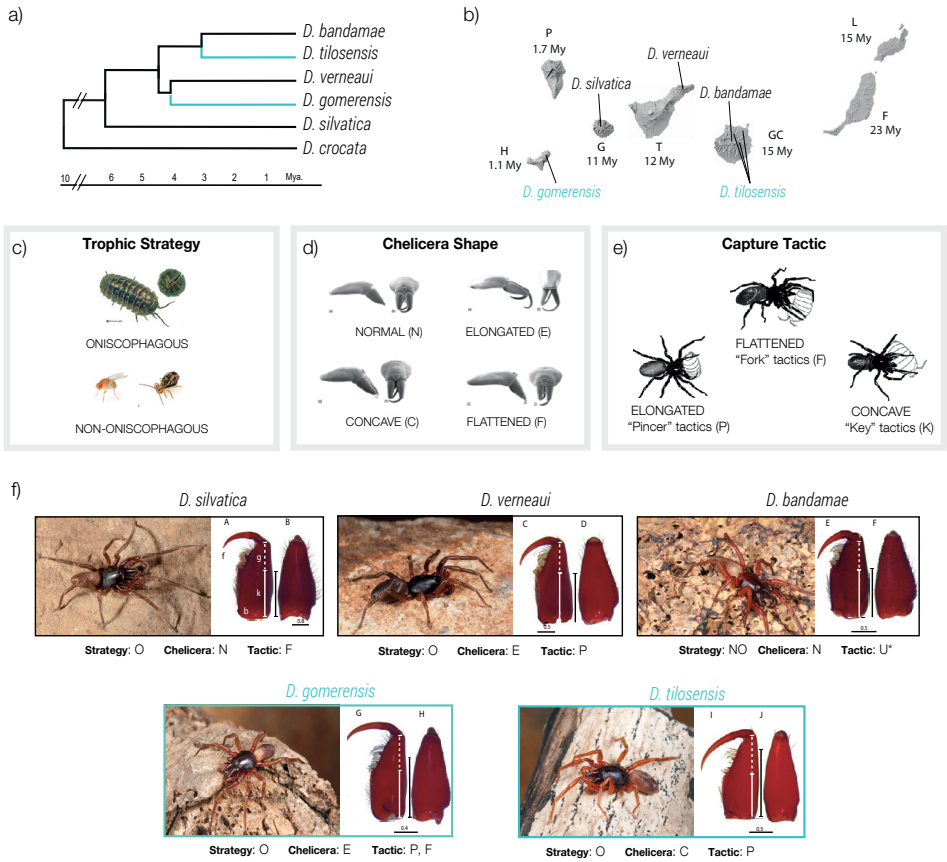


Figure 9. Phylogeny, geographical location and phenotypic variation among species of the *Dysdera* genus from Canary Islands. a) Phylogenetic relationships and divergence times between some Canary Island endemic species rooted using a continental member of the genus. *D. gomerensis* Strand, 1911 (El Hierro), *D. verneaui* Simon, 1883 (Tenerife), *D. tilosensis* Wunderlich, 1992 (Gran Canaria), *D. bandamae* Schmidt, 1973 (Gran Canaria), and D. silvatica Schmidt, 1981 (La Gomera). Species with a higher level of stenophagy (higher preference towards a woodlouse diet) are shadowed in blue-green color. b) Map of Canary Islands including the geographical localization of the species shown in panel a). Numbers indicate the approximate age of islands in million years. F: Fuerteventura, L: Lanzarote, GC: Gran Canaria, T: Tenerife, G: La Gomera, P: La Palma and H: El Hierro. c) d) and e) show trophic, phenotypic and capture tactics found in Canary Island *Dysdera* f) Images of the Canary Island endemic species shown in panel a) (photo credit: Pedro Oromí) with their corresponding chelicerae (in millimeter scale). Bars indicate relative length of the distinct parts to remark differences between spiders with different levels of oniscophagy. Trophic strategies, shape of chelicera and capture tactics according to Rezac et al.[163].

transcriptomic data of *Dysdera* species endemic from the Canary Islands to elucidate the genetic bases of the adaptive radiation of the genus in this archipelago[56,164]; the study was conducted by applying a comparative, differential gene expression analysis between species with different levels of dietary specialization (Figure 9).

Here, we provide a new resource, and new knowledge on the genomic basis of this extraordinary island diversification. Particularly, i) we have generated the first assembly of the genome of *D. silvatica*, and later upgraded to a chromosome level quality, ii) we provided a comprehensive structural and functional annotation of the genome, including an improved annotation of the chemosensory gene families in this species, iii) we also carried out the first population genomics study of an endemic *Dysdera* species to gain insights into the process underlying the adaptive radiation.

Objectives

HTS technologies are revolutionizing the field of evolutionary biology, allowing generating high quality genomic sequences in a short time and at a reasonable cost. Indeed, they have become fundamental for understanding genome structure and organization, and for determining the role of the different evolutionary forces, including adaptive and non-adaptive forces, in species diversification. In addition, the scientific-based knowledge that can be obtained thanks to these techniques is instrumental to manage and conserve biodiversity in a changing world. In this thesis, we used these technologies to study relevant evolutionary genomics questions at very different timescales, including the origin and evolution of Arthropod chemosensory gene families across Panarthropoda, and the genomic determinants underlying recent species diversification. For this last question, we used the radiation of genus *Dysdera* in the Canary Islands, and more specifically the species *D. silvatica* as the model.

The specific objectives of this thesis are:

- Generate a high-quality, chromosome-level assembly of the genome of the Canary Islands endemic spider *D. silvatica*, including functional annotations and the characterization of repetitive elements across the main chromosomes.

- Perform an evolutionary genomics study of the Arthropod chemosensory gene families across representative genomes of the Panarthropoda clade, including new data from onychophorans generated in this thesis.

- Carry out a comprehensive analysis of the number, chromosomal distribution and gene clustering of chemosensory receptors in *D. silvatica* taking benefit of the high-quality reference genome generated in this thesis.

- Explore the population genomics factors shaping intra- and interspecific variation in a natural population of *D. silvatica* from la Gomera using low-coverage whole-genome re-sequencing data from 12 individuals.

Supervisors report

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 Alejandro Sánchez-Gracia,** directors de la Tesi Doctoral elaborada per la Sra. Paula Escuer, amb el títol "**Evolutionary Genomics of Panartropoda: Study of the chemosensory gene families across phyla and the radiation of the spider genus *Dysdera* (Araneae, Dysderidae) in the Canary Islands**"

INFORMEN

Que la tesi doctoral s'ha elaborat com a un compendi de sis publicacions científiques, totes elles amb dades originals, quatre de les quals formen part del cos central de la tesi, mentre que les dues restants apareixen a l'apèndix.

Publicacions:

1. 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* **8:** 1-9. Doi: 10.1093/gigascience/giz099.
   Factor d'impacte (5 Year Impact Factor): 7.715; **Q1** (Quartil 1; dins la categoria de Multidisciplinary Sciences). Citacions: 19.

2. Vizueta, J*., **Escuer, P.***, Frías-López, C., Guirao-Rico, S., Hering, L., Mayer, G., Rozas, J., Sánchez-Gracia, A. 2020. Evolutionary history of major chemosensory gene families across Panarthropoda. *Mol. Biol. Evol*. **37:** 3601-3615. doi: 10.1093/molbev/msaa197.
   Factor d'impacte (5 Year Impact Factor): IF = 18.670; **D1** (Decil 1; dins la categoria de Genetics & Heredity). Citacions = 8.
   *, la mateixa contribució.

3. **Escuer, P.**, Pisarenco, V. A., Fernández-Ruiz, A. A., Vizueta, A. J, Sánchez-Herrero, J. F., Arnedo, M. A., Sánchez-Gracia, A., Rozas, J. 2022.Chromosome-scale assembly of the Canary Island endemic spider *Dysdera silvatica* (Arachnida, Araneae) sheds light on the origin and genome structure of chemoreceptor gene families in spiders. *Mol. Ecol. Res.* **22:** 375-380. Doi: DOI: 10.1111/1755-0998.13471.
   Factor d'impacte (5 Year Impact Factor): 8.984. **D1** (Decil 1; dins la categoria de Ecology). Citacions = 3.

4. **Escuer, P.,** Guirao-Rico, S., Arnedo, M. A., Sánchez-Gracia, A., Rozas, J. 2022. Population genomics of adaptive radiations: Exceptionally high levels of genetic diversity in a spider endemic from the Canary Islands.
   Preparat per enviar a enviar a una revista amb *peer review*.

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

5. Rispe, C. et al. (including **Escuer, P.,** Rozas J., and Sánchez-Gracia, A. 2020. The genome sequence of the grape phylloxera provides insights into the evolution, adaptation, and invasion routes of an iconic pest. *BMC Biol.* **18:** 90. doi: 10.1186/s12915-020-00820-5.
Factor d'impacte (5 Year Impact Factor): IF = 8.182; **D1** (Decil 1; dins la categoria de Biology). Citacions = 30.

6. Balart-García, P., Cieslak, A., **Escuer, P.**, Rozas, J., Ribera, I., Fernández, R. 2021. Smelling in the dark: Phylogenomic insights into the chemosensory system of a subterranean beetle. *Mol. Ecol.* **30:** 2573-2590. doi: 10.1111/mec.15921.
Factor d'impacte (5 Year Impact Factor): IF = 6.895; **Q1** (Quartil 1; dins la categoria de Ecology). Citacions = 2.

A la primera publicació, que també es va presentar com a part de la tesi doctoral del Dr. José F. Sánchez Herrero, la doctoranda va fer una part rellevant de les tasques experimentals. A la segona publicació, la doctoranda va dur a terme una part important de les tasques computacionals i analítiques, a més de col·laborar en la redacció del primer esborrany del manuscrit. A les publicacions tercera i quarta, la doctoranda va realitzar la part més rellevant del treball computacional i analític i va redactar el primer esborrany del manuscrit.

Les dues publicacions incorporades al apèndix de la tesi, han resultat de col·laboracions científiques on la doctoranda va contribuir a les anàlisis, fent servir eines computacionals o analítiques desenvolupades en la seva tesi doctoral.

ROZAS LIRAS JULIO ANTONIO - 46030727E
Firmado digitalmente por ROZAS LIRAS JULIO ANTONIO - 46030727E
Fecha: 2022.09.27 07:18:01 +02'00'

Dr. Julio Rozas Liras
Catedràtic de Genètica
Universitat de Barcelona

Firmado digitalmente por ALEJANDRO SANCHEZ GRACIA - DNI 46785605V
Fecha: 2022.09.27 11:45:50 +02'00'

Dr. Alejandro Sánchez-Gracia
Professor Agregat de Genètica
Universitat de Barcelona

Chapters

Chapter 1

The draft genome sequence of the spider *Dysdera silvatica* (Araneae, Dysderidae): A valuable resource for functional and evolutionary genomic studies in chelicerates

JF Sánchez-Herrero, C Frías-López, P Escuer, S Hinojosa-Alvarez, MA Arnedo, A Sánchez-Gracia, J Rozas

2019. GigaScience, 8, 1–9. doi: 10.1093/gigascience/giz099

En aquest article presentem el primer assemblatge del genoma de *Dysdera silvatica*, una aranya nocturna que habita al sòl, d'un gènere que ha patit una notable radiació adaptativa a les Illes Canàries. L'assemblatge es va obtenir mitjançant lectures de seqüenciació curtes (Illumina) i llargues (PacBio i Nanopore). El nostre assemblatge de novo (1,36 Gb), que representa el 80% de la mida del genoma estimada per citometria de flux (1,7 Gb), està constituït per una alta fracció d'elements repetitius intercalats (53,8%). La integritat de l'assemblatge, utilitzant BUSCO i gens eucariotes bàsics, oscil·la entre el 90% i el 96%. Les anotacions genòmiques basades tant en informació *ab initio* com en informació basada en diverses evidències (incloent-hi dades del transcriptoma de D. silvatica) van produir un total de 48.619 seqüències codificants de proteïnes, de les quals 36.398 (74,9%) tenen la marca molecular de dominis proteics coneguts o semblança de seqüències amb seqüències de Swiss-Prot. El genoma de *D. silvatica* és el primer reportat d'un representant de la superfamília Dysderoidea i tan sols el segon genoma disponible de Synspermiata, un dels principals llinatges evolutius dels aràcnids (Araneomorphae). Els disdèrids, coneguts pels seus nombrosos exemples d'adaptació a entorns subterranis, inclouen alguns dels pocs exemples d'especialització tròfica a les aranyes. Aquest recurs serà útil com a punt de partida per estudiar qüestions evolutives i funcionals fonamentals, incloses les bases moleculars de l'adaptació a ambients extrems i canvis ecològics, així com de l'origen i l'evolució de trets rellevants de les aranyes. binarias previamente identificadas como nuevas interacciones que todavía no han sido confirmadas experimentalmente.

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

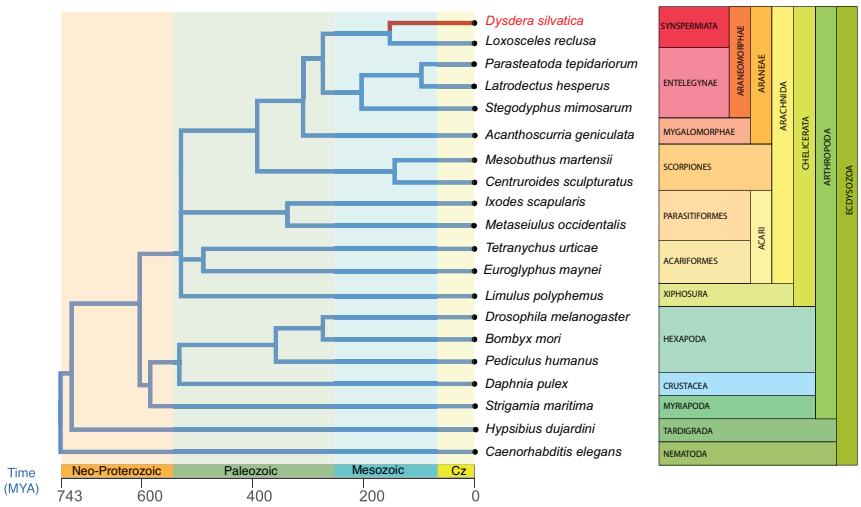


Figure 2 Phylogenetic relationships of the species used for the *D. silvatica* genome annotation (see Supplementary Table S1-11 for further details) and completeness analysis. Because the chelicerata phylogeny is controversial (e.g., [20], [21]), we set the most conflictive clades as polytomies. Divergence 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 of repetitive sequences, (i) *D. silvatica*–specific repetitive elements generated with RepeatModeler v1.0.11 [46], (ii) the Dfam_Consensus [48] (version 20170127), and (iii) 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).

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).

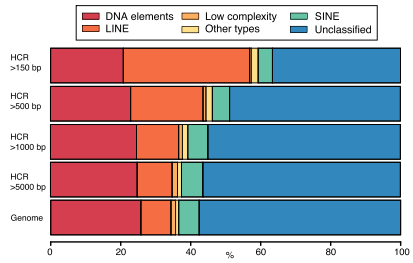


Figure 3 Bar plot of the annotation of the repetitive elements within the HCRs (2.5× threshold) at different intra-HCR length cutoffs (150, 500, 1,000, and 5,000 bp) (Supplementary Table S2-2a). Colors represent the type of repeat element identified by RepeatMasker. "Other types" class includes the LTR 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)
Arthropoda (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-

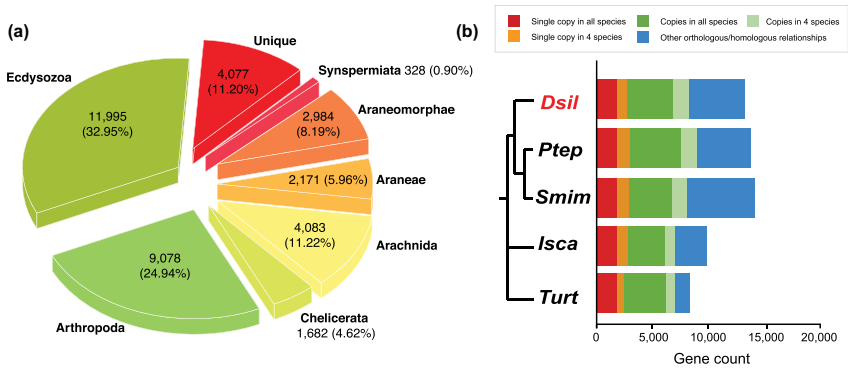


Figure 4 (a) Pie chart illustrating the taxonomic distribution of positive BLAST hits of the *D. silvatica* protein-coding genes against the sequence data of species included in Fig. 2. **(b)** Homology relationships among *D. silvatica* (Dsil) and different chelicerates genomes available in OrthoDB 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.

The draft genome sequence of the spider *Dysdera silvatica* (Araneae, Dysderidae): A valuable resource for functional and evolutionary genomic studies in chelicerates

JF Sánchez-Herrero, C Frías-López, P Escuer, S Hinojosa-Alvarez, MA Arnedo, A Sánchez-Gracia, J Rozas

2019. GigaScience, 8, 1–9. doi: 10.1093/gigascience/giz099

Supplementary Material

The draft genome sequence of the spider *Dysdera silvatica* (Araneae, Dysderidae): A valuable resource for functional and evolutionary genomic studies in chelicerates

José Francisco Sánchez-Herrero[1,2], Cristina Frías-López[1,2], Paula Escuer[1,2], Silvia Hinojosa-Alvarez[1,2,3], Miquel A. Arnedo[1,4], Alejandro Sánchez-Gracia[1,2,*] and Julio Rozas[1,2,*]

[1]Departament de Genètica, Microbiologia i Estadística, Universitat de Barcelona (UB), Barcelona, Spain
[2]Institut de Recerca de la Biodiversitat (IRBio) (UB)
[3]Jardín Botánico, Instituto de Biología, Universidad Nacional Autónoma de México, Ciudad de México, México
[4]Departament de Biologia Evolutiva, Ecologia i Ciències Ambientals (UB)

Additional Files

SUPLEMENTARY FIGURES

Supplementary Figure 1: GenomeScope *k-mer* profile plot for the *D. silvatica* genome *Dsil v1.2*, based on 21-*mers* of the PE reads. The observed *k-mer* frequency distribution is depicted in blue, whereas the GenomeScope fit model is shown as a black line. The unique and putative error k-*mer* distributions are plotted in yellow and red, respectively.

Supplementary Figure 2: Schematic representation of the hierarchical workflow used to generate the assembly of the *D. silvatica* genome.

Supplementary Figure 3: Genome coverage distribution for the different genome sequencing data used in this study. Dash lines indicate the mean genome coverage for the particular sequencing technology.

Supplementary Figure 4: Analysis of the number and distribution of the High Coverage Regions (HCR) across the genome. **a)** Schematic representation of the genome coverage distribution along a contig (~35 kb). The pink dotted line denotes the mean genome coverage estimated for PE read library (Supplementary Table S1-9) (~30X). The green and orange dotted lines reflect, 2.5x and 5x thresholds, respectively, of the average coverage (75X and 150X, respectively). The intra HCR length (in blue) reflects the physical distance fulfilling the threshold coverage (2.5x or 5x), while the Inter HCR (red) denotes the distance between HCRs. **b)** Frequency distribution of the intra-HCR length (blue) and inter-HCR (red) across the 34 937 contigs for the 2.5x (green) or 5x (orange) times the average coverage (See Supplementary Table 2-2 and Supplementary File for details). The minimum value for any inter-HCR was always >10bp. The yellow line denotes the mean distribution value.

Supplementary Figure 5: Frequency distribution of the intra-HCR length (blue) and inter-HCR (red) for different length cutoffs (150, 500, 1 000 and 5 000) across the 34 937 contigs for the 2.5x (a) or 5x (b) threshold coverage (See Supplementary Table 2-2 and Supplementary File for details). The minimum value for any inter-HCR was >10bp. The yellow line denotes the mean distribution value.

Supplementary Figure 6: Bar plot of the annotation of the repetitive elements within the HCRs (5x 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, include the LTR elements, Small RNA and Satellites information that represent a small fraction.

Supplementary Figure 7: Cumulative fraction of the frequency distribution of the Annotation Edit Distance (AED) provided by MAKER2 for different steps of the annotation process (Supplementary Table S1-12). The two iterative training rounds (R1 and R2) are shown in dashed blue. The final test (F) rounds are depicted by green lines: F1 using only *D. silvatica* transcripts and F2 using proteins from a broad taxonomic range (Figure 2; Supplementary Table S11). The F2 line shows the cumulative fraction of annotation for the final 48,619 protein-coding genes annotated with an average AED of 0.32. The red and orange dashed lines, represent the cumulative fraction of annotation for the 36,398 functionally annotated protein-coding genes (AED value of 0.268), and for the 4 077 unique *Dysdera silvatica* genes (AED of 0.4), respectively. The AED value is a direct measure of the annotation quality and its values range from 0 (high evidence and exact match based on alignment) to 1 (no evidence support).

Supplementary Figure 8: Homologous relationships between *D. silvatica* (Dsil) and five representative metazoan genomes available in OrthoDB v10 database (Kriventseva 2019): *Strigamia maritima* (Smar), *Drosophila melanogaster* (Dmel), *Limulus polyphemus* (Lpol), *Ixodes scapularis* (Isca), *Parasteatoda tepidariorum* (Ptep) and *D. silvatica* (Dsil). Red and orange bars indicate the fraction of single copy genes (1:1 orthologs) identified in all species, and in all but one (eg, missing in one species), respectively. The dark and light green bar indicates the fraction of orthologs present in all species and in all but one, respectively, that are not included previous categories. The blue bar (other orthology/homology) shows other more complex homologous relationships. The results were generated uploading *D. silvativa* proteins to the OrthoDB web server.

Supplementary Figure S9: Cumulative fraction of frequency distribution of covered exon overlap match by RNAseq or ab initio evidence at the splice site (dash lines) or exon level (solid lines) for the different datasets (in colours: red for species-specific proteins; green for functionally annotated proteins and blue for all structurally annotated proteins) (Supplementary Table S1-12b).

Supplementary Figure S10: Structure and functional annotation of the mitochondrial genome.

SUPLEMENTARY TABLES

Supplementary Table S1-1: Collection of samples used in these study.

Supplementary Table S1-2: DNA sequencing read files used in this study.

Supplementary Table S1-3: NCBI data used for the contaminant search step.

Supplementary Table S1-4: Pre-processing statistics for each library.

Supplementary Table S1-5: Samples used for the RNAseq study.

Supplementary Table S1-6: RNA sequencing read files.

Supplementary Table S1-7: Genome descriptive statistics

Supplementary Table S1-8: Coverage analysis.

Supplementary Table S1-9: RepeatMasker analysis of *D. silvatica* genome.

Supplementary Table S1-10: Reference-guided transcriptome assembly statistics.

Supplementary Table S1-11: Source of proteins to conduct the annotation of *D. silvatica* genes and completeness analysis.

Supplementary Table S1-12: Annotation statistics.

Supplementary Table S1-13: Mitochondrial assembly metrics and annotation features

Supplementary Table S2-1: Example of results for the High Coverage Region (HCR) analysis.

Supplementary Table S2-2: High Coverage Region (HCR) descriptive statistics.

Supplementary Table S2-3: Enrichment analysis of the intersection of High Coverage regions (HCRs) with RepeatMasker annotation (main repeats).

Supplementary Table S2-4: Enrichment analysis of the intersection of High Coverage regions (HCRs) (2.5x mean coverage threshold) with RepeatMasker annotation (subtypes of main repeats).

Supplementary Table S2-5: Enrichment analysis of the intersection of High Coverage regions (HCRs) (5x mean coverage threshold) with RepeatMasker annotation (subtypes of main repeats).

Supplementary Table S3-1: List of the *D. silvatica* genes identified in the OrthoDB analysis across five chelicerates.

Supplementary Table S3-2: List of the *D. silvatica* genes identified in the OrthoDB analysis across six arthropods.

SUPLEMENTARY FILES

Supplementary File SF1: Additional data and results including the mapping coverage distribution results, High Coverage Region (HCR) analysis, Annotation Edit distance (AED) statistics and OrthoDB comparative results.

SUPPLEMENTARY FIGURES

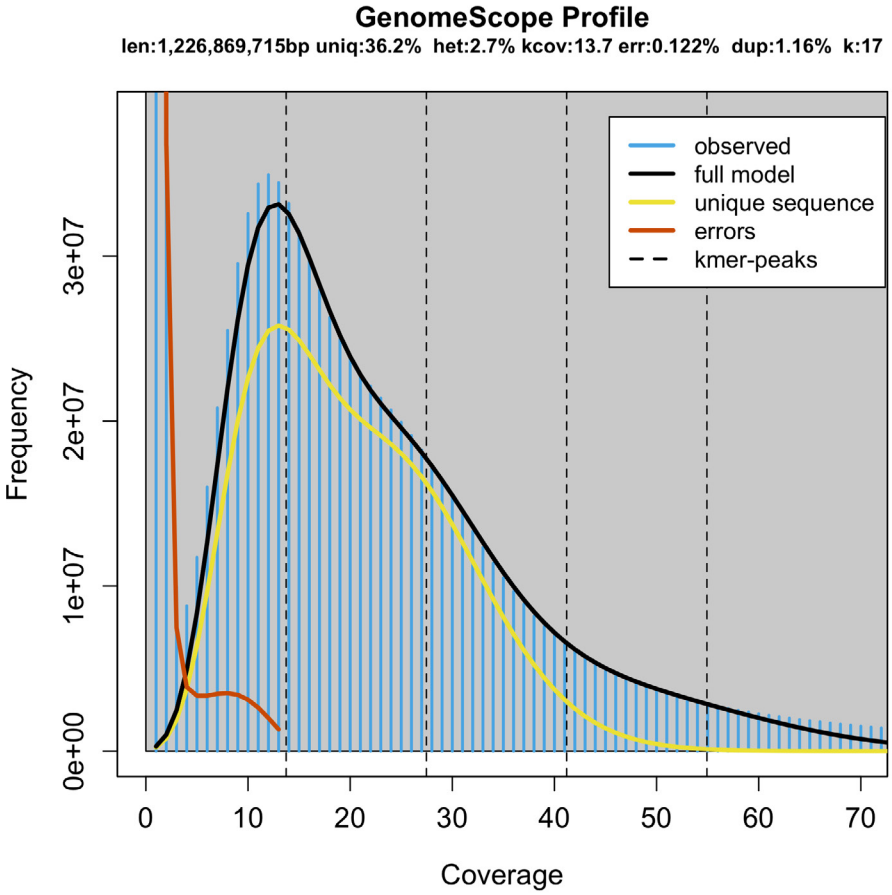

GenomeScope Profile
len:1,226,869,715bp uniq:36.2% het:2.7% kcov:13.7 err:0.122% dup:1.16% k:17
observed
full model
unique sequence
errors
kmer-peaks
Frequency
0e+00
1e+07
2e+07
3e+07
0
10
20
30
40
50
60
70
Coverage

Figure S1

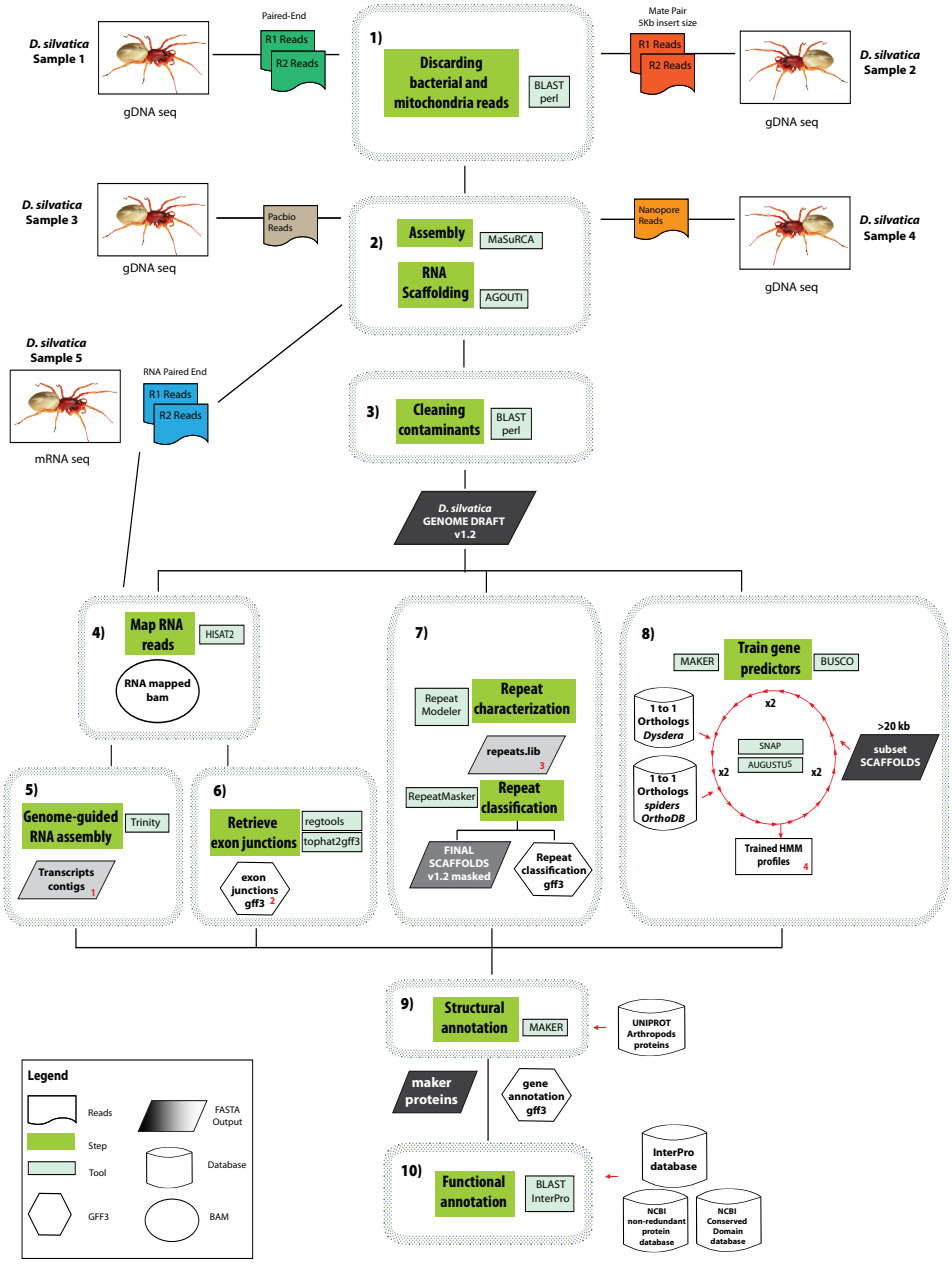

D. silvatica Sample 1
gDNA seq
Paired-End
R1 Reads
R2 Reads
1)
Discarding bacterial and mitochondria reads
BLAST perl
Mate Pair 5Kb insert size
R1 Reads
R2 Reads
D. silvatica Sample 2
gDNA seq
D. silvatica Sample 3
gDNA seq
Pacbio Reads
2)
Assembly
MaSuRCA
RNA Scaffolding
AGOUTI
Nanopore Reads
D. silvatica Sample 4
gDNA seq
D. silvatica Sample 5
mRNA seq
RNA Paired End
R1 Reads
R2 Reads
3)
Cleaning contaminants
BLAST perl
D. silvatica GENOME DRAFT v1.2
4)
Map RNA reads
HISAT2
RNA mapped bam
7)
Repeat Modeler
Repeat characterization
repeats.lib 3
RepeatMasker
Repeat classification
FINAL SCAFFOLDS v1.2 masked
Repeat classification gff3
8)
MAKER
Train gene predictors
BUSCO
1 to 1 Orthologs Dysdera
1 to 1 Orthologs spiders OrthoDB
x2
SNAP
AUGUSTUS
>20 kb
subset SCAFFOLDS
Trained HMM profiles 4
5)
Genome-guided RNA assembly
Trinity
Transcripts contigs 1
6)
Retrieve exon junctions
regtools
tophat2gff3
exon junctions gff3 2
9)
Structural annotation
MAKER
UNIPROT Arthropods proteins
maker proteins
gene annotation gff3
10)
Functional annotation
BLAST InterPro
InterPro database
NCBI non-redundant protein database
NCBI Conserved Domain database
Legend
Reads
Step
Tool
GFF3
FASTA Output
Database
BAM

Figure S2

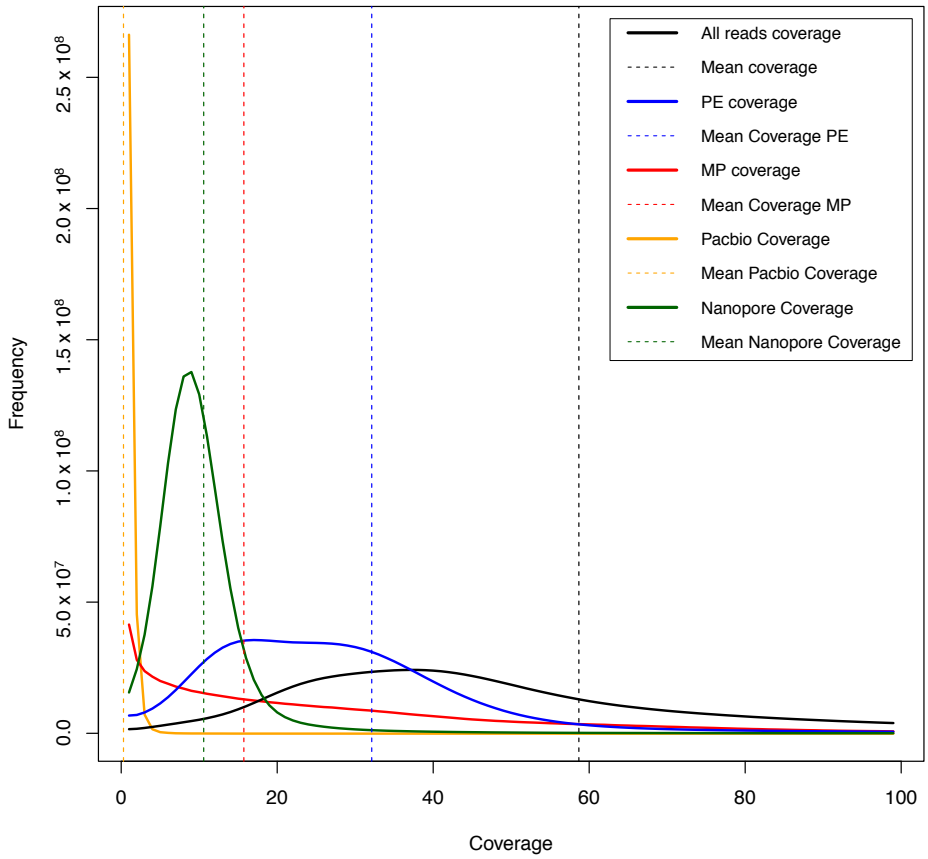

All reads coverage
Mean coverage
PE coverage
Mean Coverage PE
MP coverage
Mean Coverage MP
Pacbio Coverage
Mean Pacbio Coverage
Nanopore Coverage
Mean Nanopore Coverage
Frequency
0.0
5.0 x 10^7
1.0 x 10^8
1.5 x 10^8
2.0 x 10^8
2.5 x 10^8
0
20
40
60
80
100
Coverage

Figure S3

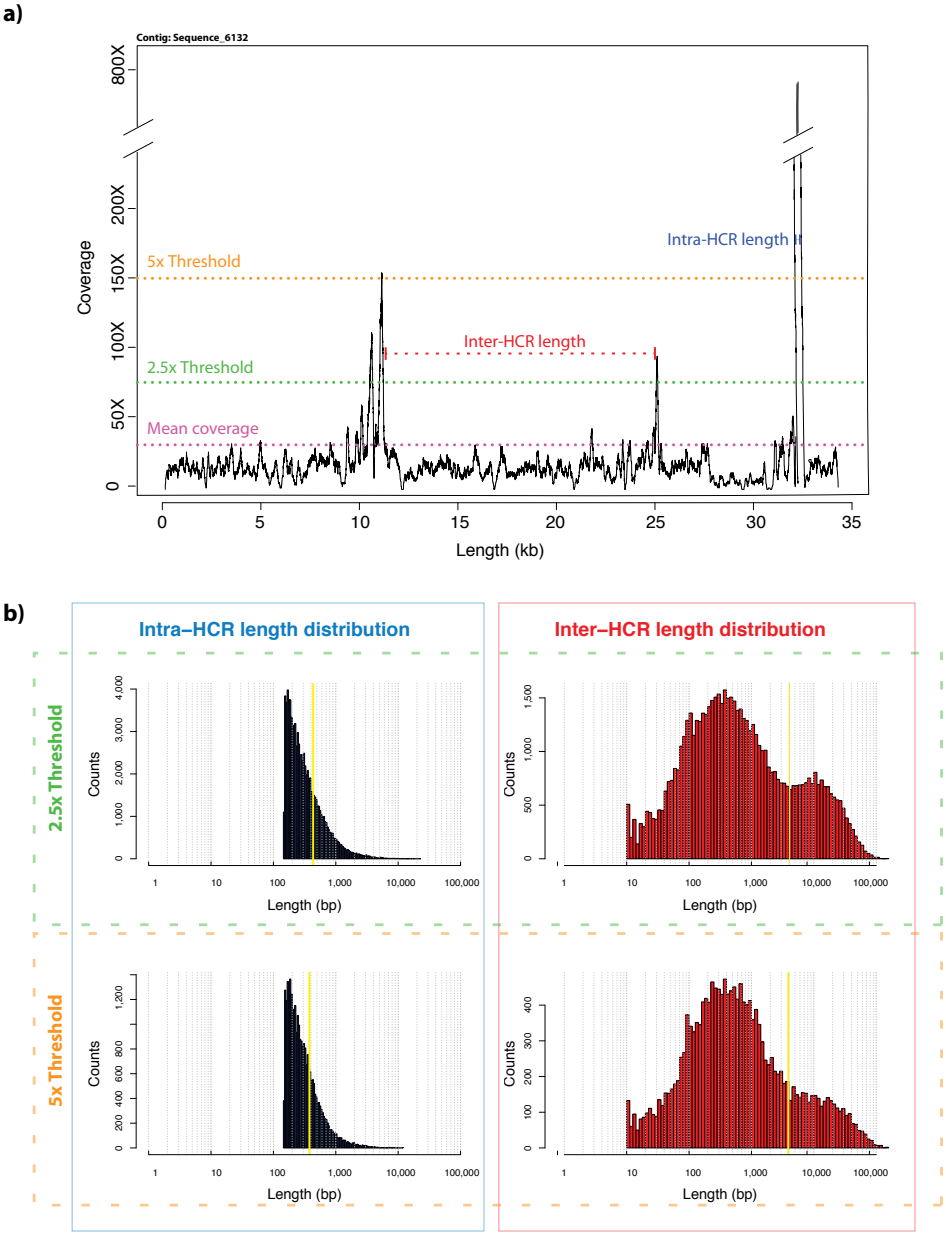

a)
Contig: Sequence_6132
Coverage
800X
200X
150X
100X
50X
0
5x Threshold
2.5x Threshold
Mean coverage
Intra-HCR length
Inter-HCR length
0
5
10
15
20
25
30
35
Length (kb)
b)
Intra–HCR length distribution
Inter–HCR length distribution
2.5x Threshold
5x Threshold
Counts
Length (bp)
1
10
100
1,000
10,000
100,000

Figure S4

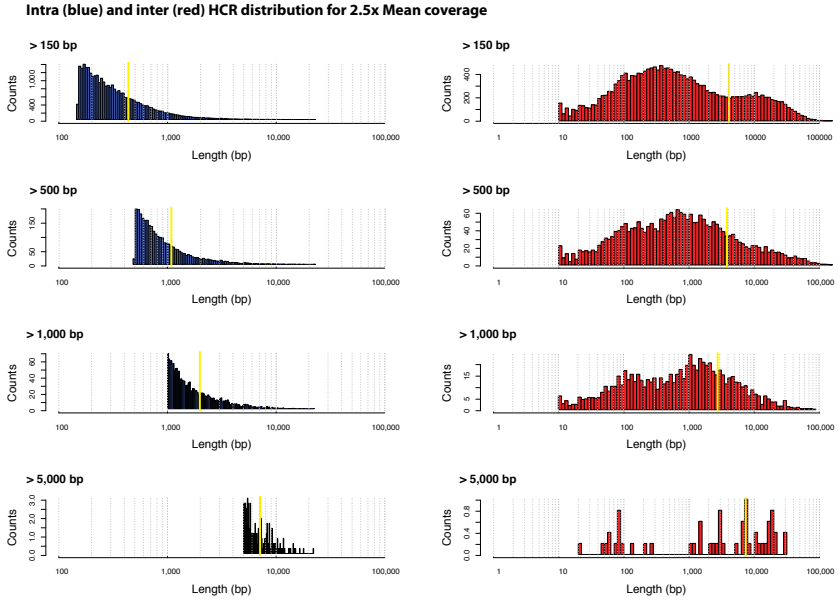


Figure S5a

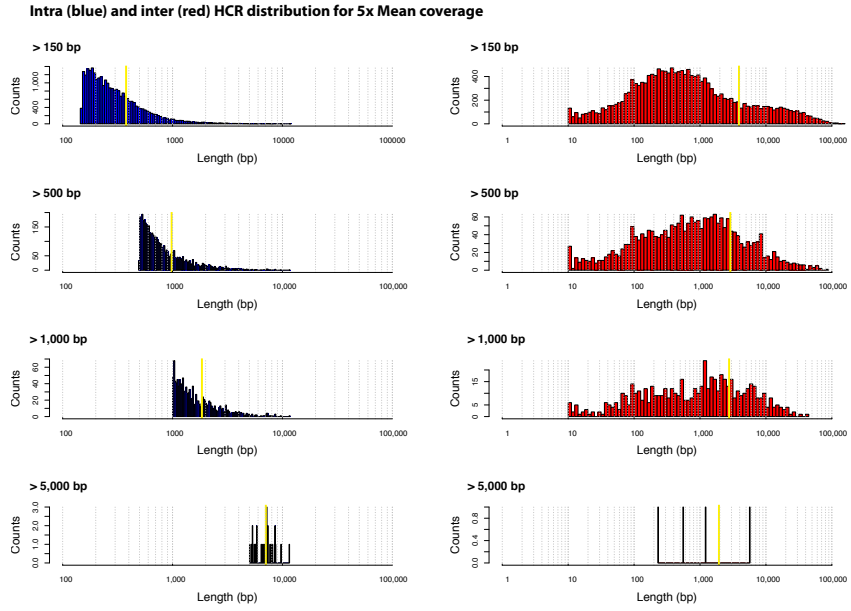


Figure S5b

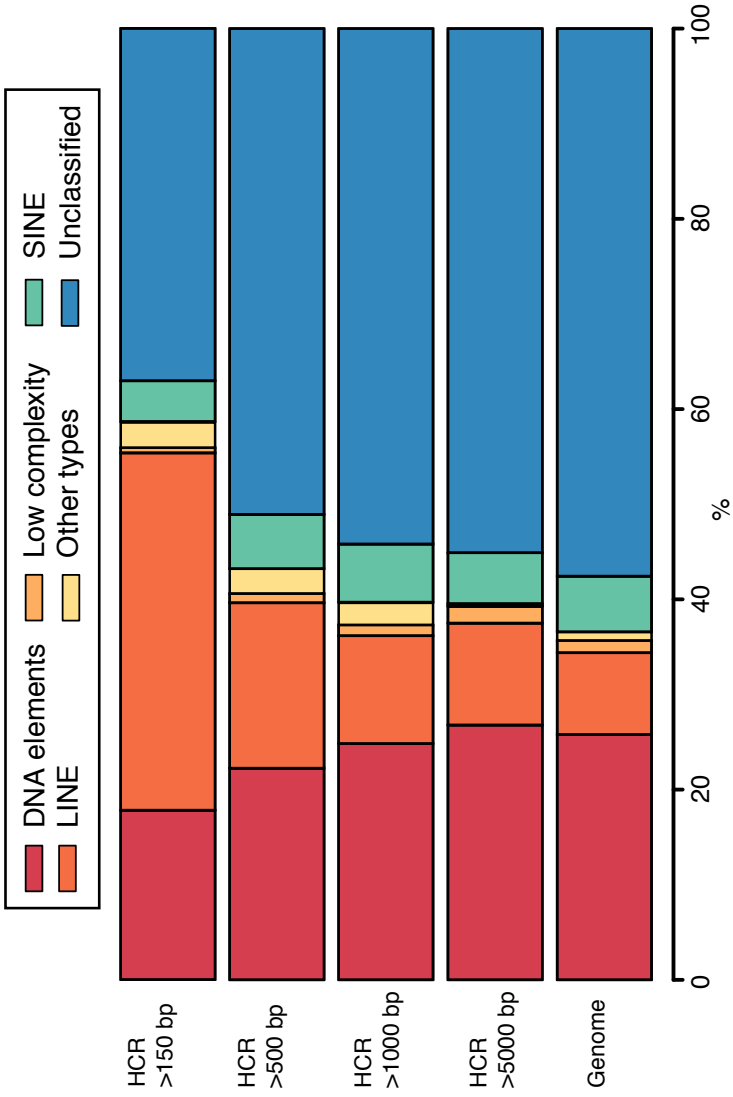

DNA elements
LINE
Low complexity
Other types
SINE
Unclassified
HCR >150 bp
HCR >500 bp
HCR >1000 bp
HCR >5000 bp
Genome
0
20
40
60
80
100
%

Figure S6

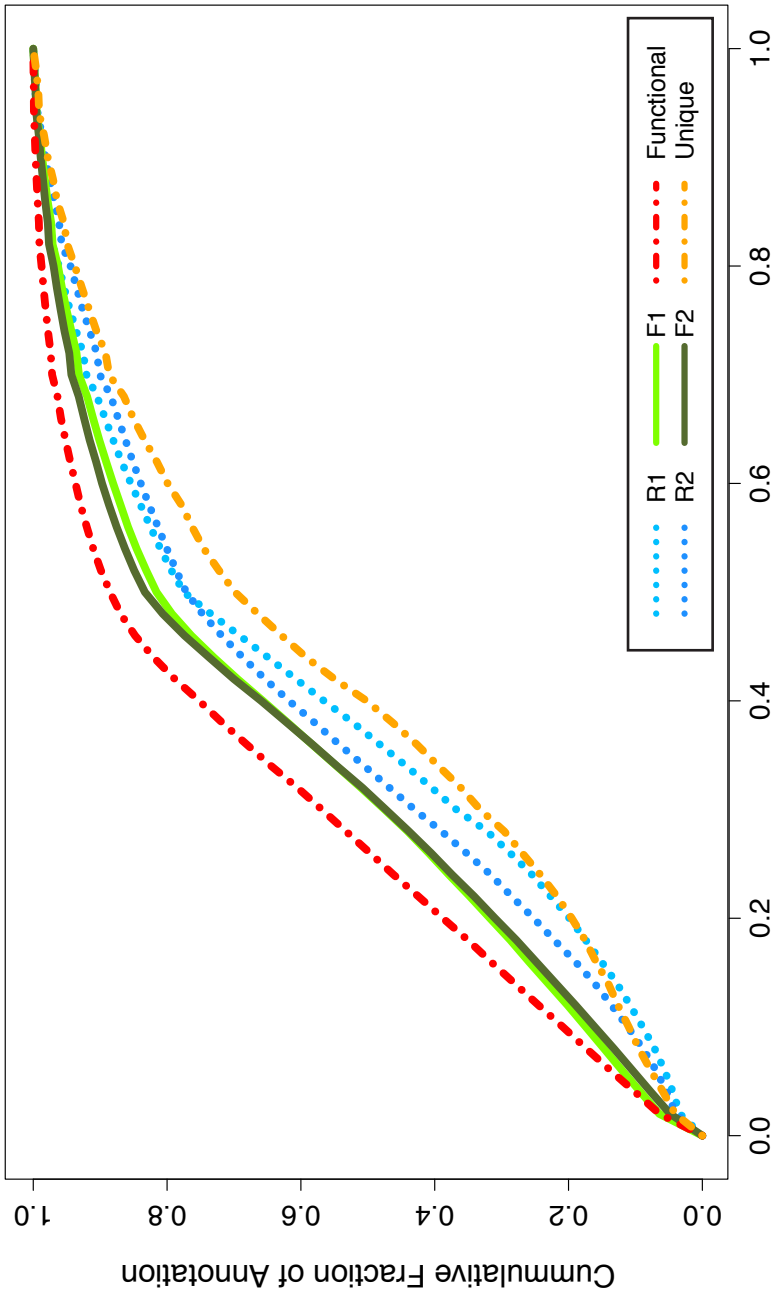

Cummulative Fraction of Annotation
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
AED value
R1
R2
F1
F2
Functional
Unique

Figure S7

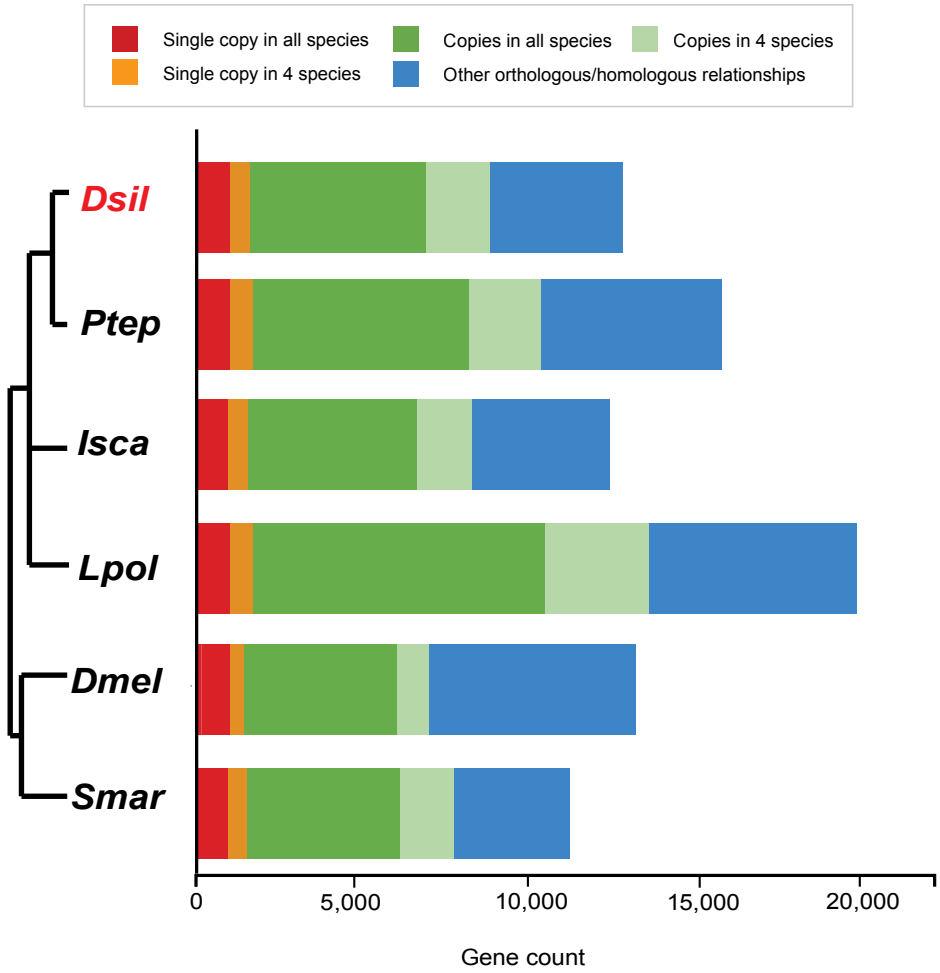

Single copy in all species
Copies in all species
Copies in 4 species
Single copy in 4 species
Other orthologous/homologous relationships
Dsil
Ptep
Isca
Lpol
Dmel
Smar
0
5,000
10,000
15,000
20,000
Gene count

Figure S8

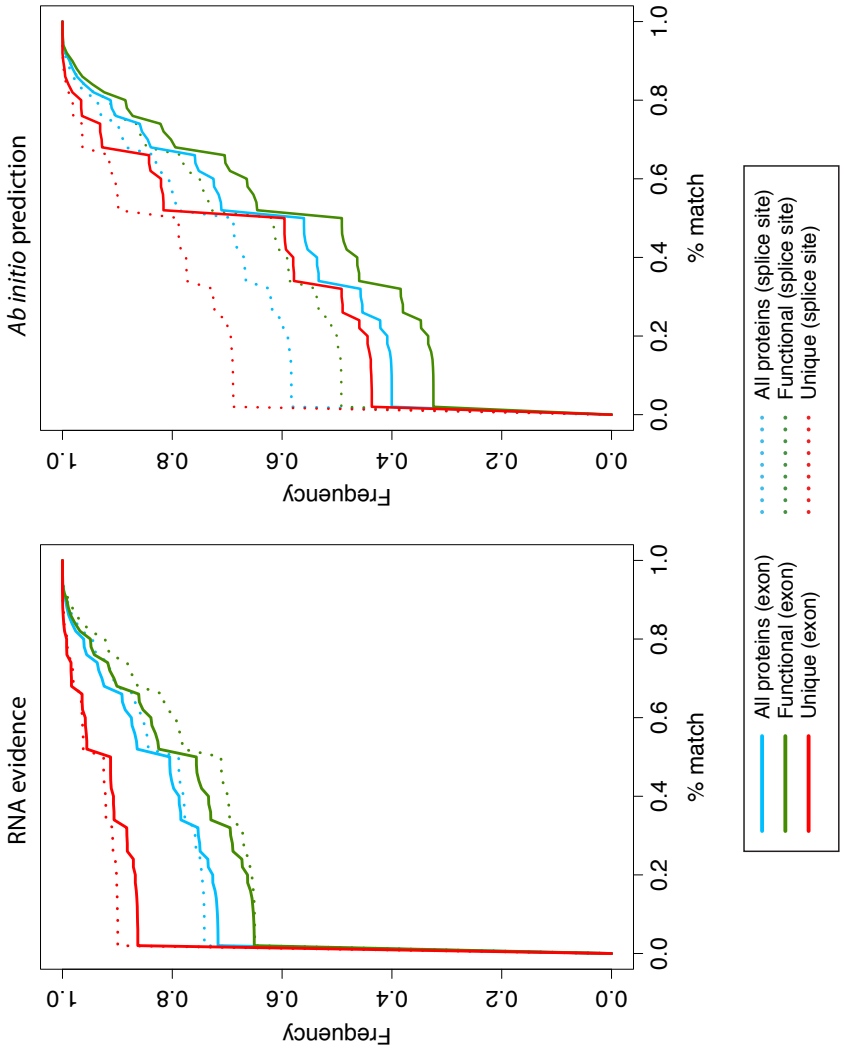

RNA evidence
Ab initio prediction
Frequency
% match
0.0
0.2
0.4
0.6
0.8
1.0
All proteins (exon)
Functional (exon)
Unique (exon)
All proteins (splice site)
Functional (splice site)
Unique (splice site)

Figure S9

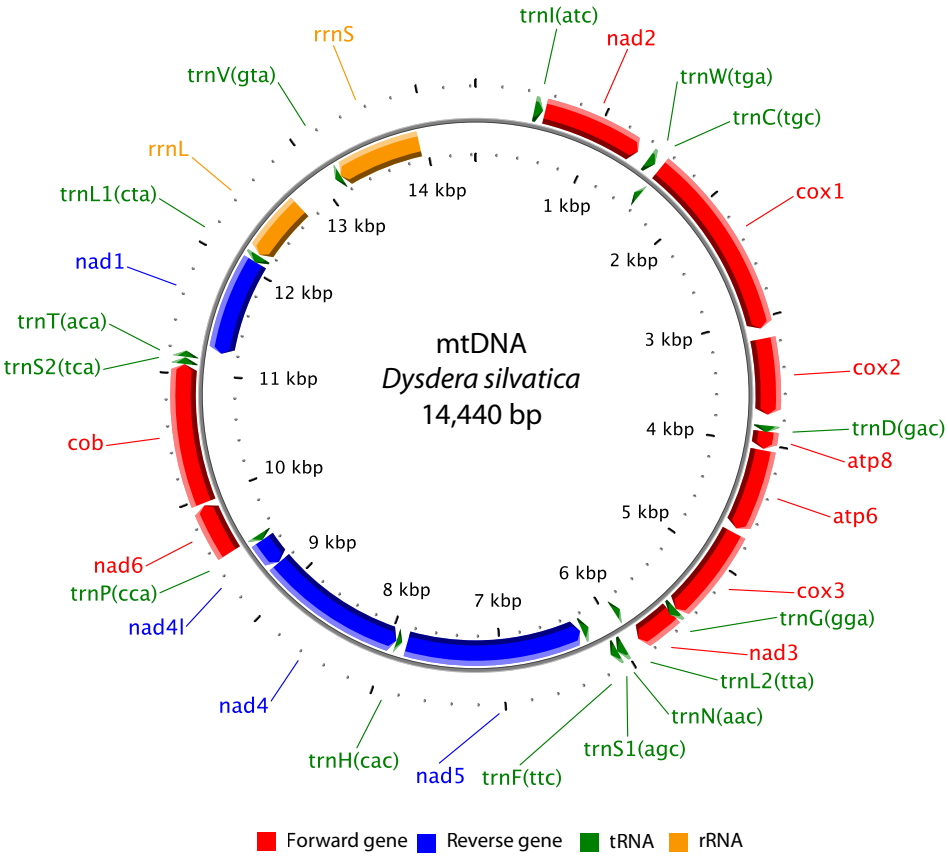

mtDNA
Dysdera silvatica
14,440 bp
1 kbp
2 kbp
3 kbp
4 kbp
5 kbp
6 kbp
7 kbp
8 kbp
9 kbp
10 kbp
11 kbp
12 kbp
13 kbp
14 kbp
trnI(atc)
nad2
trnW(tga)
trnC(tgc)
cox1
cox2
trnD(gac)
atp8
atp6
cox3
trnG(gga)
nad3
trnL2(tta)
trnN(aac)
trnS1(agc)
trnF(ttc)
nad5
trnH(cac)
nad4
nad4l
trnP(cca)
nad6
cob
trnS2(tca)
trnT(aca)
nad1
trnL1(cta)
rrnL
trnV(gta)
rrnS
Forward gene
Reverse gene
tRNA
rRNA

Figure S10

SUPPLEMENTARY TABLES

Table S1-1: **Collection of samples used in these study**

Sample ID	Species	Sex	NCBI BioSample Accession	Sampling Date	Location	Latitude	Longitude	Extraction Protocol	Extraction Date	Application	Run ID	Libraries	Machine	Read Type	Chemistry
NMH2597	*Dysdera silvatica*	Male	SAMN09381544	26-Mar-13	Las Tajoras, La Gomera, Spain	28.112 N	17.262 W	Qiagen DNeasy Blood & Tissue Kit	16-Sep-14	DNA fragment library	Dsilvatica_NMH2597_JFSH-CF_gDNA_PE	Short fragment	Illumina HiSeq 2000	Paired end	TruSeq
NMH2605	*Dysdera silvatica*	Female	SAMN09381545	26-Mar-13	Las Tajoras, La Gomera, Spain	28.112 N	17.262 W	Qiagen DNeasy Blood & Tissue Kit	15-Jan-15	DNA Nextera library	Dsilvatica_NMH2605_JFSH-CF_gDNA_5kb_MP	5 kb Nextera Mate Pair	Illumina HiSeq 2000	Paired end	Nextera
NMH2610	*Dysdera silvatica*	Male	SAMN09389329	04-Jun-13	Teselinde, La Gomera, Spain	28.196 N	17.287 W	Qiagen DNeasy Blood & Tissue Kit	15-May-16	PacBio SMRT libraries	Dsilvatica_NMH2610_JFSH_PacB	20 kb SMRTbell Templates	PacBio RSII	SMRT long reads	P6-C4
NMH1840	*Dysdera silvatica*	Female	SAMN11609372	10-Mar-12	Teselinde, La Gomera, Spain	28.196 N	17.287 W	Qiagen Blood Cell culture Minikit (modified)	16-Mar-18	Nanopore library	Dsilvatica_NMH1840_Nanopore	Nanopore 1D	-	long reads	-
NMH1829	*Dysdera silvatica*	Female	SAMN11609373	10-Mar-12	Teselinde, La Gomera, Spain	28.196 N	17.287 W	Qiagen Blood Cell culture Minikit (modified)	28-Jun-18	Nanopore library	Dsilvatica_NMH1829_Nanopore	Nanopore 1D	-	long reads	-

Table S1-2: **DNA sequencing read files used in this study**

a) Short reads libraries

Library	Run ID	Read Lengths	Read #1 Filename	Read #2 Filename	Total Bases	Raw Read Pairs	GC (%)	Q20 (%)	Q30 (%)	NCBI SRA Accession
Short fragment	Dsilvatica_NMH2597_JFSH-CF_gDNA_PE	100x100 PE	Dsilvatica_NMH2597_JFSH-CF_gDNA_PE_1	Dsilvatica_NMH2597_JFSH-CF_gDNA_PE_2	51,202,445,102	506,954,902	36.68	95.26	86.55	SRR7340408
5 kb Nextera Mate Pair	Dsilvatica_NMH2605_JFSH-CF_gDNA_5kb_MP	100x100 PE	Dsilvatica_NMH2605_JFSH-CF_gDNA_5kb_MP_1	Dsilvatica_NMH2605_JFSH-CF_gDNA_5kb_MP_2	39,609,522,995	392,173,495	35.39	98.55	93.95	SRR7340407

b) PacBio long reads libraries

Library	Run ID	Read Lengths[a]	Read #1 Filename	Total Bases	Raw Read Pairs	NCBI SRA Accession
20 kb SMRTbell Templates	Dsilvatica_NMH2610_JFSH_PacB-1	SMRT	JFSH_NMH2610_Dsil_PB_subreads	1,268,701,492	175,819	SRR7429499
20 kb SMRTbell Templates	Dsilvatica_NMH2610_JFSH_PacB-2	SMRT	JFSH_NMH2610_Dsil_PB-2_subreads	1,262,187,088	185,427	SRR7429498
20 kb SMRTbell Templates	Dsilvatica_NMH2610_JFSH_PacB-3	SMRT	JFSH_NMH2610_Dsil_PB-3_subreads	1,215,789,613	184,206	SRR7429501
20 kb SMRTbell Templates	Dsilvatica_NMH2610_JFSH_PacB-4	SMRT	JFSH_NMH2610_Dsil_PB-4_subreads	1,185,146,691	179,94	SRR7429500
20 kb SMRTbell Templates	Dsilvatica_NMH2610_JFSH_PacB-5	SMRT	JFSH_NMH2610_Dsil_PB-5_subreads	1,209,437,458	186,657	SRR7429503
20 kb SMRTbell Templates	Dsilvatica_NMH2610_JFSH_PacB-6	SMRT	JFSH_NMH2610_Dsil_PB-6_subreads	1,231,887,854	194,049	SRR7429505
20 kb SMRTbell Templates	Dsilvatica_NMH2610_JFSH_PacB-7	SMRT	JFSH_NMH2610_Dsil_PB-7_subreads	1,240,408,666	193,592	SRR7429502
20 kb SMRTbell Templates	Dsilvatica_NMH2610_JFSH_PacB-8	SMRT	JFSH_NMH2610_Dsil_PB-8_subreads	1,039,286,018	155,598	SRR7429504
[a]SMRT: Single Molecule Real Time			Total PB subreads:	9,652,844,880	1,455,288	

c) Nanopore long read libraries

Library	Run ID	Read Lengths[b]	Read #1 Filename	Total Bases	Raw Read Pairs	NCBI SRA Accession
ONT_1	Dsilvatica_NMH1840_Nanopore_1	Nanopore	Dsilvatica_NMH1840_Nanopore_1.fastq	8.070.627.302	3517656	SRR9033390
ONT_2	Dsilvatica_NMH1840_Nanopore_2	Nanopore	Dsilvatica_NMH1840_Nanopore_2.fastq	5.672.731.340	2156287	SRR9033391
ONT_3	Dsilvatica_NMH1840_Nanopore_3	Nanopore	Dsilvatica_NMH1840_Nanopore_3.fastq	6.393.750.434	7407703	SRR9033392
ONT_4	Dsilvatica_NMH1829_Nanopore_4	Nanopore	Dsilvatica_NMH1829_Nanopore_4.fastq	1.808.992.949	1891482	SRR9033393
ONT_5	Dsilvatica_NMH1829_Nanopore_5	Nanopore	Dsilvatica_NMH1829_Nanopore_5.fastq	1.247.255.456	1312358	SRR9033389
[b]Nanopore variable lengths			Total Nanopore subreads:	23,193,357,481	16,285,486	

Table S1-3: **NCBI Data used for the contaminant search step**

Taxon	Genomes retrieved	FTP site
Archae	683	ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/archaea/assembly_summary.txt
Virus	7.538	ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/viral/assembly_summary.txt
Bacteria	95.973	ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/bacteria/assembly_summary.txt
Mitochondria	5.883	ftp://ftp.ncbi.nlm.nih.gov/refseq/release/mitochondrion/

Downloaded on October, 2015

Table S1-4: **Pre-processing statistics for each library**

a) Trimming of PE reads using Trimmomatic

Class	Nº reads	Percentage (%)
Input read pairs	253.445.279	100
Both surviving	243.763.790	96,18
Forward only	8.031.776	3,17
Reverse only	1.312.067	0,52
Dropped	337.646	0,13

b) Trimming of MP reads using NxTrim

Nº reads	Percentage (%)	
195.996.473	100	reads passed chastity/purity filters
189	0,00	reads had two copies of adapter
166.336	0,08	reads pairs were ignored because a template length appeared less than read length
195.829.948	99,91	remaining reads were trimmed
62.192.088	31,76	read pairs had MP orientation
50.195.582	25,63	read pairs had PE orientation
81.678.049	41,71	read pairs had Unknown orientation
1.764.229	0,90	were single end reads
23.410.880	11,95	extra single end reads were generated from overhangs

c) Circularizing Pacbio reads using SMRT analysis software

Lane	Raw Bases	Subreads	Circularized Bases	CCS reads*
1	1.268.701.492	175.819	46.040.000	5.184
2	1.262.187.088	185.427	62.310.000	7.364
3	1.215.789.613	184.206	65.070.000	7.966
4	1.185.146.691	179.940	64.960.000	7.796
5	1.209.437.458	186.657	73.500.000	8.772
6	1.231.887.854	194.049	58.240.000	7.954
7	1.240.408.666	193.592	52.170.000	7.437
8	1.039.286.018	155.598	33.740.000	4.824
TOTAL	9.652.844.880	1.455.288	456.030.000	57.297

CCS: Circularized Consensus Sequences

Table S1-5: **Samples used for the RNAseq study**

Species	Sex	Sampling date	Latitude	Longitude	Location	Tissue	Library	Machine	NCBI BioSample Accession
Dysdera silvatica	Male	2013	28.112 N	17.262 W	Las Tajoras, La Gomera, Spain	First pair of legs	RNA TrueSeq	Illumina HiSeq 2000	SAMN04527047
Dysdera silvatica	Male	2013	28.112 N	17.262 W	Las Tajoras, La Gomera, Spain	Pedipalps	RNA TrueSeq	Illumina HiSeq 2000	SAMN04527048
Dysdera silvatica	Male	2013	28.112 N	17.262 W	Las Tajoras, La Gomera, Spain	Remaining legs	RNA TrueSeq	Illumina HiSeq 2000	SAMN04527049
Dysdera silvatica	Male	2013	28.112 N	17.262 W	Las Tajoras, La Gomera, Spain	Remaining parts of the spider	RNA TrueSeq	Illumina HiSeq 2000	SAMN04527050

All samples belong to the NCBI project PRJNA313901 and they are all a mixed of the RNA extraction from four different *D. silvatica* individuals with code: NMH2597, NMH2598, NMH2599, NMH2601

Reference:
Joel Vizueta, Cristina Frías-López, Nuria Macías-Hernández, Miquel A. Arnedo, Alejandro Sánchez-Gracia, Julio Rozas;
Evolution of Chemosensory Gene Families in Arthropods: Insight from the First Inclusive Comparative Transcriptome Analysis across Spider Appendages.
Genome Biol Evol 2017; 9 (1): 178-196. doi: 10.1093/gbe/evw296

Table S1-6: **RNA sequencing read files**

Library	Read Lengths	Tissue	Read #1 Filename	Read #2 Filename	# Read Pairs	NCBI SRA Accession
RNA TrueSeq	100x100 PE	First pair of legs	CF1Leg_1	CF1Leg_2	59.008.693	SRX1612801
RNA TrueSeq	100x100 PE	Pedipalps	CF2Palp_1	CF2Palp_2	57.493.091	SRX1612802
RNA TrueSeq	100x100 PE	Remaining legs	CF3Rest_1	CF3Rest_1	51.932.520	SRX1612803
RNA TrueSeq	100x100 PE	Remaining parts of the spider	CF4Body_1	CF4Body_2	52.483.628	SRX1612804

All samples belong to the NCBI project PRJNA313901 and they are all a mixed of the RNA extraction from four different *D. silvatica* individuals with code: NMH2597, NMH2598, NMH2599, NMH2601

Reference:
Joel Vizueta, Cristina Frías-López, Nuria Macías-Hernández, Miquel A. Arnedo, Alejandro Sánchez-Gracia, Julio Rozas;
Evolution of Chemosensory Gene Families in Arthropods: Insight from the First Inclusive Comparative Transcriptome Analysis across Spider Appendages.
Genome Biol Evol 2017; 9 (1): 178-196. doi: 10.1093/gbe/evw296

***Table S1-7*: Genome descriptive statistics**

a) Summary of the comparison of genome assembly steps

		MaSurCA Assembly	AGOUTI RNA scaffolding	Clean contaminants (v.1.2)
Assembly Statistics	Sequences	66.830	65.221	65.205
	Total Length (bp)	1.359.506.718	1.361.115.718	1.359.336.805
	A	32,484	32,445	32,445
	T	32,504	32,468	32,467
	C	17,431	17,412	17,411
	G	17,44	17,417	17,417
	A+T	64,989	64,912	64,913
	C+G	34,87	34,829	34,828
	N	0,14	0,258	0,258
	Length (no Ns)	1.357.599.280	1.357.599.280	1.355.820.767
	Capture Gaps[a]	19.075	20.684	20.680
	Capture Gaps Length	1.907.438	3.516.438	3.516.038
	MinLen	1.001	1.001	1.001
	MaxLen	381.119	381.119	340.047
	Average Len	20.342,76	20.869,29	20.847,13
	Median Len	12.186	12.176	12.171
	n50	36.307	38.050	38.017
	L50	11.019	10.428	10.436

Completeness statistics	BUSCO Arthropoda[b] [*n* = 1066]	Identified (I)	879 (82.46%)	916 (85.93%)	921 (86.4%)
		Complete (C)	739 (69.3%)	786 (73.7%)	782 (73.4%)
		Complete and single copy (S)	709 (66.5%)	753 (70.6%)	750 (70.4)
		Complete and duplicated (D)	30 (2.8%)	33 (3.1%)	32 (3%)
		Fragmented (F)	140 (13.1%)	130 (12.2%)	139 (13%)
		Missing (M)	187 (17.6%)	150 (14.1%)	145 (13.6%)
	BUSCO Metazoa[b] [*n* = 978]	Identified (I)	818 (83.64%)	855 (87.42%)	855 (87.42%)
		Complete (C)	689 (70.5%)	739 (75.6%)	734 (75.1%)
		Complete and single copy (S)	661 (67.6%)	705 (72.1%)	706 (72.2%)
		Complete and duplicated (D)	28 (2.9%)	34 (3.5%)	28 (2.9%)
		Fragmented (F)	129 (13.2%)	116 (11.9%)	121 (12.4%)
		Missing (M)	160 (16.3%)	123 (12.5%)	123 (12.5%)
	BLAST[f]	1 to 1 Dysdera[c] [9473]	9,121 (96.28%)	9,121 (96.28%)	9,120 (96.27%)
		CEGMA[d] [457]	435 (95.19%)	435 (95.19%)	434 (94.97%)
		Dysdera silvatica Transcripts[e] [58966]	40,778 (69.16%)	40,771 (69.14%)	40,499 (68.68%)

[a] Capture gaps are defined as regions containing a minimum of 25 consecutive base pairs codified as "N".

[b] BUSCO analysis using default parameters for different datasets: arthropoda (*n* = 1066 genes) and metazoa (*n* = 978 genes)

[c] TBLASTX search statistics (*evalue* 1e-30) of 1:1 Ortholog proteins (9,473) identified in the species *D. silvatica*, *D. gomerensis* Strand, 1911; *D. verneaui* Simon, 1883; *D. tilosensis* Wunderlich, 1992 and *D. bandamae* Schmidt, 1973 (Joel Vizueta et al., 2019, unpublished results)

[d] TBLASTX search statistics (evalue 1e-5) of CEGMA proteins from *Drosophila melanogaster* (*n* = 457).

[e] TBLASTX search statistics (evalue: 1e-30) of proteins from *Dysdera silvatica* transcriptome (*n* = 58,966) (Vizueta et. al 2017).

[f] Identified hits within the genome at the give e-value for each dataset at any coverage

b) Genome annotation statistics[a]

Protein coding genes	48.619
Functionally annotated	36,398 (74.86%)
Swissprot	17,225 (35.43%)
InterPro[b]	32,322 (66.48%)
Without functional annotation	12,221 (25.14%)
tRNA genes	33.934

[a] Values are refered to final assembly version (v1.2)

[b] Analysis using the following databases (included in InterProScan): PANTHER (Mi, Muruganujan, and Thomas 2013), Pfam (Punta et al. 2012), PRINTS (Attwood et al. 2012), PrositePatterns & PrositeProfiles (Gattiker, Gasteiger, and Bairoch 2002; Sigrist et al. 2012), SMART (Letunic, Doerks, and Bork 2012), SUPERFAMILY (de Lima Morais et al. 2011), TIGRAM (Haft et al. 2012), SFLD (Akiva et al. 2014), Gene3D (Lees et al. 2012), Hamap (Pedruzzi et al. 2015), ProDom (Bru et al. 2004), PIRSF (Nikolskaya et al. 2007), MobiDBLite (Necci et al. 2017).

Table S1-8: Coverage analyses

a) Estimated coverage based on the total sequencing data

Run ID	Total Bases	Coverage[a]
PE	51.202.445.102	30X
MP	39.609.522.995	23X
ONT	23.193.357.481	14X
Pacbio	9.652.844.880	6X
Pacbio CSS	456.030.000	0.27X
All	123.658.170.458	73X

[a]Based on the genome size estimated by flow cytometry (1.7 Gb)
ONT: Oxford Nanoporte Technologies

b) Estimated coverage based on mapped data

Run ID	Mean Coverage[a]	Mean Coverage[b]
PE	25.7X	32.12X
MP	12.58X	15.72X
ONT	8.46X	10.57X
Pacbio CSS	0.23X	0.23X
All	46.97X	58.64X

[a]Based on the genome size estimated by flow cytometry (1.7 Gb)
[b]Based on the genome size assembled (1.36 Gb)
ONT: Oxford Nanoporte Technologies

***Table S1-9*: RepeatMasker analysis of *D. silvatica* genome**

a) RepeatMasker analyses

Type	Number of elements[a]	Length (bp)	Sequence (%)
SINEs[b]	170.863	25.173.255	1,85
LINEs[c]	256.091	145.456.128	10,7
LINE1	608	152.963	0,01
LINE2	37.773	21.216.392	1,56
L3/CR1	2.525	1.461.178	0,11
LTR elements[d]	20.021	8.746.088	0,64
ERV_classI	171	155.175	0,01
ERV_classII	7.604	940.768	0,07
DNA elements	779.436	227.842.465	16,76
hAT-Charlie	119.132	24.177.681	1,78
TcMar-Tigger	67.086	12.923.815	0,95
Unclassified:	1.729.161	306.272.536	22,53
Total interspersed repeats		713.490.472	52,48
Small RNA	665	960.355	0,07
Satellites	4.653	1.491.285	0,11
Simple repeats	282.910	13.440.971	0,99
Low complexity	41.169	2.157.298	0,16
Total Bases masked (bp)	**3.284.969**	**731.540.381**	**53,81**

[a]Most repeats fragmented by insertions or deletions have been counted as one element
[b] SINEs: short interspersed nuclear element
[c] LINEs: long interspersed nuclear element
[d] LTR: long terminal repeat

b) Top 10 Repeat families analysis[a]

Family name	Count hits[b]	Type of family	Count total hits (%)[c]	Mean length[d]	Standard Deviation	Mean length Unknown[d]	Mean length SINEs[d]	Mean length LINEs[d]
rnd-2_family-13	28.831	Unknown	0,915	179,53	108,72	179,53	-	-
rnd-6_family-515	25.708	Unknown	0,816	257,79	203,56	257,79	-	-
rnd-1_family-23	23.184	Unknown	0,736	122,18	40,65	122,18	-	-
rnd-3_family-426	22.537	Unknown	0,715	158,04	88,64	158,04	-	-
rnd-1_family-35	20.382	LINE/RTE-RTE	0,647	1412,43	1047,77	-	-	1412,43
rnd-4_family-547	20.258	SINE/ID	0,643	119,51	64,39	-	119,51	-
rnd-1_family-36	18.802	Unknown	0,597	251,98	325,65	251,98	-	-
rnd-1_family-7	18.350	SINE/ID	0,582	212,24	49,16	-	212,24	-
rnd-3_family-155	17.829	LINE/RTE-RTE	0,566	665,48	654,55	-	-	665,48
rnd-6_family-1201	16.467	SINE/ID	0,523	152,32	97,43	-	152,32	-
Total	212.348		6.741			193,91	161,36	1038,96

[a] Top 10 most common repeats across the 2,604 families identified
[b] The number of times (hits) identified across the genome
[c] Percentage of hits for each family identified among the 3,150,262 total hits identified
[d] Mean repeat length

***Table S1-10*: Reference guided transcriptome assembly statistics**

Sequences	245.905
Total Length (bp)	145.213.967
A (%)	31,867
T (%)	31,599
C (%)	18,049
G (%)	18,483
A+T (%)	63,467
C+G (%)	36,532

Set	>150 bp	>500 bp	>1000 bp
Number Seqs	245.905	79.722	33.778
% Seqs	100	32,419	13,736
Total Length (bp)	145.213.967	94.768.228	63.221.799
% Bases	100	65,261	43,537
Minimum Length (bp)	201	501	1.001
Maximun Length (bp)	16.350	16.350	16.350
Average Length (bp)	590,53	1.188,73	1.871,69
Median Length (bp)	355	875	1.547
N50	804	1.392	1.956

***Table S1-11*: Source of proteins to conduct the annotation of *D. silvatica* genes and completeness analysis (Figure 2).**

Species	Taxonomic range	Source	Downloaded from (id)
Drosophila melanogaster	hexapoda	UNIPROT	ftp uniprot.org (UP000000803)
Bombyx mori	hexapoda	UNIPROT	ftp uniprot.org (UP000005204)
Pediculus humanus	hexapoda	UNIPROT	ftp uniprot.org (UP000009046)
Daphnia pulex	crustacea	UNIPROT	ftp uniprot.org (UP000000305)
Strigamia maritima	myriapoda	UNIPROT	ftp uniprot.org (UP000014500)
Hypsibius dujardini	tardigrada	UNIPROT	ftp uniprot.org (UP000192578)
Caenorhabditis elegans	nematoda	UNIPROT	ftp uniprot.org (UP000001940)
Tetranychus urticae	acari	UNIPROT	ftp uniprot.org (UP000015104)
Euroglyphus maynei	acari	UNIPROT	ftp uniprot.org (UP000194236)
Ixodes scapularis	acari	UNIPROT	ftp uniprot.org (UP000001555)
Metaseiulus occidentalis	acari	NCBI	ftp ncbi.nlm.nih.gov (GCF_000255335.1_Mocc_1.0)
Stegodyphus mimosarum	aranae	UNIPROT	ftp uniprot.org (UP000054359)
Acanthoscurria geniculata	aranae	Original paper[a]	Additional information original paper
Lactodrectus hesperus	aranae	i5K	i5K Project
Loxosceles reclusa	aranae	i5K	i5K Project
Parasteatoda tepidariorum	aranae	i5K	i5K Project
Mesobuthus martensii	scorpion	NCBI	ftp ncbi.nlm.nih.gov (GCA_000484575.1)
Centruroides sculpturatus	scorpion	i5K	i5K Project
Limulus polyphemus	Xiphosura	RyanLab website	ryanlab.whitney.ufl.edu (GCF_000517525.1_Limulus_polyphemus-2.1.2)

[a] https://doi.org/10.1038/ncomms4765

***Table S2-1*: Example of results for the HCR analysis**

a) Example data file including the high coverage regions (HCRs) data[a]

Scaffold ID	Min Intra length[b]	Scaffold Length	# HCR[c]	HCR Id[d]	Inter Start[e]	Intert End[e]	Inter Length[f]	Intra Start[g]	Intra End[g]	Intra Length[h]
sequence_10003	150	24752	4	repeat_1	-	-	-	6338	6508	170
sequence_10003	150	24752	4	repeat_2	6509	11083	4574	11084	11776	692
sequence_10003	150	24752	4	repeat_3	11777	20259	8482	20260	21098	838
sequence_10003	150	24752	4	repeat_4	21099	22039	940	22040	22497	457
sequence_10005	150	13790	3	repeat_1	-	-	-	9668	10197	529
sequence_10005	150	13790	3	repeat_2	10198	10335	137	10336	11212	876
sequence_10005	150	13790	3	repeat_3	11213	11250	37	11251	11749	498
sequence_10011	150	14044	1	repeat_1	-	-	-	8016	8201	185
sequence_10010	150	12713	2	repeat_1	-	-	-	449	1347	898
sequence_10010	150	12713	2	repeat_2	1348	1484	136	1485	1715	230
sequence_10008	150	78454	2	repeat_1	-	-	-	56694	56874	180
sequence_10008	150	78454	2	repeat_2	56875	74155	17280	74156	74466	310
...	...	...	...	...	...	...	...	...	...	...

[a] Example of the first entries of the data file (subset_10kb-2.5x_coverage.HCR_150.txt). This data file contains full information and is provided as Supplementary Files.

[b] Minimum intra-repeat length cutoff

[c] Number of High Coverage regions (HCR) in this scaffold

[d] HCR identifier within scaffold

[e] Inter-repeat start and end coordinates within the scaffold

[f] Inter-repeat length

[g] Intra-repeat start and end coordinates within the scaffold

[h] Intra-repeat length

b) Example in BED format of the intersection of the annotation between RepeatMasker and HCR data[a]

Scaffold ID	Intra HCR Start	Intra HCR End	HCR ID	HCR Score[b]	HCR Strand[c]	Scaffold ID	Repeat Start	Repeat End	Repeat (Family == Type)	Repeat Score	Repeat Strand	Overlapping Base Pairs[d]
sequence_10003	11084	11776	repeat_2	1	.	sequence_10003	10627	11807	rnd-4_family-152==LINE/RTE-RTE	10056	+	692
sequence_10003	20260	21098	repeat_3	1	.	sequence_10003	19683	22063	rnd-1_family-685==LINE/L2	13611	+	838
sequence_10003	22040	22497	repeat_4	1	.	sequence_10003	19683	22063	rnd-1_family-685==LINE/L2	13611	+	23
sequence_10005	10336	11212	repeat_2	1	.	sequence_10005	9655	10691	rnd-1_family-678==Unknown	8946	+	355
sequence_10005	10336	11212	repeat_2	1	.	sequence_10005	10283	11237	rnd-1_family-10==Unknown	8156	+	876
sequence_10005	10336	11212	repeat_2	1	.	sequence_10005	10828	11768	rnd-1_family-364==Unknown	7436	+	384
sequence_10005	11251	11749	repeat_3	1	.	sequence_10005	10828	11768	rnd-1_family-364==Unknown	7436	+	498
sequence_10010	1485	1715	repeat_2	1	.	sequence_10010	1	1804	rnd-1_family-84==LINE/Dong-R4	15595	+	230
sequence_10008	74156	74466	repeat_2	1	.	sequence_10008	74123	74520	rnd-1_family-84==LINE/Dong-R4	3457	+	310
...	...	...	...	...	...	...	...	...	...	...	...	...

[a] Example of the firsts entries of the data file (HCR_annotation_2.5_150-intersection.bed). This data file was generated using Bedtools (intersectBed -wao option) to calculate the intersection of the annotation by repeatmasker and the HCR annotation previously generated. The whole data file is provided as Supplementary Files

[b] HCR Score: BED format attribute. Check for further details https://genome.ucsc.edu/FAQ/FAQformat.html#format1

[c] HCR Strand: No information available.

[d] Overlapping base pairs among both features (Repeatmasker annotation and HCR element)

***Table S2-2*: High coverage regions (HCRs) analysis**

a) Descriptive statistics in the subset (34,937 contigs)

	Length	Number HCR[a]	Contigs containing HCR (%)	HCR/Contig[b]	HCR/Total Contig[c]	Number HCR/Total Mb
2.5x	150	86.641	21,614 (61.87)	4,01	2,48	77,7
	500	18.375	7,764 (22.22)	2,37	0,53	16,48
	1.000	5.637	2,605 (7.45)	2,16	0,16	5
	5.000	251	183 (0.52)	1,37	0,01	0,23
5x	150	28.512	10,604 (30.35)	2,69	0,81	25,57
	500	4,690	2,277 (6.51)	2,06	0,12	4,21
	1.000	1,248	628 (1.79)	1,98	0,03	1,12
	5.000	31	26 (0.07)	1,19	0	0,03

[a] High Cover

[b] Median number of HCR per contig containing ≥1 HCR

[c] Median number of HCR per contig.

b) Inter and Intra HCR lengths distribution

Intra-HCR

	Length	Min.	1st Quartile	Median	Mean	3rd Quartile	Max.
2.5x	150	150	195	274	433,6	450	21.901
	500	500	597	755	1.079	1.120	21.901
	1.000	1.000	1.170	1.467	1.985	2.195	21.901
	5.000	5.016	5.447	6.243	7.119	8.200	21.901
5x	150	150	192	261	377,7	398	11.598
	500	500	580	712,5	980,9	1029,8	11.598
	1.000	1.000	1.162	1.436	1.850	2.048	11.598
	5.000	5.013	5.770	7.170	7.060	7.826	11.598

Inter-HCR

	Length	Min.	1st Quartile	Median	Mean	3rd Quartile	Max.
2.5x	150	10	149	523	4.081	2.691	144.900
	500	10	161	666	3.817	2.483	150.879
	1.000	10	163	855,5	2.782,30	2.577,50	82.821
	5.000	20	272	3.375	7.389	12.348	31.256
5x	150	10	161	504	3.901	1.896	146.087
	500	10	189,2	768,5	2.882,20	2.462,80	81.329
	1.000	10	197,8	1.037	2.762,30	3.035	41.244
	5.000	232	476,5	892,5	1.938,50	2.354,50	5.737

***Table S2-3*: Enrichment analysis of the intersection of High Coverage regions (HCRs) with RepeatMasker annotation (main repeats)[a]**

a) 2.5x Mean Coverage

Type	Average Length (bp)	Genome[b]	> 150 bp	> 500 bp	> 1000 bp	> 5000 bp
SINEs[d]	157,13	187240 (5.87 %)[a]	3552 (4.07 %) [0/1][c]	1820 (4.93 %) [0/1]	1238 (5.75 %) [0.389/0.622]	218 (6.14 %) [0.769/0.253]
LINEs[e]	541,12	274771 (8.62 %)	31478 (36.12 %) [1/0]	7651 (20.74 %) [1/0]	2601 (12.08 %) [1/0]	355 (10.01 %) [0.998/0.002]
LTR elements[f]	388,72	22969 (0.72 %)	1288 (1.48 %) [1/0]	493 (1.34 %) [1/0]	238 (1.11 %) [1/0]	34 (0.96 %) [0.957/0.062]
DNA elements	283,14	824501 (25.86 %)	18126 (20.8 %) [0/1]	8466 (22.94 %) [0/1]	5297 (24.61 %) [0/1]	881 (24.84 %) [0.085/0.921]
Unclassified	176,24	1840236 (57.72 %)	31928 (36.63 %) [0/1]	18022 (48.84 %) [0/1]	11847 (55.04 %) [0/1]	2004 (56.5 %) [0.073/0.931]
Small RNA[g]	1.128,62	851 (0.03 %)	261 (0.3 %) [1/0]	92 (0.25 %) [1/0]	56 (0.26 %) [1/0]	0 (0 %) [0.388/1]
Satellites	307,13	4890 (0.15 %)	166 (0.19 %) [0.9974/0.0034]	76 (0.21 %) [0.995/0.008]	39 (0.18 %) [0.87/0.168]	6 (0.17 %) [0.696/0.461]
Low complexity	51,2	41405 (1.3 %)	422 (0.48 %) [0/1]	314 (0.85 %) [0/1]	229 (1.06 %) [0.001/0.999]	54 (1.52 %) [0.893/0.136]
	Total	3,188,449[h]	87.159	36.898	21.526	3.547

b) 5x Mean Coverage

Type	Average Length (bp)	Genome[b]	> 150 bp	> 500 bp	> 1000 bp	> 5000 bp
SINEs[d]	157,13	187240 (5.87 %)	948 (4.29 %) [0/1]	471 (5.72 %) [0.282/0.734]	251 (6.11 %) [0.756/0.265]	21 (5.36 %) [0.383/0.698]
LINEs[e]	541,12	274771 (8.62 %)	8305 (37.6 %) [1/0]	1434 (17.4 %) [1/0]	466 (11.35 %) [1/0]	42 (10.71 %) [0.938/0.085]
LTR elements[f]	388,72	22969 (0.72 %)	255 (1.15 %) [1/0]	72 (0.87 %) [0.953/0.06]	35 (0.85 %) [0.862/0.18]	0 (0 %) [0.059/1]
DNA elements	283,14	824501 (25.86 %)	3939 (17.83 %) [0/1]	1835 (22.27 %) [0/1]	1020 (24.84 %) [0.07/0.935]	105 (26.79 %) [0.686/0.356]
Unclassified	176,24	1840236 (57.72 %)	8191 (37.08 %) [0/1]	4212 (51.12 %) [0/1]	2226 (54.21 %) [0/1]	216 (55.1 %) [0.16/0.864]
Small RNA[g]	1.128,62	851 (0.03 %)	243 (1.1 %) [1/0]	95 (1.15 %) [1/0]	44 (1.07 %) [1/0]	0 (0 %) [0.901/1]
Satellites	307,13	4890 (0.15 %)	107 (0.48 %) [1/0]	47 (0.57 %) [1/0]	19 (0.46 %) [1/0]	1 (0.26 %) [0.878/0.452]
Low complexity	51,2	41405 (1.3 %)	123 (0.56 %) [0/1]	81 (0.98 %) [0.005/0.996]	47 (1.14 %) [0.213/0.826]	7 (1.79 %) [0.858/0.25]
	Total	3,188,449[h]	22.089	8.240	4.106	392

[a]Results generated using Bedtools (analysis of the intersection of the annotation by repeatmasker and the HCR annotations). The number of hits for each element could slightly differ from those results reported in Table S1-9a, as in this analysis we have not take into account overlapping annotations, or multiple annotations in the same region.

[b]Count -number of elements- and percentage across repeat elements (%)

[c]Count (%) [CDF/SF]; where CDF and SF are the P-values

CDF = *P*-value of observing (obtaining by random) a number of repeat elements less than or equal to the observed

SF = *P*-value of observing (obtaining by random) a number of repeat elements greater than or equal to the observed

[d]SINEs: short interspersed nuclear element

[e]LINEs: long interspersed nuclear element

[f]LTR: long terminal repeat

[g]Small RNA referes to rRNA

[h]Total amount of identified repeats could slightly differ from those reported by RepeatMasker. In the current analysis we have not take into account overlapping annotations, or multiple annotations in the same region.

Table S2-4: Enrichment analysis of the intersection of High Coverage regions (HCRs) (2.5 x mean coverage) with RepeatMasker annotation (subtypes of main repeats)[a]

				> 150 bp					> 500 bp					> 1000 bp					> 5000 bp				
	Total Genome[b]	Total Genome type[c]	%	Total Count Subset[d]	Total Subset Type[e]	%	CDF[f]	SF[g]	Total Count Subset	Total Subset Type	%	CDF	SF	Total Count Subset	Total Subset Type	%	CDF	SF	Total Count Subset	Total Subset Type	%	CDF	SF
DNA/CMC-Chapaev-3	3188449	5416	0,170	87159	127	0,146	0,041	0,966	36898	67	0,182	0,734	0,308	21526	38	0,177	0,635	0,428	3547	10	0,282	0,957	0,085
DNA/CMC-En5pm	3188449	10378	0,32549	87159	198	0,22717	3,08514E-08	0,999999975	36898	100	0,27102	0,033011047	0,973556235	21526	48	0,22299	0,003260787	0,997855522	3547	7	0,19735	0,1108	0,9416
DNA/CMC-Transib	3188449	1359	0,043	87159	40	0,046	0,718	0,340	36898	17	0,046	0,685	0,407	21526	12	0,056	0,863	0,214	3547	1	0,028	0,554	0,780
DNA/Crypton-H	3188449	1430	0,04485	87159	27	0,03098	0,02519494	0,983784669	36898	15	0,04065	0,412505741	0,682982943	21526	12	0,05575	0,823886197	0,264128302	3547	2	0,05639	0,7858	0,4722
DNA/Ginger	3188449	9241	0,290	87159	159	0,182	0,000	1,000	36898	81	0,220	0,005	0,996	21526	40	0,186	0,002	0,999	3547	11	0,310	0,665	0,452
DNA/hAT	3188449	364071	11,4184	87159	5891	6,75891	0	0,999999992	36898	3216	8,71592	2,40773E-65	0,999999998	21526	2113	9,81604	1,96349E-14	0,999999995	3547	380	10,7133	0,0969	0,912
DNA/hAT-Ac	3188449	7724	0,242	87159	153	0,176	0,000	1,000	36898	80	0,217	0,172	0,854	21526	47	0,218	0,263	0,781	3547	6	0,169	0,246	0,858
DNA/hAT-Blackjack	3188449	14738	0,46223	87159	282	0,32355	6,60647E-11	0,999999991	36898	103	0,27915	1,29213E-08	0,999999991	21526	57	0,2648	2,32585E-06	0,999998695	3547	19	0,53566	0,7842	0,2908
DNA/hAT-Charlie	3188449	127020	3,984	87159	1899	2,179	0,000	1,000	36898	1058	2,867	0,000	1,000	21526	682	3,168	0,000	1,000	3547	118	3,327	0,023	0,982
DNA/hAT-hAT19	3188449	3988	0,12508	87159	52	0,05966	6,03549E-10	0,999999991	36898	31	0,08402	0,011405698	0,992757131	21526	21	0,09756	0,145786098	0,859643132	3547	5	0,14096	0,7139	0,4559
DNA/hAT-hAT5	3188449	21404	0,671	87159	243	0,279	0,000	1,000	36898	150	0,407	0,000	1,000	21526	116	0,539	0,008	0,994	3547	30	0,846	0,912	0,123
DNA/hAT-hATm	3188449	10907	0,34208	87159	222	0,2547	1,69768E-06	0,999998756	36898	111	0,30083	0,091445787	0,922080642	21526	75	0,34842	0,559487043	0,452184155	3547	17	0,47928	0,9322	0,1082
DNA/hAT-Tag1	3188449	11250	0,353	87159	124	0,142	0,000	1,000	36898	56	0,152	0,000	1,000	21526	32	0,149	0,000	1,000	3547	4	0,113	0,005	0,998
DNA/hAT-Tip100	3188449	145979	4,57837	87159	2489	2,8557	1,0694E-153	1	36898	1393	3,77527	1,19641E-14	1	21526	953	4,4272	0,147119073	0,860359567	3547	159	4,48266	0,4125	0,6187
DNA/Kolobok-Hydra	3188449	2865	0,090	87159	409	0,469	1,000	0,000	36898	96	0,260	1,000	0,000	21526	37	0,172	1,000	0,000	3547	1	0,028	0,173	0,959
DNA/Maverick	3188449	8167	0,25614	87159	197	0,22602	0,038096883	0,967404124	36898	88	0,2385	0,270108809	0,763777093	21526	58	0,26944	0,681691697	0,342131725	3547	11	0,31012	0,795	0,304
DNA/Merlin	3188449	2046	0,064	87159	37	0,042	0,004	0,997	36898	13	0,035	0,012	0,994	21526	5	0,023	0,006	0,998	3547	1	0,028	0,336	0,898
DNA/MuLE-MuDR	3188449	917	0,02876	87159	54	0,06196	0,999999901	2,26243E-07	36898	17	0,04607	0,976801363	0,041941771	21526	10	0,04646	0,949612157	0,096943096	3547	0	0	0,3603	1
DNA/P	3188449	6726	0,211	87159	152	0,174	0,008	0,994	36898	60	0,163	0,021	0,985	21526	37	0,172	0,117	0,911	3547	4	0,113	0,133	0,940
DNA/PIF-ISL2EU	3188449	1382	0,04334	87159	28	0,03213	0,058006629	0,960216238	36898	5	0,01355	0,001326385	0,999619247	21526	1	0,00465	0,000831768	0,999914238	3547	0	0	0,2147	1
DNA/PIF-Spy	3188449	14582	0,457	87159	230	0,264	0,000	1,000	36898	132	0,358	0,002	0,999	21526	79	0,367	0,024	0,981	3547	9	0,254	0,038	0,981
DNA/PiggyBac	3188449	5290	0,16591	87159	122	0,13997	0,028633665	0,976725285	36898	51	0,13822	0,103149647	0,919289247	21526	33	0,1533	0,363907124	0,698595968	3547	9	0,25374	0,9241	0,1408
DNA/Sola	3188449	7995	0,251	87159	179	0,205	0,003	0,998	36898	65	0,176	0,002	0,999	21526	45	0,209	0,121	0,905	3547	6	0,169	0,216	0,878
DNA/Sola-2	3188449	2932	0,09196	87159	58	0,06655	0,00529747	0,996334276	36898	21	0,05691	0,011555206	0,993302615	21526	17	0,07897	0,312021657	0,766377268	3547	3	0,08458	0,5888	0,6331
DNA/TcMar-Fot1	3188449	3291	0,103	87159	107	0,123	0,967	0,041	36898	46	0,125	0,912	0,115	21526	29	0,135	0,934	0,094	3547	10	0,282	0,999	0,004
DNA/TcMar-m44	3188449	788	0,02471	87159	52	0,05966	0,999999994	1,06035E-08	36898	18	0,04878	0,997376117	0,005749807	21526	8	0,03716	0,9097387	0,168163487	3547	0	0	0,4159	1
DNA/TcMar-Mariner	3188449	20902	0,656	87159	1078	1,237	1,000	0,000	36898	335	0,908	1,000	0,000	21526	204	0,948	1,000	0,000	3547	35	0,987	0,992	0,013
DNA/TcMar-Pogo	3188449	53692	1,68395	87159	788	0,90409	1,05159E-87	0,999999995	36898	448	1,21416	7,45388E-14	1	21526	291	1,35185	4,89743E-05	0,999961355	3547	48	1,35326	0,0676	0,9488
DNA/TcMar-Tc1	3188449	139366	4,371	87159	5164	5,925	1,000	0,000	36898	1980	5,366	1,000	0,000	21526	1122	5,212	1,000	0,000	3547	157	4,426	0,585	0,447
DNA/TcMar-Tc2	3188449	1325	0,04156	87159	17	0,0195	0,000259446	0,999885386	36898	8	0,02168	0,030771792	0,985594283	21526	6	0,02787	0,210818091	0,881834158	3547	0	0	0,2287	1
DNA/TcMar-Tc4	3188449	21816	0,684	87159	659	0,756	0,995	0,005	36898	262	0,710	0,740	0,281	21526	155	0,720	0,755	0,272	3547	18	0,507	0,117	0,922
DNA/TcMar-Tigger	3188449	71621	2,24626	87159	1065	1,2219	5,364E-113	0,999999993	36898	634	1,71825	4,34921E-13	0,999999998	21526	423	1,96507	0,002387133	0,997948043	3547	70	1,9735	0,1486	0,8771
DNA/Zator	3188449	16524	0,518	87159	220	0,252	0,000	1,000	36898	128	0,347	0,000	1,000	21526	93	0,432	0,040	0,968	3547	16	0,451	0,341	0,743
DNA/Zisupton	3188449	1953	0,06125	87159	40	0,04589	0,032588999	0,977044849	36898	15	0,04065	0,059838573	0,963868058	21526	10	0,04646	0,235153731	0,847037246	3547	2	0,05639	0,63	0,639
LINE/CR1	3188449	3223	0,101	87159	164	0,188	1,000	0,000	36898	56	0,152	0,998	0,002	21526	23	0,107	0,657	0,423	3547	3	0,085	0,518	0,695
LINE/CR1-Zenon	3188449	378	0,01186	87159	45	0,05163	0,99999999	2,91168E-16	36898	17	0,04607	0,999999311	2,98011E-06	21526	5	0,02323	0,955009232	0,115184166	3547	0	0	0,6565	1
LINE/Dong-R4	3188449	21921	0,688	87159	4158	4,771	1,000	0,000	36898	980	2,656	1,000	0,000	21526	222	1,031	1,000	0,000	3547	20	0,564	0,218	0,840
LINE/I	3188449	34972	1,09683	87159	1779	2,041	0,999999998	6,05681E-131	36898	645	1,74806	1	5,08839E-29	21526	357	1,65846	1	8,79876E-14	3547	66	1,86073	1	4E-05
LINE/I-Nimb	3188449	2695	0,085	87159	499	0,573	1,000	0,000	36898	144	0,390	1,000	0,000	21526	58	0,269	1,000	0,000	3547	4	0,113	0,816	0,352
LINE/Jockey	3188449	5625	0,17642	87159	373	0,4275	0,999999997	1,42228E-52	36898	135	0,36587	1	1,54331E-14	21526	59	0,27409	0,999449947	0,000894702	3547	6	0,16916	0,5649	0,595
LINE/L1	3188449	1028	0,032	87159	42	0,048	0,995	0,008	36898	27	0,073	1,000	0,000	21526	15	0,070	0,998	0,005	3547	4	0,113	0,994	0,029
LINE/L1-Tx1	3188449	374	0,01173	87159	6	0,00688	0,113299027	0,943338937	36898	5	0,01355	0,732648589	0,435354133	21526	2	0,00929	0,536963787	0,718849153	3547	1	0,02819	0,9342	0,3405
LINE/L2	3188449	42322	1,327	87159	1520	1,744	1,000	0,000	36898	569	1,542	1,000	0,000	21526	322	1,496	0,985	0,018	3547	63	1,776	0,990	0,015
LINE/Penelope	3188449	62053	1,94618	87159	3029	3,47526	0,999999996	1,5711E-197	36898	1083	2,93512	1	3,07011E-38	21526	581	2,69906	0,999999938	1,8315E-14	3547	82	2,31181	0,9461	0,0676
LINE/R1	3188449	1976	0,062	87159	69	0,079	0,981	0,026	36898	28	0,076	0,880	0,164	21526	13	0,060	0,536	0,574	3547	1	0,028	0,355	0,889
LINE/R1-LOA	3188449	714	0,02239	87159	25	0,02868	0,911019808	0,128061051	36898	8	0,02168	0,555715353	0,584076758	21526	4	0,01858	0,472032971	0,709809782	3547	0	0	0,4517	1
LINE/RTE-BovB	3188449	20952	0,657	87159	707	0,811	1,000	0,000	36898	290	0,786	0,999	0,001	21526	176	0,818	0,998	0,003	3547	25	0,705	0,685	0,390
LINE/RTE-RTE	3188449	78216	2,45131	87159	19574	22,4578	0,999999998	0	36898	3810	10,3258	1	0	21526	816	3,79076	1	2,05391E-32	3547	82	2,31181	0,3167	0,7216
LINE/RTE-X	3188449	690	0,022	87159	14	0,016	0,153	0,899	36898	8	0,022	0,595	0,546	21526	6	0,028	0,811	0,324	3547	3	0,085	0,992	0,043
low_complexity	3188449	41405	1,29859	87159	422	0,48417	1,0413E-133	0,999999996	36898	314	0,85099	2,54411E-16	1	21526	229	1,06383	0,000941023	0,999244795	3547	54	1,52241	0,8926	0,1357
LTR	3188449	22969	0,720	87159	1288	1,478	1,000	0,000	36898	493	1,336	1,000	0,000	21526	238	1,106	1,000	0,000	3547	34	0,959	0,957	0,062
LTR/Copia	3188449	1775	0,05567	87159	77	0,08834	0,999953854	7,68684E-05	36898	22	0,05962	0,678688784	0,402337763	21526	11	0,0511	0,4630774	0,651798823	3547	0	0	0,1386	1
LTR/ERV1	3188449	211	0,007	87159	17	0,020	1,000	0,000	36898	6	0,016	0,988	0,037	21526	2	0,009	0,828	0,417	3547	0	0,000	0,791	1,000
LTR/ERVK	3188449	7990	0,25059	87159	147	0,16866	1,2467E-07	0,999999912	36898	81	0,21952	0,124158589	0,896844982	21526	60	0,27873	0,816427104	0,220694024	3547	6	0,16916	0,2166	0,8778
LTR/Gypsy	3188449	7031	0,221	87159	467	0,536	1,000	0,000	36898	184	0,499	1,000	0,000	21526	86	0,400	1,000	0,000	3547	16	0,451	0,997	0,007
LTR/Pao	3188449	3390	0,10632	87159	530	0,60808	0,999999996	1,59495E-227	36898	170	0,46073	1	8,17136E-55	21526	56	0,26015	0,999999996	3,04683E-09	3547	12	0,33831	0,9998	0,0005
RC/Helitron	3188449	2444	0,077	87159	66	0,076	0,493	0,557	36898	42	0,114	0,994	0,009	21526	34	0,158	1,000	0,000	3547	4	0,113	0,860	0,290
rRNA	3188449	851	0,02669	87159	261	0,2995	0,999999994	1,16291E-189	36898	92	0,24934	1	1,75145E-57	21526	56	0,26015	0,999999995	3,50901E-36	3547	0	0	0,3878	1
Satellite	3188449	4890	0,153	87159	166	0,190	0,997	0,003	36898	76	0,206	0,995	0,008	21526	39	0,181	0,870	0,168	3547	6	0,169	0,696	0,461
SINE?	3188449	12984	0,40722	87159	401	0,46008	0,993233303	0,007812063	36898	243	0,65857	0,999999999	1,46766E-12	21526	148	0,68754	0,999999993	2,17254E-09	3547	22	0,62024	0,9775	0,0379
SINE/ID	3188449	143670	4,506	87159	2754	3,160	0,000	1,000	36898	1326	3,594	0,000	1,000	21526	918	4,265	0,044	0,959	3547	162	4,567	0,590	0,441
SINE/tRNA-Core	3188449	5423	0,17008	87159	80	0,09179	3,68199E-10	1	36898	55	0,14906	0,178910356	0,853640582	21526	37	0,17189	0,569054079	0,496529315	3547	9	0,25374	0,9141	0,1559
SINE/tRNA-Deu	3188449	14305	0,449	87159	189	0,217	0,000	1,000	36898	118	0,320	0,000	1,000	21526	82	0,381	0,072	0,942	3547	16	0,451	0,575	0,525
SINE/tRNA-Deu-CR1	3188449	10858	0,34054	87159	128	0,14686	3,73048E-29	1	36898	78	0,21139	2,86752E-06	0,999998827	21526	53	0,24621	0,007713463	0,994654093	3547	9	0,25374	0,235	0,8507
Unknown	3188449	1840236	57,716	87159	31928	36,632	0,000	1,000	36898	18022	48,843	0,000	1,000	21526	11847	55,036	0,000	1,000	3547	2004	56,498	0,073	0,931

[a]Results generated using Bedtools (analysis of the intersection of the annotation by repeatmasker and the HCR annotations). The number of hits for each element could slightly differ from those results reported in Table S1-9a, as in this analysis we have not take into account overlapping annotations, or multiple annotations in the same region.

[b]Total number of repeat annotations for the whole genome (n = 3,188,449). This value is slightly different from those reported by RepeatMasker (n = 3,284,969) . In the current analysis we have not take into account overlapping annotations, or multiple annotations in the same region.

[c]Total number of hits for every type of repeat class for the whole genome

[d]Total number of annotation for this subset

[e]Total number of hit for every type of repeat class for this subset

[f]CDF = P-value of observing (obtaining by random) a number of repeat elements less than or equal to the observed

[g]SF = P-value of observing (obtaining by random) a number of repeat elements greater than or equal to the observed

Table S2-5: Enrichment analysis of the intersection of High Coverage regions (HCRs) (5 x mean coverage) with RepeatMasker annotation (subtypes of main repeats)[a]

				> 150 bp					> 500 bp					> 1000 bp					> 5000 bp				
	Total Genome[b]	Total Genome type[c]	%	Total Count Subset[d]	Total Subset Type[e]	%	CDF[f]	SF[g]	Total Count Subset	Total Subset Type	%	CDF	SF	Total Count Subset	Total Subset Type	%	CDF	SF	Total Count Subset	Total Subset Type	%	CDF	SF
DNA/CMC-Chapaev-3	3188449	5416	0,170	22089	28	0,127	0,065	0,955	8240	15	0,182	0,670	0,429	4106	10	0,244	0,904	0,167	392	0	0,000	0,514	1,000
DNA/CMC-EnSpm	3188449	10378	0,32549	22089	39	0,17656	1,47531E-05	0,999992252	8240	15	0,18204	0,009480618	0,995055159	4106	9	0,21919	0,14247753	0,916262668	392	0	0	0,2786	1
DNA/CMC-Transib	3188449	1359	0,043	22089	5	0,023	0,092	0,958	8240	1	0,012	0,134	0,970	4106	0	0,000	0,173	1,000	392	0	0,000	0,846	1,000
DNA/Crypton-H	3188449	1430	0,04485	22089	6	0,02716	0,135185158	0,930021458	8240	3	0,03641	0,494853581	0,714256235	4106	2	0,04871	0,71951717	0,549614786	392	0	0	0,8387	1
DNA/Ginger	3188449	9241	0,290	22089	19	0,086	0,000	1,000	8240	5	0,061	0,000	1,000	4106	1	0,024	0,000	1,000	392	1	0,255	0,686	0,679
DNA/hAT	3188449	364071	11,4184	22089	1483	6,71375	3,876E-124	0,999999998	8240	751	9,11408	6,25797E-12	0,999999996	4106	414	10,0828	0,003374031	0,997109321	392	45	11,4796	0,5547	0,5083
DNA/hAT-Ac	3188449	7724	0,242	22089	45	0,204	0,134	0,895	8240	17	0,206	0,299	0,777	4106	4	0,097	0,030	0,989	392	1	0,255	0,754	0,614
DNA/hAT-Blackjack	3188449	14738	0,46223	22089	55	0,24899	2,09331E-07	0,999999888	8240	23	0,27913	0,005793546	0,996694161	4106	11	0,2679	0,034625838	0,981733261	392	0	0	0,1626	1
DNA/hAT-Charlie	3188449	127020	3,984	22089	431	1,951	0,000	1,000	8240	228	2,767	0,000	1,000	4106	132	3,215	0,005	0,996	392	14	3,571	0,401	0,698
DNA/hAT-hAT19	3188449	3988	0,12508	22089	11	0,0498	0,000277036	0,999889861	8240	10	0,12136	0,544704156	0,580061989	4106	7	0,17048	0,852343985	0,257742234	392	0	0	0,6122	1
DNA/hAT-hAT5	3188449	21404	0,671	22089	78	0,353	0,000	1,000	8240	34	0,413	0,001	0,999	4106	16	0,390	0,012	0,993	392	0	0,000	0,071	1,000
DNA/hAT-hATm	3188449	10907	0,34208	22089	49	0,22183	0,000697413	0,999563784	8240	31	0,37621	0,740489516	0,32204425	4106	12	0,29226	0,353499367	0,744399973	392	1	0,2551	0,6122	0,739
DNA/hAT-Tag1	3188449	11250	0,353	22089	41	0,186	0,000	1,000	8240	17	0,206	0,011	0,994	4106	10	0,244	0,145	0,912	392	0	0,000	0,250	1,000
DNA/hAT-Tip100	3188449	145979	4,57837	22089	585	2,64838	2,31977E-50	1	8240	323	3,9199	0,001883805	0,998423 47	4106	190	4,62737	0,578693776	0,450616667	392	24	6,12245	0,938	0,0393 7
DNA/Kolobok-Hydra	3188449	2865	0,090	22089	65	0,294	1,000	0,000	8240	10	0,121	0,871	0,212	4106	2	0,049	0,287	0,883	392	0	0,000	0,703	1,000
DNA/Maverick	3188449	8167	0,25614	22089	39	0,17656	0,00844164	0,99445601	8240	31	0,37621	0,984094392	0,025277457	4106	17	0,41403	0,978053801	0,039798146	392	2	0,5102	0,9192	0,2657
DNA/Merlin	3188449	2046	0,064	22089	9	0,041	0,101	0,944	8240	5	0,061	0,566	0,609	4106	0	0,000	0,072	1,000	392	0	0,000	0,778	1,000
DNA/MuLE-MuDR	3188449	917	0,02876	22089	16	0,07243	0,999686234	0,000874339	8240	6	0,07282	0,989235786	0,033665936	4106	5	0,12177	0,998633181	0,007204291	392	0	0	0,8934	1
DNA/P	3188449	6726	0,211	22089	33	0,149	0,023	0,985	8240	14	0,170	0,251	0,824	4106	7	0,170	0,365	0,761	392	1	0,255	0,799	0,563
DNA/PIF-ISL2EU	3188449	1382	0,04334	22089	2	0,00905	0,001831823	0,999139343	8240	0	0	0,027961105	1	4106	0	0	0,168420964	1	392	0	0	0,8437	1
DNA/PIF-Spy	3188449	14582	0,457	22089	54	0,244	0,000	1,000	8240	22	0,267	0,004	0,998	4106	13	0,317	0,106	0,934	392	2	0,510	0,733	0,536
DNA/PiggyBac	3188449	5290	0,16591	22089	32	0,14487	0,250197404	0,801554873	8240	15	0,18204	0,701744266	0,394674516	4106	8	0,19484	0,753433854	0,373303939	392	1	0,2551	0,8614	0,4785
DNA/Sola	3188449	7995	0,251	22089	44	0,199	0,067	0,950	8240	29	0,352	0,969	0,048	4106	10	0,244	0,546	0,579	392	1	0,255	0,742	0,626
DNA/Sola-2	3188449	2932	0,09196	22089	19	0,08602	0,442293659	0,645257069	8240	11	0,1335	0,916520205	0,144103211	4106	3	0,07306	0,478282243	0,727484613	392	1	0,2551	0,9488	0,3028
DNA/TcMar-Fot1	3188449	3291	0,103	22089	17	0,077	0,130	0,913	8240	8	0,097	0,522	0,615	4106	6	0,1466	0,863	0,253	392	1	0,255	0,937	0,333
DNA/TcMar-m44	3188449	788	0,02471	22089	11	0,0498	0,989854067	0,023655609	8240	3	0,03641	0,850737962	0,333202454	4106	1	0,02435	0,730323097	0,637794974	392	0	0	0,9076	1
DNA/TcMar-Mariner	3188449	20902	0,656	22089	159	0,720	0,889	0,127	8240	60	0,728	0,814	0,224	4106	35	0,852	0,947	0,075	392	5	1,276	0,954	0,118
DNA/TcMar-Pogo	3188449	53692	1,68395	22089	170	0,76961	1,84998E-32	1	8240	95	1,15291	4,56463E-05	0,999969569	4106	53	1,29079	0,025230609	0,981612574	392	5	1,27551	0,3526	0,7898
DNA/TcMar-Tc1	3188449	139366	4,371	22089	970	4,391	0,567	0,446	8240	370	4,490	0,713	0,305	4106	204	4,968	0,970	0,035	392	17	4,337	0,551	0,548
DNA/TcMar-Tc2	3188449	1325	0,04156	22089	3	0,01358	0,018391335	0,994718361	8240	3	0,03641	0,552927719	0,665320127	4106	3	0,07306	0,906031693	0,244374443	392	0	0	0,8496	1
DNA/TcMar-Tc4	3188449	21816	0,684	22089	111	0,503	0,000	1,000	8240	47	0,570	0,115	0,910	4106	25	0,609	0,320	0,747	392	3	0,765	0,718	0,503
DNA/TcMar-Tigger	3188449	71621	2,24626	22089	309	1,39889	3,76277E-20	0,999999999	8240	158	1,91748	0,021868647	0,981911426	4106	93	2,26498	0,559650078	0,482048014	392	12	3,06122	0,892	0,1762
DNA/Zator	3188449	16524	0,518	22089	57	0,258	0,000	1,000	8240	37	0,449	0,215	0,829	4106	27	0,658	0,908	0,130	392	3	0,765	0,852	0,332
DNA/Zisupton	3188449	1953	0,06125	22089	7	0,03169	0,040281574	0,981410 49	8240	4	0,04854	0,431960574	0,741987542	4106	2	0,04871	0,539840548	0,716054095	392	0	0	0,7865	1
LINE/CR1	3188449	3223	0,101	22089	35	0,158	0,995	0,008	8240	10	0,121	0,782	0,325	4106	7	0,170	0,900	0,127	392	0	0,000	0,673	1,000
LINE/CR1-Zenon	3188449	378	0,01186	22089	11	0,0498	0,999982321	8,4576E-05	8240	4	0,04854	0,996740628	0,017443075	4106	1	0,02435	0,913886948	0,385607239	392	0	0	0,9546	1
LINE/Dong-R4	3188449	21921	0,688	22089	1111	5,030	1,000	0,000	8240	156	1,893	1,000	0,000	4106	35	0,852	0,912	0,120	392	0	0,000	0,067	1,000
LINE/I	3188449	34972	1,09683	22089	485	2,19566	1	6,56172E-44	8240	157	1,90534	1	9,97095E-11	4106	85	2,07014	0,999999968	5,81639E-08	392	9	2,29592	0,9876	0,0309
LINE/I-Nimb	3188449	2695	0,085	22089	154	0,697	1,000	0,000	8240	34	0,413	1,000	0,000	4106	14	0,341	1,000	0,000	392	1	0,255	0,956	0,282
LINE/Jockey	3188449	5625	0,17642	22089	109	0,49346	1	1,77939E-20	8240	33	0,40049	0,999991211	2,10756E-05	4106	12	0,29226	0,966228413	0,064840 11	392	1	0,2551	0,8472	0,4995
LINE/L1	3188449	1028	0,032	22089	15	0,068	0,997	0,006	8240	8	0,097	0,998	0,006	4106	6	0,146	1,000	0,002	392	0	0,000	0,881	1,000
LINE/L1-Tx1	3188449	374	0,01173	22089	2	0,00905	0,520164794	0,731964509	8240	2	0,02427	0,925997925	0,251916884	4106	2	0,04871	0,987043398	0,084557 58	392	0	0	0,9551	1
LINE/L2	3188449	42322	1,327	22089	261	1,182	0,029	0,975	8240	113	1,371	0,659	0,376	4106	62	1,510	0,862	0,169	392	7	1,786	0,846	0,267
LINE/Penelope	3188449	62053	1,94618	22089	654	2,96075	1	1,33016E-24	8240	220	2,6699	0,999997425	3,61558E-06	4106	113	2,75207	0,999835318	0,000241286	392	8	2,04082	0,6447	0,4951
LINE/R1	3188449	1976	0,062	22089	9	0,041	0,124	0,929	8240	4	0,049	0,422	0,750	4106	1	0,024	0,278	0,922	392	0	0,000	0,784	1,000
LINE/R1-LOA	3188449	714	0,02239	22089	3	0,01358	0,271620507	0,871683518	8240	2	0,02427	0,718545269	0,550825729	4106	0	0	0,398452994	1	392	0	0	0,9159	1
LINE/RTE-BovB	3188449	20952	0,657	22089	139	0,629	0,322	0,708	8240	55	0,667	0,582	0,472	4106	32	0,779	0,856	0,189	392	2	0,510	0,524	0,729
LINE/RTE-RTE	3188449	78216	2,4531	22089	5475	24,7861	1	0	8240	672	8,15534	1	1,2446E-155	4106	111	2,70336	0,861125693	0,161672004	392	14	3,57143	0,9374	0,1065
LINE/RTE-X	3188449	690	0,022	22089	4	0,018	0,479	0,704	8240	4	0,049	0,965	0,106	4106	1	0,024	0,777	0,589	392	1	0,255	0,997	0,081
Low_complexity	3188449	41405	1,29859	22089	123	0,55684	2,88353E-28	1	8240	81	0,98301	0,004986284	0,996356056	4106	47	1,14467	0,212384838	0,826067363	392	7	1,78571	0,8584	0,2503
LTR	3188449	22969	0,720	22089	255	1,154	1,000	0,000	8240	72	0,874	0,953	0,060	4106	35	0,852	0,862	0,180	392	0	0,000	0,059	1,000
LTR/Copia	3188449	1775	0,05567	22089	11	0,0498	0,427555552	0,683965191	8240	7	0,08495	0,906386933	0,179986021	4106	4	0,09742	0,918083181	0,197638272	392	0	0	0,8039	1
LTR/ERV1	3188449	211	0,007	22089	0	0,000	0,231	1,000	8240	0	0,000	0,579	1,000	4106	0	0,000	0,762	1,000	392	0	0,000	0,974	1,000
LTR/ERVK	3188449	7990	0,25079	22089	16	0,07243	4,12864E-10	1	8240	7	0,08495	0,000490313	0,999844602	4106	6	0,14613	0,112472716	0,943425566	392	0	0	0,374	1
LTR/Gypsy	3188449	7031	0,221	22089	113	0,512	1,000	0,000	8240	35	0,425	1,000	0,000	4106	13	0,317	0,924	0,128	392	0	0,000	0,421	1,000
LTR/Pao	3188449	3390	0,10632	22089	104	0,47082	1	9,51983E-35	8240	20	0,24272	0,999697603	0,000749463	4106	11	0,2679	0,998143111	0,005320013	392	0	0	0,659	1
RC/Helitron	3188449	2444	0,077	22089	11	0,050	0,086	0,950	8240	6	0,073	0,556	0,604	4106	5	0,122	0,901	0,210	392	0	0,000	0,740	1,000
rRNA	3188449	851	0,02669	22089	243	1,1001	0,999999999	2,6054E-308	8240	95	1,15291	0,999999997	1,1683E-119	4106	44	1,0716	0,999999992	1,98124E-54	392	0	0	0,9006	1
Satellite	3188449	4890	0,153	22089	107	0,484	1,000	0,000	8240	47	0,570	1,000	0,000	4106	19	0,463	1,000	0,000	392	1	0,255	0,878	0,452
SINE?	3188449	12984	0,40722	22089	182	0,82394	0,999999999	6,41903E-18	8240	89	1,0801	0,999999995	1,17641E-15	4106	40	0,97418	0,999999644	8,75723E-07	392	1	0,2551	0,5257	0,798
SINE/ID	3188449	143670	4,506	22089	656	2,970	0,000	1,000	8240	330	4,005	0,014	0,988	4106	180	4,384	0,371	0,658	392	17	4,337	0,499	0,598
SINE/tRNA-Core	3188449	5423	0,17008	22089	19	0,08602	0,000614837	0,999704458	8240	12	0,14563	0,356577486	0,741751591	4106	8	0,19484	0,731425548	0,398837982	392	1	0,2551	0,8558	0,4869
SINE/tRNA-Deu	3188449	14305	0,449	22089	58	0,263	0,000	1,000	8240	27	0,328	0,054	0,963	4106	16	0,390	0,338	0,746	392	2	0,510	0,742	0,525
SINE/tRNA-Deu-CR1	3188449	108158	3,40454	22089	33	0,1494	3,17019E-08	0,999999993	8240	13	0,15777	0,000121 2041	0,999471017	4106	7	0,17048	0,031635811	0,985784466	392	0	0	0,2626	1
Unknown	3188449	1840236	57,716	22089	8191	37,082	0,000	1,000	8240	4212	51,117	0,000	1,000	4106	2226	54,213	0,000	1,000	392	216	55,102	0,160	0,864

[a] Results generated using Bedtools (analysis of the intersection of the annotation by repeatmasker and the HCR annotations). The number of hits for each element could slightly differ from those results reported in Table S1-9a, as in this analysis we have not take into account overlapping annotations, or multiple annotations in the same region.

[b] Total number of repeat annotations for the whole genome (*n* = 3,188,449). This value is slightly different from those reported by RepeatMasker (*n* = 3,284,969) . In the current analysis we have not take into account overlapping annotations, or multiple annotations in the same region.

[c] Total number of hits for every type of repeat class for the whole genome

[d] Total number of annotation for this subset

[e] Total number of hit for every type of repeat class for this subset

[f] CDF = a *P*-value of observing (obtaining by random) a number of repeat elements less than or equal to the observed

[g] SF = a *P*-value of observing (obtaining by random) a number of repeat elements greater than or equal to the observed

Chapter 2

The chromosome-scale assembly of the Canary Islands endemic spider *Dysdera silvatica* (Arachnida, Araneae) sheds light on the origin and genome structure of chemoreceptor gene families in chelicerates

P Escuer, VA Pisarenco, AA Fernández-Ruiz, J Vizueta, JF Sánchez-Herrero, MA Arnedo, A Sánchez-Gracia, J Rozas

2022. Molecular Ecology Resources, 22:375–390. doi: 10.1111/1755-0998.13471

En aquest article presentem l'assemblatge genòmic a escala cromosòmica de *Dysdera silvatica* Schmidt, 1981, una aranya nocturna endèmica de les Illes Canàries. El gènere *Dysdera* ha experimentat una notable diversificació en aquest arxipèlag, majoritàriament associada a canvis en el nivell d'especialització tròfica, convertint-se en un excel·lent model per estudiar els les causes genòmiques de les radiacions adaptatives. El nou assemblatge (1,37 Gb; *scaffold* N50 de 174,2 Mb), realitzat mitjançant la tècnica de captura de la conformació cromosòmica, representa una millora de més de 4500 vegades en la continuïtat genòmica respecte a la versió anterior. El conjunt de *scaffolds* o pseudocromosomes més grans, que cobreixen el 87% de la mida total del genoma, probablement representen el set cromosomes del cariotip d'aquesta espècie, inclòs el gran i característic cromosoma X. Hem utilitzat aquest nou recurs per fer una anàlisi exhaustiva de les dues grans famílies de gens quimioreceptors d'artròpodes (els receptors gustatius i ionotròpics). Hem identificat 545 quimioreceptors distribuïts per tots els pseudocromosomes, amb una notable infrarepresentació al cromosoma X. Almenys el 54% d'ells es localitzen en 83 clústers genòmics, amb distàncies evolutives significativament menors entre ells que la mitjana de la família, cosa que suggereix un origen recent de molts d'ells. Aquest genoma a escala cromosòmica és el primer genoma d'alta qualitat representatiu del clade Synspermiata, i només el tercer entre les aranyes. Aquest nou recurs és molt valuós per obtenir informació sobre l'estructura i l'organització dels genomes de quelicerats, inclòs el paper que tenen les variants estructurals, elements repetitius i grans famílies de gens en l'extraordinària biologia de les aranyes.

Received: 12 April 2021 | Revised: 21 June 2021 | Accepted: 12 July 2021

DOI: 10.1111/1755-0998.13471

RESOURCE ARTICLE

MOLECULAR ECOLOGY RESOURCES WILEY

The chromosome-scale assembly of the Canary Islands endemic spider *Dysdera silvatica* (Arachnida, Araneae) sheds light on the origin and genome structure of chemoreceptor gene families in chelicerates

Paula Escuer[1] | **Vadim A. Pisarenco**[1] | **Angel A. Fernández-Ruiz**[1] | **Joel Vizueta**[1,2] | **Jose F. Sánchez-Herrero**[3] | **Miquel A. Arnedo**[4] | **Alejandro Sánchez-Gracia**[1] | **Julio Rozas**[1]

[1]Departament de Genètica, Microbiologia i Estadística & Institut de Recerca de la Biodiversitat (IRBio), Universitat de Barcelona, Barcelona, Spain

[2]Section for Ecology and Evolution, Department of Biology, Villum Centre for Biodiversity Genomics, University of Copenhagen, Copenhagen, Denmark

[3]High Content Genomics and Bioinformatics Unit, Institut d'Investigació en Ciències de la Salut Germans Trias i Pujol, Badalona, Spain

[4]Departament de Biologia Evolutiva, Ecologia i Ciències Ambientals & Institut de Recerca de la Biodiversitat (IRBio), Universitat de Barcelona, Barcelona, Spain

Correspondence
Julio Rozas and Alejandro Sánchez-Gracia, Departament de Genètica, Microbiologia i Estadística & Institut de Recerca de la Biodiversitat (IRBio), Universitat de Barcelona, Barcelona, Spain.
Emails: jrozas@ub.edu (J.R.); elsanchez@ub.edu (A.S.)

Funding information
Ministerio de Economía y Competitividad of Spain, Grant/Award Number: CGL2016-75255, CGL2016-80651, PID2019-103947GB, PID2019-105794GB and BES-2017-081740; Comissió Interdepartamental de Recerca I Innovació Tecnològica of Catalonia, Spain, Grant/Award Number: 2017SGR83 and 2017SGR1287

Abstract

Here, we present the chromosome-level genome assembly of *Dysdera silvatica* Schmidt, 1981, a nocturnal ground-dwelling spider endemic from the Canary Islands. The genus *Dysdera* has undergone a remarkable diversification in this archipelago mostly associated with shifts in the level of trophic specialization, becoming an excellent model to study the genomic drivers of adaptive radiations. The new assembly (1.37 Gb; scaffold N50 of 174.2 Mb), was performed using the chromosome conformation capture scaffolding technique, represents a continuity improvement of more than 4500 times with respect to the previous version. The seven largest scaffolds or pseudochromosomes, which cover 87% of the total assembly size, probably correspond with the seven chromosomes of the karyotype of this species, including a characteristic large X chromosome. To illustrate the value of this new resource we performed a comprehensive analysis of the two major arthropod chemoreceptor gene families (i.e., gustatory and ionotropic receptors). We identified 545 chemoreceptor sequences distributed across all pseudochromosomes, with a notable underrepresentation in the X chromosome. At least 54% of them localize in 83 genomic clusters with a significantly lower evolutionary distances between them than the average of the family, suggesting a recent origin of many of them. This chromosome-level assembly is the first high-quality genome representative of the Synspermiata clade, and just the third among spiders, representing a new valuable resource to gain insights into the structure and organization of chelicerate genomes, including the role that structural variants, repetitive elements and large gene families played in the extraordinary biology of spiders.

Paula Escuer and Vadim A. Pisarenco contributed equally.

KEYWORDS
Araneae, chemoreceptors, chromosome-level assembly, Dysdera, gene clusters, gene family, Hi-C

1 | INTRODUCTION

Arthropoda is the most species-rich phylum on Earth, virtually found in all ecosystems, and includes about 80% of known animal species (Zhang, 2011, 2013). It encompasses four extant subphyla, namely Chelicerata, Myriapoda, Crustacea (paraphyletic) and Hexapoda, and the extinct Trilobita. Among them, Chelicerata accounts for about 10% of arthropod species, most of which belong to the class Arachnida (about 130,000 species; Coddington et al., 2004; Garb et al., 2018; Sharma et al., 2014). Despite this great representation and special interest in many basic and translational research areas, such as material sciences (silk from spiders), bioactive natural compounds (venom toxins from spiders and scorpions), or pest control (mites, Acari), currently are few available genome sequences of species from this group (Garb et al., 2018; Thomas et al., 2020; Vizueta et al., 2018), being most of them highly fragmented and incomplete. In fact, only two chromosome-level spider genomes have been reported to date (Fan et al., 2021; Sheffer et al., 2021).

Spiders (order Araneae), the largest group within Arachnida, with at least ~49,000 known species (World Spider Catalogue, 2021), is a highly diverse group of predators that can be found in nearly all terrestrial ecosystems (Figure 1). Recent studies have greatly helped to elucidate their phylogeny and delimitate its main evolutionary lineages (Kallal et al., 2020; Wheeler et al., 2017). The nocturnal ground-dwelling genus *Dysdera* Latreille 1804, which contains 286 species (World Spider Catalogue, 2021), mostly with a circum-Mediterranean distribution, represents nearly half of the diversity of the Dysderidae family. Approximately 50 species of this genus are endemic from the Canary Islands archipelago, representing one of the most spectacular examples of diversification on islands within spiders (Arnedo et al., 2001, 2007; Macías-Hernández et al., 2016; Vizueta et al., 2019). Shifts in dietary preferences have been identified as one of the main drivers of island diversification in this group (Řezáč et al., 2021). Indeed, *Dysdera* includes some of the few reported cases of stenophagy (i.e., prey specialization) across the mostly generalists spiders (Pekár et al., 2016), with some species (both continental and island species) facultatively or even obligatorily specialized in feeding on terrestrial woodlice (Crustacea: Isopoda). This trophic specialization was accompanied by morphological (modifications of mouthparts), behavioural (unique hunting strategies) and physiological adaptations to capture woodlice and to assimilate the toxic substances and heavy metals accumulated in these usually rejected prey (Hopkin & Martin, 1985; Řezáč & Pekár, 2007; Řezáč et al., 2008; Toft & Macías-Hernández, 2017). In the Canary Islands, as in continental species, these diet shifts have occurred recurrently in different geographic areas.

The high rates of species proliferation coupled with multiple independent ecophenotypic shifts make *Dysdera* an excellent model for understanding the genomic basis of adaptive radiations (Vizueta et al., 2019). With the aim of obtaining a reference genome for this genus, we sequenced the genome of *Dysdera silvatica* Schmidt, 1981 (~1.37 Gb) and generated the first de novo genome assembly of this species using a hybrid strategy (Sánchez-Herrero et al., 2019). Nevertheless, most of the assembly was based on short reads, which, added to the high repetitive nature of the genome sequences of this species (53.8%), resulted in a very fragmented genome draft (N50 of 38 kb). While this first draft has been a fruitful research resource, it prevented the study of genomic aspects requiring greater continuity, such as gene mapping across the chromosomes, the comprehensive annotation of very long genes and gene clusters, or the identification of structural variation. These features are fundamental for understanding the biological and evolutionary meaning of the genome structure and gene organization. Some clear examples of the benefits of having a highly continuous chromosome-level assembly are the study of the genome structure and evolution of gene families, the impact of a number genome features (e.g., recombination, base content, distribution of genes and repetitive regions, etc) on adaptive processes, the analysis of impact of hybridization and divergence between populations, or the role of chromosomal evolution in speciation (Bleidorn, 2016; Pollard et al., 2018; Saha, 2019).

Here, we present the first high-quality chromosome-level assembly of the species *D. silvatica* (*D. silvatica* version 2.0). Using the first version of the genome assembly of this species (Sánchez-Herrero et al., 2019) as a starting point, we used proximity ligation libraries (Chicago and Hi-C libraries; Dovetail genomics), and the HiRise pipeline (Putnam et al., 2016) to obtain an improved, highly continuous assembly of this genome. As an example of the enhanced utility of the version 2.0, we have identified and annotated the members of the two major arthropod chemoreceptor gene families in this genome, and performed a comprehensive analysis of the physical clustering of all family members at chromosome-level scale. This new genomic resource will foster further studies of the molecular basis underlying rapid diversification in islands and ecological shifts by allowing comparative genomic analyses based on variation data that is inaccessible in currently available fragmented genomes, such as structural variants, repetitive elements, and large gene families. Additionally, this high-quality genomic data will contribute to improve our understanding of the structure, organization, and genome evolution in chelicerates.

2 | MATERIALS AND METHODS

2.1 | Sampling, DNA sequencing and genome assembly

The version 1 of *D. silvatica* was generated using a hybrid de novo genome assembly by combining information from five individuals and four types of sequencing libraries; see Table 1 in Sánchez-Herrero et al. (2019). We later identified that one these individuals was in fact

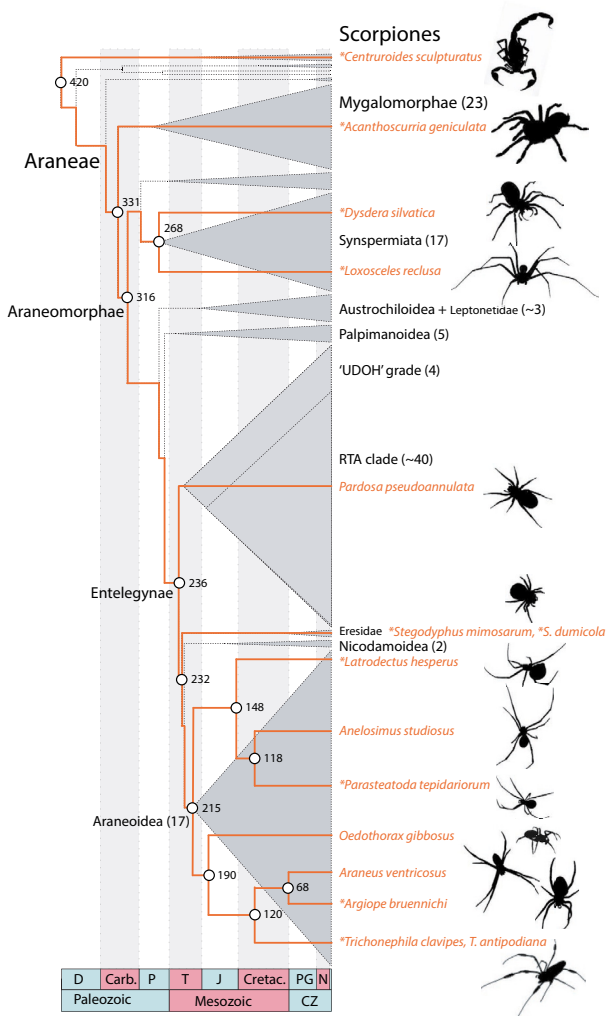


FIGURE 1 Tree of life of the order Araneae. Relationships and divergence time estimates followed (Kallal et al., 2020). Main evolutionary lineages and taxonomic groups within spiders indicated as grey clades; the number in brackets indicates the estimated number of included families. Phylogenetic relationships of species with genomic information are highlighted (orange lines). Species used for the annotation and homology-based searches of *Dysdera silvatica* genome are denoted with an asterisk

Dysdera enghoffi Arnedo, Oromí, & Ribera, 1997, a phylogenetically close relative to *D. silvatica* also endemic from La Gomera (Arnedo et al., 2007). The new genome assembly *D. silvatica* reported here (version 2) was obtained using the previous genome assembly, which was further scaffolded using Chicago and Hi-C libraries. We generated the latest sequencing libraries using a single female of the genus *Dysdera*, sampled in March 2012 in La Gomera (Canary Islands), identified in the laboratory as *D. silvatica*, frozen in nitrogen liquid and stored at −80°C until its use.

DNA extraction and sequencing were performed in Dovetail Genomics. The newly generated sequence data was obtained as a combination of Chicago (one library), Hi-C (one library) and Illumina 150-bp paired-end (HiSeq X ten platform; one lane) sequencing libraries. The final assembly (*D. silvatica* reference genome version 2)

TABLE 1 Genome assembly statistics

Assembly size (bp)	1,365,686,336
GC/N Content (%)	34.75/0.70
Number of scaffolds	15,360
Longest scaffold (Mb)	317.95
Scaffold N50 (Mb)/L50	174.19/4 scaffolds
Sequence in chromosomes (%)[a]	86.57
BUSCO analysis	
Arachnida data set	2532 (86.3%)
Arthropoda data set	818 (80.8%)
Eukaryota data set	237 (92.9%)

[a]Sequences in the seven described pseudochromosomes.

was generated using the HiRise scaffolding software (Putnam et al., 2016), and further polished with NextPolish (Hu et al., 2020) using Illumina short-read data from a single individual unequivocally identified as *D. silvatica* (see Section 3). We determined the completeness of the new assembly by applying the pipeline of Benchmarking Universal Single Copy Orthologues (BUSCO; v. 5.0.0; Seppey et al., 2019), searching the arachnida (odb10; 2934 genes), arthropoda (odb10; 1013 genes) and eukaryota (odb10; 255 genes) data sets.

2.2 | Repetitive elements identification

We analysed repetitive regions in the new assembly using a combination of a de novo transposable elements identification with RepeatModeler v1.0.11 (Smit & Hubley, 2008-2015; RepeatModeler), and a similarity-based search with RepeatMasker v.4.0.7 (Smit et al., 2013-2015, RepeatMasker). For the analysis we used the repetitive elements database built for the version 1 of the *D. silvatica* draft genome (see Sánchez-Herrero et al., 2019 for more details), generated with RepeatModeler, plus Dfam_Consensus-20171107 and RepBase-20181026 databases. The new assembly was masked for these repetitive sequences using the `(-xsmall)` option of RepeatMasker to be used in the genome annotation steps.

2.3 | Structural and functional genome annotation

We performed a completely new genome annotation on the *D. silvatica* v2 assembly. Structural annotation of the soft-masked genome was accomplished with BRAKER2 (Brůna et al., 2021; Hoff et al., 2019; Stanke et al., 2006, 2008), using both *D. silvatica* RNAseq data (Vizueta et al., 2017) and orthologous sequences from five species of this genus (including *D. silvatica*) (Vizueta et al., 2019) as evidence (`--etpmode`) (Hoff et al., 2019). To perform the functional annotation of the gene models, we used BLASTP v2.4 searches (*E*-value = 10^{-5}) against NCBI-nr, Swiss-Prot, and an updated version of the ArthropodDB databases (see Vizueta et al., 2017, for a detailed description of ArthropodDB). Furthermore, we also searched the predicted peptides for specific protein domain signatures in InterProScan v5.31.70.0 (Jones et al., 2014) and integrated all functional evidence.

Gene ontology (GO) terms (Ashburner et al., 2000) were obtained inherited from the BLAST (using top five significant hits) and InterProScan results. We used Blast2GO to associate KEGG enzymes and pathways to the annotated genes (Kanehisa & Goto, 2000). We detected the transfer RNA genes (tRNAs) encoded in the genomic sequence of *D. silvatica* using the tRNAscan-SE v2.0.7 software (Chan et al., 2019).

2.4 | Homology relationships with other taxa

We searched for homologues of *D. silvatica* in the genome data of representatives of a broad taxonomic range across the order Araneae, namely *Acanthoscurria geniculata* (Koch, 1841) (Theraphosidae), in the suborder Mygalomorphae, and the representatives of the suborder Araneomorphae *Loxosceles reclusa* Gertsch & Mulaik, 1940 (Sicariidae), together with *D. silvatica* member of the clade Synspermiata, *Stegodyphus mimosarum* Pavesi, 1883 and *Stegodyphus dumicola* Pocock, 1898 (Eresidae), and the Araneoidea *Parasteatoda tepidariorum* (Koch, 1841) and *Latrodectus hesperus* Chamberlin & Ivie, 1935, both in the family Theridiidae, and the Araneidae *Trichonephila clavipes* (Linnaeus, 1767) and *Argiope bruennichi* (Scopoli, 1772) (Figure 1; Table S1), as well as all other arachnids, arthropods and ecdysozoa species surveyed in Sánchez-Herrero et al. (2019). The search was conducted by a series of BLASTP searches (*E*-value cutoff $<10^{-3}$; we also applied a filter of >30% alignment length to consider a hit as a positive). Finally, we also searched for orthogroups and establish homology relationships among *D. silvatica* and the arachnids included in the OrthoDB (v10) (Kriventseva et al., 2019) database, namely the tick *Ixodes scapularis* Say, 1821 (Ullmann et al., 2005), the mite *Tetranychus urticae* Koch, 1836 (Grbić et al., 2011), and the spiders *Stegodyphus mimosarum* (Sanggaard et al., 2014) and *Parasteatoda tepidariorum* (Schwager et al., 2017).

2.5 | Annotation of the chemoreceptor genes

We performed a comprehensive curation of all members of the major chemoreceptor gene families encoded in the genome of *D. silvatica*, namely the Gustatory-receptor (*Gr*) and Ionotropic (glutamate) receptor (*Ir/iGluR*) families (Vizueta, Escuer, Frías-López, et al., 2020; Vizueta et al., 2018). For this task, we used the pipeline BITACORA (Vizueta et al., 2020; Vizueta et al., 2020), along with the homologous sequence data set and hidden Markov model (HMM) profiles used in Vizueta et al. (2018) and using the annotated gene models and genome sequence as input. The resulting identified proteins were validated, and reannotated when necessary, in the Apollo genome browser (Lee et al., 2013). We classified a gene as "complete" if the length of the encoded protein contains, at least, 80% of the protein domain length characteristic of the family (235 and 180 amino

acids, for the GR and IR/iGluR proteins, respectively). The remaining incomplete gene models that could not be recovered using Apollo were classified as "partial" fragments. For each chemoreceptor family, we computed the minimum number of chemoreceptor sequences that could be unequivocally attributed to different genes (S_{MIN}) as in Vizueta et al. (2018).

2.6 | Genomic clusters of chemosensory genes: Definition, chromosomal distribution and evolutionary distances

We determined whether the members of a given chemoreceptor gene family are physically closer (forming a cluster) in the pseudochromosomes than expected by chance by analysing the distribution of pairwise physical distances between the members of a particular gene family and scaffold. We classified the paralogous copies as "clustered" and "nonclustered". Operationally we consider that *n* closely linked genes from a gene family are clustered if they are arranged within a genomic region that spans less than certain cutoff C_L value following Vieira et al. (2007):

$$C_L = g(n-1)$$

where C_L is the maximum length of a cluster that contains two or more copies of the same family, and *g* is the maximum distance between two copies of a given family to consider that they are clustered. Here, we set the value of *g* to 100 kb. The gene density of the *Ir* family in the *D. silvatica* genome is about one copy every 3.32 Mb (and even lower in the *Gr* family). Assuming a uniform distribution of gene family members across the genome, the probability of finding by chance two (or more) *Ir* genes in a 100 kb stretch is $p = .0004$ (Poisson distribution, $\lambda = 0.0301$); this *p*-value is even lower for the *Gr* family. Thus, the selected *g* guarantees conservative C_L lengths for the two chemoreceptor families. Pairwise physical distances between gene family copies were processed with the R package ComplexHeatmap (Gu et al., 2016), and plotted as heatmaps to facilitate the visualization of gene clustering across scaffolds.

2.7 | Phylogenetic and evolutionary analyses

2.7.1 | Multiple sequence alignments

We used Mafft v. 7.475 (Katoh & Standley, 2013) to build three multiple sequence alignments (MSA) per each gene family. First, we generated a MSA (MSAc) with only complete sequences identified in *D. silvatica* using the L-INS-i algorithm with options `--localpair --maxiterate` 1000. Then, we built a second MSA (MSAf) for all sequences (the full data set, comprise the complete and partial copies identified in *D. silvatica*) using the Mafft `--addfragments` option (`--keeplength`) to align fragment (partial) sequences to the previous computed MSAc. Finally, we also used the L-INS-i algorithm to generate a third MSA for each family (MSAp), which included the complete sequences from *D. silvatica* and the curated members of the same family annotated in the genome of *Drosophila melanogaster* Meigen, 1830.

2.7.2 | Physical vs. evolutionary distances

We used the best-fit amino acid substitution model found by IQ-TREE software v. 2.1.2 (Minh et al., 2020) to estimate the evolutionary distances (measured as the number of amino acid replacements per amino acid site) across all pairwise comparisons. The analysis was performed with MEGA-CC 10.2.4 software (command-line version) (Kumar et al., 2012), using the JTT substitution model (Jones et al., 1992), with gamma-distributed heterogeneous rate variation among sites (5 and 7 discrete classes for the Gr and Ir families, respectively).

We investigated the relationship between physical and evolutionary distances by means of the C_{ST} statistic, a new index which measures the proportion of the evolutionary distance that is attributable to unclustered genes. We computed C_{ST} independently in the two chemoreceptor families, and separately for each scaffold (or for the whole genome). C_{ST} is estimated as:

$$C_{ST} = \frac{D_T - D_C}{D_T}$$

where D_T, the average of the pairwise evolutionary (amino acid replacements per site) distances between gene family copies, is estimated as:

$$D_T = \frac{2}{n(n-1)} \sum_{i<j} d_{ij}$$

where *n* is the number of gene family members in the surveyed scaffold (or in the entire genome), and d_{ij} is the evolutionary distance between sequences *i* and *j*.

And D_C, the average of the pairwise evolutionary distance between copies from within a cluster, averaged across all clusters of the same scaffold (or across the genome), is estimated as:

$$D_C = \frac{1}{m} \sum_{k=1}^{m} D_{Ck}$$

where *m* is the number of clusters (in a specific scaffold or in the entire genome), and D_{Ck} the average of pairwise evolutionary distances within the cluster *k*, is computed as:

$$D_{Ck} = \frac{2}{n(n-1)} \sum_{i<j} d_{ij}$$

where *n* is the number of copies in cluster *k*, and d_{ij} is the amino acid-based distance between sequences *i* and *j*.

We used the Mann–Whitney *U* test to determine whether the evolutionary distances between copies of the same family in

genomic clusters (in a particular scaffold) are significantly different from those estimated across unclustered genes.

2.7.3 | Phylogenetic analyses

We inferred the phylogenetic relationships among the members of the chemoreceptor gene families identified in the genome of *D. silvatica* by the maximum likelihood method implemented in IQ-TREE. We also included *Drosophila melanogaster* chemoreceptors in the analysis as evolutionary references (*D. melanogaster* is the best annotated genome so far for these families in Arthropods), and only complete proteins from both species (i.e., using MSAp; see above). For the analysis we used the best-fit amino acid substitution model found by IQ-TREE, and estimated node support from 1000 ultrafast bootstrap replicates. To test whether amino acid variation in the ligand-binding domain (LBD: PF00060; Mistry et al., 2021) can be used to distinguish between *iGluR* and *Ir* subfamilies, we built an additional MSA (MSAp2) including only the sequences encoding this domain in *D. silvatica* and in *D. melanogaster*. We used HMMSEARCH (Eddy, 2011) and in-house *Perl* scripts to identify and cut the sequences encoding this domain in all complete *Ir*/*iGluR* genes. We used the iTOL web server to visualize and manipulate the trees (Letunic & Bork, 2007).

3 | RESULTS AND DISCUSSION

3.1 | A new, high-quality genome assembly and annotation

The new genome of *D. silvatica*, which has an assembly size of 1.37 Gb (Table 1), shows a high completeness, with the detection of 86.3% and 92.9% of the BUSCO genes across the arachnida and eukaryota data sets, respectively, in their genome sequence (Table 1; Table S2). Despite having 15,360 scaffolds, the N50 and L50 values, 174.2 Mb and four scaffolds, respectively, also demonstrate the high continuity of the assembly. The seven largest scaffolds (or pseudochromosomes), including the larger scaffold that probably corresponds to the X chromosome (317.9 Mb long, nearly twice the size of the second largest scaffold), represent ~87% of total assembly size, matching perfectly with the haploid component of this species (6 autosomes and the X chromosome; Bellvert & Arnedo, unpublished data; Figure 2).

The structural annotation shows that the genome of *D. silvatica* encodes 33,275 coding-protein genes (35,370 transcripts) and includes the 90% and 95% of the BUSCO arthropoda and eukaryote data sets, respectively (Table 2). We have also identified 37,198 putative tRNAs applying the tRNAscan-SE with default parameters; this number drops to 1028 using the stringent EukHighConfidenceFilter. Sequence similarity-based searches uncovered 28,904 sequences with positive hits against the surveyed protein databases (16,241, 25,917 and 22,604 against Swiss-Prot, ArthropodDB and InterPro databases, respectively). Furthermore 22,093 of these functionally annotated sequences also have at least one associated GO term.

We identified ~3.2 millions of repetitive sequences, which encompass 53.0% of the total assembly size (Table 2; Tables S3). The great majority of these sequences (51.6%) correspond to transposable elements, many of them (22.1%) without detectable homologues in known databases; class II elements are the most abundant type (16.5%), followed by retrotransposons (class I), including LINEs (10.6%) and SINEs (1.8%).

The great majority of structurally annotated genes in the new genome assembly are shared across Arthropoda (63.7%), being 36.3% of them also present in Ecdysozoa (Figure S1; Table S1). Besides, 25.0% of these genes are spider-specific (order Araneae), and 17.4% were identified as lineage-specific in *D. silvatica*. When considering only functionally annotated genes (n = 28,904), this analysis yields

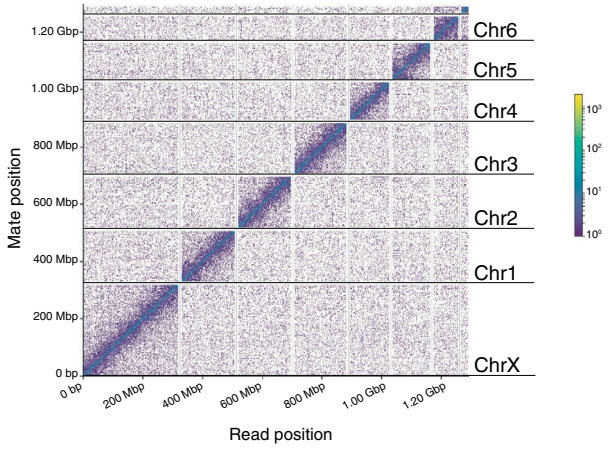


FIGURE 2 Long-range contact heatmap of paired-end Hi-C reads. The *x* and *y* axes show the mapping positions of the first and second read in the read pair, respectively, grouped into bins; the colour of each square gives the number of read pairs within that bin. The seven largest scaffolds, which probably correspond to the seven pseudochromosomes described in this species, represent the ~87% of total assembly. The large scaffold, which corresponds to the X chromosome, is 317.9 Mb long. In order of length, and after chromosome 6, the next scaffold is the ChrU1 (22.3 Mb long)

equivalent results (Figure 3a), although a slightly higher fraction of genes shared within Arthropoda (72.8%) and less *D. silvatica* lineage-specific genes were detected (2200 genes: 7.6%). Remarkably, the number of lineage-specific genes is nearly the half of those initially reported by Sánchez-Herrero et al. (2019); this feature could be partly explained by the fact that here we have used a broader Araneae data set for the searches, although the higher quality of the new assembly would have allowed to identify much more proteins accurately annotated. Homology-search results based on OrthoDB (Figure 3b) were similar to those obtained in Sánchez-Herrero et al. (2019).

TABLE 2 Genome annotation statistics

Repetitive regions (%)	52.95
Transposable elements (bp)[a]	704,860,272
Number of elements	2,829,393
Other repetitive elements (bp)	18,316,112
Number of elements	336,883
Genome annotation	
Protein-coding genes	33,275
Number of transcripts	35,370
Coding sequence (%)[b]	2.58
Functionally annotated	28,904
BUSCO analysis	
Arachnida data set	2539 (86.5%)
Arthropoda data set	912 (90.0%)
Eukaryota data set	241 (94.5%)

[a]Transposable elements (LINEs, SINEs, LTRs and other DNA elements).
[b]Fraction of the genome representing protein-coding regions.

Globally, the new chromosome-scale assembly of *D. silvatica* represents a huge improvement compared with our previous draft assembly. In terms of continuity, it implies an improvement of more than 4500 times (the N50 value) yielding a scaffold N50 of 174.2 Mb from the 38 kb in the previous assembly. This improvement is also reflected in the high number of annotated genes (Table 2, Table S2), despite the number the gene models in current version drops from 48,619 (75% of them with functional annotation) to 33,275 (87% with functional annotation). On the other hand, the new reference sequence encompass sequence data exclusively from *D. silvatica*, while that reported in version 1 was generated using information from various individuals one of them now identified as *D. enghoffi* Arnedo et al., 1997, a phylogenetically close relative to *D. silvatica* also endemic from La Gomera (Arnedo et al., 2007; see also Adrián-Serrano et al., 2021 for the new mtDNA data). The new chromosome-level assembly of *D. silvatica* is the first highly continuous genome of a representative of the spider clade Synspermiata, which currently includes 17 families, and the third within the Araneae order (the other two are members of the superfamily Araneoidea), an extremely poor and biased genomic representation of the taxonomic and evolutionary diversity of the spider tree of life (Figure 1). Our assembly, therefore, represents a valuable resource to further conduct molecular evolutionary and functional studies in spiders and their relatives.

3.2 | The chemoreceptor repertoire of *D. silvatica*

The chemosensory repertoire of chelicerates, particularly of spiders, is characterized by a high diversity in terms of copy number and sequence divergence, probably resulted from a constant and prolonged

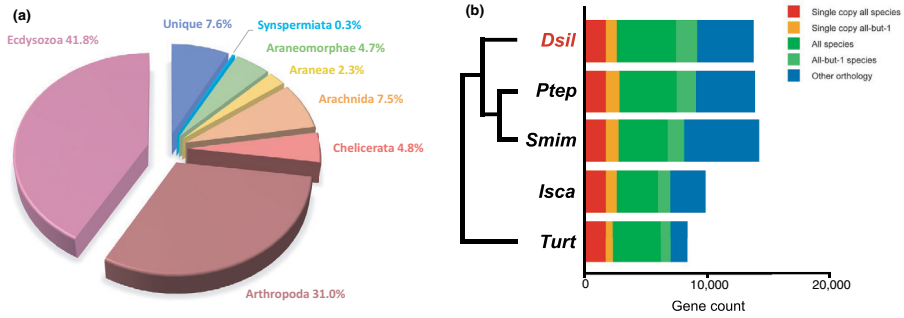


FIGURE 3 Homology-based search across different ecdysozoa species. (a) Pie chart showing the taxonomic distribution of positive BLAST hits of the functional annotation set of *Dysdera silvatica* (n = 28,904 genes) across the Araneae species illustrated in Figure 1 and the rest of the arachnida, arthropoda and ecdysozoa also analysed in Sánchez-Herrero et al. (2019). (b) Homology relationships across *D. silvatica* (Dsil) and chelicerates genomes available in OrthoDB v10, *Parasteatoda tepidariorum* (Ptep), *Stegodyphus mimosarum* (Smim), *Ixodes scapularis* (Isca), and *Tetranychus urticae* (Turt). Red and orange bars indicate the fraction of single-copy genes identified in all species (1:1 orthologues), and those identified in four species (missing in one species), respectively. The dark and light green bars show orthologous relationships present in all, or in four species, respectively, that are not included in the two previous categories. The blue bar shows other more complex homologous relationships

gene birth and death process, only altered by episodic bursts of gene duplication yielding lineage-specific expansions (Vizueta et al., 2018). Here, our comprehensive analysis using the BITACORA pipeline allowed the identification in the genome of *D. silvatica* of 545 chemoreceptor genes, corresponding to 134 *Gr* and 411 *Ir* (plus one homologue to the insect *Ir25a* and 24 *iGluR*) sequences (Figure 4; Tables 3, S4 and S5). Although these family sizes are in agreement with the variability observed across spider genomes, the 411 *Ir* reported here represents the largest repertoire found for this family in chelicerates, only surpassed by the extraordinary chemoreceptor repertoire found in the German cockroach genome (Robertson et al., 2018). On the other hand, the number of Grs in *D. silvatica* is relatively low compared to other spiders (i.e., S_{MIN} of 634 Grs in *Parasteatoda tepidariorum*), which could suggest a different contribution of the gustatory repertories to the evolutionary, and probably adaptive history of these different spider lineages.

Noticeably, most identified chemoreceptor genes are complete (459 complete genes under the criterion considered in this article; see methods), yielding a S_{MIN} of 540 genes (Table 3). This large proportion of complete copies is highly unusual in reports on these chemoreceptor families across arthropod genomes (Vizueta et al., 2018), clearly demonstrating the benefits of having a chromosome-level assembly to annotate, characterize and study multigene families in animal genomes. The new high-quality assembly has uncovered, for example, that the two studied gene families are unevenly distributed across pseudochromosomes. The X chromosome, representing the 23.3% of the genome assembly, harbors only 3.0% and 3.2% of the *Gr* and *Ir* identified copies, respectively (Table 3); the members of these families, nevertheless, are more uniformly distributed across the other major pseudochromosomes. Even so, the number of chemoreceptors located in the small scaffolds, which only represent 11.8% of the total assembly, represents 23.1% and 59.6% of the *Gr* and *Ir* repertories, respectively. This uneven distribution, however, is not observed when considering all genes identified in the genome (Table 3). While the first feature probably hides some (unknown) particularities of the sex chromosome, the high representation of gene family members across the minor scaffolds could be explained by the assembly difficulties of these repetitive regions. Further studies including gene families with different family sizes will get insights into this genomic feature.

3.3 | Chemoreceptor genes are unevenly distributed across the genome of *D. silvatica*

Despite that not all chemoreceptor genes could be mapped on the scaffolds corresponding to the main cytological described chromosomes, the new high-quality assembly allowed us to study the genomic organization and evolution of a great number of paralogous copies of each family. According to our criterion (see Methods), we identified 83 genomic clusters, 17 and 66 of them including *Gr* and *Ir* genes, respectively (Figures 5a,c, S2A and S2C). These clusters, which harbour up to 10 copies of the same family, were found in all major scaffolds of *D. silvatica*.

To gain insights into the evolutionary meaning of such gene clustering structure we investigated the relationship between pairwise evolutionary divergences, measured as d_{ij} (the number of amino acid substitutions per site between two sequences), and physical distances (in kb). We found that C_{ST} values are high in all pseudochromosomes, ranging from 0.418 to 0.982 and from 0.428 to 0.894 for the Gr and Ir gene families, respectively, considering all identified sequences (Table 3); these values are similar when using only the complete data set (Table S4). These high C_{ST} values translate into statistically lower evolutionary distances among family copies included in clusters than those dispersed along the genome, both at the chromosome (Mann–Whitney *U* test, *p*-values < .05 for nearly all pseudochromosomes) but also at the whole genome levels (*p*-values < .001 in all cases) (Tables 3 and S4; Figures 5 and S2). This result, jointly with the large number of genomic clusters found across the *D. silvatica* genome, point to the recent origin of many of the chemoreceptors in this species, and to the unequal crossing-over as a major mechanism accounting for this origin. After gene duplication, the paralogues that are retained long enough (i.e., those that are not lost by genetic drift or purifying selection), continuously diverged at the sequence, and probably at the functional level (at least in terms of ligand specificity or signalling characteristics). We expect, therefore, that over time, evolutionary distances of these retained copies increase with physical distance, just as we have found (Figure 5b,d, Figures S2B,D). This genomic architecture could have relevant functional and evolutionary implications. For instance, the presence of distantly related family members within the same genomic cluster, could be the hallmark of the interaction between functional and gene regulation constraints preventing cluster breaking. A more comprehensive analysis of these specific cases deserves to be further evaluated.

3.4 | Phylogenetic analysis of the chemoreceptor genes in arthropods

As commented above, having a very continuous assembly opens the door of annotating as "complete genes" most copies of a medium to large-sized multigene families, almost outside of the scope in most of the available, highly fragmented non-model chelicerates genomes. These improved annotations (with the inclusion of more and longer family copies in the multiple sequence alignments) in turn, yield to much more accurate phylogenetic analyses, increasing the evolutionary signal, and improving the tree node support. In many cases, furthermore, these new complete copies could add very valuable information about, for instance, recent bursts of duplication and gene retention.

Current phylogenetic analysis of the *Gr* and *Ir* families, which are based on the high-quality annotations from the new assembly of *D. silvatica*, are clear examples of these benefits. Our analysis undoubtedly reflects the high gene turnover rates of these families in chelicerates (and, in general, in panarthropods, Vizueta, Escuer, Frías-López, et al., 2020; Vizueta et al., 2018). However, after

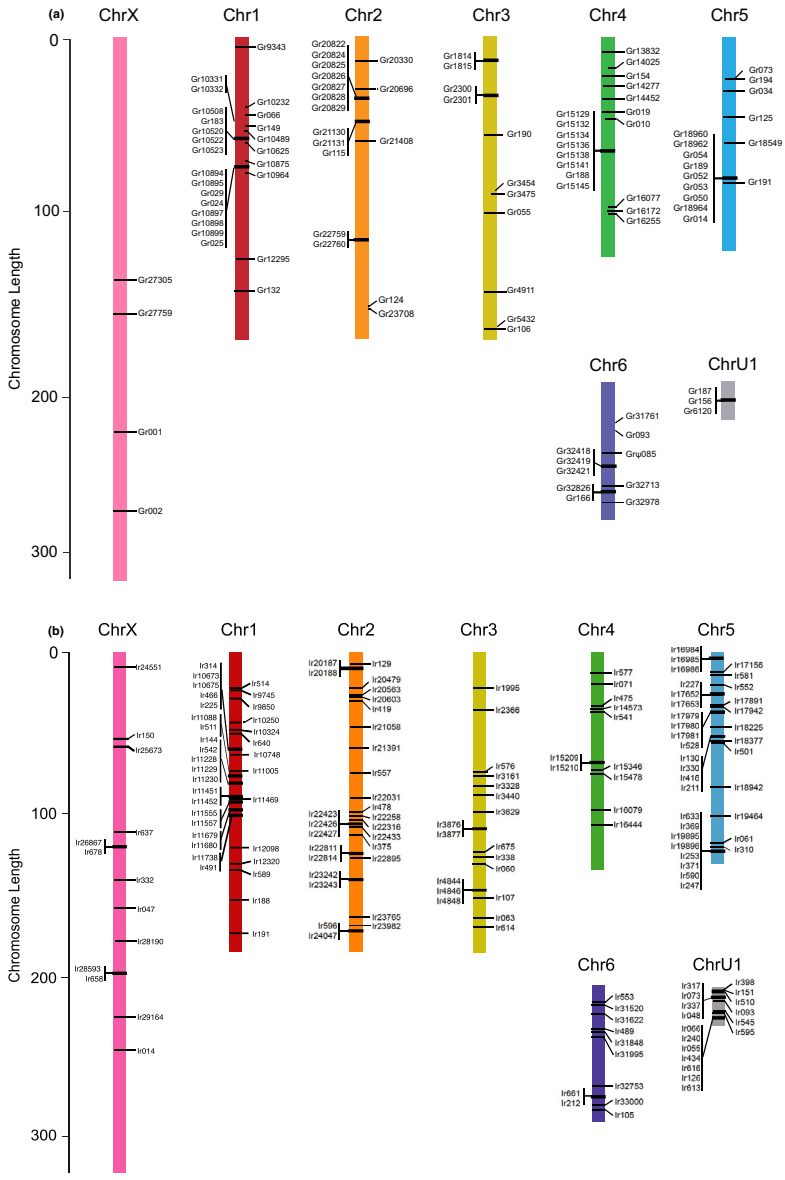


FIGURE 4 Distribution of the *Gr* (panel a) and *Ir* (panel b) family members across the seven pseudochromosomes and the scaffold ChrU1 of *D. silvatica*. Genes in clusters are shown to the left of pseudochromosomes

TABLE 3 Chromosome level statistics

	ChrX	Chr1	Chr2	Chr3	Chr4	Chr5	Chr6	ChrU1	Minor scaffolds	Total
Length (Mb)	317.95	177.17	176.73	174.24	129.24	125.97	80.96	22.26	161.17	1365.69
Genome fraction	23.28%	12.97%	12.94%	12.76%	9.46%	9.22%	5.93%	1.63%	11.80%	100%
Total number of genes	6812	4139	4227	4005	3226	3018	1904	513	5431	33,275
Gene density[a]	20.47%	12.44%	12.70%	12.04%	9.69%	9.07%	5.72%	1.54%	16.32%	100.00%
Total number of chemoreceptors[b]	26 (6)	62 (12)	46 (5)	32 (2)	32 (13)	50 (8)	25 (3)	20 (8)	277 (31)	570 (88)
Chromosomal gene density[c]	4.56%	10.88%	8.07%	5.61%	5.61%	8.77%	4.39%	3.51%	48.60%	100%
Gr	4 (1)	25 (10)	17 (3)	11 (0)	18 (9)	15 (4)	10 (2[d])	3 (1)	31 (6)	134 (36[d])
IR	13 (3)	34 (2)	28 (2)	18 (2)	11 (4)	34 (4)	11 (1)	17 (7)	245 (25)	411 (50)
iGluR	9 (2)	3 (0)	1 (0)	3 (0)	3 (0)	1 (0)	4 (0)	0 (0)	1 (0)	25 (2)
Clusters of *Gr* genes	0	3	3	2	1	1	2	1	4	17
Analysed[e]	0 (0, 4)	3 (15, 25)	3 (12, 17)	2 (4, 11)	1 (8, 18)	1 (9, 15)	2 (5, 9)	1 (3, 3)	4 (17, 30)	17 (73, 132)[f]
C_{ST}	na	0.757	0.844	0.982	0.418	0.462	0.616	na	0.704	0.769
Mann-Whitney test; *p*-value		***	***	*	***	***	*		***	***
Clusters of *Ir* genes	2	7	5	2	1	5	1	2	41	66
Analysed	0 (0, 10)	7 (19, 32)	5 (11, 28)	2 (5, 16)	1 (2, 9)	5 (21, 32)	1 (2, 11)	2 (8, 13)	34 (83, 229)	57 (151, 380)[g]
C_{ST}	na	0.5790	0.428	0.847	0.894	0.687	0.690	0.530	0.679	0.670
Mann-Whitney test; *p*-value		***	***	***	ns	***	ns	***	***	***

Note: na, not applicable; *, $p < .05$; **, $p < .01$; ***, $p < .001$; ns, not significant.

[a]Fraction of genes in the chromosome.

[b]Gr, Ir and iGluR genes identified in *Dysdera silvatica* genome; in parenthesis the number of incomplete genes.

[c]Fraction of chemoreceptor genes in the chromosome.

[d]One incomplete copy is a pseudogene.

[e]Number of clusters analysed (some short incomplete genes were excluded from the analysis). The values in parenthesis indicate the number of family members in clusters and the total number of family members in the scaffold.

[f]Two partial *Gr* copies were excluded from the analyses.

[g]Thirty-one partial *Ir* copies were excluded from the analyses.

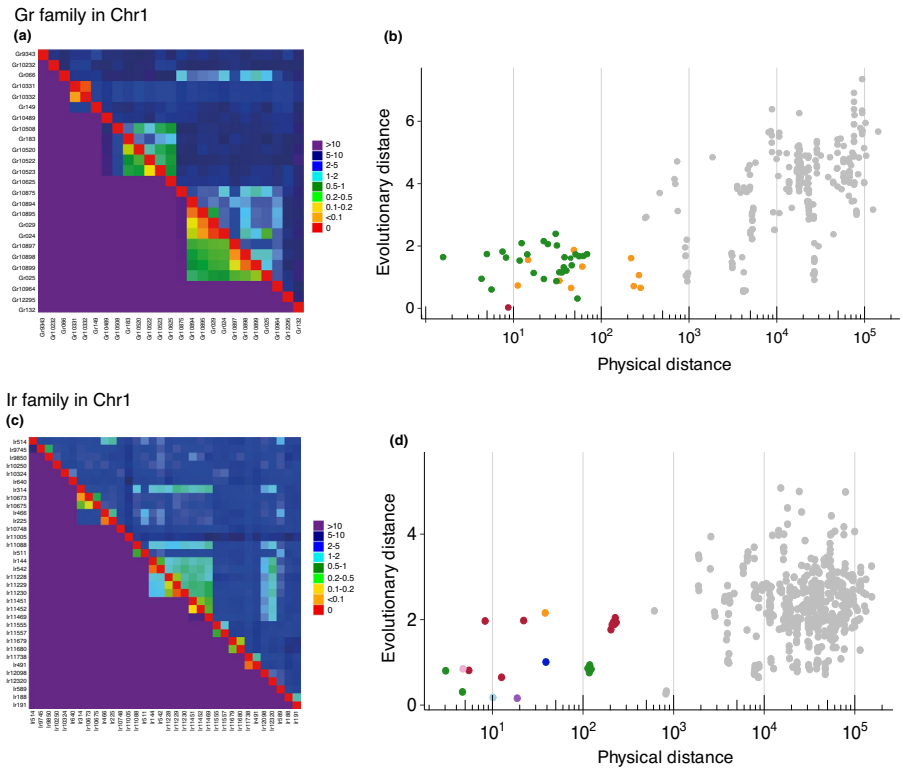


FIGURE 5 Genome organization and relationships between physical and evolutionary distances of the members of the *Gr* and *Ir* families on the *Dysdera silvatica* pseudochromosome 1. (a and c) Heatmaps illustrating the distribution of physical (below the diagonal; in units of 100 kb) and evolutionary (above the diagonal; in units of amino acid substitutions per site) distances, along the pseudochromosome. (b and d) Plots comparing pairwise amino acid and physical (on a logarithmic scale) distances between *Gr* or *Ir* copies in the *D. silvatica* pseudochromosome. Coloured and grey points show distances within and outside genomic clusters, respectively. Different clusters are shown in different colours

including the complete chemoreceptor set, a remarkable evolutionary hallmark emerges in the *D. silvatica* lineage (Figures 6, 7, S3 and S4). Only a small group of *Ir* genes, probably involved in some essential animal chemoreception functions, such as coreceptors (*Ir25a/8a* related sequences), and the receptors involved in thermosensation and hygrosensation, and in amino acid taste in *Drosophila* (i.e., *Ir93a* and *Ir76b* related sequences) (Ni, 2021), seem to be fairly conserved between insects and spiders. Although this extreme feature is well known in arthropods, where most family copies cluster in species-specific clades in the phylogenetic trees, current analysis is the first that incorporate nearly complete information of most copies of these two families in a chelicerate. The quality of the data allowed us to explore the origin and diversification trends of *D. silvatica* chemoreceptors with unprecedented precision and robustness. We found, for instance, that the distribution of gene ages in the *Gr* family is similar in *D. silvatica* and *D. melanogaster*, with most family members being old (probably during the early diversification of these subphyla). In the *Gr* family of *D. silvatica*, however, we uncovered very recent duplication events that created (at least) one new genomic cluster in a very short period in the scaffold U29 (with at least 10 genes in the cluster).

The contrasting pattern between *D. silvatica* and *D. melanogaster* is much more pronounced in the *Ir* family. Particularly noteworthy is the presence of two very recent bursts of gene duplication that originated 116 new *Ir* genes (83 and 33 copies, respectively; Figures 7 and S4A). Interestingly, most of these novel nearly identical

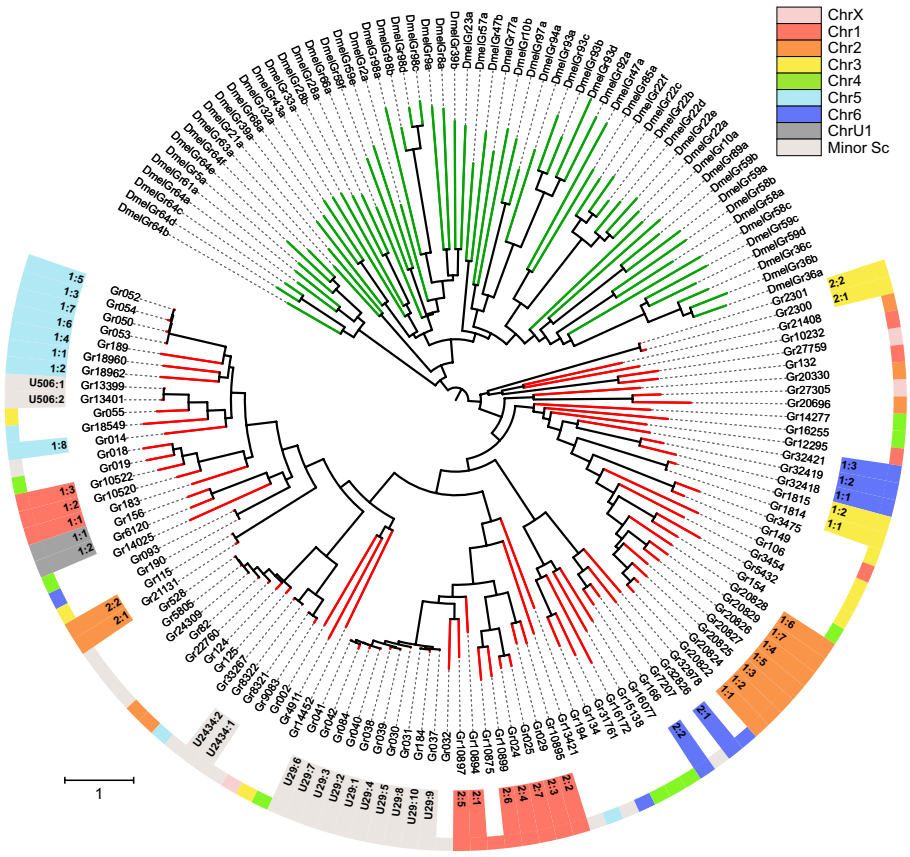


FIGURE 6 Phylogenetic relationships among the members of the *Gr* family of *Dysdera silvatica* and *Drosophila melanogaster*. The tree only includes the copies of *D. silvatica* classified as complete genes. The outer ring indicates the chromosome (Chr; in different colours) in which the genes included in the tree are located. The inner ring shows information about genomic clusters (chromosome, in the same color scale that in outer ring, genomic cluster number: member number in the cluster). The scale bar refers to 1 amino acid substitution per site. The tree was rooted in its midpoint. Minor SC, minor scaffolds. Red and green terminal branches correspond to *D. silvatica* and *D. melanogaster Gr*, respectively. Given that the tree in this figure only includes complete copies, some genomic clusters in Table 3 and Figure 4a do not appear here

receptors map in multiple clusters in many of the smallest scaffolds; this feature suggest that they could be indeed part of much large genomic clusters, which were not well assembled due to the high number of repetitive *Ir* copies arranged in tandem in the same genomic region (Clifton et al., 2020). Such duplication burst generating many new chemoreceptor genes could reflect some relevant evolutionary events related to the chemosensory biology of these organisms and, consequently, they deserve to be investigated more in depth, especially in relation to the role of selective and nonselective forces in their retention and divergence. In fact, since current available chelicerate genomes do not allow detecting such copies accurately or they are just annotated as different partial sequences (Vizueta, Escuer, Frías-López, et al., 2020; Vizueta et al., 2018), they have been scarcely added to phylogenetic analyses, thus preventing the precise understanding of the relevant events that shaped the repertoire size of large gene families at a very short time scale. Our results demonstrate that new high-quality data are especially useful to conduct comprehensive studies of the evolution of large multigene families.

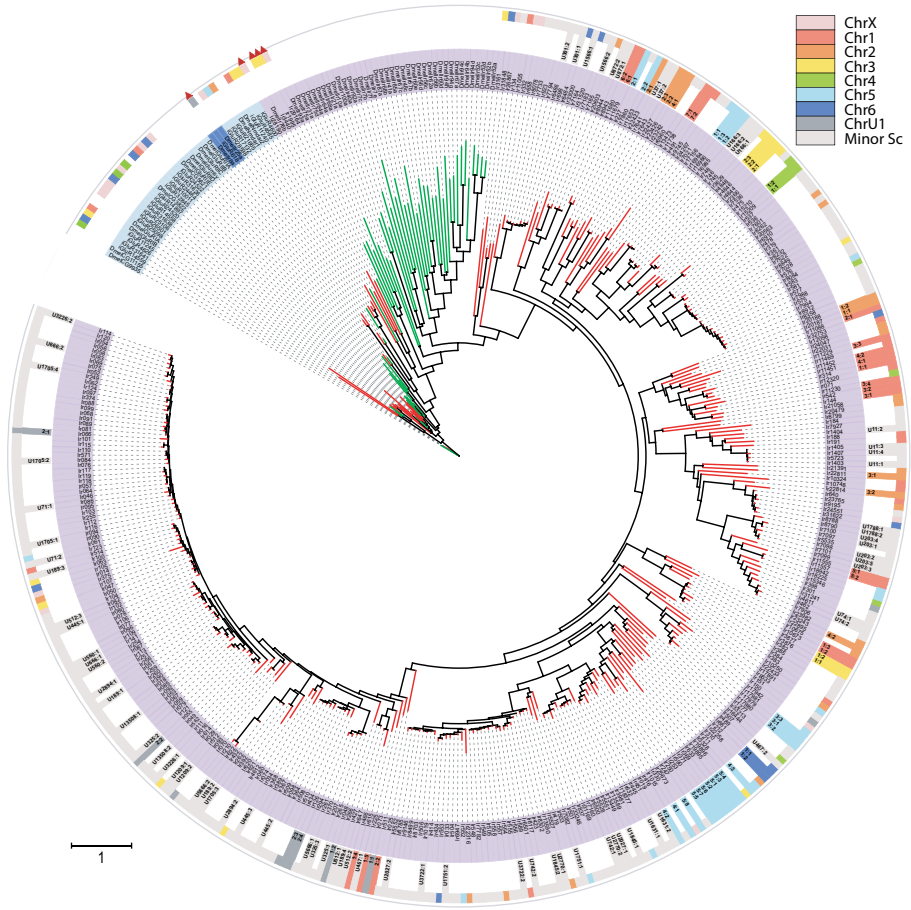


FIGURE 7 Phylogenetic relationships among the members of the *Ir*/*IGluR* family of *Dysdera silvatica* and *Drosophila melanogaster*. The tree only includes the copies classified as complete genes in this work. The light blue, dark blue and purpura shading of gene names designate the members of the *iGluR*, *Ir25a/8a* and *Ir* subfamilies, respectively. The tree was rooted considering NMDAR clade as the outgroup (Croset et al., 2010). The outer ring indicates the chromosome (Chr) in which the genes included in the tree are located. The inner ring shows information about genomic clusters (chromosome, in the same colour scale that in outer ring, genomic cluster number: member number in the cluster). The scale bar refers to one amino acid substitution per site. Minor SC, minor scaffolds. Red and green terminal branches correspond to *D. silvatica* and *D. melanogaster*, respectively. The red triangles mark the *D. silvatica* genes with putative distant homologues in *D. melanogaster*. Given that the tree in this figure only includes complete copies, some genomic clusters in Table 3 and Figure 4b do not appear here

We have also used our new assembly to test whether the LBD domain (PF00060) has enough phylogenetic signal to classify the different subfamilies within the *Ir/GluR* superfamily (a strategy that we previously used in fragmented assemblies, for example, Vizueta et al., 2018). Here, we take advantage that the high continuity of the assembly permitted the complete annotation of many *iGluR* genes, with both the ANF-receptor (PF01094) and the LBD domains in the same gene model. The latter combination, that is characteristic of the *iGluR* subfamily, is never found in *Ir* genes, which lack the ANF-receptor domain (Croset et al., 2010). This genomic structure makes it possible to unequivocally distinguish *iGluR* from

Ir genes. Our phylogenetic trees based on the complete sequences of this superfamily was fully consistent with those built using only the LBD domains identified in *D. silvatica* (Figure S4C). This feature demonstrates that this domain, by itself holds enough subfamily-specific information to place correctly the proteins having the ANF domain in the phylogenetic tree (i.e., close to the *D. melanogaster iGluRs* and separated from the *Ir* sequences of both species). In fact, the information of the LBD domain allowed us to classify correctly as *iGluR* some copies of this superfamily for which we were not able to identify an ANF domain in the genomic sequences (and that, in principle, would have been annotated as *Ir*).

4 | CONCLUSIONS

The chromosome-level assembly of *D. silvatica* is the first high-quality continuous genome of a representative of the Synspermiata clade, one of the major evolutionary lineages within spiders. This new assembly will contribute to alleviate the scarce representation of spiders and chelicerates genomes within the tree of life while represents a very useful resource to rightly characterize structural variants, repetitive elements and large gene families involved in relevant biological functions in spiders, such as, for example, those encoding chemosensory system proteins and venom components. An immediate application will be the comprehensive evolutionary analysis of these genomic variants beyond single nucleotide changes to elucidate the genomic regions and the mechanisms underlying the remarkable adaptive radiation of the genus *Dysdera* in Canary Islands.

ACKNOWLEDGEMENTS

This work was supported by the Ministerio de Economía y Competitividad of Spain (CGL2016-75255; CGL2016-80651; PID2019-103947GB, PID2019-105794GB) and the Comissió Interdepartamental de Recerca I Innovació Tecnològica of Catalonia, Spain (2017SGR83, 2017SGR1287). A.S.-G. is a Serra Húnter Fellow. P. E. was supported by an FPI grant (Ministerio de Economía y Competitividad of Spain, BES-2017-081740). We acknowledge the Garajonay National Parks for granting collection permits and helping with lodging and logistics during fieldwork.

CONFLICT OF INTEREST

The authors declare no conflicts of interest.

AUTHOR CONTRIBUTIONS

J.R., A.S.-G., and M.A.A. conceived the study. P.E. and J.R. drafted the manuscript. P.E., V.A.P., A.A.F.-R., J.V. and J.F.S.-H. performed the bioinformatic analysis. P.E., V.A.P., A.S.-G. and J.R. interpreted the data. All authors revised and approved the final manuscript.

OPEN RESEARCH BADGES

This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available at https://github.com/molevol-ub/Dysdera_silvatica_genome.

DATA AVAILABILITY STATEMENT

The whole-genome shotgun project has been deposited at DDBJ/ENA/GenBank under accession number QLNU00000000 and project ID PRJNA475203. The genome version described in this article is version QLNU02000000. The Hi-C (SRR13907633) and Chicago libraries raw reads (SRR13907634) are also included in the Bioproject repository. It also includes the raw data and sequencing libraries information from the assembly version 1. Other relevant data sets for this new version assembly, such as those including the structural and functional annotations, are available in https://github.com/molevol-ub/Dysdera_silvatica_genome. Protein sequence data of all chemoreceptor proteins (including incomplete fragments) identified in this study are provided in the Supporting Information online.

ORCID

Paula Escuer https://orcid.org/0000-0002-5941-0106
Vadim A. Pisarenco https://orcid.org/0000-0002-4968-4090
Angel A. Fernández-Ruiz https://orcid.org/0000-0003-1495-0923
Joel Vizueta https://orcid.org/0000-0003-0139-3013
Jose F. Sánchez-Herrero https://orcid.org/0000-0001-6771-4807
Miquel A. Arnedo https://orcid.org/0000-0003-1402-4727
Alejandro Sánchez-Gracia https://orcid.org/0000-0003-4543-4577
Julio Rozas https://orcid.org/0000-0002-6839-9148

REFERENCES

Adrián-Serrano, S., Lozano-Fernandez, J., Pons, J., Rozas, J., & Arnedo, M. A. (2021). On the shoulder of giants: Mitogenome recovery from non-targeted genome projects for phylogenetic inference and molecular evolution studies. *Journal of Zoological Systematics and Evolutionary Research*, *59*(1), 5–30. https://doi.org/10.1111/jzs.12415

Arnedo, M. A., Oromí, P., Múrria, C., Macías-Hernández, N., & Ribera, C. (2007). The dark side of an island radiation: Systematics and evolution of troglobitic spiders of the genus *Dysdera* Latreille (Araneae: Dysderidae) in the Canary Islands. *Invertebrate Systematics*, *21*(6), 623–660. https://doi.org/10.1071/IS07015

Arnedo, M. A., Oromí, P., & Ribera, C. (2001). Radiation of the spider genus *Dysdera* (Araneae, Dysderidae) in the Canary Islands: Cladistic assessment based on multiple data sets. *Cladistics*, *17*(4), 313–353. https://doi.org/10.1006/clad.2001.0168

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., & Sherlock, G. (2000). Gene Ontology: Tool for the unification of biology. *Nature Genetics*, *25*(1), 25–29. https://doi.org/10.1038/75556

Bleidorn, C. (2016). Third generation sequencing: Technology and its potential impact on evolutionary biodiversity research. *Systematics and Biodiversity*, *14*(1), 1–8. https://doi.org/10.1080/14772000.2015.1099575

Brůna, T., Hoff, K. J., Lomsadze, A., Stanke, M., & Borodovsky, M. (2021). BRAKER2: Automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein

database. *NAR Genomics and Bioinformatics*, *3*(1), 1–11. https://doi.org/10.1093/nargab/lqaa108

Chan, P. P., Lin, B. Y., Mak, A. J., & Lowe, T. M. (2019). TRNAscan-SE 2.0: Improved detection and functional classification of transfer RNA genes. *BioRxiv*, (6). https://doi.org/10.1101/614032

Clifton, B. D., Jimenez, J., Kimura, A., Chahine, Z., Librado, P., Sanchez-Gracia, A., Abbassi, M., Carranza, F., Chan, C., Marchetti, M., Zhang, W., Shi, M., Vu, C., Yeh, S., Fanti, L., Xia, X.-Q., Rozas, J., & Ranz, J. M. (2020). Understanding the early evolutionary stages of a tandem *Drosophila melanogaster*-specific gene family: A structural and functional population study. *Molecular Biology and Evolution*, *37*(9), 2584–2600. https://doi.org/10.1093/molbev/msaa109

Coddington, J. A., Giribet, G., Harvey, M. S., Prendini, L., & Walter, D. E. (2004). Arachnida. In J. Cracraft & M. J. Donoghue (Eds.), *Assembling the tree of life* (pp. 296–318). Oxford University Press.

Croset, V., Rytz, R., Cummins, S. F., Budd, A., Brawand, D., Kaessmann, H., Gibson, T. J., & Benton, R. (2010). Ancient protostome origin of chemosensory ionotropic glutamate receptors and the evolution of insect taste and olfaction. *PLoS Genetics*, *6*(8), e1001064. https://doi.org/10.1371/journal.pgen.1001064

Eddy, S. R. (2011). Accelerated profile HMM searches. *PLoS Computational Biology*, *7*(10), e1002195. https://doi.org/10.1371/journal.pcbi.1002195

Fan, Z., Yuan, T., Liu, P., Wang, L.-Y., Jin, J.-F., Zhang, F., & Zhang, Z.-S. (2021). A chromosome-level genome of the spider *Trichonephila antipodiana* reveals the genetic basis of its polyphagy and evidence of an ancient whole-genome duplication event. *GigaScience*, *10*(3), 1–15. https://doi.org/10.1093/gigascience/giab016

Garb, J. E., Sharma, P. P., & Ayoub, N. A. (2018). Recent progress and prospects for advancing arachnid genomics. *Current Opinion in Insect Science*, *25*, 51–57. https://doi.org/10.1016/j.cois.2017.11.005

Grbić, M., Van Leeuwen, T., Clark, R. M., Rombauts, S., Rouzé, P., Grbić, V., Osborne, E. J., Dermauw, W., Thi Ngoc, P. C., Ortego, F., Hernández-Crespo, P., Diaz, I., Martinez, M., Navajas, M., Sucena, É., Magalhães, S., Nagy, L., Pace, R. M., Djuranović, S., … Van de Peer, Y. (2011). The genome of *Tetranychus urticae* reveals herbivorous pest adaptations. *Nature*, *479*(7374), 487–492. https://doi.org/10.1038/nature10640

Gu, Z., Eils, R., & Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*, *32*(18), 2847–2849. https://doi.org/10.1093/bioinformatics/btw313

Hoff, K. J., Lomsadze, A., Borodovsky, M., & Stanke, M. (2019). Whole-genome annotation with BRAKER. *Methods in Molecular Biology*, *1962*, 65–95. https://doi.org/10.1007/978-1-4939-9173-0_5

Hopkin, S. P., & Martin, M. H. (1985). Assimilation of zinc, cadmium, lead, copper, and iron by the spider *Dysdera crocata*, a predator of woodlice. *Bulletin of Environmental Contamination and Toxicology*, *34*(1), 183–187. https://doi.org/10.1007/BF01609722

Hu, J., Fan, J., Sun, Z., & Liu, S. (2020). NextPolish: A fast and efficient genome polishing tool for long-read assembly. *Bioinformatics*, *36*(7), 2253–2255. https://doi.org/10.1093/bioinformatics/btz891

Jones, D. T., Taylor, W. R., & Thornton, J. M. (1992). The rapid generation of mutation data matrices from protein sequences. *Bioinformatics*, *8*(3), 275–282. https://doi.org/10.1093/bioinformatics/8.3.275

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., Pesseat, S., Quinn, A. F., Sangrador-Vegas, A., Scheremetjew, M., Yong, S.-Y., Lopez, R., & Hunter, S. (2014). InterProScan 5: Genome-scale protein function classification. *Bioinformatics*, *30*(9), 1236–1240. https://doi.org/10.1093/bioinformatics/btu031

Kallal, R. J., Kulkarni, S. S., Dimitrov, D., Benavides, L. R., Arnedo, M. A., Giribet, G., & Hormiga, G. (2020). Converging on the orb: Denser taxon sampling elucidates spider phylogeny and new analytical methods support repeated evolution of the orb web. *Cladistics*, *37*(3), 1–19. https://doi.org/10.1111/cla.12439

Kanehisa, M., & Goto, S. (2000). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, *28*(1), 27–30. https://doi.org/10.1093/nar/28.1.27

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution*, *30*(4), 772–780. https://doi.org/10.1093/molbev/mst010

Kriventseva, E. V., Kuznetsov, D., Tegenfeldt, F., Manni, M., Dias, R., Simão, F. A., & Zdobnov, E. M. (2019). OrthoDB v10: Sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Research*, *47*(D1), D807–D811. https://doi.org/10.1093/nar/gky1053

Kumar, S., Stecher, G., Peterson, D., & Tamura, K. (2012). MEGA-CC: Computing core of molecular evolutionary genetics analysis program for automated and iterative data analysis. *Bioinformatics*, *28*(20), 2685–2686. https://doi.org/10.1093/bioinformatics/bts507

Lee, E., Helt, G. A., Reese, J. T., Munoz-Torres, M. C., Childers, C. P., Buels, R. M., Stein, L., Holmes, I. H., Elsik, C. G., & Lewis, S. E. (2013). Web Apollo: A web-based genomic annotation editing platform. *Genome Biology*, *14*(8), R93. https://doi.org/10.1186/gb-2013-14-8-r93

Letunic, I., & Bork, P. (2007). Interactive Tree Of Life (iTOL): An online tool for phylogenetic tree display and annotation. *Bioinformatics*, *23*(1), 127–128. https://doi.org/10.1093/bioinformatics/btl529

Macias-Hernandez, N., de la Cruz López, S., Roca-Cusachs, M., Oromí, P., & Arnedo, M. A. (2016). A geographical distribution database of the genus *Dysdera* in the Canary Islands (Araneae, Dysderidae). *ZooKeys*, *2016*(625), 11–23. https://doi.org/10.3897/zookeys.625.9847

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., Von Haeseler, A., & Teeling, E. (2020). IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Molecular Biology and Evolution*, *37*(5), 1530–1534. https://doi.org/10.1093/molbev/msaa015

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G. A., Sonnhammer, E. L. L., Tosatto, S. C. E., Paladin, L., Raj, S., Richardson, L. J., Finn, R. D., & Bateman, A. (2021). Pfam: The protein families database in 2021. *Nucleic Acids Research*, *49*(D1), D412–D419. https://doi.org/10.1093/nar/gkaa913

Ni, L. (2021). The structure and function of ionotropic receptors in *Drosophila*. *Frontiers in Molecular Neuroscience*, *13*(February), 1–11. https://doi.org/10.3389/fnmol.2020.638839

Pekár, S., Líznarová, E., & Řezáč, M. (2016). Suitability of woodlice prey for generalist and specialist spider predators: A comparative study. *Ecological Entomology*, *41*(2), 123–130. https://doi.org/10.1111/een.12285

Pollard, M. O., Gurdasani, D., Mentzer, A. J., Porter, T., & Sandhu, M. S. (2018). Long reads: Their purpose and place. *Human Molecular Genetics*, *27*(R2), R234–R241. https://doi.org/10.1093/hmg/ddy177

Putnam, N. H., Connell, B. O., Stites, J. C., Rice, B. J., Blanchette, M., Calef, R., Troll, C. J., Fields, A., Hartley, P. D., Sugnet, C. W., Haussler, D., Rokhsar, D. S., & Green, R. E. (2016). Chromosome-scale shotgun assembly using an in vitro method for long-range linkage arXiv: 1502. 05331v1 [q-bio. GN] 18 Feb 2015. *Genome Research*, *26*, 342–350. https://doi.org/10.1101/gr.193474.115.Freely

Řezáč, M., & Pekár, S. (2007). Evidence for woodlice-specialization in Dysdera spiders: Behavioural versus developmental approaches. *Physiological Entomology*, *32*(4), 367–371. https://doi.org/10.1111/j.1365-3032.2007.00588.x

Řezáč, M., Pekár, S., Arnedo, M., Macías-Hernández, N., & Řezáčová, V. (2021). Evolutionary insights into the eco-phenotypic diversification of *Dysdera* spiders in the Canary Islands. *Organisms Diversity and Evolution*, *21*(1), 79–92. https://doi.org/10.1007/s13127-020-00473-w

Řezáč, M., Pekár, S., & Lubin, Y. (2008). How oniscophagous spiders overcome woodlouse armour. *Journal of Zoology*, *275*(1), 64–71. https://doi.org/10.1111/j.1469-7998.2007.00408.x

Robertson, H. M., Baits, R. L., Walden, K. K. O., Wada-Katsumata, A., & Schal, C. (2018). Enormous expansion of the chemosensory gene repertoire in the omnivorous German cockroach *Blattella germanica*.

Journal of Experimental Zoology Part B: Molecular and Developmental Evolution, *330*(5), 265–278. https://doi.org/10.1002/jez.b.22797

Saha, S. (2019). Long range sequencing and validation of insect genome assemblies. In S. Brown & M. Pfrender (Eds.), *Methods in molecular biology* (Vol. *1858*, pp. 33–44). Springer New York. https://doi.org/10.1007/978-1-4939-8775-7_4

Sánchez-Herrero, J. F., Frías-López, C., Escuer, P., Hinojosa-Alvarez, S., Arnedo, M. A., Sánchez-Gracia, A., & 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*, *8*(8), 1–9. https://doi.org/10.1093/gigascience/giz099

Sanggaard, K. W., Bechsgaard, J. S., Fang, X., Duan, J., Dyrlund, T. F., Gupta, V., Jiang, X., Cheng, L., Fan, D., Feng, Y., Han, L., Huang, Z., Wu, Z., Liao, L. I., Settepani, V., Thøgersen, I. B., Vanthournout, B., Wang, T., Zhu, Y., … Wang, J. (2014). Spider genomes provide insight into composition and evolution of venom and silk. *Nature Communications*, *5*(May), 3765. https://doi.org/10.1038/ncomms4765

Schwager, E. E., Sharma, P. P., Clarke, T., Leite, D. J., Wierschin, T., Pechmann, M., Akiyama-Oda, Y., Esposito, L., Bechsgaard, J., Bilde, T., Buffry, A. D., Chao, H., Dinh, H., Doddapaneni, H. V., Dugan, S., Eibner, C., Extavour, C. G., Funch, P., Garb, J., … McGregor, A. P. (2017). The house spider genome reveals an ancient whole-genome duplication during arachnid evolution. *BMC Biology*, *15*(1), 1–27. https://doi.org/10.1186/s12915-017-0399-x

Seppey, M., Manni, M., & Zdobnov, E. M. (2019). BUSCO: Assessing genome assembly and annotation completeness. In Kollmar, M. (2019). *BUSCO: Assessing genome assembly and annotation completeness BT - Gene prediction: Methods and protocols* (Vol. *1962*, pp. 227–245). https://doi.org/10.1007/978-1-4939-9173-0_14

Sharma, P. P., Kaluziak, S. T., Pérez-Porro, A. R., González, V. L., Hormiga, G., Wheeler, W. C., & Giribet, G. (2014). Phylogenomic interrogation of arachnida reveals systemic conflicts in phylogenetic signal. *Molecular Biology and Evolution*, *31*(11), 2963–2984. https://doi.org/10.1093/molbev/msu235

Sheffer, M. M., Hoppe, A., Krehenwinkel, H., Uhl, G., Kuss, A. W., Jensen, L., Jensen, C., Gillespie, R. G., Hoff, K. J., & Prost, S. (2021). Chromosome-level reference genome of the European wasp spider *Argiope bruennichi*: A resource for studies on range expansion and evolutionary adaptation. *GigaScience*, *10*(1), 1–12. https://doi.org/10.1093/gigascience/giaa148

Smit, A. F., & Hubley, R. (2008-2015). *RepeatModeler-1.0*. Retrieved from http://www.repeatmasker.org

Smit, A. F., Hubley, R., & Green, P. (2013-2015). *RepeatMasker-4.0*. http://www.repeatmasker.org

Stanke, M., Diekhans, M., Baertsch, R., & Haussler, D. (2008). Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics*, *24*(5), 637–644. https://doi.org/10.1093/bioinformatics/btn013

Stanke, M., Schöffmann, O., Morgenstern, B., & Waack, S. (2006). Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics*, *7*, 1–11. https://doi.org/10.1186/1471-2105-7-62

Thomas, G. W. C., Dohmen, E., Hughes, D. S. T., Murali, S. C., Poelchau, M., Glastad, K., Anstead, C. A., Ayoub, N. A., Batterham, P., Bellair, M., Binford, G. J., Chao, H., Chen, Y. H., Childers, C., Dinh, H., Doddapaneni, H. V., Duan, J. J., Dugan, S., Esposito, L. A., … Richards, S. (2020). Gene content evolution in the arthropods. *Genome Biology*, *21*(15), 1–14. https://doi.org/10.1186/s13059-019-1925-7

Toft, S., & Macías-Hernández, N. (2017). Metabolic adaptations for isopod specialization in three species of *Dysdera* spiders from the Canary Islands. *Physiological Entomology*, *42*(2), 191–198. https://doi.org/10.1111/phen.12192

Ullmann, A. J., Lima, C. M. R., Guerrero, F. D., Piesman, J., & Black, W. C. IV (2005). Genome size and organization in the blacklegged tick, *Ixodes scapularis* and the Southern cattle tick, *Boophilus microplus*. *Insect Molecular Biology*, *14*(2), 217–222. https://doi.org/10.1111/j.1365-2583.2005.00551.x

Vieira, F. G., Sánchez-Gracia, A., & Rozas, J. (2007). Comparative genomic analysis of the odorant-binding protein family in 12 *Drosophila* genomes: Purifying selection and birth-and-death evolution. *Genome Biology*, *8*(11), R235. https://doi.org/10.1186/gb-2007-8-11-r235

Vizueta, J., Escuer, P., Frías-López, C., Guirao-Rico, S., Hering, L., Mayer, G., Rozas, J., & Sánchez-Gracia, A. (2020). Evolutionary history of major chemosensory gene families across panarthropoda. *Molecular Biology and Evolution*, *37*(12), 3601–3615. https://doi.org/10.1093/molbev/msaa197

Vizueta, J., Escuer, P., Sánchez-Gracia, A., & Rozas, J. (2020). Genome mining and sequence analysis of chemosensory soluble proteins in arthropods. In P. Pelosi & W. Knoll (Eds.), *Methods in enzymology* (Vol. *642*, pp. 1–20). Elsevier Inc. https://doi.org/10.1016/bs.mie.2020.05.015

Vizueta, J., Frías-López, C., Macías-Hernández, N., Arnedo, M. A., Sánchez-Gracia, A., & Rozas, J. (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), 178–196. https://doi.org/10.1093/gbe/evw296

Vizueta, J., Macías-Hernández, N., Arnedo, M. A., Rozas, J., & Sánchez-Gracia, A. (2019). Chance and predictability in evolution: The genomic basis of convergent dietary specializations in an adaptive radiation. *Molecular Ecology*, *28*(17), 4028–4045. https://doi.org/10.1111/mec.15199

Vizueta, J., Rozas, J., & Sánchez-Gracia, A. (2018). Comparative genomics reveals thousands of novel chemosensory genes and massive changes in chemoreceptor repertories across chelicerates. *Genome Biology and Evolution*, *10*(5), 1221–1236. https://doi.org/10.1093/gbe/evy081

Vizueta, J., Sánchez-Gracia, A., & Rozas, J. (2020). bitacora: A comprehensive tool for the identification and annotation of gene families in genome assemblies. *Molecular Ecology Resources*, *20*(5), 1445–1452. https://doi.org/10.1111/1755-0998.13202

Wheeler, W. C., Coddington, J. A., Crowley, L. M., Dimitrov, D., Goloboff, P. A., Griswold, C. E., Hormiga, G., Prendini, L., Ramírez, M. J., Sierwald, P., Almeida-Silva, L., Alvarez-Padilla, F., Arnedo, M. A., Benavides Silva, L. R., Benjamin, S. P., Bond, J. E., Grismado, C. J., Hasan, E., Hedin, M., … Zhang, J. (2017). The spider tree of life: Phylogeny of Araneae based on target-gene analyses from an extensive taxon sampling. *Cladistics*, *33*(6), 574–616. https://doi.org/10.1111/cla.12182

World Spider Catalogue (2021). World spider catalog. https://doi.org/10.24436/2. Accessed March 26, 2021.

Zhang, Z. Q. (2011). Animal biodiversity: An outline of higher-level classification and survey of taxonomic richness. *Zootaxa*, *3148*, 165–191. https://doi.org/10.11646/zootaxa.3148.1.2

Zhang, Z. Q. (2013). Phylum arthropoda. *Zootaxa*, *3703*(1), 17–26. https://doi.org/10.11646/zootaxa.3703.1.6

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Escuer, P., Pisarenco, V. A., Fernández-Ruiz, A. A., Vizueta, J., Sánchez-Herrero, J. F., Arnedo, M. A., Sánchez-Gracia, A., & Rozas, J. (2022). The chromosome-scale assembly of the Canary Islands endemic spider *Dysdera silvatica* (Arachnida, Araneae) sheds light on the origin and genome structure of chemoreceptor gene families in chelicerates. *Molecular Ecology Resources*, 22, 375–390. https://doi.org/10.1111/1755-0998.13471

The chromosome-scale assembly of the Canary Islands endemic spider *Dysdera silvatica* (Arachnida, Araneae) sheds light on the origin and genome structure of chemoreceptor gene families in chelicerates

P Escuer, VA Pisarenco, AA Fernández-Ruiz, J Vizueta, JF Sánchez-Herrero, MA Arnedo, A Sánchez-Gracia, J Rozas

2022. Molecular Ecology Resources, 22:375–390. doi: 10.1111/1755-0998.13471

Supplementary Material

MOLECULAR ECOLOGY RESOURCES

Supplemental Information for:

The chromosome-scale assembly of the Canary Islands endemic spider *Dysdera silvatica* (Arachnida, Araneae) sheds light on the origin and genome structure of chemoreceptor gene families in chelicerates

Paula Escuer[1,#], Vadim A. Pisarenco[1,#], Angel A. Fernández-Ruiz[1], Joel Vizueta[1,2], Jose F. Sánchez-Herrero[3], Miquel A. Arnedo[4], Alejandro Sánchez-Gracia[1*], Julio Rozas[1*]

Supplementary Material online

Protein sequence data of all chemoreceptor proteins (including incomplete fragments) identified in this study (**Seqs_Chemoreceptors_MSA.zip file**)

Supplementary Tables

The tables are included in the file: **20210409_SupTables_S1-S5.xlsx**

Table S1. Species analyzed in this study

Table S2. Summary of BUSCO results

Table S3. Summary of repetitive elements identified in the *D. silvatica* genome

Table S4. Genome organization of the members of the *Gr* and *Ir* gene families across the *D. silvatica* pseudochromosomes

Table S5. Gr and Ir/iGluR sequences identified in the new *D. silvatica* genome assembly. All protein sequences (including those encoding partial genes) identified in this study and the

MOLECULAR ECOLOGY
RESOURCES

multiple sequence alignments used in the analyses are provided in the **Seqs_Chemoreceptors_MSA.zip file**.

Supplementary Figures

Figure S1. Homology-based search across different ecdysozoa species. Pie chart showing the taxonomic distribution of positive BLAST hits of the structural annotation of *D. silvatica* (*n* = 33,275 genes) across the Araneae species illustrated in Figure 1 and the rest of the arachnida, arthropoda and ecdysozoa also analysed in Sánchez-Herrero et al. (2019).

Figure S2. Genome organization and relationships between physical and evolutionary distances of the members of the *Gr* and *Ir* families on the *D. silvatica* pseudochromosomes. A and C) Heatmaps illustrating the distribution of physical (below the diagonal; in units of 100 kb) and evolutionary (above the diagonal; in units of amino acid substitutions per site) distances, along the pseudochromosome surveyed. White dots, evolutionary distances not available. B and D) Plots comparing pairwise amino acid and physical (on a logarithmic scale) distances between *Gr* or *Ir* copies in the *D. silvatica* pseudochromosome. Colored and grey points show distances within and outside genomic clusters, respectively. Different clusters are depicted in different colors.

.

Figure S3. Phylogenetic relationships and node support in the Gr family. A) Phylogenetic tree among all sequences encoding members of the Gr family (including both complete and incomplete or partial genes). Incomplete genes are marked with asterisks. The tree was rooted in its midpoint. The outer ring indicates the chromosome (Chr) in which the genes included in the tree are located. The inner ring shows information about genomic clusters (chromosome, in the same color scale that in outer ring, genomic cluster number: member number in the cluster).

MOLECULAR ECOLOGY
RESOURCES

The scale bar refers to 1 amino acid substitution per site. Minor SC, minor scaffolds. Red and green terminal branches correspond to *D. silvatica* and *D. melanogaster* Gr, respectively. B) Cladogram of the phylogenetic tree in Figure 6 with bootstrap node support values >90%.

Figure S4. Phylogenetic relationships and node support in the Ir/iGluR family. A) Phylogenetic tree among all sequences encoding members of the Ir/iGluR family (including both complete and incomplete or partial genes). Incomplete genes are marked with asterisks. The light blue, dark blue and purpura shading of gene names designate the members of the iGluR, Ir25a/8a and Ir subfamilies, respectively. The tree was rooted considering NMDAR clade as the outgroup (Croset et al. 2010). The outer ring indicates the chromosome (Chr) in which the genes included in the tree are located. The inner ring shows information about genomic clusters (chromosome, in the same color scale that in outer ring, genomic cluster number: member number in the cluster). The scale bar refers to 1 amino acid substitution per site. Minor SC, minor scaffolds. Red and green terminal branches correspond to *D. silvatica* and *D. melanogaster* Gr, respectively. The red triangles mark the *D. silvatica* genes with putative distant homologs in *D. melanogaster*. B) Cladogram of the phylogenetic tree in figure 7 with bootstrap node support values >90%. C) Phylogenetic relationships among the sequences encoding the LBD domain of *D. silvatica* and *D. melanogaster*. The tree only includes the LBD domains classified as complete in this work.

Species	Homology Search[a]	Order	Suborder	Section/Clade	Superfamily	Family	Assembly NCBI Accession	NCBI Genome Bioproject	Notes	Citation
Dysdera silvatica	Yes	Araneae	Araneomorphae	Synspermiata	Dysderoidea	Dysderidae	GCA_006491805.1	PRJNA475203	Version 1	1
Loxosceles reclusa	Yes	Araneae	Araneomorphae	Synspermiata	Scytodoidea	Sicariidae	GCA_001188405.1	PRJNA167404		2
Pardosa pseudoannulata	No	Araneae	Araneomorphae	Entelegyne (RTA-clade)	Lycosoidea	Lycosidae	GCA_008065355.1			3
Stegodyphus mimosarum	Yes	Araneae	Araneomorphae	Entelegyne		Eresidae	GCA_000611955.2			4
Stegodyphus dumicola	Yes	Araneae	Araneomorphae	Entelegyne		Eresidae	GCA_010614865.1			5
Latrodectus hesperus	Yes	Araneae	Araneomorphae	Entelegyne	Araneoidea	Theridiidae	GCA_000697925.2	PRJNA168123		6
Anelosimus studiosus	No	Araneae	Araneomorphae	Entelegyne	Araneoidea	Theridiidae	GCA_008297655.1			7
Parasteatoda tepidariorum	Yes	Araneae	Araneomorphae	Entelegyne	Araneoidea	Theridiidae	GCA_000365465.3	PRJNA167405		8
Oedothorax gibbosus	No	Araneae	Araneomorphae	Entelegyne	Araneoidea	Linyphiidae		PRJNA681589		9
Araneus ventricosus	No	Araneae	Araneomorphae	Entelegyne	Araneoidea	Araneidae	BGPR01000001-BGPR01300721			10
Argiope bruennichi	Yes	Araneae	Araneomorphae	Entelegyne	Araneoidea	Araneidae		PRJNA629526		11
Trichonephila clavipes	Yes	Araneae	Araneomorphae	Entelegyne	Araneoidea	Araneidae	GCA_002102615.1			12
Trichonephila antipodiana	Yes	Araneae	Araneomorphae	Entelegyne	Araneoidea	Araneidae		PRJNA627506		13
Acanthoscurria geniculata	Yes	Araneae	Mygalomorphae		Avicularioidea	Theraphosidae		PRJNA222716		14
Centruroides sculpturatus	Yes	Scorpiones				Buthidae	GCA_000671375.2			15

[a]Genome data used for the homology search analyses

Species	Citation	
Dysdera silvatica	1	Sánchez-Herrero et al. (2019). The draft genome sequence of the spider *Dysdera silvatica* (Araneae, Dysderidae): A valuable resource for functional and evolutionary genomic studies in chelicerates. GigaScience 8:1
Loxosceles reclusa	2	Unpublished. Cited in Thomas et al. (2020). Gene content evolution in the arthropods. Genome Biology 21:15.
Pardosa pseudoannulata	3	Yu et al. (2019). Genome sequencing and neurotoxin diversity of a wandering spider *Pardosa pseudoannulata* (pond wolf spider). bioRxiv preprint doi: https://doi.org/10.1101/747147
Stegodyphus mimosarum	4	Sanggaard et al. (2014). Spider genomes provide insight into composition and evolution of venom and silk. Nature communications 5:1-12.
Stegodyphus dumicola	5	Liu et al. (2019). DNA methylation patterns in the social spider, *Stegodyphus dumicola* . Genes 10:137
Latrodectus hesperus	6	Unpublished. Cited in Thomas et al. (2020). Gene content evolution in the arthropods. Genome Biology. 21:15.
		Also cited in: Sheffer et al. (2021). Chromosome-level reference genome of the European wasp spider *Argiope bruennichi*: a resource for studies on range expansion and evolutionary adaptation. GigaScience 10: giaa
Anelosimus studiosus	7	Unpublished. Sheffer et al. (2021). Chromosome-level reference genome of the European wasp spider *Argiope bruennichi:* a resource for studies on range expansion and evolutionary adaptation. GigaScience 10: giaa
Parasteatoda tepidariorum	8	Schwageret et al. (2017). The house spider genome reveals an ancient whole-genome duplication during arachnid evolution. BMC biology 15:1-27.
Oedothorax gibbosus	9	Hendrickx et al. (2021). A masculinizing supergene underlies an exaggerated malereproductive morph in a spider. bioRxiv. doi: https://doi.org/10.1101/2021.02.09.430505
Araneus ventricosus	10	Kono et al. (2019). Orb-weaving spider *Araneus ventricosus* genome elucidates the spidroin gene catalogue. Scientific Reports 9: 8380.
Argiope bruennichi	11	Sheffer et al. (2021). Chromosome-level reference genome of the European wasp spider *Argiope bruennichi*: a resource for studies on range expansion and evolutionary adaptation. GigaScience 10: giaa148.
Trichonephila clavipes	12	Babb et al. (2017). The *Nephila clavipes* genome highlights the diversity of spider silk genes and their complex expression. Nature Genetics 49: 895-903.
Trichonephila antipodiana	13	Fan et al. (2021). A chromosome-level genome of the spider *Trichonephila antipodiana* reveals the genetic basis of its polyphagy and evidence of an ancient whole-genome duplication event. GigaScience 10: giab016
Acanthoscurria geniculata	14	Sanggaard et al. (2014). Spider genomes provide insight into composition and evolution of venom and silk. Nature Communications. 5:3765.
Centruroides sculpturatus	15	Schwager et al. (2017).The house spider genome reveals an ancient whole-genome duplication during arachnid evolution. BMC Biology 15:62.

Table S2. Summary of BUSCO results

	Dysdera silvatica					
	Genome[a]			Annotation[b]		
	Arachnida (*n* = 2934)	Arthropoda (*n* = 1013)	Eukaryota (*n* = 255)	Arachnida (*n* = 2934)	Arthropoda (*n* = 1013)	Eukaryota (*n* = 255)
Complete (C)	2488 (84.8%)	738 (72.9%)	197 (77.3%)	2358 (80.4%)	784 (77.4%)	197 (77.3%)
Single Copy (S)	2397 (81.7%)	719 (71.0%)	193 (75.7%)	2007 (68.4%)	704 (69.5%)	181 (71.0%)
Duplicated (D)	91 (3.1%)	19 (1.9%)	4 (1.6%)	351 (12.0%)	80 (7.9%)	16 (6.3%)
Fragmented (F)	44 (1.5%)	80 (7.9%)	40 (15.7%)	181 (6.2%)	128 (12.6%)	40 (17.3%)
Missing (M)	402 (13.7%)	195 (19.2%)	18 (7.0%)	395 (13.4%)	101 (10.0%)	14 (5.4%)

[a]BUSCO (v5) analysis using AUGUSTUS and tBLASTn against the genome sequence
[b]BUSCO (v5) analysis using the MetaEuk gene predictor

Table S3. Summary of repetitive regions identified in the *D. silvatica* genome

	nº elements	bp	%
SINEs	169.940	25.204.935	1,85 %
LINEs	251.750	144.649.335	10,59 %
LTR elements	19.204	8.608.655	0,63 %
DNA elements	747.234	225.241.302	16,49 %
Unclassified	1.641.265	301.156.045	22,05 %
Total interespersed repeats	2.829.393	704.860.272	51,61 %
Small RNA	659	959.227	0,07 %
Satellites	4.316	1.462.447	0,11 %
Simple repeats	290.085	13.708.157	1,00 %
Low complexity	41.823	2.186.281	0,16 %
Total Other elements	336.883	18.316.112	1,34 %
Total	3.166.276	723.176.384	52,95 %

Table S4. Gene number and gene cluster analysis across *Dysdera silvatica* pseudochromosomes

	ChrX	Chr1	Chr2	Chr3	Chr4	Chr5	Chr6	ChrU1	Minor scaffolds	Total	S_{MIN}
Length (Mb)	317,95	177,17	176,73	174,24	129,24	125,97	80,96	22,26	161,17	1365,69	
Genome Fraction	23,28%	12,97%	12,94%	12,76%	9,46%	9,22%	5,93%	1,63%	11,80%	100%	
Total number of genes	6.812	4.139	4.227	4.005	3.226	3.018	1.904	513	5.431	33.275	
Gene density[a]	20,47%	12,44%	12,70%	12,04%	9,69%	9,07%	5,72%	1,54%	16,32%	100,00%	
Total number of chemoreceptors[b]	26 (6)	62 (12)	46 (5)	32 (2)	32 (13)	50 (8)	25 (3)	20 (8)	277 (31)	570 (88)	565
Chromosomal gene density[c]	4,56%	10,88%	8,07%	5,61%	5,61%	8,77%	4,39%	3,51%	48,60%	100%	
Gr	4 (1)	25 (10)	17 (3)	11 (0)	18 (9)	15 (4)	10 (2[d])	3 (1)	31 (6)	134 (36[d])	129
IR	13 (3)	34 (2)	28 (2)	18 (2)	11 (4)	34 (4)	11 (1)	17 (7)	245 (25)	411 (50)	411
iGluR	9 (2)	3 (0)	1 (0)	3 (0)	3 (0)	1 (0)	4 (0)	0 (0)	1 (0)	25 (2)	25
Clusters of *Gr* genes	0	2	2	2	0	1	2	1	3	13	
Analysed[e]	0 (0, 3)	2 (10, 15)	2 (9, 14)	2 (4, 11)	0 (0, 9)	1 (8, 11)	2 (5, 8)	1 (2, 2)	3 (14, 25)	13 (52, 98)	
C_{ST}	na	0,603	0,857	0,981	na	0,4074	0,6122	na	0,933	0,817	
Mann-Whitney test; *p*-value		***	***	*		**	ns		***	***	
Clusters of *Ir* genes	0	7	4	2	1	5	1	2	30	52	
Analysed[e]	0 (0, 10)	7 (19, 32)	4 (9, 26)	2 (5, 16)	1 (2, 7)	5 (19, 30)	1 (2, 10)	2 (6, 10)	30 (73, 220)	52 (135, 361)	
C_{ST}	na	0,579	0,429	0,847	0,895	0,716	0,617	0,596	0,679	0,674	
Mann-Whitney test; *p*-value		***	***	***	ns	***	ns	*	***	***	

[a]Fraction of genes in the chromosome

[b]Gr, Ir and iGluR genes identified in *D. silvatica* genome; in parenthesis the number of incomplete genes

[c]Fraction of chemoreceptor genes in the chromosome

[d]One incomplete copy is a pseudogene

[e]Number of clusters analyzed (only complete genes were analyzed). The values in parenthesis indicate the number of family members in clusters and the total number of family members in the scaffold

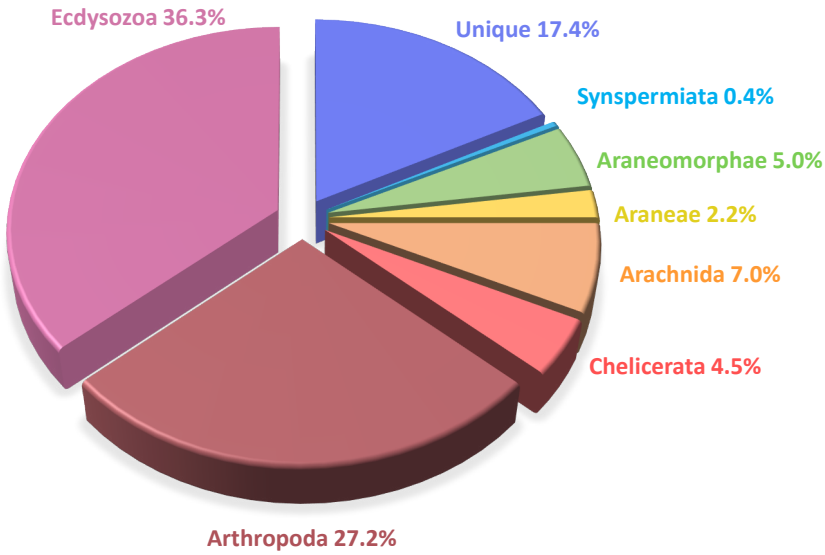
na, Not applicable; *, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$; ns, not significant

Table S5. Gr and Ir/iGluR sequences identified in the new *D. silvatica* genome assembly

#Family	Gene_name	Scaffold_name	Scaffold_Chr	Atype	Geneid_GFF	Prot_length	Status	Smin	F_clus	F_cluste	F_cme
GR	Gr27305	Scaffold_15362_HRSCAF_19823	Dsil-ChrX	GFF	g27305	328	Complete	Smin	No	-	-
GR	Gr27759	Scaffold_15362_HRSCAF_19823	Dsil-ChrX	GFF	g27759	413	Complete	Smin	No	-	-
GR	Gr001	Scaffold_15362_HRSCAF_19823	Dsil-ChrX	Bitacora	GRg001	185	Partial	Smin	No	-	-
GR	Gr002	Scaffold_15362_HRSCAF_19823	Dsil-ChrX	Bitacora	GRg002	301	Complete	Smin	No	-	-
GR	Gr9343	Scaffold_14804_HRSCAF_18385	Dsil-Chr1	GFF	g9343	206	Partial	Smin	No	-	-
GR	Gr10232	Scaffold_14804_HRSCAF_18385	Dsil-Chr1	GFF	g10232	382	Complete	Smin	No	-	-
GR	Gr066	Scaffold_14804_HRSCAF_18385	Dsil-Chr1	Bitacora	GRg066	171	Partial	Smin	No	-	-
GR	Gr10331	Scaffold_14804_HRSCAF_18385	Dsil-Chr1	GFF	g10331	138	Partial	Smin	Yes	F1	1
GR	Gr10332	Scaffold_14804_HRSCAF_18385	Dsil-Chr1	GFF	g10332	138	Partial	Smin	Yes	F1	2
GR	Gr149	Scaffold_14804_HRSCAF_18385	Dsil-Chr1	Bitacora	GRg149	359	Complete	Smin	No	-	-
GR	Gr10489	Scaffold_14804_HRSCAF_18385	Dsil-Chr1	GFF	g10489	215	Partial	Smin	No	-	-
GR	Gr10508	Scaffold_14804_HRSCAF_18385	Dsil-Chr1	GFF	g10508	185	Partial	Smin	Yes	F2	1
GR	Gr183	Scaffold_14804_HRSCAF_18385	Dsil-Chr1	Bitacora	GRg183	300	Complete	Smin	Yes	F2	2
GR	Gr10520	Scaffold_14804_HRSCAF_18385	Dsil-Chr1	GFF	g10520	256	Complete	Smin	Yes	F2	3
GR	Gr10522	Scaffold_14804_HRSCAF_18385	Dsil-Chr1	GFF	g10522	276	Complete	Smin	Yes	F2	4
GR	Gr10523	Scaffold_14804_HRSCAF_18385	Dsil-Chr1	GFF	g10523	98	Partial	Smin	Yes	F2	5
GR	Gr10625	Scaffold_14804_HRSCAF_18385	Dsil-Chr1	GFF	g10625	89	Partial	Smin	No	-	-
GR	Gr10875	Scaffold_14804_HRSCAF_18385	Dsil-Chr1	GFF	g10875	295	Complete	Smin	No	-	-
GR	Gr10894	Scaffold_14804_HRSCAF_18385	Dsil-Chr1	GFF	g10894	291	Complete	Smin	Yes	F3	1
GR	Gr10895	Scaffold_14804_HRSCAF_18385	Dsil-Chr1	GFF	g10895	388	Complete	Smin	Yes	F3	2
GR	Gr029	Scaffold_14804_HRSCAF_18385	Dsil-Chr1	Bitacora	GRg029	341	Complete	Smin	Yes	F3	3
GR	Gr024	Scaffold_14804_HRSCAF_18385	Dsil-Chr1	Bitacora	GRg024	304	Complete	Smin	Yes	F3	4
GR	Gr10897	Scaffold_14804_HRSCAF_18385	Dsil-Chr1	GFF	g10897	389	Complete	Smin	Yes	F3	5
GR	Gr10898	Scaffold_14804_HRSCAF_18385	Dsil-Chr1	GFF	g10898	180	Partial	Smin	Yes	F3	6
GR	Gr10899	Scaffold_14804_HRSCAF_18385	Dsil-Chr1	GFF	g10899	290	Complete	Smin	Yes	F3	7
GR	Gr025	Scaffold_14804_HRSCAF_18385	Dsil-Chr1	Bitacora	GRg025	343	Complete	Smin	Yes	F3	8
GR	Gr10964	Scaffold_14804_HRSCAF_18385	Dsil-Chr1	GFF	g10964	179	Partial	Smin	No	-	-
GR	Gr12295	Scaffold_14804_HRSCAF_18385	Dsil-Chr1	GFF	g12295	351	Complete	Smin	No	-	-
GR	Gr132	Scaffold_14804_HRSCAF_18385	Dsil-Chr1	Bitacora	GRg132	383	Complete	Smin	No	-	-
GR	Gr20330	Scaffold_14178_HRSCAF_16784	Dsil-Chr2	GFF	g20330	380	Complete	Smin	No	-	-

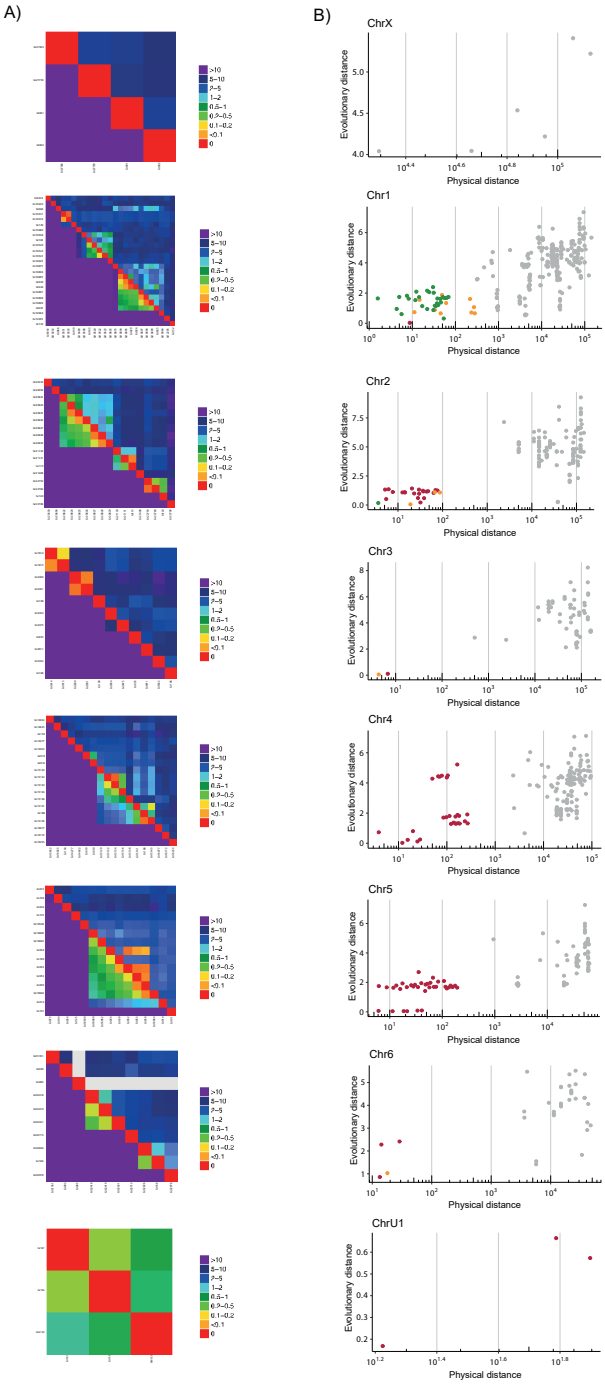
*Only showed the first 30 rows, for the complete results download the SupMaterial table from: 10.1111/1755-0998.13471

Fig. S1


Ecdysozoa 36.3%
Unique 17.4%
Synspermiata 0.4%
Araneomorphae 5.0%
Araneae 2.2%
Arachnida 7.0%
Chelicerata 4.5%
Arthropoda 27.2%

Gr family

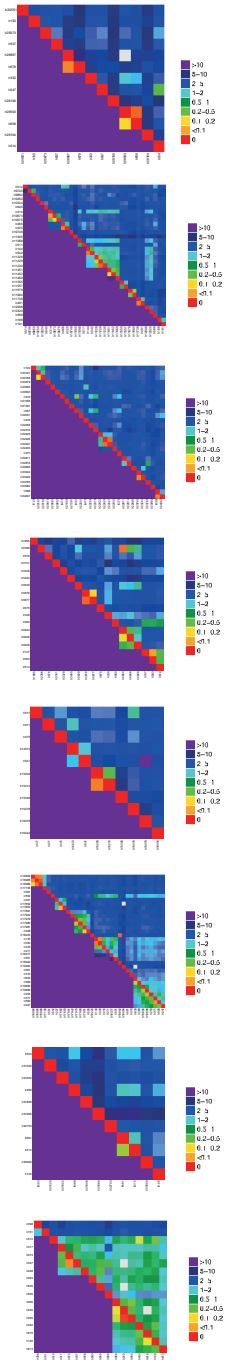
Fig. S2AB



Ir family

Fig. S2CD

C)



D)

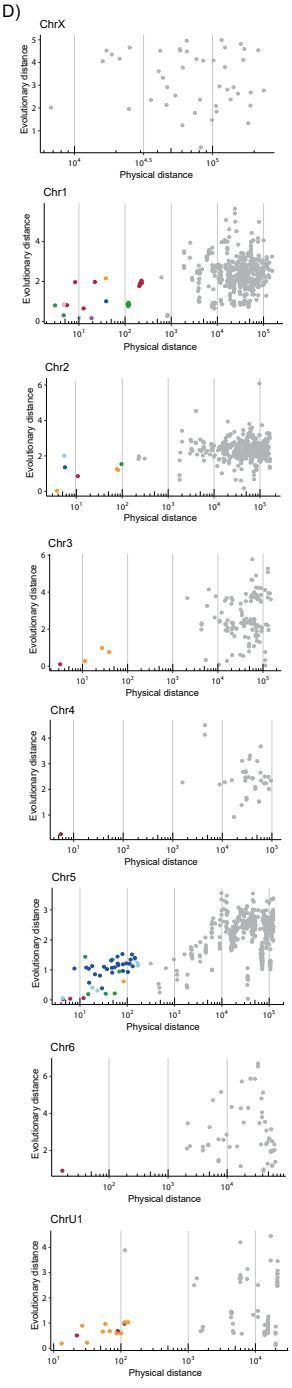


Fig. S3A

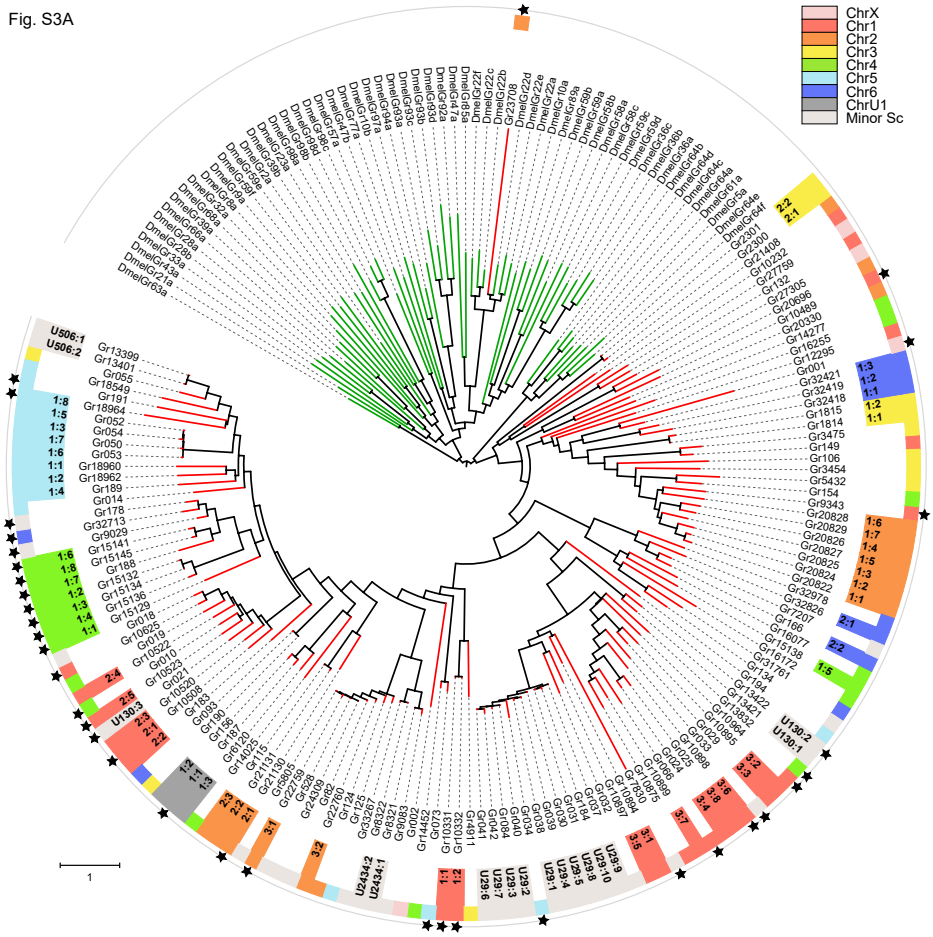


Fig. S3B

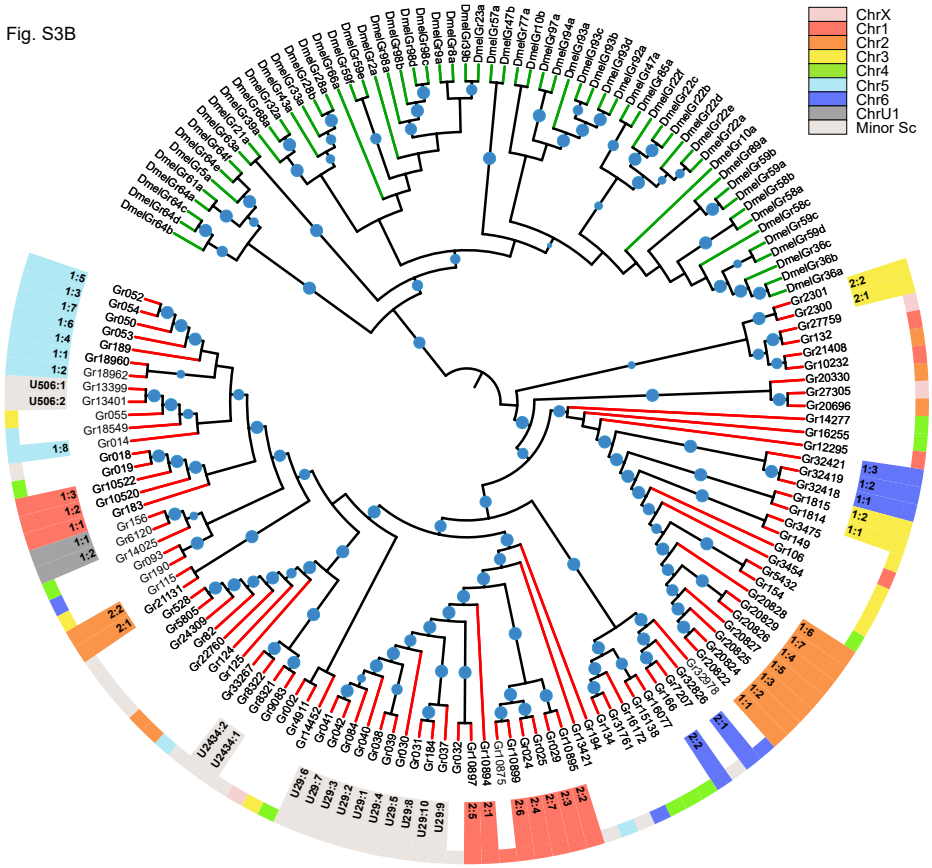


Fig. S4A

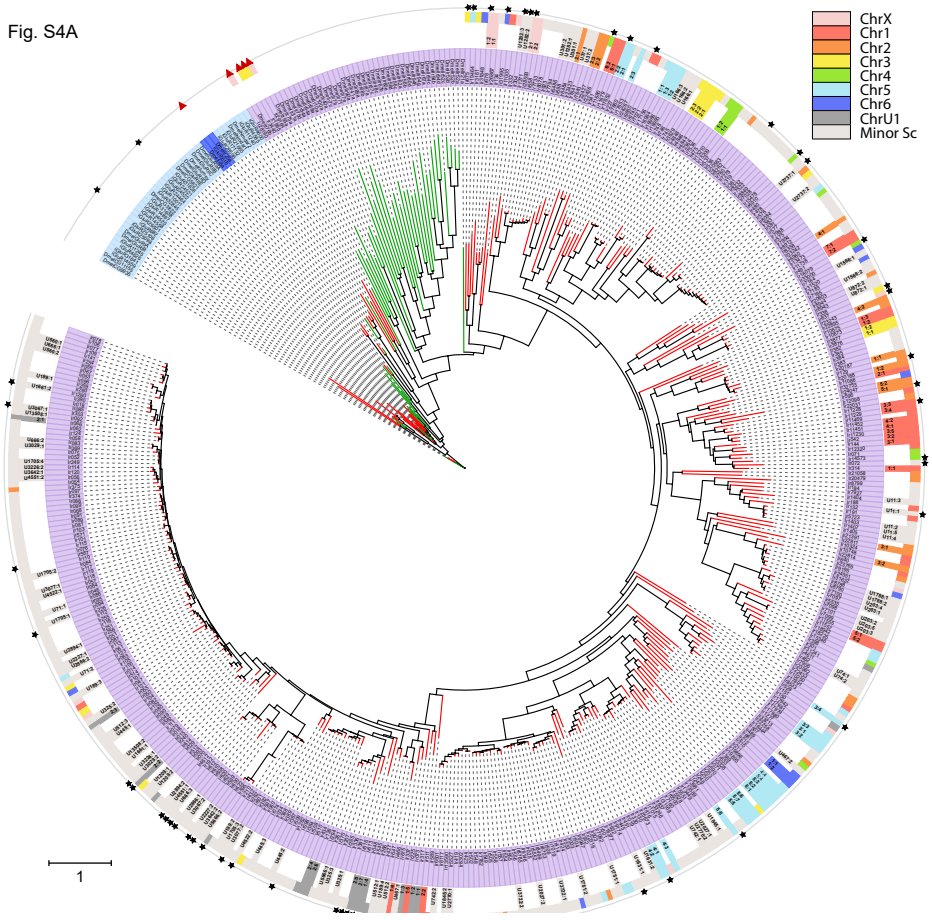


Fig. S4B

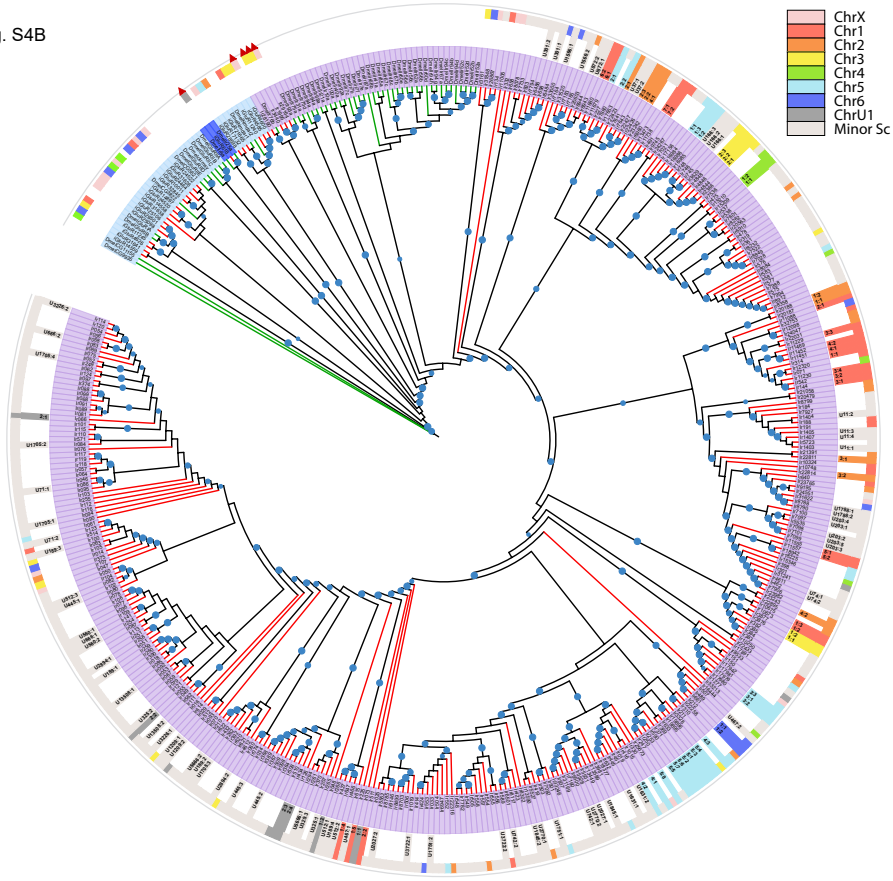
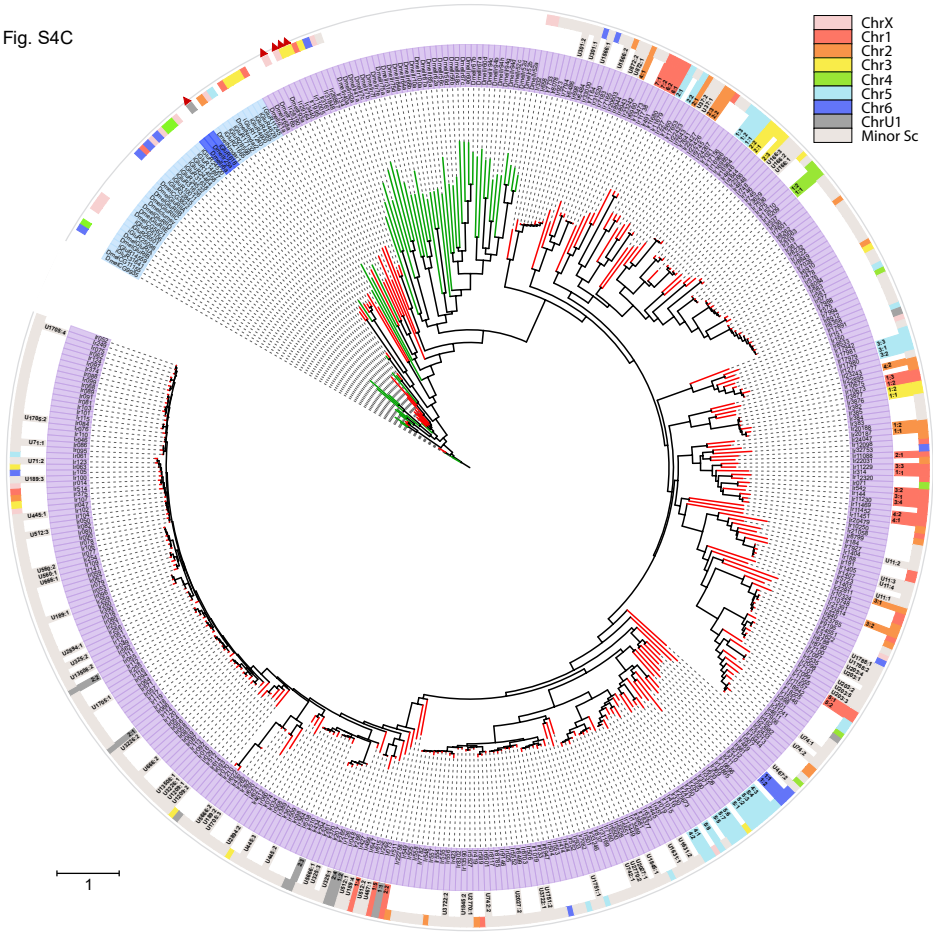


Fig. S4C



Chapter 3

Evolutionary history of major chemosensory gene families across Panarthropoda

J Vizueta, P Escuer, C Frías-López, S Guirao-Rico, L Hering, G Mayer, J Rozas, A Sánchez-Gracia

2020. Molecular Biology and Evolution, 37, 12:3601–3615. doi: https://doi.org/10.1093/molbev/msaa197

La percepció quimiosensorial és un procés biològic fonamental d'especial rellevància en l' investigació bàsica i aplicada en artròpodes. Tanmateix, a part dels insectes, hi ha poc coneixement de les molècules específiques implicades en aquest sistema, que es limita a uns pocs taxons amb mostreig filogenètic desigual entre llinatges. Des d'una perspectiva evolutiva, els onicòfors (cucs de vellut) i els tardígrads (óssos d'aigua) són d'especial interès, ja que representen els parents vius més propers dels artròpodes, que en conjunt formen els Panarthropoda. Per obtenir informació sobre l'origen evolutiu i la diversificació del repertori de gens quimiosensorials en panartròpodes, vam seqüenciar els transcriptomes específics de l'antena i el cap del cuc de vellut *Euperipatoides rowelli* i vam analitzar els membres de totes les famílies quimiosensorials principals en genomes representatius d'onicòfors, tardígrads i artròpodes. Els nostres resultats suggereixen que la família del gen NPC2 és l'única família que codificava proteïnes solubles en l'ancestre panartròpode, i que els onicòfors podrien haver perdut molts quimioreceptors que avui en dia estan als artròpodes, inclòs el receptor IR25a altament conservat en tots els protòstoms. D'altra banda, als genomes d'eutardígrads els hi manquen els gens que codifiquen les proteïnes de membrana de la neurona sensorial-CD36 i les DEG-ENaC, que també han estat retinguts en artròpodes; aquestes pèrdues poden estar relacionades amb estratègies adaptatives específiques del llinatge dels tardígrads per sobreviure a condicions ambientals extremes. Tot i que els resultats d'aquest estudi s'han de corroborar amb un augment del mostreig de taxons, els nostres descobriments aclareixen la diversificació de les famílies de gens quimiosensorials en Panarthropoda i contribueixen a una millor comprensió de l'evolució dels sentits químics dels animals.

Evolutionary History of Major Chemosensory Gene Families across Panarthropoda

Joel Vizueta,[†,1] Paula Escuer,[†,1] Cristina Frías-López,[1] Sara Guirao-Rico,[2] Lars Hering,[3] Georg Mayer,[3] Julio Rozas,[*,1] and Alejandro Sánchez-Gracia[*,1]

[1]Departament de Genètica, Microbiologia i Estadística and Institut de Recerca de la Biodiversitat (IRBio), Universitat de Barcelona, Barcelona, Spain

[2]Institute of Evolutionary Biology (CSIC-UPF), Barcelona, Spain

[3]Department of Zoology, Institute of Biology, University of Kassel, Kassel, Germany

[†]These authors contributed equally to this work.

***Corresponding authors**: E-mails: jrozas@ub.edu; elsanchez@ub.edu.

Associate editor: Meredith Yeager

The raw sequence data generated for this work have been deposited at the Sequence Read Archive (SRA) under Bioproject PRJNA607887. Additional data, including the de novo assembly and annotation of the *Euperipatoides rowelli* transcriptome, and results generated in this study have been deposited in *Figshare* (https://doi.org/10.6084/m9.figshare.12369638.v1).

Abstract

Chemosensory perception is a fundamental biological process of particular relevance in basic and applied arthropod research. However, apart from insects, there is little knowledge of specific molecules involved in this system, which is restricted to a few taxa with uneven phylogenetic sampling across lineages. From an evolutionary perspective, onychophorans (velvet worms) and tardigrades (water bears) are of special interest since they represent the closest living relatives of arthropods, altogether comprising the Panarthropoda. To get insights into the evolutionary origin and diversification of the chemosensory gene repertoire in panarthropods, we sequenced the antenna- and head-specific transcriptomes of the velvet worm *Euperipatoides rowelli* and analyzed members of all major chemosensory families in representative genomes of onychophorans, tardigrades, and arthropods. Our results suggest that the NPC2 gene family was the only family encoding soluble proteins in the panarthropod ancestor and that onychophorans might have lost many arthropod-like chemoreceptors, including the highly conserved IR25a receptor of protostomes. On the other hand, the eutardigrade genomes lack genes encoding the DEG-ENaC and CD36-sensory neuron membrane proteins, the chemosensory members of which have been retained in arthropods; these losses might be related to lineage-specific adaptive strategies of tardigrades to survive extreme environmental conditions. Although the results of this study need to be further substantiated by an increased taxon sampling, our findings shed light on the diversification of chemosensory gene families in Panarthropoda and contribute to a better understanding of the evolution of animal chemical senses.

Key words: **Onychophora, Tardigrada, chemosensory-related proteins, antenna-specific transcriptome, comparative genomics, BITACORA.**

Introduction

Major animal lineages have evolved independently strikingly similar olfactory pathways (Ache and Young 2005). Indeed, in both mammals and insects, a group of small soluble proteins is responsible for the peripheral detection and solubilization of the odorant compounds (but see Sun et al. [2018] for a broader perspective of these molecules). These proteins are secreted into the aqueous space such as the olfactory mucosa in vertebrates and the sensillar lymph in insects, which is in direct contact with the external environment (Pelosi 1994; Tegoni et al. 2000; Leal 2013). Odorants activate the highly tuned transmembrane receptors located in the dendrites of olfactory neurons, triggering electrical signals that are initially processed in intermediate brain structures (e.g., the olfactory bulb in vertebrates and olfactory glomeruli in insects) and subsequently integrated in higher brain centers (Pelosi 1996; Sánchez-Gracia et al. 2009).

Although soluble proteins and membrane receptors are encoded by large multigene families varying from tens to thousands of copies per genome in both mammals and insects, their evolutionary origin is completely different. Although the soluble proteins of insects (mostly represented by odorant-binding, OBP, and chemosensory proteins, CSP; Pelosi et al. 2014) are small, globular alpha-helix-rich proteins, those of mammals are much larger and with a typical beta barrel domain (belonging to the lipocalin family, mammalian OBP; Tegoni et al. 2000). Nonhomology of mammalian and insect olfactory receptors with their characteristic seven-transmembrane domains is also evident from their inverted

membrane topologies and different signal transduction mechanisms. In insects, the chemoreceptor superfamily, composed of olfactory (*Or*) and gustatory (*Gr*) receptor families, and the ionotropic receptor (*Ir*) subfamily (a group of highly divergent members of the ionotropic glutamate receptor superfamily, *iGluR*, involved in smell and taste) are ligand-gated ion channels (Joseph and Carlson 2015). Conversely, mammalian olfactory (OR) and taste (T1R and T2R) receptors are G protein-coupled receptors (GPCRs) that activate the second messengers indirectly through gating the corresponding ion channels (Wicher 2012). Furthermore, in mammals, salty and sour stimuli are known to be sensed by amiloride-sensitive ion channels (ENaC; Ben-Shahar 2011), a gene family that has been reported to play a role in insect pheromone perception (Lu et al. 2012), besides being involved in salt and water reception taste and osmotic stress responses (Liu et al. 2003; Chen et al. 2010). Finally, another family related to the human fatty acid transporter CD36 (Vogt et al. 2009), the sensory neuron membrane protein (SNMP) family has been also associated with chemosensory neurons in insects.

All this knowledge, however, is based only on a few invertebrate lineages, with data completely missing from many other important bilaterian clades (Eyun et al. 2017; Vizueta et al. 2018). Among these unexplored taxa, Onychophora (velvet worms) and Tardigrada (water bears) are especially relevant since they represent the closest living relatives of arthropods, with which they have been united in the so-called Panarthropoda (Nielsen 1995; Giribet and Edgecombe 2017). Onychophorans and tardigrades are, thus, key for understanding the evolutionary changes that have taken place in the arthropod lineage. It remains unknown, for instance, whether the chemosensory gene repertories found in arthropods were the result of specific adaptations to the extraordinarily range of environments they inhabit (both aquatic and terrestrial) or whether they were already present in the last common ancestor of Panarthropoda. In other words, to what extent do onychophoran and tardigrade genomes encode members of the arthropod chemosensory families?

Onychophorans most likely originated from an aquatic ancestor over 500 Ma (Rota-Stabelli et al. 2013), although the ~200 extant species of this group are exclusively terrestrial (Oliveira et al. 2012, 2016; Murienne et al. 2014). Velvet worms are elongated, soft-bodied invertebrates that inhabit tropical and temperate forests of the southern hemisphere and around the equator. One remarkable feature of velvet worms is the high phenotypic and anatomic conservation with respect to their Cambrian ancestors (lobopodians), emerging as an important outgroup and excellent model for evolutionary studies of arthropods (e.g., Mayer et al. 2010; Ou et al. 2012; Pauli et al. 2016; Janssen 2017; Martin et al. 2017; Petersen et al. 2019). In onychophorans, the main chemosensory perception structures are located on the antennae (fig. 1A), although the lip papillae surrounding the mouth might also have sensory cells responding to chemical stimuli (Storch and Ruhberg 1977; Storch and Ruhberg 1993). The antennae and the oral lips are innervated by differentiated groups of cell bodies located in different brain regions, suggesting that these structures might have some chemosensory specialization (Martin and Mayer 2014; Martin et al. 2017). However, only the antennae are associated with the olfactory lobes, which are situated in the protocerebrum (Schürmann 1995; Mayer et al. 2010).

Tardigrades, or water bears, are represented by ~1,300 described microscopic species that inhabit marine and semi-terrestrial environments and feed on algae or plant and animal cell fluids (Degma et al. 2020). These animals are renowned for their miniaturized body and ability to survive extreme environmental conditions (Clegg 2001; Horikawa et al. 2013; Smith et al. 2016; Gross et al. 2019). Unlike onychophorans and arthropods, they do not possess modified limbs with a clear chemosensory function, which is likely performed by internal structures covered with a cuticle of variable permeability (e.g., Mayer et al. 2013; Møbjerg et al. 2018). The phylogenetic relationships between arthropods, onychophorans, and tardigrades and even the validity of Panarthropoda as a clade are still under debate, although the first two are consistently recovered as sister groups in most molecular phylogenetic analyses (Laumer et al. 2019).

Here, we present a comprehensive comparative genomics analysis across members of the chemosensory gene families of the three major subgroups of Panarthropoda. Our aim is to shed light on the origin and evolution of molecular components of the chemosensory system in these invertebrates and, more specifically, to determine which molecules (or gene families) are responsible for chemoreception in onychophorans and tardigrades and to clarify their evolutionary relationship to those characterized in arthropods. For the analyses, we obtained the specific transcriptomes from the antennae, the head, and the rest of the body of the velvet worm *Euperipatoides rowelli*. We integrated these transcriptomic data with information obtained from publicly available genomic data of this onychophoran species (i5K Consortium 2013; Thomas et al. 2020) and two tardigrades, *Hypsibius exemplaris* (formerly referred to as "*H. dujardini*") and *Ramazzottius varieornatus* (Hashimoto et al. 2016; Koutsovoulos et al. 2016; Yoshida et al. 2017), and with transcriptomic and genomic data from arthropods.

Our results uncovered striking differences in the chemosensory repertoires of panarthropods, including the absence of some key families (which do not only encode chemosensory genes) in specific lineages, and allow a more precise delimitation of their origin. These findings highlight the need for extending molecular studies to taxa that have not received much attention in order to better understand the emergence of major genetic innovations and the diversification of animals.

Results

Novel, Mostly Complete Onychophoran Reference Gene Set

The publicly available draft genome of *E. rowelli* is highly fragmented and largely incomplete; only 43.9% and 47.3% of genes conserved in Eukaryota (Eu) and Metazoa (Mt) (based on BUSCO gene collection; ran under the "genome"

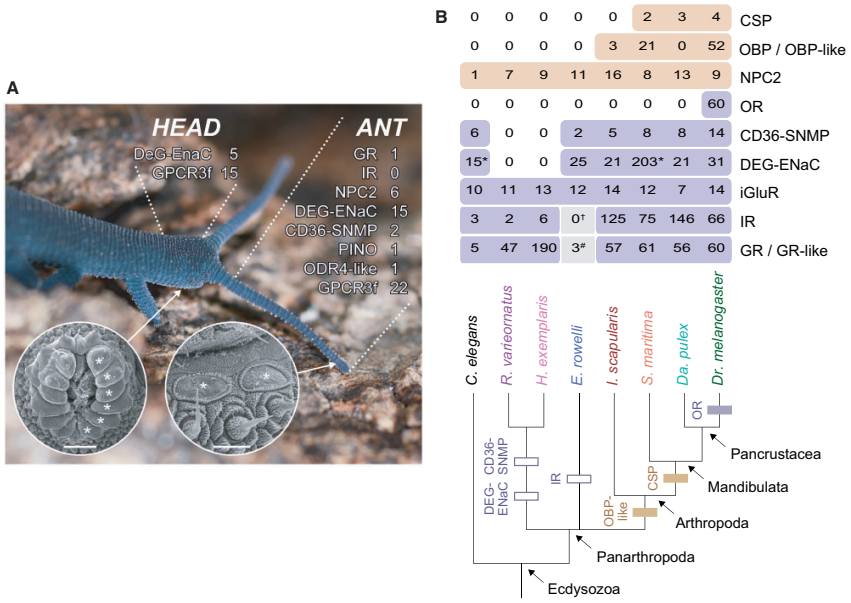


FIG. 1. Chemosensory structures in the onychophoran *Euperipatoides rowelli* and summary of major findings. (*A*) Anterior end of a specimen with anatomical compartments indicated by dotted lines. Insets illustrate scanning electron micrographs of mouth surrounded by lip papillae (asterisks in left micrograph; scale bar: 300 μm) and putative chemosensory organs situated on antennae (asterisks in right micrograph; scale bar: 20 μm). Images provided by Ivo de Sena Oliveira and Christine Martin. Note that chemosensory-related genes are expressed in the anatomical compartments with expected chemosensory function. Numbers refer to those genes specifically or differentially expressed in antenna (*ANT*) and head (*HEAD*). (*B*) Minimum estimates of gene family sizes (S_{MIN}) in the genomes from nine major ecdysozoan lineages (numbers for iGluR and IR subfamilies correspond to complete copies; see Results). Solid and empty colored boxes in the phylogeny indicate gains and losses of particular gene families, respectively. Purple and light-brown shadings denote membrane receptors and soluble proteins, respectively. †Three very short sequences encoding parts of the iGluR/IR LCD (PF00060) that, although they are phylogenetically related to IRs, could not be unambiguously assigned to this subfamily (supplementary fig. S4, Supplementary Material online). #One complete GR receptor and two sequences resulting from partial BLAST hits. *Values obtained after new BITACORA searches in these genomes.

mode; Seppey et al. 2019) are complete, respectively, whereas 30.4% of Eu and 23.3% of Mt genes are missing. On the other hand, the two genome assemblies of tardigrades show good continuity and completeness statistics. These assemblies contain a high proportion of complete genes, ranging from 85% to 95%, and only a few fragmented or missing genes (BUSCO, supplementary table S1, Supplementary Material online). Unlike the genome draft, our deep transcriptome sequencing data from *E. rowelli* (i.e., 60–80 million reads per RNA-seq experiment) allowed to obtain a mostly complete reference gene set (table 1). The final consensus transcriptome of this species consists of 1,072,091 nonredundant transcripts. Although this huge number would indicate that the transcriptome is highly fragmented, several lines of evidence suggest the opposite. On the one hand, the sequencing library was prepared using Ribo-Zero instead of the classical poly-A approach, hence our transcriptome contains all RNAs (after ribosomal RNA depletion), including short and large noncoding RNA transcripts. Indeed, only 8.9% of our consensus transcripts encode putative proteins, the rest being short noncoding sequences (supplementary table S2, Supplementary Material online). On the other hand, we identified all cluster of essential genes (CEG) members, most of them being complete (supplementary table S3, Supplementary Material online), and 99.0% and 99.1% complete BUSCO Eu and Mt genes (using the "transcriptome" mode), respectively; the remaining 1% of BUSCO genes are also present but fragmented (supplementary table S1, Supplementary Material online).

Antennal and Head-Specific Transcriptomes of *E. rowelli*

The consensus transcriptome of *E. rowelli* includes a total of 191,116 candidate protein-coding sequences (encoding 245,070 putative peptides), 95,433 of which are functionally

Table 1. Summary of the Transcriptomic Data Newly Generated for This Study and the Functional Annotation Statistics.

	ANT	*HEAD*	*REST*	Total	Covered by Reads[a]	Protein Coding
Assembled contigs	313,898	640,096	538,450	1,212,132	865,014	245,070
Unique sequences (transcripts)	246,146	541,258	448,126	1,072,091	742,596	191,116
Average length of transcripts (nt)	611	495	527	427	473	490
Longest transcript (nt)	56,010	56,010	56,010	56,010	56,010	55,107
CEG sequences	450	438	440	458	453	458
Sequences with GO annotation	22,514	43,734	41,035	69,901	55,577	69,901
Sequences with functional annotation[b]	29,156	59,247	54,911	95,433	75,060	95,433

[a]Transcripts with mapped reads CEG, cluster of essential genes.
[b]Based on Interpro and BLAST searches (include annotations without GO).

annotated. This number, although still quite high, is similar or even lower to those obtained in the other currently available transcriptomes of this and other onychophoran species (Hering et al. 2012; Mapalo et al. 2020). We found 39,128 genes (20.5%) with detectable expression in the three anatomical compartments. About 8% of the protein-coding transcripts are expressed exclusively in the antenna (*ANT*) (supplementary fig. S1, Supplementary Material online), which is in agreement with the lower number of cells and molecular functions expected in these appendages. Conversely, almost 11% of transcripts are expressed exclusively in the head (*HEAD*). Finally, we found 4,615 (~2.4%) transcripts expressed in these two compartments but not in the rest of the body (*REST*). The differential expression analysis (based on RSEM and DESeq2; Li and Dewey 2011; Love et al. 2014) revealed that 9,129 putative protein-coding transcripts are significantly overexpressed in *ANT*, 351 in *HEAD*, 352 in *REST*, and 6,722 in *HEAD* + *REST*. As expected, we found among the transcripts overexpressed in *ANT* and *HEAD* several gene ontology (GO) terms that are enriched in biological functions associated with the response to chemical and external stimuli (supplementary fig. S2, Supplementary Material online).

The Chemosensory Gene Repertoire in Panarthropoda

We identified 440 sequences encoding putative members of the major arthropod chemosensory families in the onychophoran transcriptome (and genome draft) and the two tardigrade genomes (86 in *E. rowelli*, 266 in *H. exemplaris*, and 88 in *R. varieornatus*) (supplementary table S4, Supplementary Material online). Although most of the chemosensory genes found in tardigrades (352 out of 354) had annotated structural features in the general feature format files (GFF), many of them lacked a fitting functional annotation. Using BITACORA (Vizueta, Sánchez-Gracia, et al. 2020), we were able to annotate (and in some cases curate) as chemosensory genes 310 GFF features previously labeled as hypothetical proteins, and to identify new candidate sequences (two novel genes, one in each species).

Chemoreceptors

The *Gr* family is the largest chemosensory gene family in tardigrades. We identified 192 sequences encoding GR-like proteins (the minimum number of protein-coding sequences that can be unequivocally attributed to different gene family copies, S_{MIN}, was 190, 162 of them encoding complete proteins) and 49 ($S_{MIN} = 47$, 46 complete) in *H. exemplaris* and *R. varieornatus*, respectively (fig. 1B and supplementary table S4, Supplementary Material online). In contrast, we only found three transcripts encoding putative members of this family in *E. rowelli*. One of these copies encoded a complete GR-like member with the protein domain characteristic of this family (7TM chemoreceptor; PF08395). The other two transcripts are short sequences with some similarity to the transmembrane domain of some arthropod GRs. Noticeably, one of them is expressed exclusively in *ANT* (supplementary table S5, Supplementary Material online), whereas the other one might be a pseudogene or an incorrect transcript due to assembly artifacts or sequencing errors. Both this and all the results below obtained from our compartmentalized transcriptome data were qualitatively reproduced when the other three transcriptomic sources of *E. rowelli* and the additional surveyed onychophoran species were used as the subject of our searches (supplementary table S6, Supplementary Material online).

As with arthropod copies of the same family in previously reported gene trees (Eyun et al. 2017; Vizueta et al. 2018), the newly identified tardigrade and onychophoran GR-like sequences form lineage-specific clades in the Panarthropoda tree (fig. 2 and supplementary fig. S3, Supplementary Material online). The presence of two phylogenetically unrelated tardigrade-specific clades, and three onychophoran GRs interspersed with other arthropod copies, would suggest that this family underwent an expansion in the ancestor of panarthropods, followed by a second more recent burst in a tardigrade subclade containing *H. exemplaris* with the loss of most of its members in the onychophoran lineage.

The IR/iGluR gene family is the second largest chemosensory family in the three species surveyed, with 47, 26, and 22 IR/iGluR encoding sequences in *H. exemplaris*, *R. varieornatus*, and *E. rowelli*, respectively (19, 13, and 12 of them are complete; in this case, we calculated S_{MIN} only for the whole family since the copies estimated from partial fragments could not be unambiguously assigned to one of the two subfamilies; see supplementary table S5, Supplementary Material online, for further details). The phylogenetic tree of the ligand-gated ion channel domains (LCDs) of these receptors shows a similar picture to that of the GRs, with the predominance of lineage-specific clades. According to the phylogenetic and OrthoFinder results (supplementary table S7, Supplementary Material online), *H. exemplaris* encodes seven

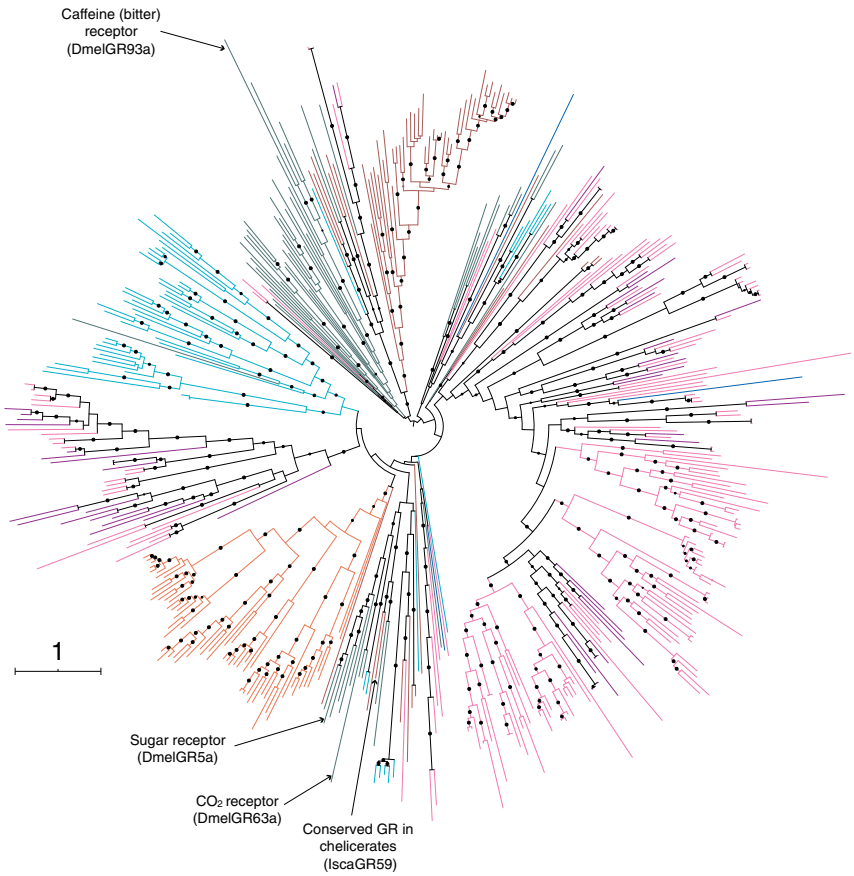


FIG. 2. Maximum likelihood phylogenetic tree of GR family in panarthropods. We excluded all partial proteins and putative pseudogenes and artifacts from the analysis. The color code for species is the same as in figure 1B. Nodes with bootstrap support values >90% are shown as solid circles. Scale bar represents one amino acid substitution per site. (See supplementary figure S3, Supplementary Material online, for a gene tree including all identified sequences.)

Kainate, two AMPA, and five NMDAR receptors, whereas the *R. varieornatus* iGluR repertoire is composed of eight Kainate, and five NMDAR receptors, with no AMPA homolog found in this species. In addition, we identified a candidate homolog of the coreceptor IR25a in both tardigrade species. Based on the phylogenetic relationships of the LCD sequences, tardigrades would encode 27 (*H. exemplaris*) and 11 (*R. varieornatus*) divergent IR proteins, thus predicting a chemosensory function of this family in this animal group. In the case of *E. rowelli*, however, we only found significant evidence for the presence of iGluR members (ten Kainate, one AMPA, and nine NMDAR receptors). Specifically, we identified two antennal expressed sequences encoding partial fragments of an IR/GluR protein that are phylogenetically related to some arthropod divergent IRs; nevertheless, the poor node support and the very short length of the aligned region preclude us from drawing firm conclusions about their subfamily identity (fig. 3 and supplementary figs. S4 and S5, Supplementary Material online). In fact, the remarkable absence of expression (but also of the signal of a gene in the genome draft) of an IR25a homolog in *E. rowelli* could point to a complete loss of this subfamily of ancient chemoreceptors in Onychophora.

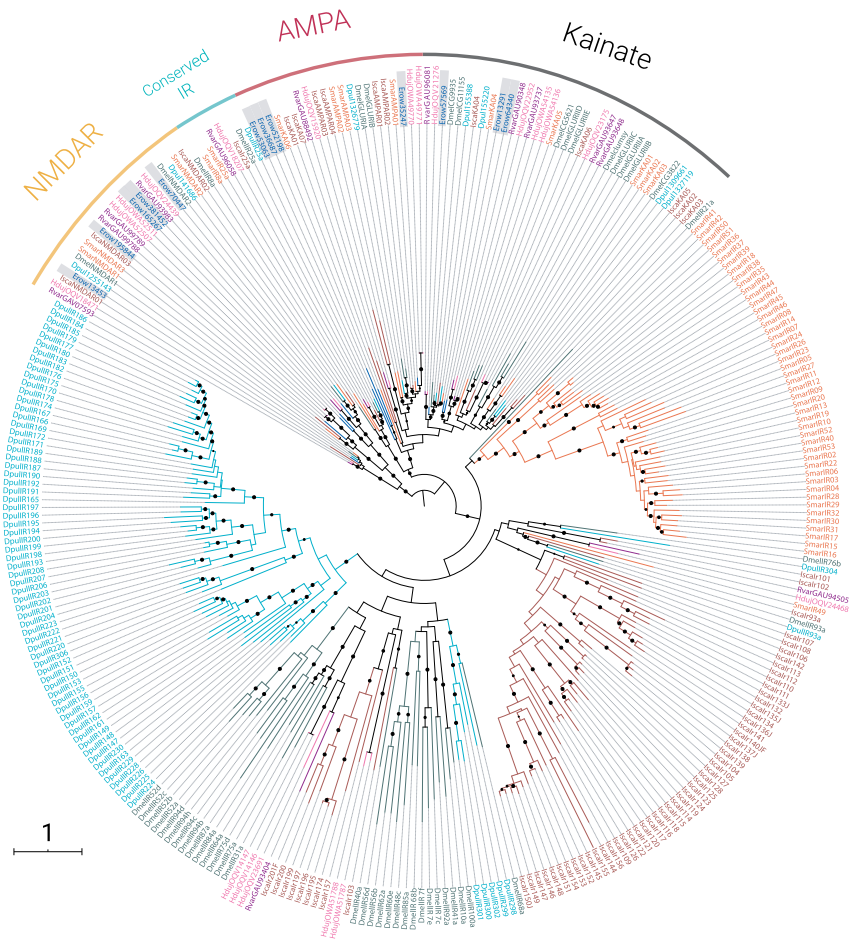


FIG. 3. Maximum likelihood phylogenetic tree of iGluR/IR LCDs (PF00060) in panarthropods. Only complete domains were used for this analysis. The color code for species is the same as in figure 1B. Nodes with bootstrap support values >90% are shown as solid circles. Scale bar represents one amino acid substitution per site. (See supplementary figure S4, Supplementary Material online, for a gene tree including all identified sequences.)

Tardigrades would also lack some highly conserved members of this subfamily occurring across arthropods, such as IR8a, IR93a, and IR76b, suggesting important changes in the chemosensory role played by this subfamily also in water bears.

Other Candidate Chemoreceptors and Related Chemosensory Genes

We identified 48 DEG-ENaC sequences in the transcriptome of *E. rowelli*. Although only ten of them encoded complete receptors, we estimated a $S_{MIN} = 25$ in this species (supplementary tables S5 and S7, Supplementary Material online), a value which is similar to that found in other arthropods (fig. 4). Surprisingly, the two tardigrades do not encode any DEG-ENaC members, suggesting a complete loss of the family. The phylogenetic tree of the DEG-ENaC family in Panarthropoda is also characterized by the presence of large lineage-specific clades, pointing to a similar mode of evolution as for the other surveyed receptors. Interestingly, many of the

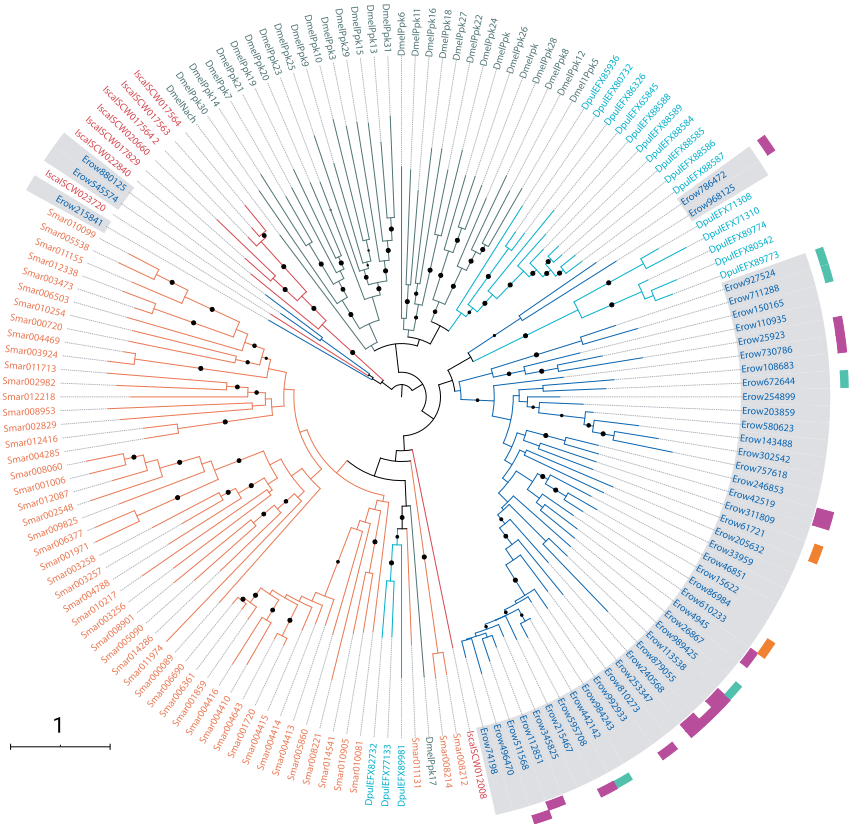


FIG. 4. Maximum likelihood phylogenetic tree of DEG-ENaC family in panarthropods. The color code for species is the same as in figure 1B. Boxes in the outer circle indicate the genes specifically (first layer) or differentially (outer layer) expressed in *ANT* (purple), *HEAD* (green), or both (orange). Nodes with bootstrap support values >90% are shown as solid circles. Scale bar represents one amino acid substitution per site.

members of this family are expressed in *ANT* and/or *HEAD* (19 out of 46 transcripts) of *E. rowelli*, being the family with the greatest number of copies expressed in the chemosensory structures of this species.

Our analysis also uncovered two transcripts encoding CD36-SNMPs in *E. rowelli*, a family phylogenetically related to the SNMPs of arthropods but missing in tardigrades. These transcripts are specific or differentially expressed in *ANT* (supplementary fig. S5 and table S5, Supplementary Material online). We also detected the expression of other genes in *ANT* and *HEAD* of *E. rowelli* that have been related directly or indirectly with arthropod chemosensory activity. For instance, we found 20 antenna-specific copies of the GPCR family 3 of receptors (out of 91 characterized in the whole transcriptome) and homologs of the ODR4-like and Pinocchio proteins with differential expression in the *ANT* compartment (supplementary table S5, Supplementary Material online).

Noticeably, the NPC2 is the only family encoding soluble proteins in tardigrades and onychophorans, which fully lack members of the OBP-like and CSP families. We identified 9, 7, and 11 complete *Npc2* genes in the genomes of *H. exemplaris* and *R. varieornatus* and the transcriptome of *E. rowelli*, respectively (supplementary table S4, Supplementary Material online). These family sizes represent a considerable increase in the number of copies with respect to nonpanarthropod invertebrates, in which this family typically consists of a single gene (Pelosi et al. 2014). These results suggest an expansion of

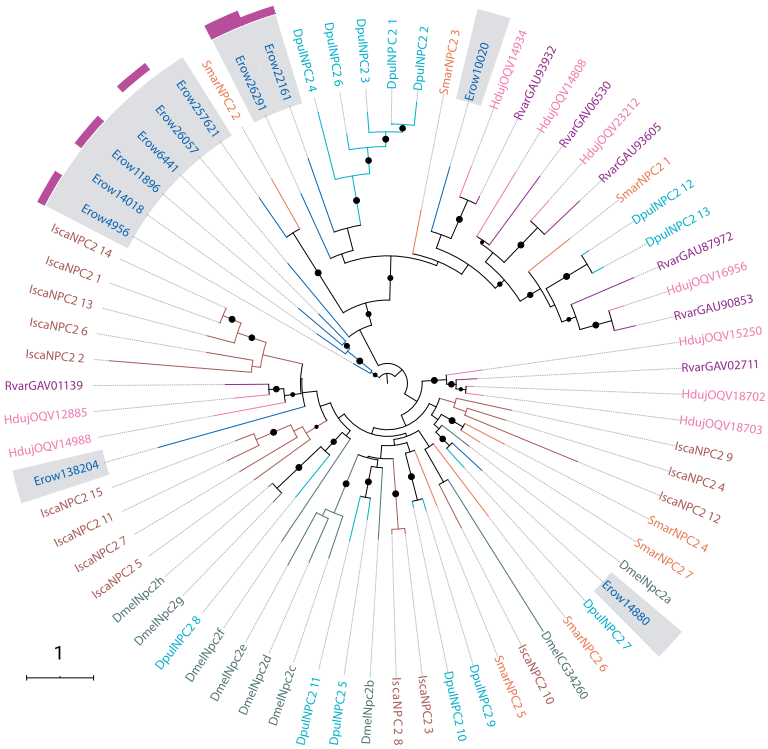


FIG. 5. Maximum likelihood phylogenetic tree of NPC2 family in panarthropods. Boxes in the outer circle indicate the genes specifically (first layer) or differentially (outer layer) expressed in *ANT* (purple), *HEAD* (green), or both (orange). Nodes with bootstrap support values >90% are shown as solid circles. Scale bar represents one amino acid substitution per site.

the NPC2 family in the last common ancestor of Panarthropoda (fig. 5). However, after the expansion, this family shows the lowest turnover rate among the surveyed chemosensory families. It is remarkable that eight NPC2 members are differentially or specifically expressed in the *ANT* compartment of *E. rowelli* (supplementary table S5, Supplementary Material online), indicating a hypothetical chemosensory role of this family in onychophorans.

Discussion

Evidence suggests that arthropods, onychophorans, and tardigrades colonized the land independently after their initial split from an aquatic ancestor (Rota-Stabelli et al. 2013). Similar processes occurred in the three major arthropod groups including pancrustaceans, myriapods, and chelicerates, which originated from an aquatic ancestor 550–450 Ma (Lozano-Fernandez et al. 2016). These terrestrialization events might have impacted many aspects of chemosensory perception in these animals. Nonetheless, the extensive comparative genomics analyses in arthropods have revealed a very similar qualitative chemosensory gene composition in all lineages (i.e., we found the same families in most of them, but see Brand et al. [2018] and Vizueta et al. [2018] for two exceptions), suggesting the presence of these proteins in the last common ancestor of Arthropoda. The analysis of representative species across bilaterians revealed that GRs and IRs involved in arthropod chemoreception might have originated through the co-option of ancient gustatory receptor-like (*Gr*-like) and ionotropic glutamate receptor (*iGlur*) genes, respectively (Croset et al. 2010; Krishnan et al. 2014; Robertson 2019). GR-like proteins might have already been present in the last common ancestor of metazoans, as they have been identified in many animal lineages (Robertson 2015; Eyun et al. 2017), and its ancestral function is still under debate (Robertson 2019). Similarly, the origin of the

chemosensory IRs, including the coreceptor IR25a and other divergent sequences that evolved independently in different lineages, has been dated back to the protostome ancestor (Croset et al. 2010; Eyun et al. 2017). Although functional evidence of the participation of these proteins in chemoreception comes from studies of insects, various tissue-specific transcriptomes from crustaceans, myriapods, and spiders have confirmed the specific or preferential expression of GR and divergent IR genes in the chemosensory structures of crustaceans (Kozma et al. 2020), spiders (Vizueta et al. 2017), and centipedes (Frías-López C, unpublished data).

Intriguingly, we found that *E. rowelli* encodes an exceptionally low number of GR-like members, the lowest reported from panarthropods, and we did not find any trace of the highly conserved coreceptor IR25a. These results are likely not caused by a lack of sensitivity since the sequencing depth of our transcriptomes should be sufficient for detecting lowly expressed genes even in the antenna, as we performed nine independent RNA-seq experiments, enriched in compartment-specific transcripts and each yielding between 60 and 80 million reads. Furthermore, we have corroborated all these findings in the full body transcriptomes from six onychophoran species (including representatives of Peripatidae and Peripatopsidae; supplementary table S6, Supplementary Material online). In fact, the marginal GR-like repertoire size detected in all these species is similar to that observed in *Caenorhabditis elegans* (and other major nematode clades; see supplementary table S6, Supplementary Material online), in which a nonchemosensory role has been established (Bhatla and Horvitz 2015). Similarly, the absence of an IR25a homolog and the doubtful presence of divergent IRs in these transcriptomes would also certainly question the role of this family in velvet worm chemoreception. Still, these results have to be approached with caution due to high fragmentation of the surveyed transcriptomes and the unavailability of well-assembled, complete genome sequences. Besides, the picture is further complicated by the observation that at least one of the two putatively functional onychophoran GR-like copies is expressed in the antenna of *E. rowelli*, precluding us from drawing a firm conclusion. Thus, it remains uncertain whether these few members of the GR-like lineage have, or ever had, a chemosensory function in onychophorans.

In light of these remarkable absences, some members of the DEG-ENaC or other receptor families, such as the GPCR family 3 or the TRP channels, might have a chemosensory function in *E. rowelli*. In fact, we detected in the antenna of this species the expression of a candidate homolog of the *C. elegans odr-4* gene, which is required for localizing a subset of odorant GPCRs in the cilia of olfactory neurons of this nematode (Dwyer et al. 1998). In *C. elegans*, the olfactory receptors are synthesized in the endoplasmic reticulum of the olfactory neurons, trafficked to the cell surface membrane, and transported to the tip of the olfactory cilium, where they bind to odorants. Interestingly, the chemoreceptors of onychophorans, which are situated on the antennal tip and covered with a specialized thin cuticle (fig. 1A), also contain receptor cells with branched cilia (Storch and Ruhberg 1977), suggesting that the onychophoran *odr-4* homolog might be expressed in these cells. TRP channels are highly conserved nonvoltage gated, cation channels with a role in insect thermosensation and mechanosensation (Venkatachalam and Montell 2007) that have been attributed to gustation and repellency (Fowler and Montell 2013). We have conducted a prospective search for members of this family in the tardigrades *H. exemplaris* and *R. varieornatus* and the onychophoran *E. rowelli*, detecting a noticeable number of gene copies. From the 67 good-quality annotated onychophoran copies, two are antenna specific, and 11 show differential expression in this compartment (supplementary table S8, Supplementary Material online); at least three of these TRP candidates show remote sequence similarity with members of the TRPA (1) and TRPM (2) subfamilies, which are involved in nociception in insects and taste and cold perception in mammals (Matsuura et al. 2009; Kang et al. 2010: Kwon et al. 2010).

On the other hand, we have found that the two tardigrade genomes could have completely lost the DEG-ENaC family, a group of metazoan-specific membrane proteins that play a role in salt taste, mechanoreception and chemoreception, among other functions (Chen et al. 2010; Ben-Shahar 2011; Lu et al. 2012), and are present in other panarthropods, including in the antenna and the head of the onychophoran *E. rowelli*. It is largely known that tardigrades are organisms extraordinarily resistant to extreme conditions, with unique features among metazoans such as surviving in space, enduring very high pressures and radiation, or surviving extreme temperatures or prolonged desiccation (Møbjerg et al. 2011; Rebecchi et al. 2011; Fernandez et al. 2016; Hashimoto et al. 2016; Hering et al. 2016; Tsujimoto et al. 2016). Recent comparative genomics and transcriptomics studies in these animals have uncovered frequent losses and expansions in stress-related gene pathways, although affecting independent genes in different lineages (Yoshida et al. 2017; Kamilari et al. 2019). Interestingly, in the fruit fly *Drosophila melanogaster*, some DEG-ENaC proteins are involved in maintaining osmotic and intestinal stem cell homeostasis (Kim et al. 2017), regulating the neuronal response to heat stress (Zheng et al. 2014) or are the target of mutants with lethal desiccation phenotypes of larvae (Johnson and Carder 2012). Moreover, the loss of function of one member of this family extends lifespan and health span, increases internal water stores due to the loss of the ability to sense external water, and exhibits significantly increased survivorship under desiccating conditions (Waterson et al. 2014). In fact, additional searches in six publicly available transcriptomes of different tardigrade species have confirmed the absence of this family in eutardigrades but not in heterotardigrades (supplementary table S6, Supplementary Material online), suggesting that the loss of DEG-ENaC family could be part of a lineage-specific adaptation of eutardigrades to survive in extreme environments. The putative loss of the CD36/SNMP family in eutardigrades is another interesting finding, as members of this conserved family play important sensory, digestive, and immune system roles in *D. melanogaster* (Nichols and Vogt 2008; Vogt et al. 2009).

Our study also revealed that NPC2 members are the only soluble proteins present in tardigrades and onychophorans, some of which are specifically or differentially expressed in the antenna of the *E. rowelli*. This result points to this family as the only panarthropod candidate to perform functions like those documented from arthropod soluble proteins. Overall, our findings, when integrated with previous studies on several arthropod lineages (Eyun et al. 2017; Vizueta et al. 2018), point to at least three evolutionarily independent co-options from ancestral, nonchemosensory soluble protein families, to further participate in chemoreception (fig. 1B), namely in the last common ancestors of 1) Panarthropoda (NPC2; members of this family are involved in the metabolism of cholesterol in *C. elegans*; Sym et al. 2000), 2) Arthropoda (OBP), and 3) Mandibulata (CSP). These staggered co-options might have served to progressively adapt peripheral chemoreception to the new chemical world.

Finally, it is worth noting that, if the true phylogeny of Panarthropoda was different from that considered here, the origin and evolutionary history of some of these families would be quite different. If we consider, for example, the phylogenetic hypothesis placing Tardigrada as sister to Nematoda (e.g., Yoshida et al. 2017; Arakawa 2018; Laumer et al. 2019), onychophorans would have lost ionotropic receptors and would have never had chemosensory GR-like proteins. In this case, the insect-type gustatory receptors would have appeared in two (or more) independent GR-like expansions in tardigrades and arthropods. Alternatively, GR-like genes would have been present in the last common ancestor of Panarthropoda but lost in velvet worms and nematodes. Nonetheless, it is worth noting that many of our conclusions are based on the lack of evidence in similarity-based searches, many of them in transcriptomic data and, therefore, must be considered with caution. Further broader taxonomic studies including complete genome assemblies, currently unavailable, and supported by functional evidence will be needed to confirm the striking absences found in this study and to determine their actual biological meaning.

Taken together, our findings shed light on the diversification of members of the chemosensory gene families across Panarthropoda, including hypothesized origin of some of the surveyed families (fig. 1B). We have found considerable differences in the chemosensory repertoires of panarthropods, including striking absences in specific lineages, which vindicates the importance of conducting evolutionary genomics studies on the closest arthropod relatives, such as onychophorans and tardigrades. Paradoxically, these clades have not received much attention since the beginning of the genomics era, although they might be crucial for understanding the emergence and diversification of major evolutionary innovations in arthropods.

Materials and Methods

Specimens

Specimens of *E. rowelli*, Reid, 1996 (Onychophora, Peripatopsidae) were obtained from decaying logs in the Tallaganda State Forest (New South Wales, Australia; 35°28′S, 149°32′E, 954 m) in October 2011 and January 2013. They were collected under the permit numbers SL100159 and SL101720 issued by the National Parks & Wildlife Service New South Wales and exported under the permit numbers PWSP104061 and PWSP208163 provided by the Department of Sustainability, Environment, Water, Population and Communities. The collected specimens were maintained in the laboratory as described previously (Baer and Mayer 2012).

Genome Data

The genome sequences, annotations, and predicted proteins of two tardigrade species, *H. exemplaris* (v3.5.1, Ensembl Tardigrades Genomes) (Koutsovoulos et al. 2016) and *R. varieornatus* (Rv101, Ensembl Tardigrades Genomes) (Hashimoto et al. 2016; Yoshida et al. 2017), and the draft assembly of the onychophoran *E. rowelli*, sequenced as part of the i5K initiative (i5K Consortium 2013; Thomas et al. 2020), were retrieved from http://ensembl.tardigrades.org and https://www.hgsc.bcm.edu/arthropods/velvet-worm-genome-project), respectively. Note that *H. exemplaris* was commonly referred to as "*H. dujardini*" before its formal description by Gąsiorek et al. (2018).

Transcriptome Data

Samples

We used four different sources of transcriptome data of *E. rowelli*. The first was obtained in our tissue-specific transcriptome sequencing experiment of three juvenile individuals (representing three biological replicates). The other three consisted in the raw data of two whole individual RNA-seq experiments (one female [ER9] and one male [ER10]) retrieved from Baylor i5K Initiative Pilot Project (HGSC) (accession numbers: SRX973445 and SRX973444, for the female and male, respectively), and the transcriptome assembly of the *E. rowelli* sample used in Hering et al. (2012).

RNA Extraction, Library Preparation, and Sequencing

We generated new transcriptomics data from three *E. rowelli* juvenile individuals (supplementary table S9, Supplementary Material online). This species does not show sexual dimorphism with respect to the structure of antennae and chemoreceptors, and juveniles are active hunters shortly after birth. For each individual, we built three separate RNA-seq libraries: the antenna (*ANT*; ensuring that the cut was below the antennal rings with chemoreceptors), the head (*HEAD*; butting behind the slime papillae), and the rest of the body (*REST*), henceforth referred to as anatomical compartments (fig. 1A). All dissections were performed after snap-freezing individuals in liquid nitrogen, which were starved for 1 week in the laboratory.

The small amount of tissue (and therefore of total RNA) contained in the antennae of a single individual led us to consider a specific extraction protocol specially designed for small amounts of starting material. For *ANT*, we used the PicoPure RNA Isolation Kit (Arcturus, Applied Biosystems, USA) and TRIzol reagent (Invitrogen, Waltham, MA),

especially designed to consistently recover high-quality total RNA from fewer cells. In the case of *HEAD* and *REST*, where the amount of tissue was not a limiting factor, we used the RNeasy Mini kit (Qiagen, Venlo, the Netherlands) and TRIzol reagent (Invitrogen). In addition, *ANT* RNA was amplified with RiboAmp HS PLUS Kit (Arcturus) to obtain the necessary amount for sequencing (two amplification rounds). We determined the amount and integrity of RNA using a Qubit Fluorometer (Life Technologies, Grand Island, NY) and an Agilent 2100 Bioanalyzer (CCiTUB, Barcelona, Spain), respectively. All library preparation steps and RNA sequencing were carried out in Macrogen Inc., Seoul, South Korea. Briefly, ribosomal RNA was depleted with Ribo-Zero Kit and fragmented into small pieces. Double-stranded cDNA was synthesized with random hexamer (N6) primers (Illumina) and Illumina PE adapters were ligated to the ends of adenylated cDNA fragments. The nine transcriptomes (e.g., three anatomical compartments in three biological replicates) were sequenced independently in the HiSeq 4000 system (100-bp paired-end reads) according to the manufacturer's instructions (Illumina, San Diego, CA).

Data Preprocessing and Transcriptome Assembly

We used NGSQCToolkit (Patel and Jain 2012) to filter low quality reads (raw reads with more than 30% of bases with quality scores <20) from raw data. Filtered reads were further corrected for sequencing errors with the program SEECER v_0.1.3 (Le et al. 2013). We generated a consensus transcriptome from the entire collection of reads of all individuals and anatomical compartments, which was used as the reference for differential expression analyses. We assembled these consensus reference transcripts using Bridger (k-mer size $= 31$; Chang et al. 2015). All contigs with contaminant sequences, that is, those matching the UniVec vector database, the genomes of *Escherichia coli*, *Pseudomonas aeruginosa*, *Staphylococcus aureus*, *Saccharomyces cerevisiae*, and *Homo sapiens* were removed from this reference using the software Seqclean (https://sourceforge.net/projects/seqclean/). Finally, clean contigs were clustered into putative transcripts defined in the assembly (analogous to the Trinity gene-isoform nomenclature).

Functional Annotation of the *E. rowelli* Transcriptome

We carried out exhaustive BLAST searches (E-value $= 10^{-5}$; using the assembled reference transcripts as a query) against NCBI-nr, Swiss-Prot, and an updated version of the ArthropodDB database (see Vizueta et al. 2019 for further details); for this study, we included in the latter database the predicted proteins and functional annotations of the tardigrade *H. exemplaris* and the nematode *C. elegans* (C. elegans Sequencing Consortium 1998; Yoshida et al. 2017). Coding sequences and their conceptual translations were inferred from TransDecoder results (Haas et al. 2013). In addition, we searched the predicted proteins for specific domain signatures with InterProScan (Jones et al. 2014), and signal peptides and transmembrane domains were predicted using SignalP and TMHMM, respectively (Krogh et al. 2001; Petersen et al. 2011).

GO terms (Ashburner et al. 2000) were inherited from the results of the BLAST and InterProScan searches (we used the top five positive hits with an E-value $<10^{-5}$). We also identified the KEGG enzymes and pathways (Kanehisa and Goto 2000), CEG members (Core Eukaryotic Genes; Parra et al. 2007, 2009), and Metazoa and Eukaryota conserved genes included in BUSCO v3.1.0 (Seppey et al. 2019).

Identification and Annotation of Chemosensory Families

We used the bioinformatics pipeline BITACORA (Vizueta, Escuer, et al. 2020; Vizueta, Sánchez-Gracia, et al. 2020) to identify and annotate the members of the major arthropod chemosensory gene families (CS) in the surveyed transcriptomes and genomes. We first built a database for each of the focal gene families, hereinafter the olfactory (OR), gustatory (GR), ionotropic (iGLuR/IR/*Ir*), and epithelial sodium channel (DEG/ENaC) receptor families, the genes encoding the CSP, OBP (this family also included the *Obp-like* family recently identified in Vizueta et al. [2017]), and Niemann-Pick C2 (NPC2) soluble protein families, and the genes encoding the SNMP family. Each database included the protein sequences of these families from different arthropod lineages, obtained from the literature (Croset et al. 2010; Colbourne et al. 2011; Vieira and Rozas 2011; Chipman et al. 2014; Robertson 2015; Gulia-Nuss et al. 2016; Vizueta et al. 2018), and was used to construct an HMM profile for each family. In addition, the repertoire of *D. melanogaster*, *Daphnia pulex*, *Strigamia maritima*, and *Ixodes scapularis* was selected as representative of each major arthropod lineage for figure 1 and to build Panarthropoda gene family trees. We also included in our searches representative members of the chemosensory families from other organisms, including OBPs and olfactory and taste receptors of vertebrates (InterPro signatures IPR002448, IPR000725, and IPR007960, respectively; see supplementary table S1B in Frías-López et al. [2015]), and *C. elegans* serpentine receptors, known to be involved in nematode chemoreception (Vidal et al. 2018). In each family specific search, we ran two iterative rounds of BITACORA. Specifically, we used the *full mode* on tardigrade genomes (taking advantage of existing GFF annotations), the *genome mode* on *E. rowelli* genomic draft, where no structural annotation is available, and the *protein mode* on the predicted peptides from the *E. rowelli* compartmentalized transcriptome.

All sequences identified in our searches as possible members of one of the focal CS families were classified in different categories based on the structural and functional criteria applied in Vizueta et al. (2018). Briefly, all coding sequences with premature stop codons were classified as nonfunctional or erroneous copies; this category would include putative pseudogenes, genes with sequencing errors or assembly artifacts. Among the remaining proteins, we distinguished between complete (>80% of the average length of the family protein domain) and incomplete copies; in addition, and only for the *Gr* and *iGluR/Ir* families, we required complete copies to contain a minimum of five of the seven-transmembrane domains (predicted with the software TMHMM version 2.0c; Krogh et al. 2001, and Phobius version 1.01; Käll et al. 2004) or the

presence of the LCD (Pfam identifier PF00060; a domain present in all subfamilies; Croset et al. 2010), respectively. We estimated the minimum number of different copies of a family, S_{MIN}, as in Vizueta et al. (2018) (this value can be interpreted as an estimate of the actual number of family copies in this species). All proteins identified in this study are provided in the Supplementary Material online.

Finally, we also performed some prospective searches to determine the presence of members of the chemosensory families in other publicly available sequences, including five additional nematode genomes and the transcriptome data from six onychophoran and six tardigrade species (see supplementary table S6, Supplementary Material online, for a detailed list of the species and sources used for these searches).

Expression Profiling in *E. rowelli*

We mapped the preprocessed reads of each individual and anatomical compartment back to the consensus reference transcriptome using Bowtie2 version 2.2.3 (set as default; Langmead and Salzberg 2012). We used RSEM 1.2.19 software to obtain read counts and TMM-normalized FPKM (Li and Dewey 2011). For the analysis, we considered that a gene is expressed when the FPKM value is higher than 0.01, a reasonable cutoff given the low expression levels reported for other arthropod chemosensory genes (Zhang et al. 2014). The differential gene expression analysis across anatomical compartments was conducted with DESeq2 (Love et al. 2014) considering the three sequenced individuals (per anatomical compartment) as biological replicates and adjusting the *P* values for the false discovery rate (Benjamini and Hochberg 1995).

Phylogenetic Analyses

We built a multiple sequence alignment per each focal CS family using MAFFT ("–auto" option; Katoh and Standley 2013) and used IQ-TREE version 1.6.5 to estimate the best fit substitution models and gene family trees (Nguyen et al. 2015). Node support was estimated from 1,000 ultrafast bootstrap replicates (Hoang et al. 2018). Tree images were drawn using the iTOL web server (Letunic and Bork 2007, 2019). We also assessed the orthologous relationships of some of the surveyed chemosensory gene family members using OrthoFinder v2.2.7 with default options (Emms and Kelly 2015, 2019).

GO Enrichment

We used R and GOstats to carry out a GO enrichment analysis (Falcon and Gentleman 2007) and REVIGO to generate a graphical representation of the results (Supek et al. 2011). We also used Blast2GO suite (Conesa et al. 2005; Götz et al. 2008) to identify KEGG pathways enriched in the list of candidates (Kanehisa and Goto 2000).

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

We are thankful to members of the Mayer laboratory for their support with animal husbandry. We gratefully acknowledge Dave M. Rowell, Ivo de Sena Oliveira, Sandra Treffkorn, Franziska Anni Franke, and Michael Gerth for their assistance with collecting the specimens and Noel N. Tait for his help with permits. Ivo de Sena Oliveira and Christine Martin kindly provided images of *Euperipatoides rowelli* for figure 1A. The staffs of the National Parks & Wildlife Service New South Wales (Australia) and the Department of Sustainability, Environment, Water, Population and Communities (Australia) are gratefully acknowledged for providing the collection and export permits. A.S.-G. is a Serra Húnter Fellow. This work was supported by the Ministerio de Economía y Competitividad of Spain (CGL2013-45211 and CGL2016-75255) and the Comissió Interdepartamental de Recerca I Innovació Tecnològica of Catalonia, Spain (2017SGR1287). J.V. and P.E. were supported by an FPI grant (Ministerio de Economía y Competitividad of Spain, BES-2014-068437 and BES-2017-081740, respectively). G.M. received support from the German Research Foundation (DFG: MA 4147/10-1).

References

Ache BW, Young JM. 2005. Olfaction: diverse species, conserved principles. *Neuron* 48(3):417–430.

Arakawa K. 2018. The complete mitochondrial genome of *Echiniscus testudo* (Heterotardigrada: Echiniscidae). *Mitochondrial DNA Part B* 3(2):810–811.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet.* 25(1):25–29.

Baer A, Mayer G. 2012. Comparative anatomy of slime glands in Onychophora (velvet worms). *J Morphol.* 273(10):1079–1088.

Benjamini YH, Hochberg Y. 1995. Controlling the false discovery rate—a practical and powerful approach to multiple testing. *J R Stat Soc.* 57(1):289–300.

Ben-Shahar Y. 2011. Sensory functions for degenerin/epithelial sodium channels (DEG/ENaC). *Adv Genet.* 76:1–26.

Bhatla N, Horvitz HR. 2015. Light and hydrogen peroxide inhibit *C. elegans* feeding through gustatory receptor orthologs and pharyngeal neurons. *Neuron* 85(4):804–818.

Brand P, Robertson HM, Lin W, Pothula R, Klingeman WE, Jurat-Fuentes JL, Johnson BR. 2018. The origin of the odorant receptor gene family in insects. *Elife* 7:e38340.

C. elegans Sequencing Consortium. 1998. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* 282:2012–2018.

Chang Z, Li G, Liu J, Zhang Y, Ashby C, Liu D, Cramer CL, Huang X. 2015. Bridger: a new framework for de novo transcriptome assembly using RNA-seq data. *Genome Biol.* 16(1):1–10.

Chen Z, Wang Q, Wang Z. 2010. The amiloride-sensitive epithelial Na+ channel PPK28 is essential for *Drosophila* gustatory water reception. *J. Neurosci.* 30(18):6247–6252.

Chipman AD, Ferrier DEK, Brena C, Qu J, Hughes DST, Schröder R, Torres-Oliva M, Znassi N, Jiang H, Almeida FC, et al. 2014. The first myriapod genome sequence reveals conservative arthropod gene content and genome organisation in the centipede *Strigamia maritima*. *PLoS Biol.* 12(11):e1002005.

Clegg JS. 2001. Cryptobiosis—a peculiar state of biological organization. *Comp Biochem Physiol B Biochem Mol Biol.* 128(4):613–624.

Colbourne JK, Pfrender ME, Gilbert D, Thomas WK, Tucker A, Oakley TH, Tokishita S, Aerts A, Arnold GJ, Basu MK, et al. 2011. The ecoresponsive genome of *Daphnia pulex*. *Science* 331(6017):555–561.
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(18):3674–3676.
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 Genet*. 6(8):e1001064.
Degma P, Bertolani R, Guidetti R. 2020. Actual checklist of *Tardigrada* species. 48 pp. Available from: http://www.tardigrada.modena.unimo.it/miscellanea/Actual%20checklist%20of%20Tardigrada.pdf (accessed February 29, 2020).
Dwyer ND, Troemel ER, Sengupta P, Bargmann CI. 1998. Odorant receptor localization to olfactory cilia is mediated by ODR-4, a novel membrane-associated protein. *Cell* 93(3):455–466.
Emms DM, Kelly S. 2015. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol*. 16(1):157.
Emms DM, Kelly S. 2019. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol*. 20(1):238.
Eyun S, Soh HY, Posavi M, Munro JB, Hughes DST, Murali SC, Qu J, Dugan S, Lee SL, Chao H, et al. 2017. Evolutionary history of chemosensory-related gene families across the Arthropoda. *Mol Biol Evol*. 34(8):1838–1862.
Falcon S, Gentleman R. 2007. Using GOstats to test gene lists for GO term association. *Bioinformatics* 23(2):257–258.
Fernandez C, Vasanthan T, Kissoon N, Karam G, Duquette N, Seymour C, Stone JR. 2016. Radiation tolerance and bystander effects in the eutardigrade species *Hypsibius dujardini* (Parachela: Hypsibiidae). *Zool J Linn Soc*. 178(4):919–923.
Fowler MA, Montell C. 2013. *Drosophila* TRP channels and animal behavior. *Life Sci*. 92(8–9):394–403.
Frías-López C, Almeida FC, Guirao-Rico S, Vizueta J, Sánchez-Gracia A, Arnedo MA, Rozas J. 2015. Comparative analysis of tissue-specific transcriptomes in the funnel-web spider *Macrothele calpeiana* (Araneae, Hexathelidae). *PeerJ* 3:e1064.
Gąsiorek P, Stec D, Morek W, Michalczyk Ł. 2018. An integrative redescription of *Hypsibius dujardini* (Doyère, 1840), the nominal taxon for *Hypsibioidea* (Tardigrada: Eutardigrada). *Zootaxa* 4415(1):45–75.
Giribet G, Edgecombe GD. 2017. Current understanding of Ecdysozoa and its internal phylogenetic relationships. *Integr Comp Biol*. 57(3):455–466.
Götz S, García-Gómez JM, Terol J, Williams TD, Nagaraj SH, Nueda MJ, Robles M, Talón M, Dopazo J, Conesa A. 2008. High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Res*. 36(10):3420–3435.
Gross V, Treffkorn S, Reichelt J, Epple L, Lüter C, Mayer G. 2019. Miniaturization of tardigrades (water bears): morphological and genomic perspectives. *Arthropod Struct Dev*. 48:12–19.
Gulia-Nuss M, Nuss AB, Meyer JM, Sonenshine DE, Roe RM, Waterhouse RM, Sattelle DB, de la Fuente J, Ribeiro JM, Megy K, et al. 2016. Genomic insights into the *Ixodes scapularis* tick vector of Lyme disease. *Nat Commun*. 7:10507.
Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, et al. 2013. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc*. 8(8):1494–1512.
Hashimoto T, Horikawa DD, Saito Y, Kuwahara H, Kozuka-Hata H, Shin IT, Minakuchi Y, Ohishi K, Motoyama A, Aizu T, et al. 2016. Extremotolerant tardigrade genome and improved radiotolerance of human cultured cells by tardigrade-unique protein. *Nat Commun*. 7.
Hering L, Bouameur JE, Reichelt J, Magin TM, Mayer G. 2016. Novel origin of lamin-derived cytoplasmic intermediate filaments in tardigrades. *Elife* 5:e11117.
Hering L, Henze MJ, Kohler M, Kelber A, Bleidorn C, Leschke M, Nickel B, Meyer M, Kircher M, Sunnucks P, et al. 2012. Opsins in Onychophora (velvet worms) suggest a single origin and subsequent diversification of visual pigments in arthropods. *Mol Biol Evol*. 29(11):3451–3458.
Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. 2018. UFBoot2: improving the ultrafast bootstrap approximation. *Mol Biol Evol*. 35(2):518–522.
Horikawa DD, Cumbers J, Sakakibara I, Rogoff D, Leuko S, Harnoto R, Arakawa K, Katayama T, Kunieda T, Toyoda A, et al. 2013. Analysis of DNA repair and protection in the tardigrade *Ramazzottius varieornatus* and *Hypsibius dujardini* after exposure to UVC radiation. *PLoS One* 8(6):e64793.
i5K Consortium. 2013. The i5K initiative: advancing arthropod genomics for knowledge, human health, agriculture, and the environment. *J Hered*. 104:595–600.
Janssen R. 2017. Comparative analysis of gene expression patterns in the arthropod labrum and the onychophoran frontal appendages, and its implications for the arthropod head problem. *EvoDevo* 8:1.
Johnson WA, Carder JW. 2012. *Drosophila* nociceptors mediate larval aversion to dry surface environments utilizing both the painless TRP channel and the DEG/ENaC subunit, PPK1. *PLoS One* 7(3):e32878.
Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, et al. 2014. InterProScan 5: genome-scale protein function classification. *Bioinformatics* 30(9):1236–1240.
Joseph RM, Carlson JR. 2015. *Drosophila* chemoreceptors: a molecular interface between the chemical world and the brain. *Trends Genet*. 31(12):683–695.
Käll L, Krogh A, Sonnhammer ELL. 2004. A combined transmembrane topology and signal peptide prediction method. *J Mol Biol*. 338(5):1027–1036.
Kamilari M, Jørgensen A, Schiøtt M, Møbjerg N. 2019. Comparative transcriptomics suggest unique molecular adaptations within tardigrade lineages. *BMC Genomics*. 20(1):607.
Kanehisa M, Goto S. 2000. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*. 28(1):27–30.
Kang K, Pulver SR, Panzano VC, Chang EC, Griffith LC, Theobald DL, Garrity PA. 2010. Analysis of *Drosophila* TRPA1 reveals an ancient origin for human chemical nociception. *Nature* 464(7288):597–600.
Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*. 30(4):772–780.
Kim K, Hung RJ, Perrimon N. 2017. miR-263a regulates ENaC to maintain osmotic and intestinal stem cell homeostasis in *Drosophila*. *Dev Cell* 40(1):23–36.
Koutsovoulos G, Kumar S, Laetsch DR, Stevens L, Daub J, Conlon C, Maroon H, Thomas F, Aboobaker AA, Blaxter M. 2016. No evidence for extensive horizontal gene transfer in the genome of the tardigrade *Hypsibius dujardini*. *Proc Natl Acad Sci U S A*. 113(18):5053–5058.
Kozma MT, Ngo-Vu H, Wong YY, Shukla NS, Pawar SD, Senatore A, Schmidt M, Derby CD. 2020. Comparison of transcriptomes from two chemosensory organs in four decapod crustaceans reveals hundreds of candidate chemoreceptor proteins. *PLoS One* 15(3):e0230266.
Krishnan A, Almén MS, Fredriksson R, Schiöth HB. 2014. Insights into the origin of nematode chemosensory GPCRs: putative orthologs of the srw family are found across several phyla of protostomes. *PLoS One* 9(3):e93048.
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(3):567–580.
Kwon Y, Kim SH, Ronderos DS, Lee Y, Akitake B, Woodward OM, Guggino WB, Smith DP, Montell C. 2010. *Drosophila* TRPA1 channel is required to avoid the naturally occurring insect repellent citronellal. *Curr Biol*. 20(18):1672–1678.
Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods*. 9(4):357–359.
Laumer CE, Fernández R, Lemer S, Combosch D, Kocot KM, Riesgo A, Andrade SCS, Sterrer W, Sørensen MV, Giribet G. 2019. Revisiting metazoan phylogeny with genomic sampling of all phyla. *Proc Biol Sci*. 286(1914):20191941.

Le H-S, Schulz MH, McCauley BM, Hinman VF, Bar-Joseph Z. 2013. Probabilistic error correction for RNA sequencing. *Nucleic Acids Res*. 41(10):e109.
Leal WS. 2013. Odorant reception in insects: roles of receptors, binding proteins, and degrading enzymes. *Annu Rev Entomol*. 58(1):373–391.
Letunic I, Bork P. 2007. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics* 23(1):127–128.
Letunic I, Bork P. 2019. Interactive Tree Of Life (iTOL) *v4*: recent updates and new developments. *Nucleic Acids Res*. 47:256–259.
Li B, Dewey CN. 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinf*. 12(1):323.
Liu L, Leonard AS, Motto DG, Feller MA, Price MP, Johnson WA, Welsh MJ. 2003. Contribution of *Drosophila* DEG/ENaC genes to salt taste. *Neuron* 39(1):133–146.
Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. 15(12):550.
Lozano-Fernandez J, Carton R, Tanner AR, Puttick MN, Blaxter M, Vinther J, Olesen J, Giribet G, Edgecombe GD, Pisani D. 2016. A molecular palaeobiological exploration of arthropod terrestrialization. *Philos Trans R Soc B Biol Sci*. 371.
Lu B, LaMora A, Sun Y, Welsh MJ, Ben-Shahar Y. 2012. ppk23-dependent chemosensory functions contribute to courtship behavior in *Drosophila melanogaster*. *PLoS Genet*. 8(3):e1002587.
Mapalo MA, Arakawa K, Baker CM, Persson DK, Mirano-Bascos D, Giribet G. 2020. The unique antimicrobial recognition and signaling pathways in tardigrades with a comparison across Ecdysozoa. *G3 (Bethesda)* 10:1137–1148.
Martin C, Gross V, Hering L, Tepper B, Jahn H, Oliveira IS, Stevenson PA, Mayer G. 2017. The nervous and visual systems of onychophorans and tardigrades: learning about arthropod evolution from their closest relatives. *J Comp Physiol A* 203(8):565–590.
Martin C, Mayer G. 2014. Neuronal tracing of oral nerves in a velvet worm—implications for the evolution of the ecdysozoan brain. *Front Neuroanat*. 8:7.
Matsuura H, Sokabe T, Kohno K, Tominaga M, Kadowaki T. 2009. Evolutionary conservation and changes in insect TRP channels. *BMC Evol Biol*. 9(1):228.
Mayer G, Kauschke S, Rüdiger J, Stevenson PA. 2013. Neural markers reveal a one-segmented head in tardigrades (water bears). *PLoS One* 8(3):e59090.
Mayer G, Whitington PM, Sunnucks P, Pflueger H-J. 2010. A revision of brain composition in Onychophora (velvet worms) suggests that the tritocerebrum evolved in arthropods. *BMC Evol Biol*. 10(1):255.
Møbjerg N, Halberg KA, Jørgensen A, Persson D, Bjørn M, Ramløv H, Kristensen RM. 2011. Survival in extreme environments—on the current knowledge of adaptations in tardigrades. *Acta Physiol*. 202(3):409–420.
Møbjerg N, Jørgensen A, Kristensen RM, Neves RC. 2018. Morphology and functional anatomy. In: Schill RO, editor. Water bears: the biology of tardigrades. Cham (Switzerland): Springer Nature Switzerland AG. p. 57–94.
Murienne J, Daniels SR, Buckley TR, Mayer G, Giribet G. 2014. A living fossil tale of Pangean biogeography. *Proc Biol Sci*. 281(1775):20132648.
Nguyen LT, Schmidt HA, Von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol*. 32(1):268–274.
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(4):398–415.
Nielsen C. 1995. Animal evolution: interrelationships of the living phyla. Oxford: Oxford University Press.
Oliveira IS, Bai M, Jahn H, Gross V, Martin C, Hammel JU, Zhang W, Mayer G. 2016. Earliest onychophoran in amber reveals Gondwanan migration patterns. *Curr Biol*. 26(19):2594–2601.
Oliveira IS, Read VM, Mayer G. 2012. A world checklist of Onychophora (velvet worms), with notes on nomenclature and status of names. *ZooKeys* 211:1–70.
Ou Q, Shu D, Mayer G. 2012. Cambrian lobopodians and extant onychophorans provide new insights into early cephalization in Panarthropoda. *Nat Commun*. 3:1261.
Parra G, Bradnam K, Korf I. 2007. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* 23(9):1061–1067.
Parra G, Bradnam K, Ning Z, Keane T, Korf I. 2009. Assessing the gene space in draft genomes. *Nucleic Acids Res*. 37(1):289–297.
Patel RK, Jain M. 2012. NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. *PLoS One* 7(2):e30619.
Pauli T, Vedder L, Dowling D, Petersen M, Meusemann K, Donath A, Peters RS, Podsiadlowski L, Mayer C, Liu S, et al. 2016. Transcriptomic data from panarthropods shed new light on the evolution of insulator binding proteins in insects. *BMC Genomics*. 17(1):861.
Pelosi P. 1994. Odorant-binding proteins. *Crit Rev Biochem Mol Biol*. 29(3):199–228.
Pelosi P. 1996. Perireceptor events in olfaction. *J Neurobiol*. 30(1):3–19.
Pelosi P, Iovinella I, Felicioli A, Dani FR. 2014. Soluble proteins of chemical communication: an overview across arthropods. *Front Physiol*. 5:320.
Petersen M, Armisen D, Gibbs RA, Hering L, Khila A, Mayer G, Richards S, Misof B. 2019. Diversity and evolution of the transposable element repertoire in arthropods with particular reference to insects. *BMC Evol Biol*. 19:11.
Petersen TN, Brunak S, von Heijne G, Nielsen H. 2011. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat Methods*. 8(10):785–786.
Rebecchi L, Altiero T, Cesari M, Bertolani R, Rizzo AM, Corsetto PA, Guidetti R. 2011. Resistance of the anhydrobiotic eutardigrade *Paramacrobiotus richtersi* to space flight (LIFE–TARSE mission on FOTON-M3). *J Zool Syst Evol Res*. 49(s1):98–103.
Robertson HM. 2015. The insect chemoreceptor superfamily is ancient in animals. *Chemse* 40(9):609–614.
Robertson HM. 2019. Molecular evolution of the major arthropod chemoreceptor gene families. *Annu Rev Entomol*. 64(1):227–242.
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(5):392–398.
Sánchez-Gracia A, Vieira FG, Rozas J. 2009. Molecular evolution of the major chemosensory gene families in insects. *Heredity* 103(3):208–216.
Schürmann FW. 1995. Common and special features of the nervous system of Onychophora: a comparison with Arthropoda, Annelida and some other invertebrates. In: Breidbach O, Kutsch W, editors. The nervous systems of invertebrates: an evolutionary and comparative approach. Experientia supplementum. Vol. 72. Basel (Switzerland): Birkhäuser. p. 139–158.
Seppey M, Manni M, Zdobnov EM. 2019. BUSCO: assessing genome assembly and annotation completeness. *Methods Mol Biol*. 1962:227–245.
Smith FW, Boothby TC, Giovannini I, Rebecchi L, Jockusch EL, Goldstein B. 2016. The compact body plan of tardigrades evolved by the loss of a large body region. *Curr Biol*. 26(2):224–229.
Storch V, Ruhberg H. 1977. Fine structure of the sensilla of *Peripatopsis moseleyi* (Onychophora). *Cell Tissue Res*. 177(4):539–553.
Storch V, Ruhberg H. 1993. Onychophora. In: Harrison FW, Rice ME, editors. Microscopic anatomy of invertebrates. New York: Wiley-Liss. p. 11–56.
Sun JS, Xiao S, Carlson JR. 2018. The diverse small proteins called odorant-binding proteins. *Open Biol*. 8(12):180208.
Supek F, Bošnjak M, Škunca N, Šmuc T. 2011. REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS One* 6(7):e21800.
Sym M, Basson M, Johnson C. 2000. A model for Niemann-Pick type C disease in the nematode *Caenorhabditis elegans*. *Curr Biol*. 10(9):527–530.
Tegoni M, Pelosi P, Vincent F, Spinelli S, Campanacci V, Grolli S, Ramoni R, Cambillau C. 2000. Mammalian odorant binding

proteins. *Biochim Biophys Acta Protein Struct Mol Enzymol*. 1482(1–2):229–240.
Thomas GWC, Dohmen E, Hughes DST, Murali SC, Poelchau M, Glastad K, Anstead CA, Ayoub NA, Batterham P, Bellair M, et al. 2020. Gene content evolution in the arthropods. *Genome Biol*. 21(1):15.
Tsujimoto M, Imura S, Kanda H. 2016. Recovery and reproduction of an Antarctic tardigrade retrieved from a moss sample frozen for over 30 years. *Cryobiology* 72(1):78–81.
Venkatachalam K, Montell C. 2007. TRP channels TRP: transient receptor potential. *Annu Rev Biochem*. 76(1):387–417.
Vidal B, Aghayeva U, Sun H, Wang C, Glenwinkel L, Bayer EA, Hobert O. 2018. An atlas of *Caenorhabditis elegans* chemoreceptor expression. *PLoS Biol*. 16(1):e2004218.
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.
Vizueta J, Escuer P, Sánchez-Gracia A, Rozas J. Forthcoming 2020. Genome mining and sequence analysis of chemosensory soluble proteins in arthropods. Methods Enzymol.
Vizueta J, Frías-López C, Macías-Hernández N, Arnedo MA, Sánchez-Gracia A, 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(1):178–196.
Vizueta J, Macías-Hernández N, Arnedo MA, Rozas J, Sánchez-Gracia A. 2019. Chance and predictability in evolution: the genomic basis of convergent dietary specializations in an adaptive radiation. *Mol Ecol*. 28(17):4028–4045.
Vizueta J, Rozas J, Sánchez-Gracia A. 2018. Comparative genomics reveals thousands of novel chemosensory genes and massive changes in chemoreceptor repertories across chelicerates. *Genome Biol Evol*. 10(5):1221–1236.
Vizueta J, Sánchez-Gracia A, Rozas J. Forthcoming 2020. BITACORA: a comprehensive tool for the identification and annotation of gene families in genome assemblies. *Mol Ecol Resour*. doi: 10:1111/175-0998.13202.
Vogt RG, Miller NE, Litvack R, Fandino RA, Sparks J, Staples J, Friedman R, Dickens JC. 2009. The insect SNMP gene family. *Insect Biochem Mol Biol*. 39(7):448–456.
Waterson MJ, Chung BY, Harvanek ZM, Ostojic I, Alcedo J, Pletcher SD. 2014. Water sensor ppk28 modulates *Drosophila* lifespan and physiology through AKH signaling. *Proc Natl Acad Sci U S A*. 111(22):8137–8142.
Wicher D. 2012. Functional and evolutionary aspects of chemoreceptors. *Front Cell Neurosci*. 6:48.
Yoshida Y, Koutsovoulos G, Laetsch DR, Stevens L, Kumar S, Horikawa DD, Ishino K, Komine S, Kunieda T, Tomita M, et al. 2017. Comparative genomics of the tardigrades *Hypsibius dujardini* and *Ramazzottius varieornatus*. *PLoS Biol*. 15(7):e2002266.
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(2):e87800.
Zheng X, Valakh V, DiAntonio A, Ben-Shahar Y. 2014. Natural antisense transcripts regulate the neuronal stress response and excitability. *Elife* 3.

Evolutionary history of major chemosensory gene families across Panarthropoda

J Vizueta, P Escuer, C Frías-López, S Guirao-Rico, L Hering, G Mayer, J Rozas, A Sánchez-Gracia

2020. Molecular Biology and Evolution, 37, 12:3601–3615. doi: https://doi.org/10.1093/molbev/msaa197

Supplementary Material

Supplementary Tables

Table S1. BUSCO analysis of the *H. exemplaris, R. varieornatus* and *E. rowelli* genomes and transcriptome

A) Genomic Data

		H. exemplaris		*R. varieornatus*		*E. rowelli (genome draft)*	
		%	*n*	%	*n*	%	*n*
Eukarya set *n* = 303	Complete (C)	93,4	283	95,7	290	43,9	133
	Complete and single-copy (S)	92,1	279	94,4	286	41,3	125
	Complete and duplicated (D)	1,3	4	1,3	4	2,6	8
	Fragmented (F)	1,7	5	0,7	2	25,7	78
	Missing (M)	4,9	15	3,6	11	30,4	92

		H. exemplaris		*R. varieornatus*		*E. rowelli (genome draft)*	
		%	*n*	%		%	*n*
Metazoa set *n* = 978	Complete (C)	85,3	834	85,1	832	47,3	463
	Complete and single-copy (S)	83,7	819	83,8	820	46,0	450
	Complete and duplicated (D)	1,5	15	1,2	12	1,3	13
	Fragmented (F)	2,4	23	2,1	21	29,3	287
	Missing (M)	12,4	121	12,8	125	23,3	228

TableS2_Length_distribution

Table S2. Functional annotation and transcript length in *E. rowelli*

Length range	Contigs[a]	Transcripts	Annotated transcripts	% of annotated transcripts[b]
<300	534.465	531.058	23.135	4,36 %
301-600	432.112	390.994	30.612	7,83 %
601-1000	135.738	98.146	16.053	16,36 %
1001-2000	70.989	38.061	14.248	37,43 %
>2000	38.828	13.832	11.385	82,31 %
Total	1.212.132	1.072.091	95.433	8,90 %

[a] Non-clustered transcripts
[b] Protein-coding transcripts

TableS3_CEG_table

Table S3. Distribution of coverage lengths in CEG blastx results

Coverage in %[a]	Number of Hits	Number of accumulated hits
100	230	230
90	150	380
80	47	427
70	16	443
60	8	451
50	5	456
40	2	458
30	0	458
20	0	458
10	0	458
0	0	458

[a]The cluster of essential genes (CEG) database contains 458 genes

Table S4. Summary of sequences identified in the *H. exemplaris*, *R. varieornatus* and *E. rowelli* genomes and transcriptome

Fig 1B only shows the 12 complete members. The remaining 3 members could not unambiously be assiged as IR or iGluR

A) Number of chemosensory gene family members in the transcriptome of *E. rowelli*

Chemosensory Family	Transcritps	Complete	Partial	S_{MIN}[a]
GR / GR-like	3	1	1	3[b]
IR / iGluR	22	12	10	15
OR	0	0	0	0
DEG-ENaC	48	10	37	25[b]
CD36-SNMP	2	1	1	2
OBP / OBP-like	0	0	0	0
CSP	0	0	0	0
NPC2	11	11	0	11
CCPs	0	0	0	0
Total	**86**	**35**	**49**	**28**

[a] Mimimum number of genes estimated from complete and partial copies (see Vizueta et al. 2018)

[b] Includes one copy that could be a pseudogene or an assembly or sequencing error

Fig 1B only shows the 19 complete members. The remaining 2 members could not unambiously be assiged as IR or iGluR

B) Number of chemosensory gene family members in the genome of *H. exemplaris*

Chemosensory Family	Genes	Complete	Partial	S_{MIN}[a]
GR / GR-like	210	163	29	190
IR / iGluR	47	19	28	21
OR	0	0	0	0
DEG-ENaC	0	0	0	0
CD36-SNMP	0	0	0	0
OBP / OBP-like	0	0	0	0
CSP	0	0	0	0
NPC2	9	9	0	9
CCPs	0	0	0	0
Total	**266**	**191**	**57**	**220**

[a] Minimum number of genes estimated from complete and partial copies (see Vizueta et al. 2018)

Fig 1B only shows the 13 complete members. The remaining 2 members could not be unambiously assiged to the IR or iGluR families

C)Number of chemosensory gene family members in the genome of *R. varieornatus*

Chemosensory Family	Genes	Complete	Partial	S_{MIN}[a]
GR / GR-like	55	34	15	47
IR / iGluR	26	13	13	15
OR	0	0	0	0
DEG-ENaC	0	0	0	0
CD36-SNMP	0	0	0	0
OBP / OBP-like	0	0	0	0
CSP	0	0	0	0
NPC2	7	7	0	7
CCPs	0	0	0	0
Total	**88**	**54**	**28**	**69**

[a] Minimum number of genes estimated from complete and partial copies (see Vizueta et al. 2018)

Table S5. Gene expression profiles of candidate chemosensory genes in different anatomical compartments of *E. rowelli*

Chemosensory gene family	Protein ID	Length (aa)	Status	Compartment	FPKM[a] Accumulated			Differential expression[b]						FPKM[a]								
					ANT	*HEAD*	*REST*	*ANT Vs HEAD log2FC*	*ANT Vs HEAD FDR*	*ANT Vs REST log2FC*	*ANT Vs REST FDR*	*HEAD Vs REST log2FC*	*HEAD Vs REST FDR*	*ANT rep1*	*ANT rep2*	*ANT rep3*	*HEAD rep1*	*HEAD rep2*	*HEAD rep3*	*REST rep1*	*REST rep2*	*REST rep3*
NPC2	Erow10020	182	Complete	HeadBody	0	0,176	3,348	2,801178245	1	6,564203129	0,0171351	-4,015294274	0,026411073	0	0	0	0	0,072	0,104	1,214	0,093	2,041
NPC2	Erow11896	163	Complete	AntHeadBody	8,214	0,777	0,8	-4,474646388	0,000556409	-4,107881223	0,0010872	-0,456619955	0,999993734	5,459	1,615	1,14	0,591	0,041	0,145	0,56	0,056	0,184
NPC2	Erow138204	143	Complete	None	0	0	0	NA	NA	NA	NA	NA	NA	0	0	0	0	0	0	0	0	0
NPC2	Erow14018	162	Complete	AntHeadBody	1,471	6,023	8,006	1,333461168	1	1,452303479	0,4002082	-0,453794249	0,999993734	0,232	0,243	0,996	0,598	0,266	5,159	0,647	0	7,359
NPC2	Erow14880	164	Complete	AntHeadBody	0,486	9,934	3,177	2,732383058	0,007836537	1,163192266	0,3086275	1,205182792	0,487412563	0,258	0,199	0,029	3,267	3,777	2,89	1,294	0,168	1,715
NPC2	Erow22161	167	Complete	AntBody	1,014	0	0,051	-6,904296167	0,014939	-4,536382138	0,0268281	NA	NA	0,591	0,155	0,268	0	0	0	0,051	0	0
NPC2	Erow257621	125	Complete	HeadBody	0	0,203	0,145	1,929512457	1	1,75377524	1	-0,233667589	1	0	0	0	0,089	0	0,114	0,145	0	0
NPC2	Erow26057	156	Complete	Ant	18,85	0	0	-9,826925214	1	-9,097668149	1	NA	NA	15,72	1,195	1,935	0	0	0	0	0	0
NPC2	Erow26291	154	Complete	Ant	13,743	0	0	-9,394591512	0,000426664	-8,708113635	0,0005001	NA	NA	10,089	2,102	1,552	0	0	0	0	0	0
NPC2	Erow4956	149	Complete	AntHeadBody	4,096	0,154	0,223	-5,381951671	5,80147E-05	-4,7831504	0,00023	-0,643072157	1	1,752	0,62	1,724	0	0,123	0,031	0,182	0	0,041
NPC2	Erow6441	154	Complete	AntHeadBody	4,16	1,97	4,149	-2,283050418	1	-1,169822665	0,4821839	-1,147337547	0,999993734	3,439	0,376	0,345	1,723	0,092	0,155	3,547	0	0,602

[a] FPKM: Fragments per kilobase per million mapped reads

[b] Differential expression analysis results (obtained with DeSeq2): log2 FC (Fold Change) and FDR (False Discovery Rate) values for each pairwise comparison

Table S6. Gene expression profiles of candidate chemosensory genes in different anatomical compartments of *E. rowelli*

Chemosensory gene family	Protein ID	Length (aa)	Status	Compartment	FPKM[a] Accumulated				
					ANT	*HEAD*	*REST*	*ANT Vs HEAD log2FC*	*ANT Vs HEAD FDR*
GR	Erow88033	532	Complete	HeadBody	0	0,089	3,488	2,223472279	1
GR	Erow497874	97	Partial	Ant	0,399	0	0	-3,649317810	1
GR	Erow5372	287	Pseudogene	AntHeadBody	3,475	6,097	4,102	0,049540650	0,966453459

[a] FPKM: Fragments per kilobase per million mapped reads

[b] Differential expression analysis results (obtained with DeSeq2): log2 FC (Fold Change) and FDR (False Discovery Rate) values for each pairwise comparison

*Only showed the first 10 columns, for the complete results download the SuppMaterial table from: https://doi.org/10.1093/molbev/msaa197

Table S6. Gene expression profiles of candidate chemosensory genes in different anatomical compartments of *E. rowelli*

Chemosensory gene family	Protein ID	Length (aa)	Status	Compartment	FPKM[a] Accumulated				
					ANT	*HEAD*	*REST*	*ANT Vs HEAD log2FC*	*ANT Vs HEAD FDR*
DEG-ENaC	Erow108683	428	Complete	AntHeadBody	0,189	0,536	1,258	0,218029765	0,907968621
DEG-ENaC	Erow112851	340	Complete	AntHeadBody	1,349	0,113	0,044	-4,325779713	0,021967505
DEG-ENaC	Erow15622	494	Complete	AntHeadBody	0,471	2,16	2,511	0,986804074	0,464264804
DEG-ENaC	Erow246853	467	Complete	HeadBody	0	1,783	0,378	6,287416409	0,024340156
DEG-ENaC	Erow33959	447	Complete	AntHeadBody	0,165	0,521	3,864	0,779602146	0,617562628
DEG-ENaC	Erow42519	515	Complete	Ant	3,907	0	0	-9,59377601	0,000242344
DEG-ENaC	Erow46851	488	Complete	AntHeadBody	0,11	1,56	3,271	2,958753198	0,065272982
DEG-ENaC	Erow61721	443	Complete	AntHead	1,349	0,038	0	-5,650481488	1
DEG-ENaC	Erow74198	484	Complete	AntHeadBody	0,224	1,201	1,421	1,159600381	0,307733066
DEG-ENaC	Erow86984	510	Complete	AntHeadBody	0,045	0,102	0,961	-0,279369053	1
DEG-ENaC	Erow110935	255	Partial	AntHeadBody	0,081	0,089	0,102	-1,0158327	1
DEG-ENaC	Erow113538	199	Partial	Ant	1,136	0	0	-6,995627351	0,013974041
DEG-ENaC	Erow143488	190	Partial	AntBody	0,396	0	0,065	-5,145754442	0,075291057
DEG-ENaC	Erow150165	97	Partial	HeadBody	0	0,349	0,051	3,538318779	1
DEG-ENaC	Erow203859	256	Partial	HeadBody	0	0,121	0,498	1,569119735	1
DEG-ENaC	Erow205632	255	Partial	AntHeadBody	0,05	0,243	0,116	0,810654611	1
DEG-ENaC	Erow215467	146	Partial	AntHeadBody	0,448	0,487	0,269	-0,660802182	0,612466854
DEG-ENaC	Erow215841	92	Partial	AntHeadBody	0,096	0,194	0,073	0,323190371	1
DEG-ENaC	Erow216390	110	Partial	None	0	0	0	NA	NA
DEG-ENaC	Erow240568	87	Partial	Ant	2,326	0	0	-6,413285879	0,033164321
DEG-ENaC	Erow253347	278	Partial	HeadBody	0	0,382	0,48	3,584434406	1
DEG-ENaC	Erow254899	87	Partial	HeadBody	0	1	0,444	4,843914616	0,080935385
DEG-ENaC	Erow25923	85	Partial	Ant	14,091	0	0	-9,359635364	1
DEG-ENaC	Erow26867	42	Partial	AntHeadBody	0,57	1,082	0,563	0,08388846	0,945990018
DEG-ENaC	Erow302542	68	Partial	None	0	0	0	NA	NA
DEG-ENaC	Erow311809	260	Partial	AntHeadBody	0,056	0,829	0,933	2,268660921	0,424609831
DEG-ENaC	Erow345825	147	Partial	AntBody	0,061	0	0,102	NA	NA
DEG-ENaC	Erow442142	182	Partial	Head	0	0,102	0	NA	NA

[a] FPKM: Fragments per kilobase per million mapped reads

[b] Differential expression analysis results (obtained with DeSeq2): log2 FC (Fold Change) and FDR (False Discovery Rate) values for each pairwise comparison

*Only showed the first 10 columns, for the complete results download the SuppMaterial table from: https://doi.org/10.1093/molbev/msaa197

Table S6. Gene expression profiles of candidate chemosensory genes in different anatomical compartments of *E. rowelli*

Chemosensory gene family	Subfamily	Protein ID	Length (aa)	Status	Compartment	FPKM[a] Accumulated			
						ANT	*HEAD*	*REST*	*ANT Vs HEAD log2FC*
IR/iGluR	NMDA	Erow13453	875	Complete	AntHeadBody	6,808	0,486	0,035	-5,164471632
IR/iGluR	NMDA	Erow70447	952	Complete	AntHeadBody	0,089	7,222	0,636	6,20649675
IR/iGluR	NMDA	Erow195844	852	Complete	AntHeadBody	0,025	2,902	0,337	5,098718053
IR/iGluR	NMDA	Erow381452	295	Complete	AntHead	0,298	0,114	0	-2,32417875
IR/iGluR	NMDA	Erow165267	553	Complete	AntHeadBody	0,23	0,427	0,589	-0,178347595
IR/iGluR	NMDA	Erow104694	324	Partial	HeadBody	0	1,908	0,304	5,993834557
IR/iGluR	NMDA	Erow477211	77	Partial	HeadBody	0	1,477	0,316	3,338532631
IR/iGluR	Kainate	Erow64340	911	Complete	AntHeadBody	0,02	10,585	1,297	7,411680953
IR/iGluR	Kainate	Erow13291	403	Complete	Ant	10,747	0	0	-10,71707304
IR/iGluR	Kainate	Erow57569	941	Complete	AntHeadBody	0,666	1,764	1,789	0,355659298
IR/iGluR	Kainate	Erow225205	494	Partial	HeadBody	0	1,603	0,612	6,217146447
IR/iGluR	AMPA	Erow35247	912	Complete	Ant	0,216	0	0	-6,320428572
IR/iGluR	Kainate/AMPA?	Erow53063	798	Complete	AntHeadBody	0,406	1,107	0,582	0,721016878
IR/iGluR	Kainate/AMPA?	Erow36687	812	Complete	AntHeadBody	0,155	2,166	0,481	2,964511351
IR/iGluR	Kainate/AMPA?	Erow52198	418	Complete	Ant	0,187	0	0	-5,02927204
IR/iGluR	Kainate/AMPA?	Erow433530	292	Partial	HeadBody	0	0,44	0,679	3,654103296
IR/iGluR	Kainate/AMPA?	Erow381046	110	Partial	HeadBody	0	0,617	0,81	2,223472792
IR/iGluR	Kainate/AMPA?	Erow621307	92	Partial	HeadBody	0	0,369	0,459	1,569119735
IR/iGluR	IR?	Erow181532	367	Partial	Ant	0,025	0	0	NA
IR/iGluR	IR?	Erow212555	148	Partial	AntHeadBody	0,357	0,261	0,095	-1,53084044
IR/iGluR	IR?	Erow257143	340	Partial	HeadBody	0	1,953	0,231	6,010910725
IR/iGluR	IR?	Erow60497	319	Partial	Ant	1,672	0	0	-8,274330776

[a] FPKM: Fragments per kilobase per million mapped reads

[b] Differential expression analysis results (obtained with DeSeq2): log2 FC (Fold Change) and FDR (False Discovery Rate) values for each pairwise comparison

*Only showed the first 10 columns, for the complete results download the SuppMaterial table from: https://doi.org/10.1093/molbev/msaa197

Table S5. Gene expression profiles of candidate chemosensory genes in different anatomical compartments of *E. rowelli*

Chemosensory gene family	Protein ID	Compartment	FPKM[a] Accumulated			Differential exp		
			ANT	*HEAD*	*REST*	ANT Vs HEAD log2FC	ANT Vs HEAD FDR	ANT Vs REST log2FC
comp100427	807	Ant	0,05	0	0	NA	NA	NA
comp1055817	208	None	0	0	0	NA	NA	NA
comp106912	647	AntHeadBody	1,782	0,538	0,888	-2,653509768	0,131685357	-1,8200652
comp110736	1450	Ant	0,22	0	0	-5,332487546	0,075491586	-4,58913671
comp13355	3254	AntHeadBody	8,044	0,73	1,824	-4,243578684	1,81471E-07	-1,13235504
comp146762	1417	Ant	1,635	0	0	-7,972487668	0,003762492	-7,29112185
comp14753	5510	AntHeadBody	4,136	0,031	0,019	-7,644280126	3,97601E-09	-8,12844738
comp147952	591	AntHeadBody	0,655	0,206	0,199	-2,247910769	0,252618196	-2,07447653
comp14981	4559	AntHeadBody	0,577	7,959	5,318	2,611892377	0,022392306	1,93004274
comp155747	1792	HeadBody	0	3,469	0,07	7,415195468	0,005540714	2,0966638
comp163923	531	Ant	0,081	0	0	NA	NA	NA
comp166654	1067	AntHead	0,479	0,044	0	-3,795932681	0,075406586	-5,15760941
comp182830	320	Head	0	0,515	0	2,341082647	1	NA
comp194790	593	AntHeadBody	0,808	0,165	0,748	-2,640059002	0,236264912	-1,08002266
comp195106	665	Ant	2,015	0	0	-6,854564423	0,022907634	-6,14571209
comp199656	457	Ant	0,332	0	0	NA	NA	NA
comp211026	848	HeadBody	0	0,222	0,058	2,223472792	1	NA
comp23865	3137	AntHeadBody	0,565	12,528	0,698	3,186434485	1,8243E-05	-0,7338927
comp241072	352	HeadBody	0	1,939	0,153	4,507202818	0,101086562	NA
comp252695	766	Ant	0,328	0	0	-4,61439557	1	-3,69583314
comp26773	1987	AntHeadBody	1,405	3,205	3,492	0,690034958	0,505887854	0,98080645
comp27632	747	AntHeadBody	0,056	9,467	0,312	5,567638863	0,001120649	1,04186661
comp279089	756	Ant	0,111	0	0	-3,205603011	1	-2,48692016
comp284103	262	Head	0	0,203	0	NA	NA	NA
comp308932	290	Ant	0,737	0	0	-3,882017049	1	-3,06486216
comp309971	556	Head	0	1,621	0	4,876029267	0,078142146	NA
comp348551	426	HeadBody	0	1,663	0,124	4,567428768	0,100668441	NA
comp356177	478	None	0	0	0	NA	NA	NA
comp359686	807	Ant	0,163	0	0	-3,351865789	1	-2,65010118
comp366553	939	None	0	0	0	NA	NA	NA
comp372456	233	None	0	0	0	NA	NA	NA
comp40462	4553	Ant	0,48	0	0	-7,9126663	0,003538616	-7,26479677
comp40475	2235	AntHeadBody	0,304	17,902	1,433	5,654108095	9,12967E-10	1,48703661
comp405835	494	AntHeadBody	0,086	0,681	0,647	1,301843212	1	1,14742412
comp416941	361	Ant	2,713	0	0	-6,502317694	0,025663015	-5,89283973
comp425892	423	HeadBody	0	0,435	0,131	2,481345917	1	NA
comp436780	630	HeadBody	0	1,446	0,214	4,959563789	0,075819969	2,0720144
comp444257	430	Head	0	1,252	0	4,209958699	0,129429764	NA
comp46029	1285	Ant	4,204	0	0	-8,982665422	0,000782413	-8,33765834
comp469357	347	None	0	0	0	NA	NA	NA
comp482240	348	HeadBody	0	1,046	0,473	3,669197749	1	2,08557883
comp502526	608	None	0	0	0	NA	NA	NA
comp505762	233	HeadBody	0	0,229	0,247	NA	NA	NA
comp507490	289	HeadBody	0	0,178	0,827	NA	NA	2,85042385
comp511906	492	Ant	0,439	0	0	-4,369969307	1	-3,46819017
comp52316	3161	AntHeadBody	0,397	0,013	0,029	-5,200482122	0,030456105	-4,04236478
comp527536	303	Head	0	0,172	0	NA	NA	NA
comp532867	405	None	0	0	0	NA	NA	NA
comp56722	4094	AntHeadBody	0,04	10,481	0,305	6,453897265	2,24942E-08	1,77662506
comp57268	4155	AntHead	0,01	0,064	0	1,12479547	1	NA
comp58138	2685	AntHead	1,901	0,021	0	-7,025425513	1,53376E-05	-8,40070200
comp58138	2685	AntHead	1,901	0,021	0	-7,025425513	1,53376E-05	-8,40070200
comp59727	718	AntBody	0,657	0	0,383	-5,59696795	0,062933553	-1,45120116
comp64834	2073	AntHeadBody	0,269	0,968	2,506	0,736091427	0,585945697	1,85938040
comp650038	300	HeadBody	0	0,673	0,189	2,712325045	1	NA
comp65404	1972	AntHeadBody	0,101	1,24	2,122	2,107539358	0,223532437	2,69946578
comp65404	1972	AntHeadBody	0,101	1,24	2,122	2,107539358	0,223532437	2,69946578
comp65404	1972	AntHeadBody	0,101	1,24	2,122	2,107539358	0,223532437	2,69946578
comp655223	304	Head	0	0,544	0	2,341082647	1	NA
comp665999	400	HeadBody	0	0,248	0,138	1,569119735	1	NA
comp672949	350	Head	0	1,158	0	3,521118982	1	NA
comp672949	350	Head	0	1,158	0	3,521118982	1	NA
comp672952	452	Head	0	0,459	0	2,64565057	1	NA
comp678553	356	HeadBody	0	0,42	0,305	2,249139166	1	1,7537752
comp68585	2943	Ant	0,571	0	0	-7,305261498	0,01444378	-6,52813555
comp689049	384	AntBody	0.116	0	0.138	NA	NA	-0.54799944

*Only showed the first 10 columns, for the complete results download the SuppMaterial table from: https://doi.org/10.1093/molbev/msaa197

Table S5. Gene expression profiles of candidate chemosensory genes in different anatomical compartments of *E. rowelli*

Chemosensory gene family	Protein ID	Length (aa)	Status	Compartment	FPKM[a] Accumulated			Differential expression[b]			
					ANT	*HEAD*	*REST*	ANT Vs HEAD log2FC	ANT Vs HEAD FDR	ANT Vs REST log2FC	ANT Vs REST FDR
CD36-SNMP	Erow47702	490	Complete	AntHead	0,918	0,041	0	-5,34577835	0,003063911	-6,898872159	0,0070012
CD36-SNMP	Erow113129	140	Partial	Ant	0,413	0	0	-4,318546938	1	-3,703552545	1

[a] FPKM: Fragments perkilo base per million mapped reads

[b] Differential expression analysis results (obtained with DeSeq2): log2 FC (Fold Change) and FDR (False Discovery Rate) values for each pairwise comparison

Table S5. Gene expression profiles of candidate chemosensory genes in different anatomical compartments of *E. rowelli*

Protein ID	Length (aa)	Location	FPKM[a] Accumulated			Differential exp		
			ANT	*HEAD*	*REST*	ANT Vs HEAD log2FC	ANT Vs HEAD FDR	ANT Vs REST log2FC
comp13154	1028	Ant	1,889	0	0	-7,693334514	0,00556291	-7,03549013

[a] FPKM: Fragments per kilobase per million mapped reads

[b] Differential expression analysis results (obtained with DeSeq2): log2 FC (Fold Change) and FDR (False Discovery Rate) values for each pairwise comparis

*To consult the rest and totality of the tables, download the Supplemental Material from https://doi.org/10.1093/molbev/msaa197

Table S6. Summary of chemosensory gene family members identified in whole body transcriptomes of onychophorans and tardigrades and in the genome of representative nematodes

Table S6.1 Additional searches in tardigrade transcriptomes

Species	Clade	BUSCO[a] % complete	BUSCO[a] % missing	Chemosensory gene families GR / GR-like	IR / iGluR	DEG-ENaC	CD36-SNMP	NPC2
Richtersius cf. coronifer [b]	Eutardigrada (Richtersiidae)	88,24	8,27	13	39	0	0	10
Mesobiotus philippinicus [c]	Eutardigrada (Macrobiotidae)	72,39	22,39	7	27	0	0	7
Paramacrobiotus richtersi [d]	Eutardigrada (Macrobiotidae)	88,14	8,79	17	45	0	0	8
Milnesium tardigradum [d]	Apotardigrada (Milnesiidae)	60,74	21,27	12	8	0	0	6
Echiniscoides cf. sigismundi [b]	Heterotardigrada (Echiniscoididae)	81,9	14,31	12	35	12	1	7
Echiniscus testudo [c]	Heterotardigrada (Echiniscidae)	75,56	11,15	10	89	12	2	20

[a] BUSCO results based the Metazoa dataset (978 BUSCO)

[b] Kamilari et al. (2019): *Echiniscoides cf. sigismundi* (SRX421163), *Richtersius cf. coronifer* (SRX4213802)

[c] Mapalo et al. (2020)

[d] NCBI TSA database (https://www.ncbi.nlm.nih.gov/genbank/tsa/): *Milnesium tardigradum* (GFGZ00000000.1), *Paramacrobiotus richtersi* (GFGY00000000.1);

Table S6.2 Additional searches in onychophorran transcriptomes

Species	Clade	BUSCO[a] % complete	BUSCO[a] % missing	Chemosensory gene families GR / GR-like	iGluR (IR)[b]	DEG-ENaC	CD36-SNMP	NPC2
Epiperipatus sp. [c]	Onychophora (Peripatidae)	77,61	2,54	2	**27 (0)**	27	2	7
Opisthopatus kwazululandi [c]	Onychophora (Peripatopsidae)	82,62	2,15	2	**51 (0)**	36	2	13
Principapillatus hitoyensis [d]	Onychophora (Peripatidae)	82,9	10,8	2	**24 (0)**	17	3	13
Ooperipatus hispidus [d]	Onychophora (Peripatopsidae)	77,2	9,1	0	**29 (0)**	19	2	4
Phallocephale tallagandensis [d]	Onychophora (Peripatopsidae)	74,1	11	0	**24 (0)**	32	1	5
Eoperipatus sp. [d]	Onychophora (Peripatidae)	77,3	10,6	0	**20 (0)**	22	1	9
Euperipatoides rowelli [d]	Onychophora (Peripatopsidae)	81,4	11,2	3	**28 (0)**	33	1	11
Euperipatoides rowelli (female)[e]	Onychophora (Peripatopsidae)	89,3	4	1	**29 (0)**	29	1	5
Euperipatoides rowelli (male)[e]	Onychophora (Peripatopsidae)	84,2	3,6	1	**28 (0)**	27	3	7

[a] BUSCO results based the Metazoa dataset (978 BUSCO)

[b] **Based on the phylogenetic tree of these IR/iGluR, no copy can be unambiguously assigned as divergent IR with node support in these species, not even IR25a**

[c] Mapalo et al. (2020)

[d] Hering et al. (2012)

[e] Raw data retrieved from the Human Genome Sequencing Center [HGSC] (References: SRX973445 f and SRX973444 g)

Table S6.3 Additional searches in publicly available nematode genomes (NCBI)

Species	Clade	Live-style	Chemosensory gene families GR / GR-like	IR / iGluR	DEG-ENaC	CD36-SNMP	NPC2
Caenorhabditis elegans	Nematoda (Rhabditina)	Free-living	5	14	31	6	1
Pristionchus pacificus	Nematoda /Rhabditina)	Free-living	10	26	68	11	1
Panagrellus redivivus [a]	Nematoda (Tylenchina)	Free-living	0	11	37	10	1
Strongyloides ratti	Nematoda (Tylenchina)	Parasite	1	10	23	5	4
Dracunculus medinensis	Nematoda (Spirurina)	Parasite	0	10	14	2	1
Trichinella spiralis	Nematoda (Dorylaimia)	Parasite	0	16	5	1	3

[a] Since no functional annotation (GFF) is available for this genome, we used BITACORA under the "genome mode"

Table S7 Orthofinder_results

Table S7. Orthofinder analysis of IR/iGluR sequences

Species \ Conserved	NDMAR1	NMDAR2	NMDAR3	Kainate	Kainate	Kainate	Kainate	AMPA	IR25a
Drosophila melanogaster	DmelNMDAR1	DmelIR8a, DmelNMDAR2		DmelCG11155, DmelCG9935		Dmelclumsy	DmelCG3822	DmelGLURIA, DmelGLURIB	DmelIR25a
Daphnia pulex	Dpul12551 43	Dpul141686		Dpul1309661, Dpul155220, Dpul155388			Dpul1327119	Dpul1326779	DpulIR25a
Ixodes scapularis	IscaNMDAR 01	IscaNMDAR02	IscaNMDAR03	IscaKA01, IscaKA03, IscaKA04, IscaKA05, IscaKA06	IscaKA07			IscaAMPAR01, IscaAMPAR02, IscaAMPAR03,	IscaIr25a
Strigamia maritima	SmarNMDAR1	SmarNMDAR2	SmarNMDAR3	SmarKA01, SmarKA02, SmarKA03, SmarKA04, SmarKA05	SmarKA06			SmarAMPA01, SmarAMPA02, SmarAMPA03	SmarIR25a, SmarIR8a
Euperipatoides rowelli	Erow13453	Erow165267, Erow381452, Erow477211p,	Erow104694p, Erow195844	Erow13291, Erow57569, Erow64340	Erow36687, Erow52198, Erow53063		Erow225205p, Erow381046p, Erow433530p,	Erow35247	
Hypsibius exemplaris	HdujOQV18471	HdujOQV24439, HdujOWA51021_1	HdujOWA52506, HdujOWA52507, HdujOWA52511	HdujOQV21276, HdujOQV22948_2, HdujOQV22952,	HdujOWA51397_1	HdujOQV23175		HdujOWA49770, HdujOWA49771, HdujOWA51209	HdujOQV18207
Ramazzottius varieornatus	RvarGAV07593	RvarGAU93983	RvarGAU99788, RvarGAU99789, RvarGAV08472	RvarGAU90348, RvarGAU93737, RvarGAU96081		RvarGAU93647, RvarGAU93648, RvarGAU93649	RvarGAU93646		RvarGAU96058

Table S8 TRP Results

Table S8. Gene expression profiles of TRP genes in different anatomical compartments of *E. rowelli*

Protein ID	Status	Length (aa)	Compartment	FPKM[a] Accumulated			ANT Vs HEAD log2FC	ANT Vs HEAD FDR
				ANT	HEAD	REST		
comp141328_1	Complete	508	Ant	121	0	0	-4.3699693065538	1
comp76561_1	Complete	431	Ant	845	0	0	-7.40298217882003	0.0145966301906053
comp15353_1	Complete	482	AntBody	8077	0	34	-11.0750607131625	6.91812107225789e-06
comp147364_1	Complete	723	AntBody	395	0	99	-6.04512907035836	0.0410815590086246
comp131699_1	Complete	462	AntBody	0.02	0	51	NA	NA
comp160630_1	Complete	746	AntHead	209	19	0	-3.74446049213964	0.0903856754209346
comp80482_1	Complete	466	AntHead	2539	57	0	-6.16336492782857	0.000623072258911358
comp89409_1	Complete	651	AntHeadBody	471	1259	211	-0.221398898554885	0.913707957484584
comp13955_1	Complete	841	AntHeadBody	522	253	221	-1.89221328128554	0.08390956409006
comp95411_1	Complete	851	AntHeadBody	0.04	0.02	35	-1.91317301647357	1
comp174007_1	Complete	342	AntHeadBody	806	228	0.14	-2.63321736934419	0.0660376786248639
comp14336_1	Complete	424	AntHeadBody	3721	436	166	-3.91573317526753	1.97532799535623e-08
comp22267_1	Complete	914	AntHeadBody	3508	331	186	-4.35608914744472	0.0172492291169296
comp74815_1	Complete	891	AntHeadBody	2726	86	151	-5.53470466215511	0.000966262867771361
comp27943_1	Complete	419	AntHeadBody	7.61	292	1251	-5.59542543448512	0.000370942350980524
comp77224_1	Complete	471	AntHeadBody	1181	25	0.16	-6.05056609871914	1
comp18534_1	Complete	611	AntHeadBody	22.92	438	164	-6.22098513635476	1
comp60177_1	Complete	425	AntHeadBody	1.22	19	22	-6.32161026930373	0.00369354511575164
comp9624_1	Complete	530	AntHeadBody	3529	0.06	93	-6.86755198106982	9.02326553357869e-08
comp7157_1	Complete	639	AntHeadBody	56.72	72	112	-9.91858426701427	6.5671537952343e-08
comp22787_1	Complete	443	AntHeadBody	0.03	102	177	0.14704257727397	1
comp437_2	Complete	426	AntHeadBody	3554	8811	12086	0.543460922951536	0.430266495973278
comp32938_1	Complete	453	AntHeadBody	678	2031	529	0.609637711748775	0.389862116679648
comp23809_1	Complete	379	AntHeadBody	15	73	7777	0.730861945637031	1
comp13211_1	Complete	841	AntHeadBody	1566	4.16	2633	0.879814129117563	0.319877851470773
comp44926_1	Complete	504	AntHeadBody	326	1245	0.32	0.94112135434903	0.262419029200126
comp120005_1	Complete	507	AntHeadBody	363	1.44	858	1.29593889889384	0.173334322883627
comp2605_1	Complete	362	AntHeadBody	1219	6586	16262	1.61154221381405	0.0937598935015493
comp57249_1	Complete	292	AntHeadBody	0.05	0.48	2507	1.71843488083632	0.313149620544787

*To consult the totality of the table, download the Supplemental Material from https://doi.org/10.1093/molbev/msaa197

[a] FPKM: Fragments per kilobase per million mapped reads

[b] Differential expression analysis results (obtained with DeSeq2): log2 FC (Fold Change) and FDR (False Discovery Rate) values for each pairwise comparison

Supplementary Figures

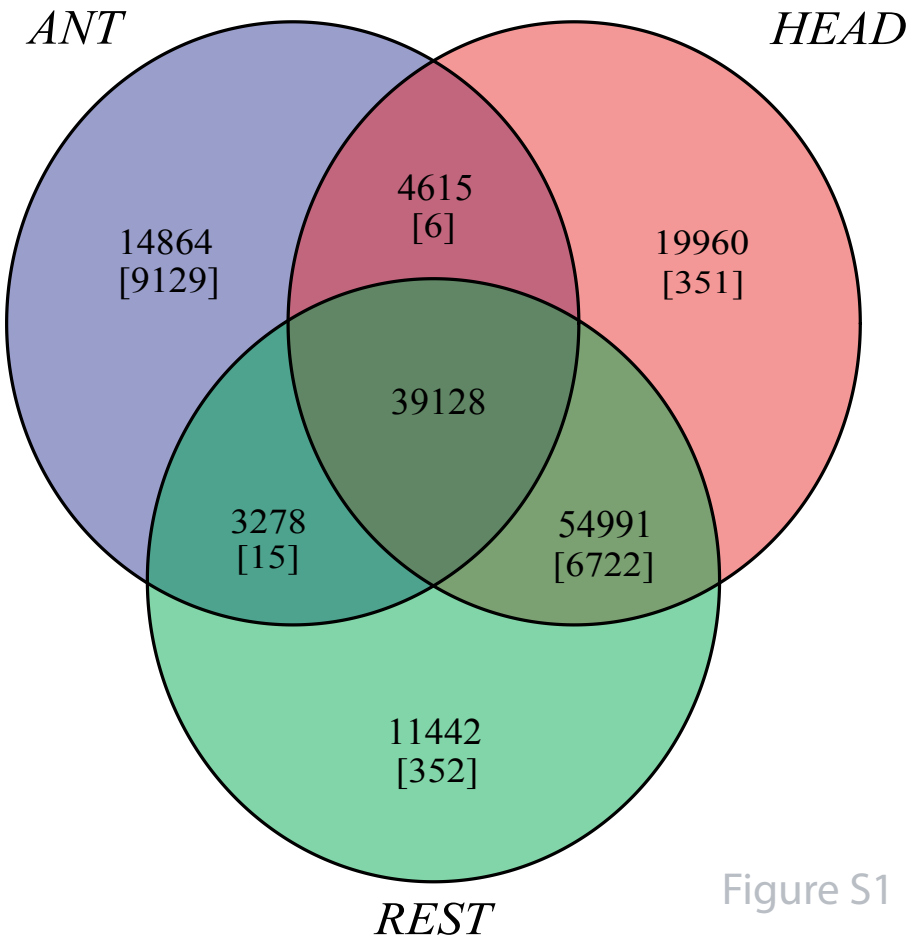
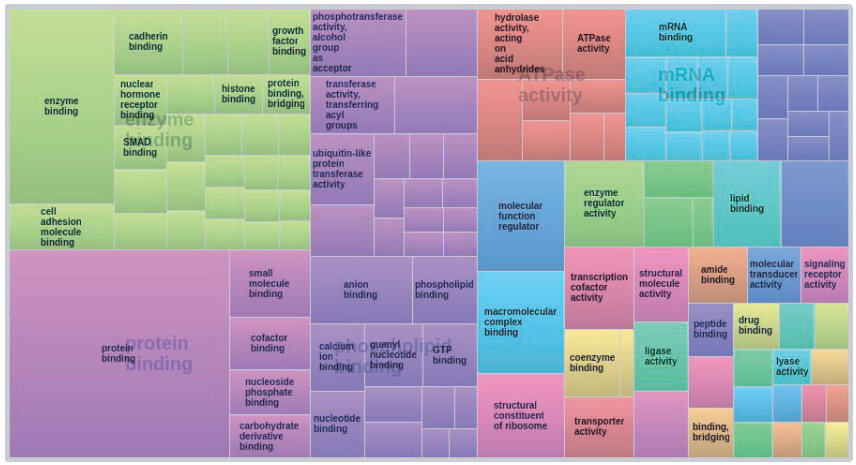

ANT
HEAD
REST
14864
[9129]
4615
[6]
19960
[351]
39128
3278
[15]
54991
[6722]
11442
[352]

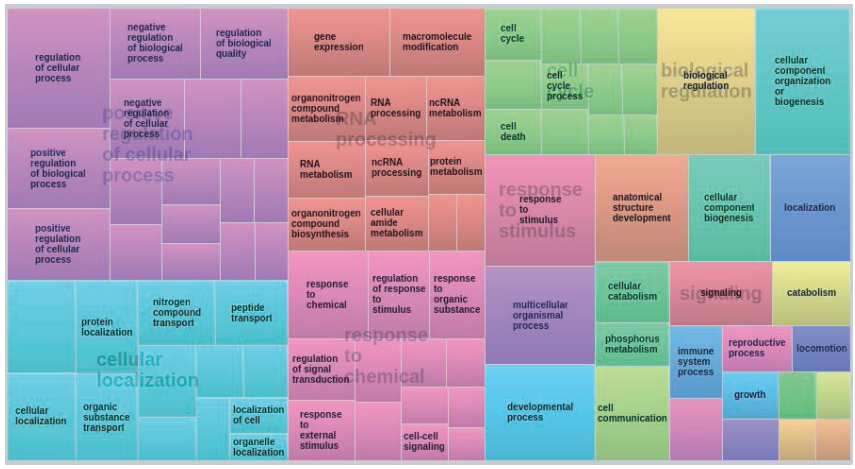
Figure S1

A



Molecular Function in ANT

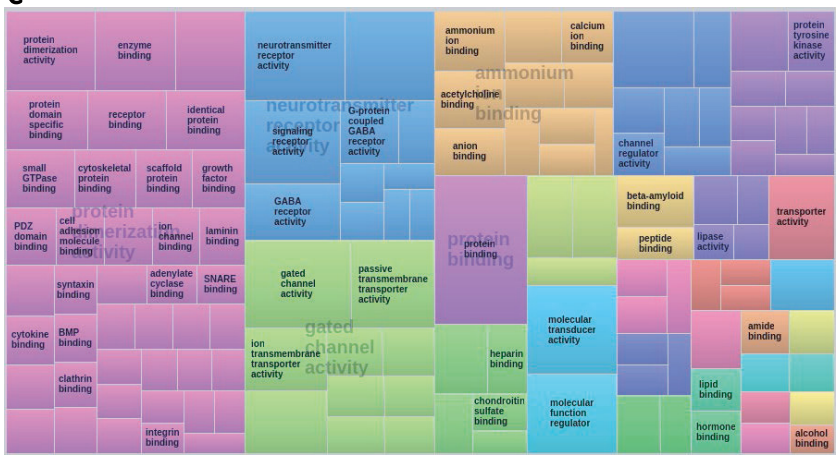
B



Biological Processin ANT

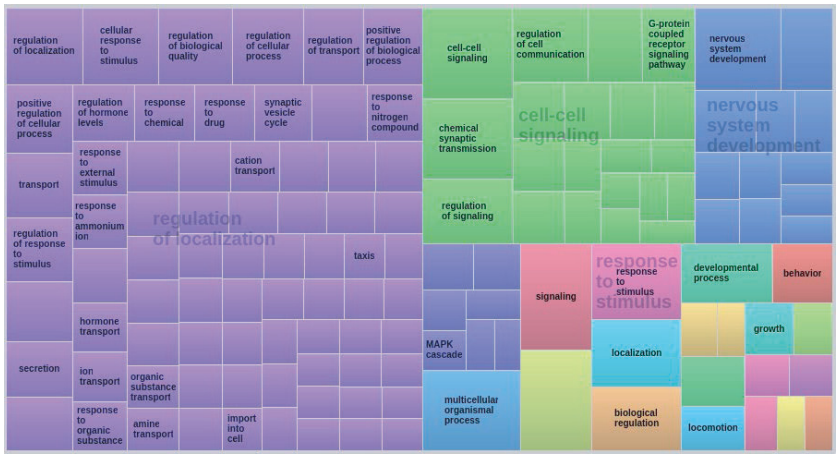
Figure S2

C



Molecular Functionin HEAD

D



Biological Processin HEAD

Figure S2-(cont)

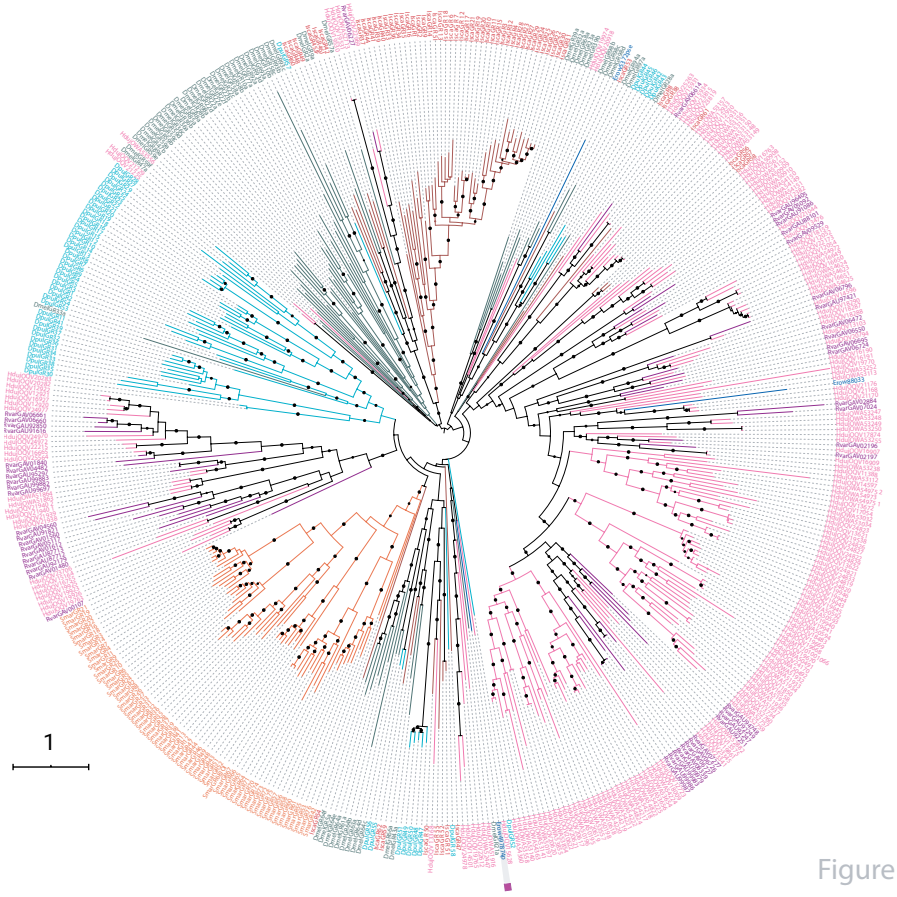


Figure S3

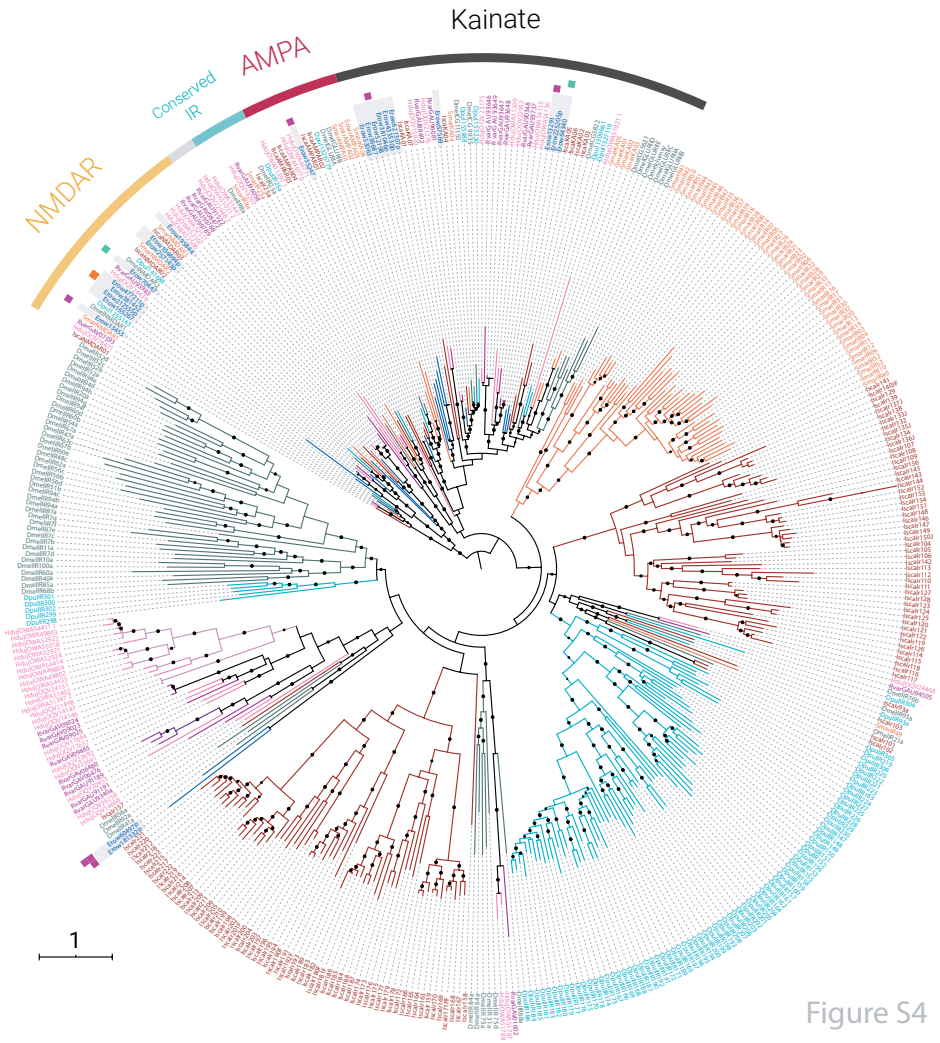


Figure S4

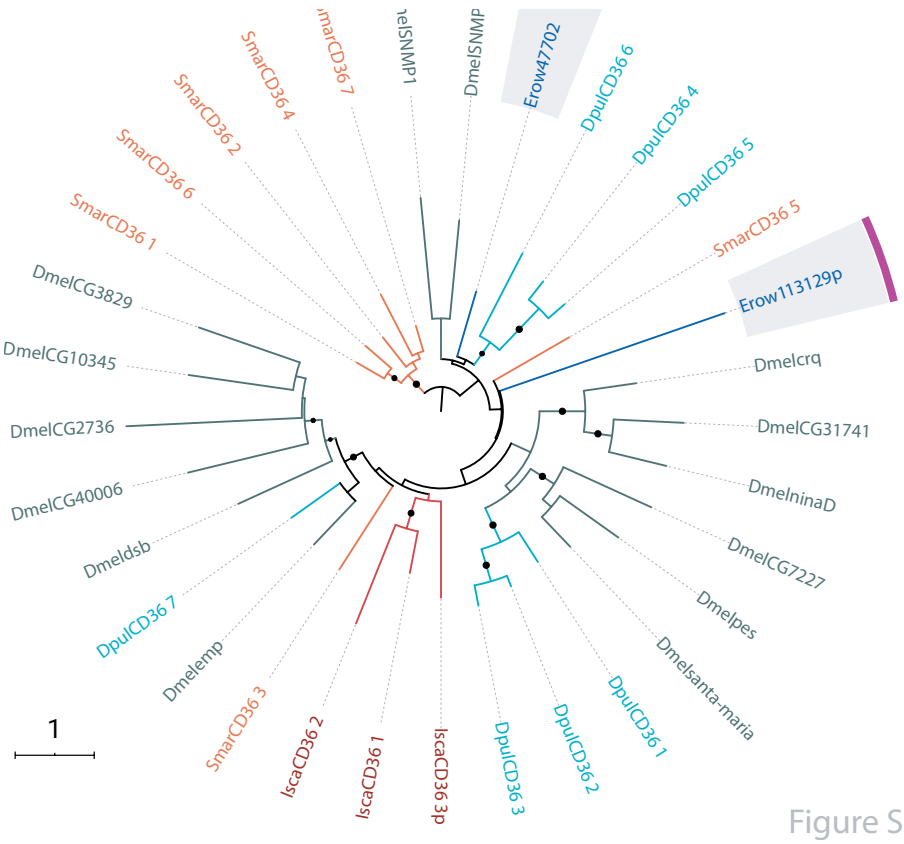

SmarCD36 7
DmelSNMP1
DmelSNMP
Erow47702
DpulCD36 6
DpulCD36 4
DpulCD36 5
SmarCD36 5
Erow113129p
SmarCD36 4
SmarCD36 2
SmarCD36 6
SmarCD36 1
DmelCG3829
DmelCG10345
DmelCG2736
DmelCG40006
Dmeldsb
DpulCD367
Dmelemp
SmarCD36 3
IscaCD36 2
IscaCD36 1
IscaCD36 3p
DpulCD36 3
DpulCD36 2
DpulCD36 1
Dmelsanta-maria
Dmelpes
DmelCG7227
DmelninaD
DmelCG31741
Dmelcrq
1

Figure S5

Chapter 4

Population genomics of adaptive radiations: Exceptionally high levels of genetic diversity in a spider endemic from the Canary Islands

P Escuer, S Guirao-Rico, MA Arnedo, A Sánchez-Gracia, J Rozas

Les aranyes endèmiques de les Illes Canàries del gènere *Dysdera* han patit una radiació d'espècies, amb ~60 espècies que han divergit en pocs milions d'anys. En aquest gènere, la diversificació de les espècies i de la dieta són processos que han estat molt associats, normalment acompanyades de modificacions fenotípiques (incloent-hi canvis morfològics, metabòlics i de comportament). Això fa que aquest gènere sigui un model excel·lent per entendre els mecanismes evolutius i per identificar les bases genètiques de les radiacions adaptatives. Recentment, hem publicat el primer assemblatge de genoma a escala cromosòmica d'una espècie endèmica de les Illes Canàries d'aquest gènere, *Dysdera silvatica*, un assemblatge de referència d'alta qualitat que ens permet realitzar estudis genòmics en aquest grup. Vam aprofitar aquest recurs genòmic per dur a terme un estudi de reseqüenciació a baixa cobertura d'una població natural de *D. silvatica* de La Gomera. Hem caracteritzat nivells de polimorfisme nucleotídic, divergència i desequilibri de lligament a tot el genoma, i hem fet exploracions exhaustives per detectar l'impacte de la selecció positiva recent. Hem trobat uns nivells excepcionalment alts de diversitat nucleotídica en aquesta espècie endèmica, cosa que apunta a una estructuració genètica ancestral en la població. També hem identificat algunes regions genòmiques, com a possible candidates a incloure els determinants moleculars subjacents a la recent diversificació d'aquesta espècie a les Illes Canàries. Aquesta anàlisi, pionera en aranyes, servirà com a pedra angular per encetar altres estudis de genòmica poblacional més amplis amb l'objectiu d'entendre les radiacions adaptatives en illes, i identificar les possibles dianes genòmiques de la selecció.en Panarthropoda i contribueixen a una millor comprensió de l'evolució dels sentits químics dels animals.

Population genomics of adaptive radiations: Exceptionally high levels of genetic diversity in a spider endemic from the Canary Islands

Paula Escuer[1,2], Sara Guirao-Rico[1,2], Miquel A. Arnedo[2,3], Alejandro Sánchez-Gracia[1,2,*], Julio Rozas[1,2,*]

1. Departament de Genètica, Microbiologia i Estadística,Universitat de Barcelona, Spain.

2. Institut de Recerca de la Biodiversitat (IRBio), Universitat de Barcelona, Spain.

3. Departament de Biologia Evolutiva, Ecologia i Ciències Ambientals, Universitat de Barcelona, Spain

*Corresponding authors

Abstract

The Canary Islands endemic spiders of the genus *Dysdera* have undergone a remarkable process of species radiation, with ~60 species that have diverged in a few millions of years. In this genus, species and dietary diversification have been repeatedly linked, typically accompanied by striking phenotypic modifications (including morphological, metabolic and behavioral changes). This makes this genus an excellent model for understanding the evolutionary drivers, and to pinpoint the genomic determinants of adaptive radiations. Recently, we have reported the first chromosome-level genome assembly of a Canary Islands endemic species of this genus, *Dysdera silvatica*, ensuring a high-quality reference sequence for performing genomics studies in this group.

Here, we took advantage of this genomic resource to perform a low-coverage based resequencing study of a natural population of *D. silvatica* from La Gomera. We have characterized genome-wide levels of nucleotide polymorphism, divergence and linkage disequilibrium, and performed exhaustive scans for recent positive selection across chromosomes. We found exceptionally high levels of nucleotide diversity in this endemic species, pointing to some ancestral genetic structure. We also identified some genomic regions candidates to be targets of positive selection that could be involved in the recent diversification of this species. This analysis, pioneering in spiders, will serve as the cornerstone of broader population genomics studies aimed to understand adaptive radiations in islands and pinpoint the genomic targets of selection.

Introduction

Understanding the origin of species is a central issue in evolutionary biology (Austin & Arnold, 2001; Ravinet et al., 2017). However, despite being instrumental to manage and conserve biodiversity in a changing world (Mergeay & Santamaria, 2012), the knowledge of the evolutionary mechanisms and the key genomic targets underlying species diversification is still limited. Although experimental evolution should be the best approach to address these questions, the direct examination of the evolutionary process in putatively evolving traits is more affordable in organisms with short generation times, such as viruses or bacteria (Coyne & Orr, 2004). Adaptive radiations, and in particular those occurring in oceanic islands offer an excellent alternative to experimental evolution in animals and plants; indeed, they include examples with enormous phenotypic diversity evolved in a very short evolutionary time (Hodges & Derieg, 2009; Schluter, 2000). For this reason, the oceanic archipelagos have long been recognized as natural laboratories to study short-term evolution (Fernández-Mazuecos et al., 2020), having received much attention over time (Carson & Kaneshiro, 1976; Emerson, 2002; Gillespie, 2004; Grant & Grant, 2008; Juan, Emerson, Oromí, & Hewitt, 2000; Machado, Rodríguez-Expósito, López, & Hernández, 2017). In recent years, the analyses based on whole-genome data, specially of non-model organisms, has further fueled the interest on adaptive radiations (Choi et al., 2021; Feder, Egan, & Nosil, 2012; Lamichhaney et al., 2015; Richards et al., 2021; Schwager et al., 2017; Wolf & Ellegren, 2017). However, the relative role of adaptive and non-adaptive forces in species diversification is still a matter of debate (Choi et al., 2020; Muschick, Indermaur, & Salzburger, 2012; Rundell & Price, 2009; Rundle & Nosil, 2005; Simões et al., 2016).

The volcanic archipelago of Canary Islands comprises eight islands, some of them differing greatly in their climate (Fernández-Palacios, 2011), and exhibits high levels of endemism reaching up to 40% in the case of Arthropods (Martín et al., 2010). Among them, spiders of the genus *Dysdera* Latreille, 1804 (Araneae: Dysderidae) represents one of the most spectacular examples of islands radiation among spiders (Arnedo, Oromí, Múrria, Macías-Hernández, & Ribera, 2007; Arnedo, Oromí, & Ribera, 2001; Macías-Hernández, López, Roca-Cusachs, Oromí, & Arnedo, 2016; Řezáč, Pekár, Arnedo, Macías-Hernández, & Řezáčová, 2021). Indeed, about 20% of the ~300 species described in this Western Palearctic genus (World Spider Catalog, 2022) are endemic from this archipelago (Macías-Hernández et al., 2016). *Dysdera* spiders are nocturnal ground-dwelling hunters that present some of the few known

cases of prey specialization (stenophagy) among spiders (Pekár, Líznarová, & Řezáč, 2016). Some species are facultatively or even obligatorily specialized in eating woodlice, a prey that is commonly avoided by most predators because, in addition of their effective curling defense strategy, they accumulate heavy metals in their exoskeleton, which makes it difficult to metabolize them (Pekár et al., 2016). These spiders have evolved morphological adaptations, and metabolic and behavioral strategies to successfully catch and digest these isopods (Hopkin & Martin, 1985; Řezáč, Pekár, & Lubin, 2008; Řezáč & Pekár, 2007; Toft & Macías-Hernández, 2017). Indeed, dietary diversification seems to be directly related with species richness and species overlapping (two or more species coexisting in the same locality; Řezáč et al., 2021). In a first attempt to determine the molecular basis of such extraordinary dietary adaptations, we carried out a comparative transcriptomic study including species with different stenophagy levels from two evolutionary independent dietary shifts to feed on woodlice (Vizueta, Macías-Hernández, Arnedo, Rozas, & Sánchez-Gracia, 2019a) (Figure 1). We uncovered several gene targets that should be directly connected with diet diversification of this genus in this archipelago. Indeed, we identified genes involved in heavy metal detoxification and homeostasis and on the metabolism of some important nutrients and venom toxins, showing parallel changes in species with a more specialized diet.

Recently, we have reported a chromosome-scale genome assembly of *D. silvatica* (Escuer et al., 2022), a species with low levels of dietary specialization that inhabits the three more western Canary Islands (La Gomera, La Palma and El Hierro; Macías-Hernández, López, Roca-Cusachs, Oromí, & Arnedo, 2016). This species has a genome size of 1.7 Gb based on flow cytometry (1.4 Gb in the published assembly), arranged in seven holocentric chromosomes (Schrader, 1935).

Here, we are benefiting from this high-quality reference genome to carry out the first population genomics study in a species of the genus *Dysdera* using low-coverage whole-genome resequencing data from a natural population. Our main goal was to quantify and describe the levels of natural genomic variation in a population of an endemic species from the Macaronesia archipelago and to determine the impact of the recent adaptive radiation directed by dietary diversification on genome-wide patterns of polymorphism and divergence. Although other population genomic studies in spiders there have been published, very few have used whole-genome data (Chen et al., 2020; Hendrickx et al., 2022; Huang et al., 2022). We have applied a low-coverage sequencing strategy, which has become a very suitable cost-effective

strategy for conducting population genomic studies in non-model species with large genome sizes and having a gold-standard reference genome (Lou, Jacobs, Wilder, & Therkildsen, 2021). We have characterized genome-wide levels and patterns of nucleotide variation (single nucleotide polymorphism -SNP- data) and linkage disequilibrium (LD) in 12 genomes of *D. silvatica* sampled from a single locality in La Gomera. Likewise, we inferred the demographic history of the surveyed population and explored the putative impact of positive selection on genomic variation. Our results represent a significant contribution to the knowledge on population genomics in spiders, especially on endemic species involved in adaptive radiations.

Material and Methods

Sampling, DNA extraction and whole-genome sequencing

We collected 12 individuals of *D. silvatica* (two males and 10 females) from the same locality of La Gomera Island (Teselinde; Ermita de Santa Clara, Vallehermoso; Supplementary Table S1). Additionally, we collected one male of *D. bandamae* (in Gran Canaria; Llanos de la Pez; Tejada; Supplementary Table S1) to be used as outgroup to polarize the nucleotide changes. For DNA extraction, we used a modification of the Gentra Puregene Cell kit (Qiagen) that increases extraction efficiency in chelicerates. The extracted DNA was sequenced using NovaSeq6000 (for all *D. silvatica* samples) and HiSeqX (for the *D. bandamae* individual) (Illumina 150 bp, paired end reads, TruSeq libraries) in Macrogen Inc. (Seoul, Korea).

Due to the large genome size of *D. silvatica* (about 1.7 Gb estimated from flow cytometry; Sánchez-Herrero et al., 2019), we decided to use a low-coverage whole-genome resequencing strategy to generate the *D. silvatica* intraspecific data (average per individual raw sequencing depth of 6.7x; Supplementary Table S1). The average raw sequencing depth in the experiment of *D. bandamae* is 31.8x (Supplementary Table S1).

Read quality assessment, trimming and mapping

We used FASTQC v0.11.9 (Andrews, 2010) to assess the quality of raw reads and Trimmomatic v0.39 (Bolger, Lohse, & Usadel, 2014) for trimming adapters. For the last, we employed the specific list of adapters used in Truseq2 libraries to filter out all reads shorter

than 50 bp or with bad quality scores (< 15) across sliding windows of 4 bp length. We also remove leading and trailing bases with low quality scores (< 3) and all missing bases.

Filtered reads were then mapped, independently for each individual of both species, against the reference genome of *D. silvatica* (Escuer et al., 2022) using bwa mem v0.7.16 (Li & Durbin, 2009) labeling shorter split hits as secondary alignments (`-M`). Then, we used Samtools v1.11 (Danecek et al., 2021) to filter out these secondary alignments. We removed duplicates from the resulting BAM files, added read groups labels, and indexed and sorted the alignments using Picard Tools v2.26.10 (Broad Institute, 2016). Finally, we estimated the average and median read depth for each scaffold using Samtools v1.11 and Bedtools v2.29.1 (Quinlan & Hall, 2010), along with Qualimap v2.2.1 (Okonechnikov, Conesa, & García-Alcalde, 2016). The average net coverage per site of the *D. silvatica* sample is 44.3x. The average net coverage per individual ranges from 4x to 5x. Since *D. silvatica* males are X0 these average sequencing depths in the X chromosome are about half that of autosomes. (Supplementary Table S2). Given their different sample size, we carried out all the analysis separately for the autosomes (n = 24) and the X chromosomes (n = 20); we excluded from the analysis the two males since ANGSD (Korneliussen, Albrechtsen, & Nielsen, 2014) cannot handle haploid data (i.e., it calculates diploid genotype likelihoods).

Nucleotide genomic variation

We used the ANGSD and related software (e. g. ngsLD (Fox, Wright, Fumagalli, & Vieira, 2019), ngsDist (Vieira, Lassalle, Korneliussen, & Fumagalli, 2016) and ngsTools (Fumagalli, Vieira, Linderoth, & Nielsen, 2014)) that take genotype uncertainty into account for the downstream population genomic analysis.

To obtain a consensus sequence of *D. bandamae* in FASTA format we mapped the filtered reads of this species against the chromosome-level reference genome of *D. silvatica* using bwa mem v0.7.16 (with the `-M` option) and using also the `-doFasta` option of ANGSD. We applied specific quality filters in ANGSD to remove non-informative positions before estimating GLs. Specifically, we filtered out those positions with very low or very high net coverage across samples, (`-setMinDepth 3, -setMaxDepth 92`). These values correspond to the 2.5th and 97.5th percentiles of the per site net coverage distribution. In addition, we only used positions with high base calling quality, and properly paired and and high mapping quality reads (`-minQ 20, -only_proper_pairs and -minMapQ 20`).

We adjusted the mapQ parameter for excessive mismatches (`-C 50`) and the qscores around indels (`-baq 1`), and discarded not primary, failure and duplicate reads (-remove_bads) and reads that do not map uniquely (`-uniqueOnly`). We set the *D. silvatica* genome as a reference genome (`-ref`) and estimated the GLs for each individual using the GATK algorithm (`-GL 2`). To control for putative bias caused by the presence of repetitive elements (that might affect the mapping results), we also analyzed the data using i) different coverage filters (`-setMinDepth 20, -setMaxDepth 72`), that would include the 66.6% of the net coverage distribution, and ii) excluding masked regions of the genome identified by RepeatMasker (Supplemental Tables S3, S4).

We analyzed the levels and patterns of nucleotide variation from the unfolded site allele frequency likelihood (SAF) using the *realSFS* program from ANGSD and the sequence of *D. bandamae* as the ancestral reference (`-doMajorMinor 5, -anc`). The SAF was estimated based on individual GL assuming HWE (`-doSaf 1`). We calculated the number of polymorphic sites (*S*), the nucleotide diversity (π; Nei, 1987), the Watterson estimator of theta ($\theta_{w;}$ Watterson, 1975) and the Tajima's *D* (Tajima, 1989), Fu and Li's *D* and *F* (Fu & Li, 1993), Fay and Wu's *H* (Fay & Wu, 2000) and Zeng's *E* (Zeng, Fu, Shi, & Wu, 2006) test from the estimated SAF using the *thetaStat* program from ANGSD. We computed these summary statistics and neutrality tests for both the entire chromosome and in a sliding window approach (in non-overlapping windows of 1 and 50 kb).

We also estimated the site pairwise linkage disequilibrium (LD) using the ngsLD v1.1.1 software, also taking genotype uncertainty into account. For that, we first generated a GLs file in BEAGLE format (`-doGlf 2`) using ANGSD and applying the same filters and options as described above. Then, LD was estimated as the average of the Pearson correlation (r^2) between genotypes (i.e., the *ZnS* statistic; Kelly, 1997) in non-overlapping windows of 1, 5 and 50 kb. To reduce the number of pairwise comparisons, we randomly sampled 0.1% of the sites to estimate the decay of LD with distance.

We used ngsDist v.1.0.10 to compute the nucleotide divergence between *D. silvatica* and *D. bandamae* based on the estimated GLs. For that, we first generated a GLs file in BEAGLE format using the BAM files of the 13 individuals (12 *D. silvatica* and one *D. bandamae*). Using these GLs, we calculated the JC69 corrected distance between the two species as the average of corrected pairwise distances between each *D. silvatica* individual and *D. bandamae* (`--`

`evol_model 2`; Jukes & Cantor, 1969). We computed the nucleotide divergence for both for the entire chromosome and using a sliding window approach (in non-overlapping windows of 1 and 50 kb).

To investigate a possible population structure in our sample, we carried out a Principal Component Analysis (PCA) using the ngsCovar software from ngsTools v3. We based this analysis on the estimated genotype posterior probabilities obtained using ANGSD and separately for the autosomes (12 *D. silvatica* and one *D. bandamae*) and the X chromosome (10 *D. silvatica* females and one *D. bandamae*). We applied the same filters and options used to estimate GL but adding `-doGeno 32, -doMaf 1`, excluding triallelic positions (`-SkipTrialletic 1`), and filtered out those SNPs with *P*-values > 1 x 10^{-6} (`-snp_pval`).

Population demographic history

We used Stairway Plot2 (Liu & Fu, 2020) and the maximum likelihood estimate of the unfolded SFS (previously obtained using ANGSD) to infer the recent demographic history of the surveyed *D. silvatica* population. The analysis was conducted separately for the autosomes and the X chromosome. We set the generation time to 1.5 years (Cooke, 1965), and the neutral mutation rate per site and per generation (μ) to 1.0 x 10^{-8}. This mutation rate was obtained from the nucleotide divergence estimated between *D. silvatica* and *D. bandamae* (K = 0.120; Table 1), and an expected divergence time for these species of 6 mya (Macías-Hernández, Oromí, & Arnedo, 2008; Macías-Hernández, Bidegaray-Batista, Emerson, Oromí, & Arnedo, 2013; M. A. Arnedo; unpublished results).

Genome scans for selection

We used the RAiSD v2.9 software (Alachiotis & Pavlidis, 2018) to search for the characteristic hallmarks of positive selection across the genome. This method relies on a composite statistic (μ-*R*), calculated by means of a SNP-driven, sliding-window algorithm, which integrates information from the different pieces of evidence that characterize selective sweeps, such as μ-*var* (reduction of nucleotide variation), μ-*sfs* (shifts to low and high-frequency derived variants), μ-*ld* (elevated levels of LD around the selective site). Since RAiSD needs a VCF file with called genotypes as the input, we first run ANGSD with the options `-doGeno 1, -doMajorMinor 5, -doPost 1, -skipTriallelic 1, -snp_pval 1e-6` and `-dobcf` and then vcfutils.pl to convert the resulting bcf file into vcf format.

We used a three-step strategy to identify outliers in the genome-wide empirical distribution of μ-*R*. First, we retained those windows (we used the default window length of 50 SNPs) with the top 0.1% values of the μ-*R* statistic. Second, we searched for consecutive windows showing μ-*R* values above 0.01% of the empirical distribution and, among them, we selected those having at least 2 consecutive windows with μ-*R* values above the 0.001%. We then used *D. silvatica* structural annotations to identify genes or other functional elements located within or near the outlier regions (up to 20 kb upstream or downstream the focal region). Genes identified that did not have any functional annotation were query to perform a Blastp (Altschul et al., 1997) search against NCBI non redundant protein database.

McDonald Kreitman test

We applied the McDonald and Kreitman test (MKT) (McDonald & Kreitman, 1991) to detect the footprint of positive selection in the *D. silvatica* protein-coding genes. We first called the genotypes of the 12 individuals of *D. silvatica* as well as the single individual of *D. bandamae* using ANGSD and applying the same filters and options to estimate GLs (see above). *D. bandamae* was used as an outgroup to calculate synonymous and nonsynonymous substitutions. We used snpEff v.5.1d (Cingolani et al., 2012) and the VCF file generated by ANGSD to predict the functional effects of genomic variants (including synonymous and nonsynonymous mutations). The annotated VCF was used to run the MKT on all *D. silvatica* functionally annotated protein-coding genes using a modification of Tomas Blankers' script (available at https://github.com/thomasblankers/popgen/blob/master/MKTtest) and the gene coordinates in the GFF file of the *D. silvatica* reference genome (available at https://github.com/molevol-ub/Dysdera_silvatica_genome/tree/master/v2). We filtered sites with a minor allele frequency < 0.1 (i.e., singletons).

Mitochondrial assembly and phylogenetic analyses

We assembled the mitochondrial genomes from the raw reads of each individual of *D. silvatica* and the outgroup *D. bandamae* with NOVOplasty v4.3 (Dierckxsens, Mardulyn, & Smits, 2017) applying a Genome Range of 14000-20000 bp and K-mer of 33. Once the sequences were obtained, we performed a multiple sequence alignment using MAFFT (Katoh & Standley, 2016) applying the options `--maxiterate 1000, --globalpair` or G-INS-i algorithm to perform an iterative alignment. Then, IQ-TREE2 v1.6.12 (Minh et al., 2020) was used with

`-m MFP -B 1000` parameters for the phylogenetic inferences, and the graphical visualization and tree editing was performed using the web interface iTOL (Letunic & Bork, 2021).

GO enrichment analysis

We used the R package GOstats (Falcon & Gentleman, 2007) to perform gene ontology (GO) enrichment analyses. Then, we also used the web server REVIGO (Supek, Bošnjak, Škunca, & Šmuc, 2011) to generate a graphical visualization of the obtained results.

Results

After excluding those positions that did not pass our strict quality filters, we analyze 632.64 Mb of autosomal DNA positions (across the six autosomes), and 219.80 Mb in the X chromosome, about 72% of the total positions (Table 1). After variant calling, we genotyped 50,518,353 and 7,718,288 SNPs in the autosomes and the X chromosome, respectively (Table 1).

Population structure analysis

To investigate the genetic structure of the surveyed population, we first inferred the phylogenetic relationships among samples using the assembled mitogenomes of all individuals, including the one from *D. bandamae*. We do not find in the phylogenetic tree any relevant population structure using information from the whole mtDNA (Supplementary Figure S1). Nevertheless, the results of the PCA analysis based on both the autosomes and X chromosome (Figure 2) show that some individuals of *D. silvatica* are slightly more differentiated than the others (explained by the PC2). However, this accounted for only 7% of the variance (in both autosomal and X chromosome). Obviously, most of the variance explained by PC1 is introduced by *D. bandamae*.

Nucleotide polymorphism, divergence and linkage disequilibrium

Autosomes display high levels of nucleotide diversity (π = 0.015; Table 1), which is homogeneously distributed among the assembled chromosomes (Table 1; Figure 3a). The X chromosome is much less variable than the autosomes in the sampled population (π = 0.008;

Table 1; Figure 3a), a feature that could be attributable by the smaller effective population size of the sex chromosome. The values of Watterson's estimator of θ (θ_w) are systematically higher than nucleotide diversity and, in concordance, Tajima's *D* (Table 2) values are consistently negative across the genome. Overall, autosomes exhibit more extreme values than the X chromosome and, moreover, are rather homogeneous across the different autosomes (Table 2). Nucleotide divergence is also homogeneously distributed among chromosomes, with an average of $K = 0.120$ (Table 1; Figure 3b). On the other hand, linkage disequilibrium, estimated as *ZnS*, decays rapidly with distance in the *D. silvatica* genome (Table 1; Figure 3d, Supplementary Table S5), pointing to high levels of genetic recombination in the genome of this spider. As expected, the *ZnS* values estimated in the X chromosome are consistently higher than in the autosomes.

Demographic inference analysis

We estimated the past demographic history of the surveyed population using a coalescent based approach (Liu, Anderson, Pearl, & Edwards, 2019; Liu, Wu, & Yu, 2015) (Figure 4). As expected, the history inferred from the autosomal and the X chromosome SNP data is greatly concordant. There are, however, some differences in the parameter estimates that could reflect distinct mutational or selective forces undergone by the X chromosome. We inferred two major demographic events (supported by bootstrap) in the recent history of *D. silvatica.* The first event reflects a population bottleneck that occurred at about 200 kya (based on autosomal data) and dropped the population size by a factor of 5, and with a duration of ~35 kya. After the complete recovery, the effective size of the population remained stable until ~10 kya, when a gradual size reduction started and lasted to the present, where we estimated a N_e of ~25 x 10^3 individuals (Figure 4a).

Genomic scans for selection

Using the RAiSD software we identified 20 genomic regions as candidates to have been the target of positive selection, 19 of them with two or more consecutive windows with μ-*R* values above the 0.001% (Supplementary Tables S6 and S7). These candidate regions were unevenly distributed across chromosomes, being overrepresented in the X chromosome (i. e., 42% of candidate regions across only 26% of the genome assembly). Nevertheless, both the highest value of the μ-*R* statistic, and the largest number of consecutive significant windows are found in chromosome 1 (Figure 5). In this chromosome we identified two highly significant regions

(labeled as ID1 and ID2; Supplementary Table S6), with a maximum μ-*R* values of 430.9 and 141.8, respectively, a region that spans over 130 kb. Moreover, these regions also exhibit the highest significant Tajima's *D* values (Supplementary Table S8). Remarkably these regions are also physically close to another candidate region (ID3), that could also be part of the same selective sweep event encompassing 257 kb. We checked some putative factors that could generate putative false positives, such as the fraction of missing data or the amount of repetitive (masked) regions on the different components of the μ-*R* statistic (see Supplementary Results for a detailed description of the analysis), and we did not find any substantial effect on the 19 outlier regions (Supplementary Tables S9 and S10).

We identified 38 protein-coding genes across these 19 candidate regions (Supplemental Table S7), having 28 of them GO annotations, one of them encoding a transposable element protein. The remaining 10 genes show no functional evidence (Supplemental Table S7). The genes included in the candidate regions at chromosome 1 have multiple GO associated with relevant adaptive functions, such as lipid biosynthetic processes and activation of appetite (ID1 region); protein binding (ID2 region), or DNA integration in viral capsid (ID3 region), among others. In ID4, we found GO related to detection of chemical stimuli involved in sensory perception of bitter taste, RNA splicing and defense response against viruses.

The GO enrichment analysis of candidate regions showed 984 biological processes (170 of them with a *P*-value of lower than 0.001) and 256 molecular functions (43 of them with a *P*-value lower than of 0.001) significantly overrepresented in this set of genes (Supplementary Tables S11 and S12). The GO terms with the highest *P*-values include molecular functions such as "protein binding", "rhodopsin kinase activity" (acyltransferase activity), "oxidoreductase activity", "metal and zinc binding", "transposase activity", "lipid metabolism", "cadherin activity and binding" and biological processes like "nitrogen compound transport", "response to chemical" and "(lipid) biosynthetic process", among others (Supplementary Figure S2). Some of these terms are also identified in candidate regions at Chromosome 1.

McDonald and Kreitman test

We applied the MKT to all protein-coding genes with functional annotations in *D. silvatica* (28,900; after discarding four putative pseudogenes). Unfortunately, 25,657 of them do not

have enough variation (in two or more variant classes), preventing the performance of the MKT analysis. Among the 3,243 analyzed tests, 104 genes showed statistically significant departures from the neutral expectations (*P*-value < 0.05), being 49 and 55 of them predicted to be under negative or positive selection, respectively (Supplementary Table S13). Only 14 of the positive selection candidates have some synonymous fixation (Supplementary Table S13). The short genetic distance of *D. silvatica* to the outgroup is the most likely cause of this lack of power of the MKT.

The GO enrichment analysis of genes encompassing significant MKT values compatible with positive selection (Supplementary Table S14) identified 1,457 biological processes and 361 molecular functions (Supplementary Tables S15 and S16; Supplementary Figure S3). Enriched GO terms included "cellular localization", "response of abiotic stimulus", "regulation of biological quality", "protein and metal binding" and "oxidoreductase activity".

Discussion

Here, we report the first comprehensive population genomics analysis of a Canary Island endemic species, the nocturnal ground-dwelling spider *D. silvatica*. We applied a low-coverage whole-genome resequencing strategy because it represents the best balance between statistical inference power and sequencing costs, in an organism with a quite large genome. Although we have not used ultra-low coverage (<0.5× coverage) in our study, there may still be considerable uncertainty in genotype calling, potentially leading to errors or biases in our downstream analyses. A growing number of tools have become available to explicitly account for genotype uncertainty. Here, we used the software ANGSD (and others related) because it not only calculates GLs but also provides various summary statistics and performs population genetic analyses all in the same suit, making the analysis more consistent.

High levels of nucleotide polymorphism in an island endemic spider

Surprisingly we have found that *D. silvatica* harbors very high nucleotide diversity ($\pi = 0.015$; for autosomes). Despite being an endemic species living in three small islands, the species shows exceptional levels of nucleotide diversity, close to the median diversity of Arthropods (Romiguier et al., 2014). It is well known that estimates of nucleotide diversity using Next Generation Sequencing (NGS) data could be inflated by the erroneous mapping to/against

repetitive regions; however, our complementary analysis filtering positions with different extreme coverage values or masking repetitive regions (identified with Repeat Masker software) demonstrated that our estimations are sound (Supplementary Tables S3, S4).

It is difficult to draw conclusions about these high levels of variability due to the paucity of population genomics studies in other spider species, or even other chelicerates. Still, current nucleotide diversity values estimated in *D. silvatica* are vastly higher than those calculated in the few available studies within this group, which include mites (π = 0.0006; Chen et al., 2020) and social spiders (e.g., *Stegodyphus sarasinorum*; π = 0.0005; Settepani et al., 2017). Moreover, these values are even higher than those reported for the cosmopolitan fruit-fly *Drosophila melanogaster* (genome-wide averages of π = 0.003-0.006) and, more notably, than that estimated in high-recombination regions (π < 0.01) of Kapun et al. (2021) report. Under the drift-neutral mutation equilibrium ($\pi = 4N_e\mu$), these high levels of nucleotide polymorphism could indicate either large historical effective population size, or high mutation rates in this species. Based on our divergence data we estimated that the neutral mutation rate was μ = 1.0 x 10^{-8}, which in addition is a typical value for eukaryotic data. Assuming a random mating population with constant population size, current nucleotide diversity values would imply an effective population size of 375,000 individuals. Indeed, our linkage disequilibrium estimates, and the demographic inference based on the coalescence, would suggest even higher N_e in the past history of this species (see below).

Besides, the levels of intraspecific variation, but not the divergence estimates, differ between the X chromosome and the autosomes of *D. silvatica*, probably reflecting the expected smaller effective size of the sex chromosome. Even so, the ratio X to autosomes (0.53) is much lower than theoretically expected (0.75) and even far from the median value estimated for Arthropods (~1; Leffler et al., 2012), which indicate that likely other mutational or selective factors are reducing the variability in the X chromosome of this population. On the other hand, our results do not support the 'faster X' hypothesis (Bechsgaard et al., 2019; Charlesworth, Campos, & Jackson, 2018) since the number of protein-coding genes with evidence of positively selected substitutions inferred by the MKT is not proportionally higher on the X than on the autosomes (Supplementary Table S13). However, the number of genes and substitutions used in the MKT analysis is too low to draw a firm conclusion. A more distant species will be necessary to obtain a large number of synonymous and nonsynonymous substitutions between species.

In line with genomic diversity estimations, recombination rate in *D. silvatica* also appears to be considerably high, with a LD decay at the same order of that found in species with large population sizes (Signor, New & Nuzhdin, 2018). Intriguingly, we have not detected a clear uneven distribution of LD, or nucleotide patterns, along the chromosomes (e.g., local drops) as that observed in *Drosophila* (MacKay et al., 2012; Figure 3). These results may be linked to the particular organization of the holokinetic chromosomes of this species, where chromosomes have multiple and diffused kinetochores, and deserve further study in the future.

Evidence of very large ancestral effective population sizes and a very recent decline

The high levels of nucleotide variability observed in *D. silvatica* could also be compatible with some putative genetic structure in the population. Although the PCA analysis uncovers some structure in both the X chromosome and the autosomes, this feature is not detected using mitogenomes. Since the potentially differentiated individuals do not match between markers, we could exclude a recent introgression or a population admixture event, as the main cause explaining these high levels of intraspecific polymorphism. Besides, the results from the demographic inference analysis based on the SFS indicate a very large ancestral population size of *D. silvatica* (~2.0 x 10^6 individuals) that just started to decline ~10 kya ago. This effective population size is surprisingly large, and perhaps could reflect some ancestral population structure that would have been maintained for a long time (~200 kya), ending a few thousands of years ago. After the admixture, the sorting of the alleles in loci that were differentiated during the previous structuring phase, would have been uneven distributed in different (independent) genomic regions. This would explain that the individuals that show genetic differentiation in the present are different when the analysis was based on different markers (i.e., alleles from different markers would coalesce in different ancestral populations).

Current data does not allow to determine the origin of this potential ancestral genetic structure. Even though individuals of *D. silvatica* were sampled in a single locality in La Gomera (a very small island of only 370 km²), this species can be also found in other Canary Islands such as La Palma and El Hierro. Indeed, results based on mitochondrial data show that the populations from these three islands would be differentiated, even being sibling species (M. A. Arnedo, unpublished results). It is conceivable that this large ancestral effective population size reflects a recent secondary contact between these differentiated populations or sibling species from

different islands (some thousands of years ago). Furthermore, other studies have also uncovered a clear intra-island genetic structure between populations of *Dysdera vernaui* (another *Dysdera* species endemic from Tenerife), likely driven by rounds of volcanic events and sudden climatic changes (Macías-Hernández et al., 2013). This successive series of geographic isolation events within the same island could also have occurred in *D. silvatica* from la Gomera. Further analyses including *D. silvatica* population genomic data from different localities from La Gomera, Hierro and La Palma islands, as well as samples from closely related species (to exclude other competitive factors such as ghost introgression from a closely related species), are required to determine the impact of a genetic structure as the main responsible for the unexpected high genetic diversity levels. This issue is becoming an increasingly relevant question to understand the origin, maintenance and fate of biodiversity, and therefore is instrumental to manage and conserve biodiversity in a changing world (Mergeay & Santamaria, 2012).

Recent positive selection in *D. silvatica*

We have pinpointed some recent selective sweeps in the genome of *D. silvatica*, most of them located in the chromosome X (Supplemental Table S7); however, the most significant hallmark is in chromosome 1, with a putative selective sweep encompassing up to 257 kb (Figure 5; Supplemental Table S7). Moreover, we have also identified a protein-coding gene encoded by a transposable element (in Chromosome 3). In the candidate regions, we identified 38 protein-coding genes, 28 of them with GO annotation. We performed some additional Blastp searches to determine the function of the 10 additional ones; we found homologs of 5 of them in Arthropods (including 3 spider-specific). Therefore, spider-specific genes seem to have been frequently the targets of selection in the recent past of this species. Moreover, we have also identified a protein-coding gene encoded by a transposable element (in Chromosome 3).

Results of the GO enrichment analysis of the selective sweep candidates reveals that in *D. silvatica,* recent (or ongoing) molecular adaptation seems to be associated with dietary diversification (Supplementary Table S11, S12). For instance, we found more genes than expected by chance involved in “nitrogen transport”, “secretion”, “nitrogen metabolic process” and “metal/zinc binding”, functions that are directly related with the detoxification and nutrient assimilation processes associated with feeding on woodlice, a dietary specialization that appear to be directly related with the high species richness in this genus (Řezáč et al., 2021).

Additionally, some enriched GO terms among the candidates from the RAiSD and MKT analyses are associated with other functions also identified in Vizueta, Macías-Hernández, Arnedo, Rozas, & Sánchez-Gracia (2019), such as "iron binding", activities of enzymes such as "phosphatase and hydrolase" and "transposase activity" (Supplementary Tables S12, S16 and Supplementary Figures S2, S3). It is largely known that in animals transposable elements are involved in the gene regulation of biological relevant genes (O'Brochta & Atkinson, 1996), so, we can hypothesize that some *D. silvatica* genes are probably the target of adaptive transpositions in their regulatory regions. Moreover, another enriched term, "lipid metabolism", would suggest that recent adaptation in this species could involve aspects of reproduction, starvation, or immunity (Toprak, 2020). Interestingly, "rhodopsin kinase activity" is also strongly enriched (Supplementary Table S12). Since *Dysdera* spiders are nocturnal hunters, without clearly developed eyes, (Morehouse, Buschbeck, Zurek, Steck, & Porter, 2017), some of the detected selective sweeps could have been caused by beneficial mutations related to dark adaptation. Indeed, in darkness the exposure to intense illumination bleaches the fraction of the rhodopsin photopigment hindering a clear vision for some time; therefore, it is conceivable that positive selection could shape genes involved in rhodopsin biosynthesis to facilitate the dark adaptation (Leibrock, Reuter, & Lamb, 1998). Finally, the "response and detection of chemical stimulus" is also enriched among candidates, clearly suggesting that the chemosensory system could have played a relevant role in the recent evolution of the *D. silvatica*.

Acknowledgements

This work was supported by the Ministerio de Economía y Competitividad of Spain (PID2019-103947GB, PID2019-105794GB) and the Comissió Interdepartamental de Recerca I Innovació Tecnològica of Catalonia, Spain (2017SGR83, 2017SGR1287). P. E. was supported by an FPI grant (Ministerio de Economía y Competitividad of Spain, BES-2017-081740). We acknowledge the Garajonay National Parks for granting collection permits and helping with lodging and logistics during fieldwork.

ORCID

Paula Escuer https://orcid.org/0000-0002-5941-0106

Sara Guirao-Rico https://orcid.org/0000-0001-9896-4665

Miquel A. Arnedo https://orcid.org/0000-0003-1402-4727

Alejandro Sánchez-Gracia https://orcid.org/0000-0003-4543-4577

Julio Rozas https://orcid.org/0000-0002-6839-9148

Availability of data

Raw sequence reads have been deposited at the DDBJ/ENA/GenBank under Bioproject XXX, with Sequence Read Archive (SRA) accession numbers YYYY. Other relevant results are available in the Supplementary Information online.

Bibliography

1. Alachiotis, N., & Pavlidis, P. (2018). RAiSD detects positive selection based on multiple signatures of a selective sweep and SNP vectors. *Communications Biology 2018 1:1*, *1*(1), 1–11. doi: 10.1038/s42003-018-0085-8

2. Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research*, Vol. 25, pp. 3389–3402. doi: 10.1093/nar/25.17.3389

3. Andrews, S. (2010). *FastQC: A Quality Control Tool for High Throughput Sequence Data*.

4. Arnedo, M. A., Oromí, P., Múrria, C., Macías-Hernández, N., & Ribera, C. (2007). The dark side of an island radiation: Systematics and evolution of troglobitic spiders of the genus Dysdera Latreille (Araneae:Dysderidae) in the Canary Islands. *Invertebrate Systematics*, *21*(6), 623–660. doi: 10.1071/IS07015

5. Arnedo, M. A., Oromí, P., & Ribera, C. (2001). Radiation of the spider genus Dysdera (Araneae, Dysderidae) in the Canary Islands: Cladistic assessment based on multiple data sets. *Cladistics*, *17*(4), 313–353. doi: 10.1006/clad.2001.0168

6. Austin, J. J., & Arnold, E. N. (2001). Ancient mitochondrial DNA and morphology elucidate an extinct island radiation of Indian Ocean giant tortoises (Cylindraspis). *Proceedings of the Royal Society B: Biological Sciences*, *268*(1485), 2515–2523. doi: 10.1098/rspb.2001.1825

7. Bechsgaard, J., Schou, M. F., Vanthournout, B., Hendrickx, F., Knudsen, B., Settepani, V., … Bilde, T. (2019). Evidence for faster X chromosome evolution in spiders. *Molecular Biology and Evolution*, *36*(6), 1281–1293. doi: 10.1093/molbev/msz074

8. Bolger, A. M.; Lohse, M.; Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina Sequence Data. *Bioinformatics*, *btu170*.

9. Carson, H. L., & Kaneshiro, K. Y. (1976). Drosophila of Hawaii: Systematics and Ecological Genetics. *Annual Review of Ecology and Systematics*, *7*(1), 311–345. doi: 10.1146/annurev.es.07.110176.001523

10. Charlesworth, B., Campos, J. L., & Jackson, B. C. (2018). Faster-X evolution: Theory and evidence from Drosophila. *Molecular Ecology*, *27*(19), 3753–3771. doi: 10.1111/mec.14534

11. Chen, L., Sun, J. T., Jin, P. Y., Hoffmann, A. A., Bing, X. L., Zhao, D. S., … Hong, X. Y. (2020). Population genomic data in spider mites point to a role for local adaptation in shaping range shifts. *Evolutionary Applications*, *13*(10), 2821–2835. doi: 10.1111/eva.13086

12. Choi, J. Y., Dai, X., Alam, O., Peng, J. Z., Rughani, P., Hickey, S., … Stacy, E. A. (2021). Ancestral polymorphisms shape the adaptive radiation of Metrosideros across the Hawaiian Islands. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(37). doi: 10.1073/pnas.2023801118

13. Choi, Y. J., Fontenla, S., Fischer, P. U., Le, T. H., Costábile, A., Blair, D., … Mitreva, M. (2020). Adaptive Radiation of the Flukes of the Family Fasciolidae Inferred from Genome-Wide Comparisons of Key Species. *Molecular Biology and Evolution*, *37*(1), 84–99. doi: 10.1093/molbev/msz204

14. Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., … Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly*, *6*(2), 80–92. doi: 10.4161/fly.19695

15. Cooke, J. A. L. (1965). Spider Genus Dysdera (Araneae, Dysderidae). *Nature 1965 205:4975*, *205*(4975), 1027–1028. doi: 10.1038/2051027b0

16. Coyne, J. A., & Orr, H. A. (2004). *Speciation: a catalogue and critique of species concepts.* (P. of biology: an Anthology, Ed.).

17. Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., … Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*(2), 1–4. doi: 10.1093/gigascience/giab008

18. Dierckxsens, N., Mardulyn, P., & Smits, G. (2017). NOVOPlasty: De novo assembly of organelle genomes from whole genome data. *Nucleic Acids Research*, *45*(4), e18–e18. doi: 10.1093/nar/gkw955

19. Emerson, B. C. (2002). Evolution on oceanic islands: molecular phylogenetic approaches to understanding pattern and process. *Molecular Ecology*, *11*(6), 951–966. doi: 10.1046/J.1365-294X.2002.01507.X

20. Escuer, P., Pisarenco, V. A., Fernández-Ruiz, A. A., Vizueta, J., Sánchez-Herrero, J. F., Arnedo, M. A., … Rozas, J. (2022). The chromosome-scale assembly of the Canary Islands endemic spider Dysdera silvatica (Arachnida, Araneae) sheds light on the origin and genome structure of chemoreceptor gene families in chelicerates. *Molecular Ecology Resources*, *22*(1), 375–390. doi: 10.1111/1755-0998.13471

21. Falcon, S., & Gentleman, R. (2007). Using GOstats to test gene lists for GO term association. *Bioinformatics*, *23*(2), 257–258. doi: 10.1093/bioinformatics/btl567

22. Fay, J. C., & Wu, C. I. (2000). Hitchhiking Under Positive Darwinian Selection. *Genetics*, *155*(3), 1405–1413. doi: 10.1093/GENETICS/155.3.1405

23. Feder, J. L., Egan, S. P., & Nosil, P. (2012, July). The genomics of speciation-with-gene-flow. *Trends in Genetics*, Vol. 28, pp. 342–350. Trends Genet. doi: 10.1016/j.tig.2012.03.009

24. Fernández-Mazuecos, M., Vargas, P., McCauley, R. A., Monjas, D., Otero, A., Chaves, J. A., … Rivas-Torres, G. (2020). The Radiation of Darwin's Giant Daisies in the Galápagos Islands. *Current Biology*, *30*(24), 4989-4998.e7. doi: 10.1016/J.CUB.2020.09.019

25. Fernández-Palacios, J. M. (2011). The islands of Macaronesia. *Terrestrial Arthropods of Macaronesia - Biodiversity, Ecology and Evolution*, 1–30.

26. Fox, E. A., Wright, A. E., Fumagalli, M., & Vieira, F. G. (2019). NgsLD: Evaluating linkage disequilibrium using genotype likelihoods. *Bioinformatics*, *35*(19), 3855–3856. doi: 10.1093/bioinformatics/btz200

27. Fu, Y. X., & Li, W. H. (1993). Statistical tests of neutrality of mutations. *Genetics*, *133*(3), 693–709. doi: 10.1093/GENETICS/133.3.693

28. Fumagalli, M., Vieira, F. G., Linderoth, T., & Nielsen, R. (2014). NgsTools: Methods for population genetics analyses from next-generation sequencing data. *Bioinformatics*, *30*(10), 1486–1487. doi: 10.1093/bioinformatics/btu041

29. Gillespie, R. (2004). Community Assembly through Adaptive Radiation in Hawaiian Spiders. *Science*, *303*(5656), 356–359. doi: 10.1126/science.1091875

30. Grant, B. R., & Grant, P. R. (2008, September 9). Fission and fusion of Darwin's finches populations. *Philosophical Transactions of the Royal Society B: Biological Sciences*, Vol. 363, pp. 2821–2829. The Royal Society. doi: 10.1098/rstb.2008.0051

31. Hendrickx, F., De Corte, Z., Sonet, G., Van Belleghem, S. M., Köstlbacher, S., & Vangestel, C. (2022). A masculinizing supergene underlies an exaggerated male reproductive morph in a spider. *Nature Ecology and Evolution*, *6*(2), 195–206. doi: 10.1038/s41559-021-01626-6

32. Hodges, S. A., & Derieg, N. J. (2009, June 16). Adaptive radiations: from field to genomic studies. *Proceedings of the National Academy of Sciences of the United States of America*, pp. 9947–9954. National Academy of Sciences. doi: 10.1073/pnas.0901594106

33. Hopkin, S. P., & Martin, M. H. (1985). Assimilation of zinc, cadmium, lead, copper, and iron by the spider Dysdera crocata, a predator of woodlice. *Bulletin of Environmental Contamination and Toxicology*, *34*(1), 183–187. doi: 10.1007/BF01609722

34. Huang, X., Wang, W., Gong, T., Wickell, D., Kuo, L. Y., Zhang, X., … Li, Q. (2022). The flying spider-monkey tree fern genome provides insights into fern evolution and arborescence. *Nature Plants 2022 8:5*, *8*(5), 500–512. doi: 10.1038/s41477-022-01146-6

35. Institute, B. (2016). "Picard tools." *Broad Institute, GitHub Repository*. Retrieved from https://broadinstitute.github.io/picard/

36. Juan, C., Emerson, B. C., Oromí, P., & Hewitt, G. M. (2000, March 1). Colonization and diversification: Towards a phylogeographic synthesis for the Canary Islands. *Trends in Ecology and Evolution*, Vol. 15, pp. 104–109. Elsevier Ltd. doi: 10.1016/S0169-5347(99)01776-0

37. Jukes, T. H., & Cantor, C. R. (1969). Evolution of protein molecules. *Mammalian Protein Metabolism*, *3*, 21–132.

38. Kapun, M., Nunez, J. C. B., Bogaerts-Márquez, M., Murga-Moreno, J., Paris, M., Outten, J., … Bergland, A. O. (2021). Drosophila Evolution over Space and Time (DEST): A New Population Genomics Resource. *Molecular Biology and Evolution*, *38*(12), 5782–5805. doi: 10.1093/molbev/msab259

39. Katoh, K., & Standley, D. M. (2016). A simple method to control over-alignment in the MAFFT multiple sequence alignment program. *Bioinformatics*, btw108. doi: 10.1093/bioinformatics/btw108

40. Kelly, J. K. (1997). A Test of Neutrality Based on Interlocus Associations. *Genetics*, *146*(3), 1197–1206. doi: 10.1093/GENETICS/146.3.1197

41. Korneliussen, T. S., Albrechtsen, A., & Nielsen, R. (2014). ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinformatics*, *15*(1), 1–13. doi: 10.1186/s12859-014-0356-4

42. Lamichhaney, S., Berglund, J., Almén, M. S., Maqbool, K., Grabherr, M., Martinez-Barrio, A., … Andersson, L. (2015). Evolution of Darwin's finches and their beaks revealed by genome sequencing. *Nature*, *518*(7539), 371–375. doi: 10.1038/nature14181

43. Leffler, E. M., Bullaughey, K., Matute, D. R., Meyer, W. K., Ségurel, L., Venkat, A., … Przeworski, M. (2012). Revisiting an Old Riddle: What Determines Genetic Diversity Levels within Species? *PLOS Biology*, *10*(9), e1001388. doi: 10.1371/JOURNAL.PBIO.1001388

44. Leibrock, C. S., Reuter, T., & Lamb, T. D. (1998). Molecular basis of dark adaptation in rod photoreceptors. *Eye (Basingstoke)*, *12*(3), 511–520. doi: 10.1038/eye.1998.139

45. Letunic, I., & Bork, P. (2021). Interactive tree of life (iTOL) v5: An online tool for phylogenetic tree display and annotation. *Nucleic Acids Research*, *49*(W1), W293–W296. doi: 10.1093/nar/gkab301

46. Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760. doi: 10.1093/bioinformatics/btp324

47. Liu, L., Anderson, C., Pearl, D., & Edwards, S. V. (2019). Modern phylogenomics: Building phylogenetic trees using the multispecies coalescent model. *Methods in Molecular Biology*, *1910*, 211–239. doi: 10.1007/978-1-4939-9074-0_7/FIGURES/5

48. Liu, L., Wu, S., & Yu, L. (2015, September 1). Coalescent methods for estimating species trees from phylogenomic data. *Journal of Systematics and Evolution*, Vol. 53, pp. 380–390. John Wiley & Sons, Ltd. doi: 10.1111/jse.12160

49. Liu, X., & Fu, Y. X. (2020). Stairway Plot 2: demographic history inference with folded SNP frequency spectra. *Genome Biology*, *21*(1), 1–9. doi: https://doi.org/10.1186/s13059-020-02196-9

50. Lou, R. N., Jacobs, A., Wilder, A. P., & Therkildsen, N. O. (2021). A beginner's guide to low-coverage whole genome sequencing for population genomics. *Molecular Ecology*, *30*(23), 5966–5993. Mol Ecol. doi: 10.1111/mec.16077

51. Machado, A., Rodríguez-Expósito, E., López, M., & Hernández, M. (2017). Phylogenetic analysis of the genus laparocerus, with comments on colonisation and diversification in macaronesia (coleoptera, curculionidae, entiminae). *ZooKeys*, *2017*(651), 1–77. doi: 10.3897/zookeys.651.10097

52. Macías-Hernández, N., Bidegaray-Batista, L., Emerson, B. C., Oromí, P., & Arnedo, M. (2013). The Imprint of Geologic History on Within-Island Diversification of Woodlouse-Hunter Spiders (Araneae, Dysderidae) in the Canary Islands. *Journal of Heredity*, *104*(3), 341–356. doi: 10.1093/JHERED/EST008

53. Macías-Hernández, N., López, S. de la C., Roca-Cusachs, M., Oromí, P., & Arnedo, M. A. (2016). A geographical distribution database of the genus Dysdera in the Canary Islands (Araneae, Dysderidae). *ZooKeys*, *2016*(625), 11. doi: 10.3897/ZOOKEYS.625.9847

54. Macías-Hernández, N., Oromí, P., & Arnedo, M. A. (2008). Patterns of diversification on old volcanic islands as revealed by the woodlouse-hunter spider genus Dysdera (Araneae, Dysderidae) in the eastern Canary Islands. *Biological Journal of the Linnean Society*, *94*(3), 589–615. doi: 10.1111/j.1095-8312.2008.01007.x

55. MacKay, T. F. C., Richards, S., Stone, E. A., Barbadilla, A., Ayroles, J. F., Zhu, D., … Gibbs, R. A. (2012). The Drosophila melanogaster Genetic Reference Panel. *Nature 2012 482:7384*, *482*(7384), 173–178. doi: 10.1038/nature10811

56. Martín, J. L., Cardoso, P., Arechavaleta, M., Borges, P. A. V., Faria, B. F., Abreu, C., … Mendonça, E. (2010). Using taxonomically unbiased criteria to prioritize resource allocation for oceanic island species conservation. *Biodiversity and Conservation*, *19*(6), 1659–1682. doi: 10.1007/s10531-010-9795-z

57. McDonald, J. H., & Kreitman, M. (1991). Adaptive protein evolution at the Adh locus in Drosophila. *Nature*, *351*(6328), 652–654. doi: 10.1038/351652a0

58. Mergeay, J., & Santamaria, L. (2012, February). Evolution and Biodiversity: The evolutionary basis of biodiversity and its potential for adaptation to global change. *Evolutionary Applications*, Vol. 5, pp. 103–106. Wiley-Blackwell. doi: 10.1111/j.1752-4571.2011.00232.x

59. Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., Von Haeseler, A., … Teeling, E. (2020). IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution*, *37*(5), 1530–1534. doi: 10.1093/molbev/msaa015

60. Morehouse, N. I., Buschbeck, E. K., Zurek, D. B., Steck, M., & Porter, M. L. (2017). Molecular evolution of spider vision: New opportunities, familiar players. *Biological Bulletin*, *233*(1), 21–38. doi: 10.1086/693977

61. Muschick, M., Indermaur, A., & Salzburger, W. (2012). Convergent Evolution within an Adaptive Radiation of Cichlid Fishes. *Current Biology*, *22*(24), 2362–2368. doi: 10.1016/J.CUB.2012.10.048

62. Nei, M. (1987). Molecular Evolutionary Genetics. *Columbia University Press, New York.*

63. O'Brochta, D. A., & Atkinson, P. W. (1996). Transposable elements and gene transformation in non-drosophilid insects. *Insect Biochemistry and Molecular Biology*, *26*(8–9), 739–753. doi: 10.1016/S0965-1748(96)00022-7

64. Okonechnikov, K., Conesa, A., & García-Alcalde, F. (2016). Qualimap 2: Advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics*, *32*(2), 292–294. doi: 10.1093/bioinformatics/btv566

65. Pekár, S., Líznarová, E., & Řezáč, M. (2016). Suitability of woodlice prey for generalist and specialist spider predators: a comparative study. *Ecological Entomology*, *41*(2), 123–130. doi: 10.1111/EEN.12285

66. Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*(6), 841–842. doi: 10.1093/bioinformatics/btq033

67. Ravinet, M., Faria, R., Butlin, R. K., Galindo, J., Bierne, N., Rafajlović, M., … Westram, A. M. (2017). Interpreting the genomic landscape of speciation: a road map for finding barriers to gene flow. *Journal of Evolutionary Biology*, *30*(8), 1450–1477. doi: 10.1111/JEB.13047

68. Řezáč, M., Pekár, S., & Lubin, Y. (2008). How oniscophagous spiders overcome woodlouse armour. *Journal of Zoology*, *275*(1), 64–71. doi: 10.1111/j.1469-7998.2007.00408.x

69. Řezáč, M. & Pekár, S. (2007). Evidence for woodlice-specialization in Dysdera spiders: Behavioural versus developmental approaches. *Physiological Entomology*, *32*(4), 367–371. doi: 10.1111/j.1365-3032.2007.00588.x

70. Řezáč, M., Pekár, S., Arnedo, M., Macías-Hernández, N., & Řezáčová, V. (2021). Evolutionary insights into the eco-phenotypic diversification of Dysdera spiders in the Canary Islands. *Organisms Diversity and Evolution*, 79–92. doi: 10.1007/s13127-020-00473-w

71. Richards, E. J., McGirr, J. A., Wang, J. R., St. John, M. E., Poelstra, J. W., Solano, M. J., … Martin, C. H. (2021). A vertebrate adaptive radiation is assembled from an ancient and disjunct spatiotemporal landscape. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(20), e2011811118. doi: https://doi.org/10.6078/D1C12S

72. Romiguier, J., Gayral, P., Ballenghien, M., Bernard, A., Cahais, V., Chenuil, A., … Galtier, N. (2014). Comparative population genomics in animals uncovers the determinants of genetic diversity. *Nature 2014 515:7526*, *515*(7526), 261–263. doi: 10.1038/nature13685

73. Rundell, R. J., & Price, T. D. (2009). Adaptive radiation, nonadaptive radiation, ecological speciation and nonecological speciation. *Trends in Ecology and Evolution*, *24*(7), 394–399. doi: 10.1016/j.tree.2009.02.007

74. Rundle, H. D., & Nosil, P. (2005). Ecological speciation. *Ecology Letters*, *8*(3), 336–352. doi: 10.1111/J.1461-0248.2004.00715.X

75. Sánchez-Herrero, J. F., Frías-López, C., Escuer, P., Hinojosa-Alvarez, S., Arnedo, M. A., Sánchez-Gracia, A., & 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*, *8*(8), 1–9. doi: 10.1093/gigascience/giz099

76. Schluter, D. (2000). Introduction to the Symposium: Species Interactions and Adaptive Radiation. *The American Naturalist*, *156*(S4), S1–S3. doi: 10.1086/303411

77. Schrader, F. (1935). Notes on the Mitotic Behavior of Long Chromosomes. *CYTOLOGIA*, *6*(4), 422–430. doi: 10.1508/cytologia.6.422

78. Schwager, E. E., Sharma, P. P., Clarke, T., Leite, D. J., Wierschin, T., Pechmann, M., … McGregor, A. P. (2017). The house spider genome reveals an ancient whole-genome duplication during arachnid evolution. *BMC Biology*, *15*(1), 1–27. doi: 10.1186/s12915-017-0399-x

79. Settepani, V., Schou, M. F., Greve, M., Grinsted, L., Bechsgaard, J., & Bilde, T. (2017). Evolution of sociality in spiders leads to depleted genomic diversity at both population and species levels. *Molecular Ecology*, *26*(16), 4197–4210. doi: 10.1111/MEC.14196

80. Signor, S. A., New, F. N., & Nuzhdin, S. (2018). A large panel of drosophila simulans reveals an abundance of common variants. *Genome Biology and Evolution*, *10*(1), 189–206. doi: 10.1093/gbe/evx262

81. Simões, M., Breitkreuz, L., Alvarado, M., Baca, S., Cooper, J. C., Heins, L., … Lieberman, B. S. (2016, January 1). The Evolving Theory of Evolutionary Radiations. *Trends in Ecology and Evolution*, Vol. 31, pp. 27–34. Trends Ecol Evol. doi: 10.1016/j.tree.2015.10.007

82. Supek, F., Bošnjak, M., Škunca, N., & Šmuc, T. (2011). Revigo summarizes and visualizes long lists of gene ontology terms. *PLoS ONE*, *6*(7), e21800. doi: 10.1371/journal.pone.0021800

83. Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, *123*(3), 585–595. doi: 10.1093/genetics/123.3.585

84. Toft, S., & Macías-Hernández, N. (2017). Metabolic adaptations for isopod specialization in three species of Dysdera spiders from the Canary Islands. *Physiological Entomology*, *42*(2), 191–198. doi: 10.1111/phen.12192

85. Toprak, U. (2020, May 7). The Role of Peptide Hormones in Insect Lipid Metabolism. *Frontiers in Physiology*, Vol. 11, p. 434. Frontiers Media S.A. doi: 10.3389/fphys.2020.00434

86. Vieira, F. G., Lassalle, F., Korneliussen, T. S., & Fumagalli, M. (2016). Improving the estimation of genetic distances from Next-Generation Sequencing data. *Biological Journal of the Linnean Society*, *117*(1), 139–149. doi: 10.1111/bij.12511

87. Vizueta, J., Macías-Hernández, N., Arnedo, M. A., Rozas, J., & Sánchez-Gracia, A. (2019a). Chance and predictability in evolution: The genomic basis of convergent dietary specializations in an adaptive radiation. *Molecular Ecology*, *28*(17), 4028–4045. doi: 10.1111/mec.15199

88. Vizueta, J., Macías-Hernández, N., Arnedo, M. A., Rozas, J., & Sánchez-Gracia, A. (2019b). Chance and predictability in evolution: The genomic basis of convergent dietary specializations in an adaptive radiation. *Molecular Ecology*, *28*(17), 4028–4045. doi: 10.1111/mec.15199

89. Watterson, G. A. (1975). On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology*, *7*, 256–276.

90. Wolf, J. B. W., & Ellegren, H. (2017). Making sense of genomic islands of differentiation in light of speciation. *Nature Reviews. Genetics*, *18*(2), 87–100. doi: 10.1038/NRG.2016.133

91. World Spider Catalog. Version 23.5. (2022). doi: 10.24436/2

92. Zeng, K., Fu, Y. X., Shi, S., & Wu, C. I. (2006). Statistical Tests for Detecting Positive Selection by Utilizing High-Frequency Variants. *Genetics*, *174*(3), 1431–1439. doi: 10.1534/GENETICS.106.061432

Tables

Table 1. Summary statistics of population genomic parameters

	Chr1	Chr2	Chr3	Chr4	Chr5	Chr6	ChrX[a]	Total
n	12	12	12	12	12	12	10	
L	177,130,251	176,685,884	174,193,557	129,208,650	125,942,283	80,935,077	317,833,791	1,181,929,493
Polymorphism								
Sites	129,374,671	129,641,630	127,754,355	94,246,118	92,862,055	58,761,896	219,789,797	852,430,522
% analysed sites	73.04%	73.37%	73.34%	72.94%	73.73%	72.60%	69.15%	72.12%
S	9,971,265	10,155,556	10,511,148	7,598,590	7,385,301	4,896,494	7,718,288	58,236,642
π	0.0147	0.0145	0.0152	0.0148	0.0148	0.0151	0.0080	
Divergence								
Sites	124,507,188	125,033,364	122,943,662	91,492,164	89,021,276	56,912,792	261,430,051	871,340,497
K	0.119	0.119	0.118	0.121	0.119	0.120	0.121	
Linkage Disequilibrium								
Average *ZnS* [b]	0.095	0.095	0.095	0.094	0.095	0.095	0.116	

[a] using information of only the 10 females

[b] calculated in windows of 50 kb

n , sample size;
L, chromosome length in bp;
Sites, number of analysed sites (after excluding filtereing positions) in bp
S , number of segregating sites
K , nucleotide divergence

Table 2. Results of the neutrality tests

	Chr1	Chr2	Chr3	Chr4	Chr5	Chr6	ChrX[a]
Tajima's *D*	-1.170	-1.251	-1.269	-1.273	-1.248	-1.304	-0.781
Fu and Li's *F*	-1.592	-1.771	-1.786	-1.805	-1.757	-1.854	-0.997
Fu and Li's *D*	-1.358	-1.540	-1.550	-1.572	-1.524	-1.616	-0.820
Fay and Wu's *H*	0.206	0.197	0.210	0.205	0.205	0.206	0.147
Zeng's *E*	-0.401	-0.417	-0.428	-0.426	-0.420	-0.435	-0.286

[a] using information of only the 10 females

Figure legends

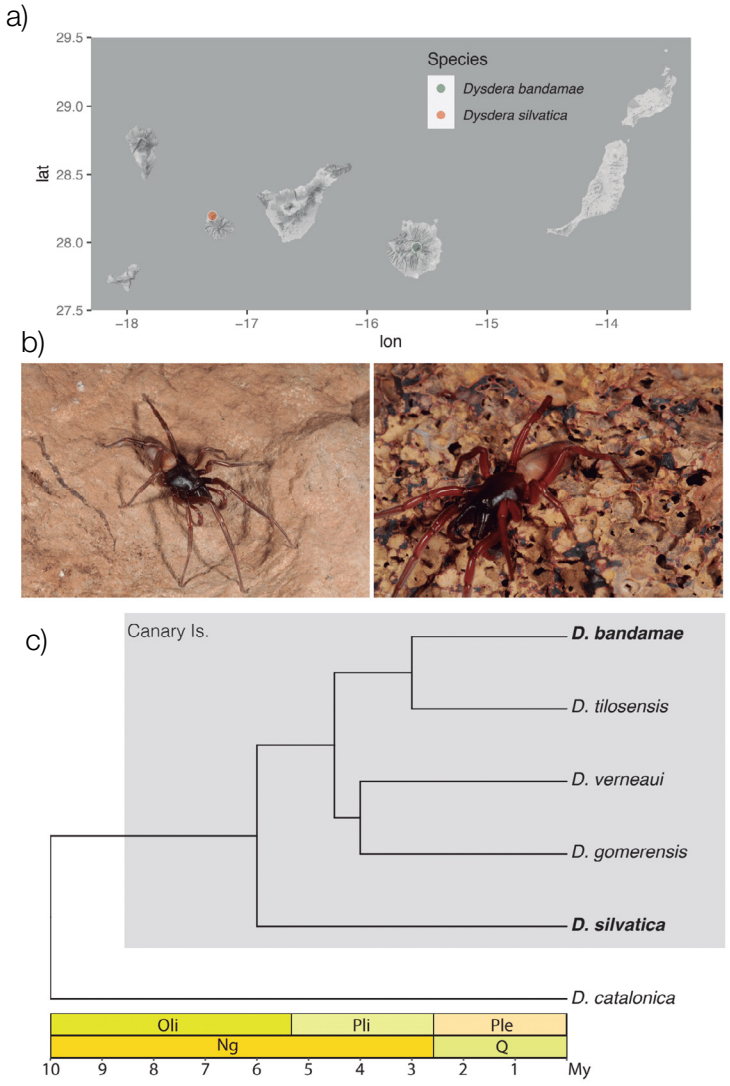
Figure 1. Geographical location, phenotypic features and phylogenetic relationships of some *Dysdera* species from the Canary Islands. a) Map of Canary Islands showing the geographical localization of the species used in this project. b) Images of the studied *Dysdera* species; left, *D. silvatica*; right, *D. bandamae*. c) Phylogenetic relationships among some *Dysdera* species.

Figure 2. Principal Component Analysis (PCA) using data from a) autosomes and b) X-chromosomes (data from only the 10 females).

Figure 3. Distribution of the summary population genomic statistics across chromosomes. a) Nucleotide diversity (π); b) Nucleotide divergence per site (K); c) Ratio π/K; d) Linkage disequilibrium (ZnS). Each point depicts the value calculated in a window of 50 kb.

Figure 4. Inferred population demographic history using the information on a) autosomal data ($n = 24$), and b) X-chromosomes ($n = 20$). The red line depicts the median over 200 estimated bootstrapped values, while the dark-gray and the light gray lines depict the 75% and the 95% of confidence intervals, respectively. Inference based on the unfolded SFS, a mutation rate per site of = 1.0×10^{-8}, and a generation time of 1.5 years.

Figure 5. Results of genome scan of selection across chromosomes. The dots depict those values of the μ-R statistic above the 0.1% of their empirical genomic distribution. Dotted and dashed lines indicate the cut-off values at 0.0001% and 0.00001%, respectively. The cut-off values were applied separately for autosomes and X chromosomes (Supplementary Table S6). Outlier regions are shaded in gray.


a)
Species
Dysdera bandamae
Dysdera silvatica
lat
lon
29.5
29.0
28.5
28.0
27.5
−18
−17
−16
−15
−14
b)
c)
Canary Is.
D. bandamae
D. tilosensis
D. verneaui
D. gomerensis
D. silvatica
D. catalonica
Oli
Pli
Ple
Ng
Q
10
9
8
7
6
5
4
3
2
1
My

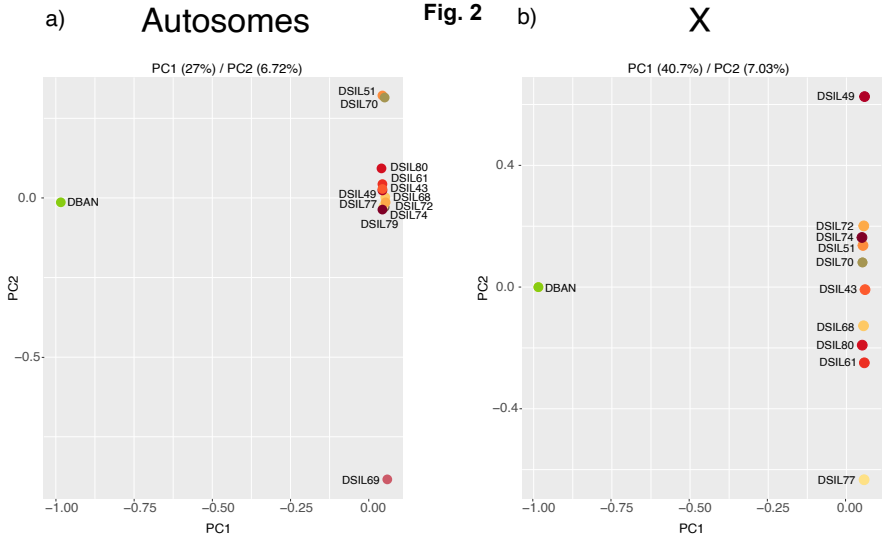

a)
Autosomes
PC1 (27%) / PC2 (6.72%)
DSIL51
DSIL70
DSIL80
DSIL61
DSIL43
DSIL49
DSIL68
DSIL77
DSIL72
DSIL74
DSIL79
DBAN
DSIL69
PC2
PC1
b)
X
PC1 (40.7%) / PC2 (7.03%)
DSIL49
DSIL72
DSIL74
DSIL51
DSIL70
DSIL43
DSIL68
DSIL80
DSIL61
DBAN
DSIL77
PC2
PC1

Fig. 2

Fig 3

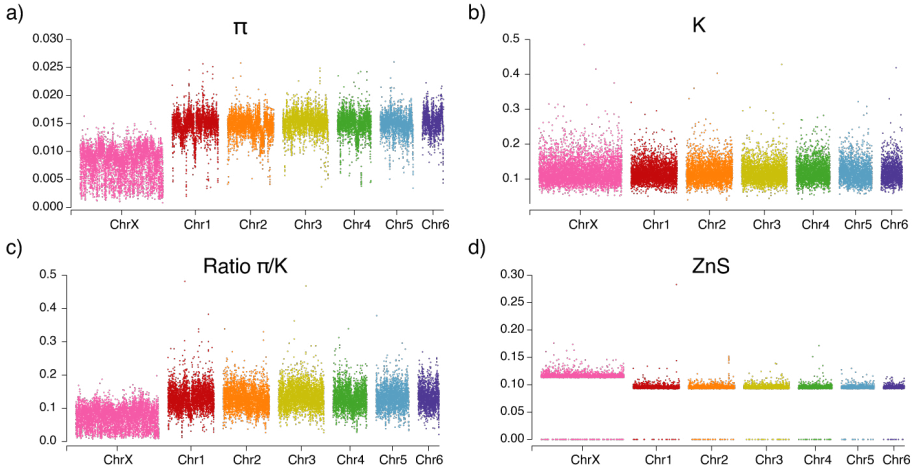
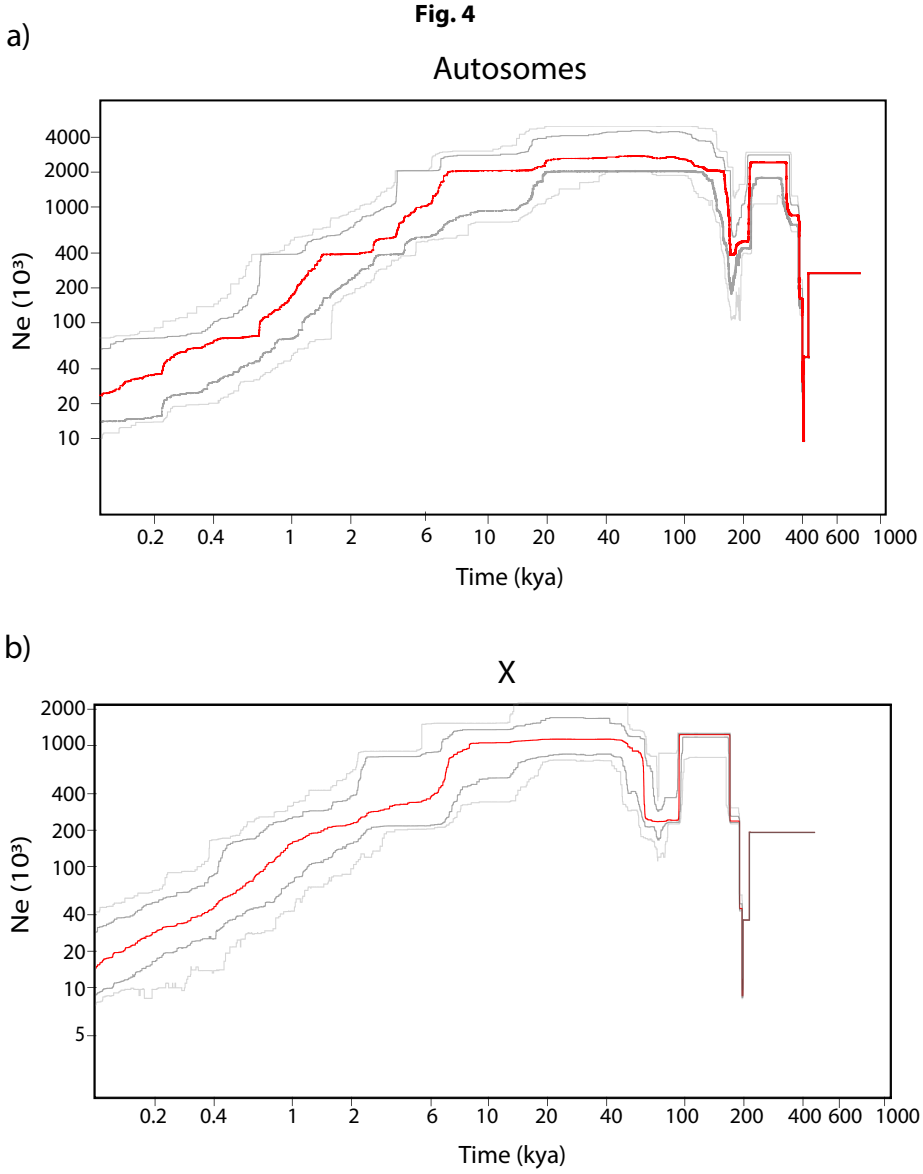

a)
π
b)
K
c)
Ratio π/K
d)
ZnS
ChrX
Chr1
Chr2
Chr3
Chr4
Chr5
Chr6

Fig. 4


a)
Autosomes
Ne (10³)
4000
2000
1000
400
200
100
40
20
10
0.2
0.4
1
2
6
10
20
40
100
200
400
600
1000
Time (kya)
b)
X
Ne (10³)
2000
1000
400
200
100
40
20
10
5
0.2
0.4
1
2
6
10
20
40
100
200
400
600
1000
Time (kya)

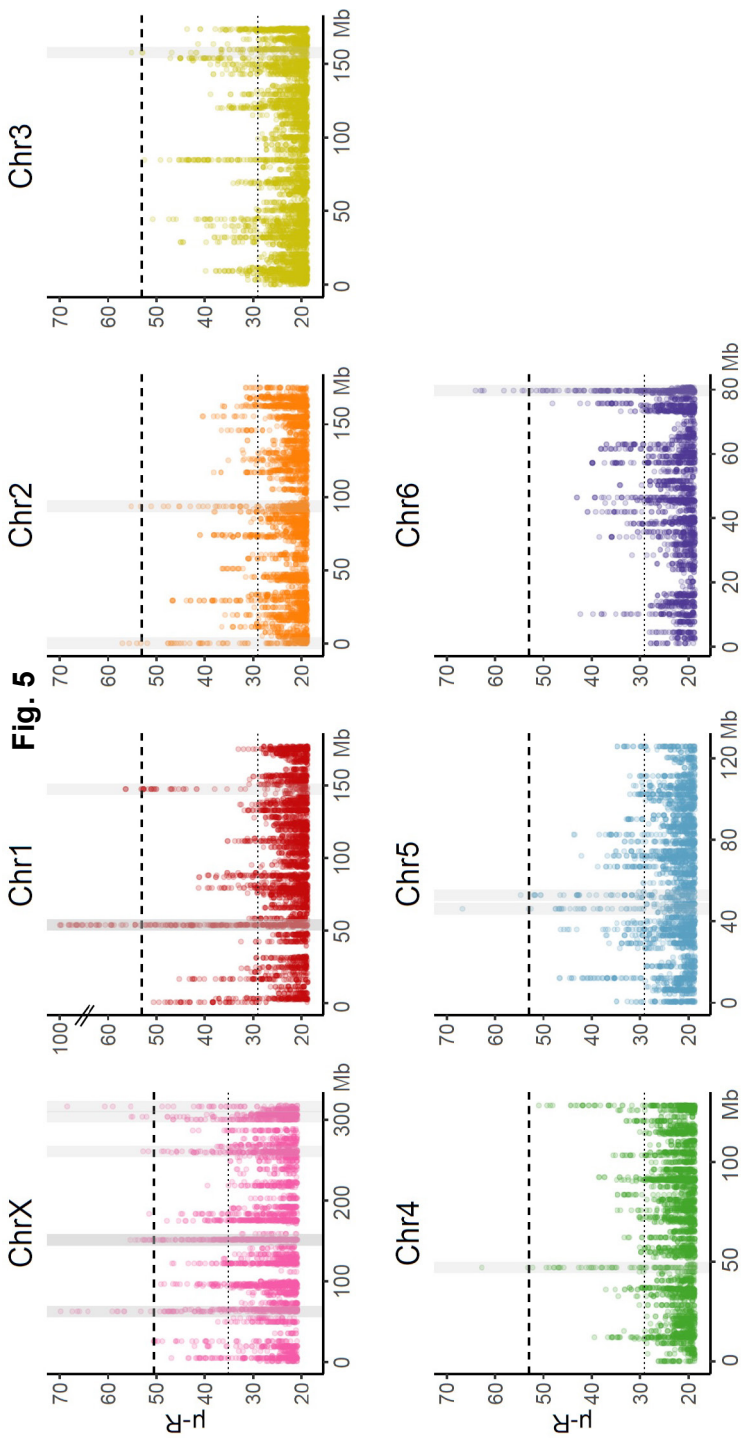


Fig. 5

Population genomics of adaptive radiations: Exceptionally high levels of genetic diversity in a spider endemic from the Canary Islands

P Escuer, S Guirao-Rico, MA Arnedo, A Sánchez-Gracia, J Rozas

Supplementary Material

SUPPLEMENTAL INFORMATION FOR:

Population genomics of adaptive radiations: Exceptionally high levels of genetic diversity in an endemic spider species from the Canary Islands

Paula Escuer, Sara Guirao-Rico, Miquel A. Arnedo, Alejandro Sánchez-Gracia, Julio Rozas

Supplementary Results

We checked some putative factors that could generate putative false positives in the RAiSD analysis. Specifically, we have controlled for the putative effect of the fraction of missing data, or masked regions (repetitive sequences identified using RepeatMasker in Escuer et al. 2022), on the estimated values of the different components of the *μ-R* statistic. Specifically, we made a prospective analysis of two candidate regions (in chromosomes 1 and 6), and observed that these regions have in fact, a low number of SNPs and a high fraction of missing data compared to regions of the same size in the genome. In fact, inspecting windows of the same size as our candidate regions of chromosome 1 and 6, only 0.001% and 0.002% of them show a lower number of SNPs, and 0.01% and 0.11% showed a higher fraction of missing data, than our candidate regions (Supplementary Table S10). Obviously, we expect that our "true" candidate regions exhibit lower levels of variation compared to other regions, but it is not clear the effect of the fraction of missing data on detecting these regions. To asses that, we performed a spearman non-parametric correlation analysis (using the function *cor.test* in R) between the different *μ-R* statistics (*μ-var*, *μ-sfs*, *μ-ld)* and also the composite *μ-R*, and also between them and the fraction of missing data across windows. We observed that both, the *μ-var and μ-R* are strongly correlated with the fraction of missing data ($\rho > 0.7$ in all cases), but the correlation between *μ-sfs* and *μ-ld* are much less pronounced (Supplementary Table S9). We found that despite there is a clear effect of missing data on the estimation of the composite *μ-R*, our

candidate regions are also outliers for the two other μ statistics that are insensitive to the effect of missing data (Supplementary Table S10). Therefore, we are confident that our candidate regions are not caused by this putative confounding factor. Additionally, we also found that the fraction of masked regions is also not a confounding factor since this was negatively correlated with the fraction of missing data (Supplementary Table S9).

Supplementary Tables

Supplementary Table S1. Sampling and sequencing information.

Supplementary Table S2. Sequence coverage and mapping information.

Supplementary Table S3. Summary statistics and neutrality tests. Values obtained after filtering out positions outside the 20-72 coverage range (representing the 66.6% of the distribution of read coverage across sites).

Supplementary Table S4. Summary statistics and neutrality tests. Values obtained after filtering out masked positions (identified by RepeatMasker) and those outside the 3-92 coverage range (representing the 95% of the distribution of read coverage across sites).

Supplementary Table S5. Average ZnS values across slinding windows of 1 kb, 5kb and 500 pb.

Supplementary Table S6. RAisD results and cut-offs.

Supplementary Table S7. Selective sweeps regions identified with RAiSD.

Supplementary Table S8. Results of the neutralitys on the selective sweep candidate regions.

Supplementary Table S9. Correlation analysis between the different components of the μ-R statistic.

Supplementary Table S10. Summary of the window analysis across chromosomes 1 and 6.

Supplementary Table S11. GO enrichment of Biological Process from RAiSD analyses.

Supplementary Table S12. GO enrichment of Molecular Function from RAiSD analyses.

Supplementary Table S13. Genes under selection identified by the MK test.

Supplementary Table S14. Positively selected genes in the MK test and their respective GO. In bold, genes that are the strong candidates.

Supplementary Table S15. GO enrichment of Biological Process from MK analyses.

Supplementary Table S16. GO enrichment of Molecular Function from MK analyses.

Supplementary Figures

Supplementary Figure S1. Phylogenetic tree using mitochondrial genomes assembled from raw reads of the *D. silvatica* population. The scale bar represents 0.01 amino acid substitutions per site.

Supplementary Figure S2. Tree maps for the GO enrichment results of GO terms of genes located in candidate selective sweep regions. Figure generated with Revigo.

Supplementary Figure S3. Tree maps for the GO enrichment results of GO terms of putatively selected genes identified in the MK analysis. Figure generated with Revigo.

Supplementary Table 1. Sampling and sequencing information

Sample ID[a]	Species	Sex	Sampling Date	Location	Latitude	Longitude	Machine	Read Type	Chemistry	Total bp	Expected coverage (x)[b]	Captured by
DSIL68	*Dysdera silvatica*	Female	21/03/2019	Teselinde (Gomera)	28.1963	-17.2875	NovaSeq6000	150PE	TruSeq	10,155,357,624	5.97	Arnedo, Bellvert, Domènech & Oromí
DSIL69	*Dysdera silvatica*	Male	21/03/2019	Teselinde (Gomera)	28.1963	-17.2875	NovaSeq6000	150PE	TruSeq	12,602,828,742	7.41	Arnedo, Bellvert, Domènech & Oromí
DSIL70	*Dysdera silvatica*	Female	21/03/2019	Teselinde (Gomera)	28.1963	-17.2875	NovaSeq6000	150PE	TruSeq	12,030,124,096	7.08	Arnedo, Bellvert, Domènech & Oromí
DSIL72	*Dysdera silvatica*	Female	21/03/2019	Teselinde (Gomera)	28.1963	-17.2875	NovaSeq6000	150PE	TruSeq	11,698,025,870	6.88	Arnedo, Bellvert, Domènech & Oromí
DSIL51	*Dysdera silvatica*	Female	21/03/2019	Teselinde (Gomera)	28.1963	-17.2875	NovaSeq6000	150PE	TruSeq	11,147,331,588	6.56	Arnedo, Bellvert, Domènech & Oromí
DSIL79	*Dysdera silvatica*	Male	21/03/2019	Teselinde (Gomera)	28.1963	-17.2875	NovaSeq6000	150PE	TruSeq	10,382,227,574	6.11	Arnedo, Bellvert, Domènech & Oromí
DSIL61	*Dysdera silvatica*	Female	21/03/2019	Teselinde (Gomera)	28.1963	-17.2875	NovaSeq6000	150PE	TruSeq	11,367,950,742	6.69	Arnedo, Bellvert, Domènech & Oromí
DSIL43	*Dysdera silvatica*	Female	21/03/2019	Teselinde (Gomera)	28.1963	-17.2875	NovaSeq6000	150PE	TruSeq	10,137,689,114	5.96	Arnedo, Bellvert, Domènech & Oromí
DSIL49	*Dysdera silvatica*	Female	21/03/2019	Teselinde (Gomera)	28.1963	-17.2875	NovaSeq6000	150PE	TruSeq	13,059,529,752	7.68	Arnedo, Bellvert, Domènech & Oromí
DSIL80	*Dysdera silvatica*	Female	21/03/2019	Teselinde (Gomera)	28.1963	-17.2875	NovaSeq6000	150PE	TruSeq	10,577,144,414	6.22	Arnedo, Bellvert, Domènech & Oromí
DSIL77	*Dysdera silvatica*	Female	21/03/2019	Teselinde (Gomera)	28.1963	-17.2875	NovaSeq6000	150PE	TruSeq	13,285,537,492	7.82	Arnedo, Bellvert, Domènech & Oromí
DSIL74	*Dysdera silvatica*	Female	21/03/2019	Teselinde (Gomera)	28.1963	-17.2875	NovaSeq6000	150PE	TruSeq	10,704,482,412	6.30	Arnedo, Bellvert, Domènech & Oromí
											80.68	
DBAN	*Dysdera bandamae*	Male	12/04/2014	Tejada (Gran Canaria)	27.9643	-15.5855	HiSeqX	150PE	TruSeq	54,096,438,258	31.82	M. Arnedo & N. Macías

[a] sample ID used in figures

[b] expected coverage assuming a genome size of 1.7 Gb (results from the flow citometry)

Supplementary Table 2. Sequence coverage and mapping information

Sample	Name	Species	Sex	Raw sequencing (bp)	Coverage (x)[a]	Coverage (x)[b]	True Mapped reads
CRBA003668_181	DSIL68	*Dysdera silvatica*	Female	10,155,357,624	5.97	3.30	38,145,007
CRBA003669_182	DSIL69	*Dysdera silvatica*	Male	12,602,828,742	7.41	4.31	49,987,849
CRBA003670_164	DSIL70	*Dysdera silvatica*	Female	12,030,124,096	7.08	3.64	41,959,251
CRBA003672_183	DSIL72	*Dysdera silvatica*	Female	11,698,025,870	6.88	3.66	42,677,141
CRBA003651_191	DSIL51	*Dysdera silvatica*	Female	11,147,331,588	6.56	3.27	38,375,997
CRBA003679_192	DSIL79	*Dysdera silvatica*	Male	10,382,227,574	6.11	4.04	48,432,562
CRBA003661_193	DSIL61	*Dysdera silvatica*	Female	11,367,950,742	6.69	3.93	45,398,455
CRBA003643_194	DSIL43	*Dysdera silvatica*	Female	10,137,689,114	5.96	3.82	44,095,352
CRBA003649_195	DSIL49	*Dysdera silvatica*	Female	13,059,529,752	7.68	3.36	40,108,300
CRBA003680_196	DSIL80	*Dysdera silvatica*	Female	10,577,144,414	6.22	4.30	50,370,508
CRBA003677_197	DSIL77	*Dysdera silvatica*	Female	13,285,537,492	7.82	3.35	39,868,756
CRBA003674_198	DSIL74	*Dysdera silvatica*	Female	10,704,482,412	6.30	3.31	38,080,506

[a] Coverage estimated before using trimmomatic and samtools filters

[b] Coverge estimated after using trimmomatic and samtools filters

Supplementary Table 3. Summary statistics and neutrality tests[a]

	Chr1	Chr2	Chr3	Chr4	Chr5	Chr6	ChrX[b]
L	177,130,251	176,685,884	174,193,557	129,208,650	125,942,283	80,935,077	317,833,791
Sites	93,197,670 (52.62 %)	93652992 (53.01 %)	92,336,237 (53.01 %)	67,968,772 (52.60 %)	6,6697,729 (52.96 %)	42,188,607 (52.13 %)	183,022,599 (57.58 %)
π	0.0150	0.0148	0.0155	0.0152	0.0151	0.0155	0.0078
Tajima's *D*	-1.165	-1.247	-1.263	-1.270	-1.244	-1.297	-0.842
Fu and Li's *F*	-1.558	-1.740	-1.751	-1.776	-1.724	-1.816	-1.093
Fu and Li's *D*	-1.318	-1.502	-1.508	-1.535	-1.484	-1.571	-0.907
Fay and Wu's *H*	0.209	0.201	0.213	0.209	0.207	0.210	0.161
Zeng's *E*	-0.402	-0.418	-0.428	-0.428	-0.420	-0.435	-0.309

[a] Values obtained after filtering out positions outside the 20-72 coverage range (66.6% of the distribution of read coverage across sites)

[b] using information of only the 10 females

L, chromosome length in bp

Sites, number of analysed sites after excluding filtereing positions in bp (in parenthesis, the % of the total length)

Supplementary Table 4. Summary statistics and neutrality tests[a]

	Chr1	Chr2	Chr3	Chr4	Chr5	Chr6	ChrX[b]
L	177,130,251	176,685,884	174,193,557	129,208,650	125,942,283	80,935,077	317,833,791
Sites	72,687,692 (41.04 %)	72,393,801 (40.97 %)	71,310,173 (40.94 %)	52,214,410 (40.41 %)	51,951,005 (41.25 %)	32,714,879 (40.42 %)	121,410,121 (38.20 %)
π	0.0139	0.0137	0.0144	0.0140	0.0139	0.0144	0.0075
Tajima's *D*	-1.153	-1.234	-1.251	-1.254	-1.230	-1.283	-0.792
Fu and Li's *F*	-1.576	-1.757	-1.769	-1.786	-1.741	-1.830	-1.039
Fu and Li's *D*	-1.348	-1.532	-1.538	-1.558	-1.514	-1.597	-0.867
Fay and Wu's *H*	0.193	0.183	0.197	0.192	0.191	0.194	0.144
Zeng's *E*	-0.390	-0.405	-0.416	-0.414	-0.408	-0.423	-0.286

[a] Values obtained after filtering out masked positions (identified by RepeatMasker) and those outside the 3-92 coverage range (95% of the distribution of read coverage across sites)

[b] using information of only the 10 females

L, chromosome length in bp

Sites, number of analysed sites after excluding filtereing positions in bp (in parenthesis, the % of the total length)

Supplementary Table 5. Average ZnS values across slinding windows of 1 kb, 5kb and 500 pb

	Chr1	Chr2	Chr3	Chr4	Chr5	Chr6	ChrX[a]
Average ZnS (500 bp)	0.189	0.188	0.188	0.187	0.187	0.189	0.220
Average ZnS (1 kb)	0.149	0.148	0.148	0.147	0.147	0.149	0.176
Average ZnS (5 kb)	0.106	0.106	0.106	0.106	0.106	0.106	0.129

[a] using information of only the 10 females

Supplementary Table 6. RAisD results and cut-offs

	Chr1	Chr2	Chr3	Chr4	Chr5	Chr6	ChrX
Total Length	177,130,251	176,685,884	174,193,557	129,208,650	125,942,283	80,935,077	317,833,791
Total SNPs	8,606,328	8,553,370	8,782,293	6,304,815	6,175,198	4,055,455	9,282,206
Analyzed SNPs	4,434,267	4,410,896	4,609,728	3,255,714	3,202,736	2,123,567	3,884,445
Number of RAiSD windows	4,434,218	4,410,847	4,609,679	3,255,665	3,202,687	2,123,518	3,884,396
0.001	4,434	4,411	4,610	3,252,409	3,199,484	2,121,394	3,884
First cut-off μ-R	**18.36**	**18.13**	**18.94**	**18.95**	**18.59**	**19.64**	**20.84**
0.0001	443	441	461	326	320	212	388
Second cut-off μ-R	**30.43**	**27.32**	**29.07**	**27.52**	**29.32**	**31.93**	**35.51**
0.00001	44	44	46	33	32	21	39
Third cut-off μ-R	**95.45**	**40.35**	**42.23**	**39.66**	**43.7**	**48.04**	**50.56**
Cut-offs used							
	4,086	3,783	4,650	3,516	3,126	2,693	3,884
First cut-off	18.68	18.68	18.68	18.68	18.68	18.68	20.84
	524	314	459	174	329	303	388
Second cut-off	29.08	29.08	29.08	29.08	29.08	29.08	35.51
	133	6	2	2	3	10	39
Third cut-off	**52.96**	**52.96**	**52.96**	**52.96**	**52.96**	**52.96**	**50.56**

Supplementary Table 7. Selective sweeps regions identified with RAiSD

	region ID	Start	End	Length	Max μ-R	No. of gene anotations	GO term	gID1	GO term	gID2	GO term	gID3	GO term	gID4	GO term
	1	62,464,226	62,485,254	21,028	64.2	1	1	g25671	GO:0006508; GO:0016579; GO:0043248; GO:0043312; GO:0043687; GO:0048863; GO:0004175; GO:0005198; GO:0005515;						
	2	62,504,764	62,562,199	57,435	71.9	1	1	g25672	GO:0003676;						
	3	151,011,709	151,079,064	67,355	52.4	2	1	g27524	GO:0000226; GO:0007091	g27525					
	4	151,024,536	151,118,952	94,416	55.4	1	0	g27526							
ChrX (n = 9)	4-5[a]					4	1	g27527	GO:0004349,GO:0006561,	g27528		g27529		g27530	
	5	151,316,519	151,374,453	57,934	52.8	1	1	g27531	GO:0005086,GO:0032012						
	6[b]	183,370,813	183,418,510	47,697	51.5	na[c]									
	7	260,834,022	260,883,821	49,799	52.7	2	1	g29899	GO:0008802; GO:0008812	g29900					
	8	303,626,273	303,677,478	51,205	55.3	2	2	g30811	GO:0004190; GO:0006508	g30812	GO:0005219; GO:0005262; GO:0005515; GO:0006816; GO:0016020; GO:0070588				
	9	316,337,494	316,401,873	64,379	71.2	4	3	g31096	GO:0005515	g31097		g31098	GO:0000981; GO:000563	g31099	GO:0005765;
	1	53,772,934	53,885,077	112,143	430.9	1	1	g10564	GO:0006089; GO:0008152; GO:0008610; GO:0006631; GO:0006633; GO:0008611; GO:0009058; GO:0009888; GO:0048468; GO:0031325; GO						
	2	53,884,694	53,903,438	18,744	141.8	1	0	g10565							
Chr1 (n = 4)	2-3[a]					1	1	g10566	GO:0005515						
	3	53,966,131	54,029,897	63,766	54.4	1	1	g10567	GO:0003676; GO:0003700; GO:0015074; GO:0005198; GO:0019028;						
	4	147,604,625	147,653,024	48,399	56.5	3	3	g12740	GO:0004725,GO:0006470,	g12741	GO:0001580; GO:0005737	g12742	GO:0007264; GO:0008380; GO:0048026		
	1	31,314	55,068	23,754	57.1	1	1	g19989	GO:0016021,GO:0055085						
Chr2 (n = 2)	2	93,820,292	93,844,164	23,872	55.2	2	2	g22075	GO:0005086; GO:0005543; GO:0032012; GO:0005622;	g22076	GO:0007165; GO:0016192; GO:0030182; GO:0032011				
Chr3 (n = 1)	1	157,606,626	157,615,217	8,591	55.2	4	3	g5127		g5128	GO:0003735,GO:0005840,	g5129	GO:0003735,GO:000584	g5130	GO:0000943; G
Chr4 (n = 1)	1	47,101,957	47,110,295	8,338	62.8	2	2	g14786	GO:0003824,GO:0004056,	g14785	GO:0000139; GO:0005794; GO:0006478; GO:0008476; GO:0009306; GO:0016021; GO:0005783; G				
Chr5 (n = 2)	1	46,104,844	46,121,227	16,383	66.8	1	1	g18210	GO:0002092; GO:0006954; GO:0007155; GO:0007159; GO:0007202; GO:0007568; GO:0010898; GO:0019722; GO:0030029; GO:0031100; GO						
	2	52,830,441	52,867,164	36,723	54.8	1	1	g21196	GO:0006457; GO:0006662; GO:0009306; GO:0034976; GO:0045454; GO:0061077; GO:1903334; GO:0003723; GO:0003756; GO:0015035; GO						
Chr6 (n = 1)	1	79,695,691	79,730,171	34,480	73.7	2	1	g33168		g33169	GO:0003676; GO:0003677; GO:0015074; GO:0004803; GO:0006313;				

[a] regions not identified as outliers that were consider because are flanking two identified regions

[b] selective sweep region having only a single window with μ-R values above the 0.001%

[c] region not analyzed

Analysis of the ChrX based on information of the 10 females

Total number of identified regions: 19 (plus 2 flanking). Excluding the region with only a single window with μ-R values above the 0.001%

Total number of included genes: 38

Total number of identified GO terms: 188

Note: Only the first columns are displayed. The complete data set is available in:

http://www.ub.edu/molevol/EGB/Escuer_Supplementary_Tables.zip

Supplementary Table 8. Results of the neutrality tests on the selective sweep candidate regions

Chr	Start	End	Length	SW[a]	Tajima (obs)	1% left[b] Tajima Cutoff	FayWu (obs)	1% left[b] FayWu Cutoff	Zeng (obs)	1% left[b] Zeng Cutoff	1% right[b] Zeng Cutoff
	62,464,226	62,485,254	21,028	62,450,001-62,500,001	-1.270		0.240		-0.444		
	62,504,764	62,562,199	57,435	62,500,001-62,550,001	-1.319		0.202		-0.436		
	151,011,709	151,079,064	67,355	151,000,001-151,050,001	-1.238		-0.220		-0.204		
	151,024,536	151,118,952	94,416	151,050,001-151,100,001	-1.165		-0.156		-0.220		
ChrX	151,316,519	151,374,453	57,934	151,300,001-151,350,001	-0.907	-1.478	-0.310	-0.537	-0.068	-0.484	0.128
	183,370,813	183,418,510	47,697	183,350,001-183,400,001	-0.947		0.005		-0.250		
	260,834,022	260,883,821	49,799	260,800,001-260,850,001	-0.285		-0.145		0.005		
	303,626,273	303,677,478	51,205	303,600,001-303,650,001	-1.402		-0.169		-0.274		
	316,337,494	316,401,873	64,379	316,350,001-316,400,001	**-1.641**		**-0.809**		0.013		
	53,772,197	53,885,077	112,880	53,800,001-53,850,001	**-1.841**		-0.225		-0.337		
	53,884,694	53,903,438	18,744	53,850,001-53,900,001	**-1.858**		**-0.287**		-0.307		
Chr1	53,966,131	54,029,897	63,766	53,950,001-54,000,001	-0.975	-1.465	0.067	-0.257	-0.278	-0.489	-0.098
				54,000,001-54,050,001	-0.613		**-0.335**		**0.028**		
	147,604,625	147,653,024	48,399	147,600,001-147,650,001	-1.148		-0.226		-0.164		
Chr2	31,314	55,068	23,754	1-50,001	-1.084	-1.536	**-0.252**	-0.083	**-0.134**	-0.497	-0.156
	93,820,292	93,844,164	23,872	93,800,001-93,850,001	-1.234		-0.077		-0.265		
Chr3	157,606,626	157,615,217	8,591	157,600,001-157,650,001	-0.941	-1.503	0.113	-0.095	-0.294	-0.502	-0.197
Chr4	47,101,957	47,110,295	8,338	47,100,001-47,150,001	**-1.577**	-1.561	**-0.766**	-0.081	**0.021**	-0.503	-0.185
	46,104,844	46,121,227	16,383	46,100,001-46,150,001	-1.164		-0.053		-0.260		
Chr5	52,830,441	52,867,164	36,723	52,800,001-52,850,001	-0.936	-1.479	0.233	-0.106	-0.358	-0.496	-0.172
				52,850,001-52,900,001	-1.152		0.233		-0.411		
Chr6	79,695,691	79,730,171	34,480	79,700,001-79,750,001	-1.473	-1.558	**-0.367**	-0.021	-0.300	-0.498	-0.231

[a] Sliding window used to compute the neutrality test

[b] Confident interval obtained from the empirical distribution of values across the complete chromosome. Shaded bold cells show the significant values

Analysis of the ChrX based on information of the 10 females

Supplementary Table 9. Correlation analysis between the different components of the μ-R statistic

A) Correlation between the different components of the μ-R statistic.

	chr1		chr6	
	r	***P*-value**	***r***	***P*-value**
μ-all vs μ-var	0.8	< 2.2e−16	0.92	< 2.2e−16
μ-all vs μ-sfs	0.13	1.7e−07	0.087	2.5e−05
μ-all vs μ-ld	0.027	0.29	0.08	0.00011

r , Spearman correlation coefficient; *P*-value, *P*-value of the Spearman correlation.

B) Correlation between the each component of the μ-R statistic and other features.

	chr1		chr6	
	r	***P*-value**	***r***	***P*-value**
μ-all vs Number of SNPs	-0.76,	< 2.2e−16	-0.81	< 2.2e−16
μ-all vs Mean % missing	0.73	< 2.2e−16	0.71	< 2.2e−16
μ-all vs Mean % masked	-0.43	< 2.2e−16	-0.4	< 2.2e−16
μ-var vs Number of SNPs	-0.96	< 2.2e−16	-0.89	< 2.2e−16
μ-var vs Mean % missing	0.8	< 2.2e−16	0.76	< 2.2e−16
μ-var vs Mean % masked	-0.5	< 2.2e−16	-0.44	< 2.2e−16
μ-sfs vs Number of SNPs	0.39	< 2.2e−16	0.19	< 2.2e−16
μ-sfs vs Mean % missing	-0.22	< 2.2e−16	-0.16	1e−15
μ-sfs vs Mean % masked	0.15	5.8e−09	0.06	0.0034
μ-ld vs Number of SNPs	-0.43	< 2.2e−16	-0.21	< 2.2e−16
μ-ld vs Mean % missing	0.27	< 2.2e−16	0.18	< 2.2e−16
μ-ld vs Mean % masked	-0.11	6e−06	-0.016	0.45

r , Spearman correlation coefficient; *P*-value, *P*-value of the Spearman correlation.

C) Correlation between the mean fraction of missing and the mean fraction of masked data.

	chr1		chr6	
	r	***P*-value**	***r***	***P*-value**
Mean % missing vs Mean %	-0.55	< 2.2e−16	-0.47	< 2.2e−16

r , Spearman correlation coefficient; *P*-value, *P*-value of the Spearman correlation.

Supplementary Table 10. Summary of the window analysis across chromosomes 1 and 6

			Fraction of windows with...							Candidate region						
Chromosome	Number of windows	Windows size (bp)	Lower number of SNPs of than the candidate region	Higher proportion of missing data than the candidate region	Higher proportion of masked data than the candidate region	Higher values of μ-*var* statistic than the candidate region	Higher values of μ-*sfs* statistic than the candidate region	Higher values of μ-*ld* statistic than the candidate region	Higher values of μ-*R* statistic than the candidate region	Number of SNPs	Proportion of missing data	Proportion of masked data	μ-*var*	μ-*sfs*	μ-*ld*	μ-*R*
Chr1	1.58	112.047	0.001	0.01	0.20	0.02	0	0	0	2.17	0.11	0.54	17.92	2.22	3.06	99.94
Chr6	2.347	34.48	0.002	0.11	1	0	0	0	0.008	663	0.10	0.38	9.27	2.33	2.08	44.11

Candidate region, for each chromosome, a genomic region with μ statistic above 0.001% of its empirical distribution.

Supplementary Table 11. GO enrichment of Biological Process from RAiSD analyses

Number	GOBPID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
1	GO:0055114	1.90E-13	25.8375	7.36E-01	13	59	oxidation-reduction process
2	GO:0050896	3.12E-10	3.8585	2.05E+01	48	1642	response to stimulus
3	GO:1901564	3.11E-08	3.7797	1.16E+01	32	929	organonitrogen compound metabolic process
4	GO:0006278	3.29E-08	104.4886	1.12E-01	5	9	RNA-dependent DNA biosynthetic process
5	GO:0044267	3.70E-08	4.0536	9.17E+00	28	735	cellular protein metabolic process
6	GO:0051716	7.48E-08	3.5638	1.27E+01	33	1017	cellular response to stimulus
7	GO:0008152	1.10E-07	3.0725	3.59E+01	61	2879	metabolic process
8	GO:0030003	1.23E-07	22.9597	4.12E-01	7	33	cellular cation homeostasis
9	GO:0006873	1.23E-07	22.9597	4.12E-01	7	33	cellular ion homeostasis
10	GO:0007165	1.66E-07	3.9133	8.61E+00	26	690	signal transduction
11	GO:0071897	2.04E-07	31.6552	2.75E-01	6	22	DNA biosynthetic process
12	GO:0051649	2.11E-07	6.1264	2.98E+00	15	239	establishment of localization in cell
13	GO:1901576	2.15E-07	3.6139	1.06E+01	29	849	organic substance biosynthetic process
14	GO:0019538	2.26E-07	3.6760	9.98E+00	28	800	protein metabolic process
15	GO:0042221	2.82E-07	4.1168	7.08E+00	23	567	response to chemical
16	GO:0009058	3.82E-07	3.5030	1.09E+01	29	872	biosynthetic process
17	GO:0006810	4.29E-07	3.7038	9.03E+00	26	724	transport
18	GO:0006793	4.37E-07	4.2757	6.13E+00	21	491	phosphorus metabolic process
19	GO:0071705	5.01E-07	6.1238	2.76E+00	14	221	nitrogen compound transport
20	GO:0006629	5.91E-07	6.0338	2.80E+00	14	224	lipid metabolic process
21	GO:0007154	6.10E-07	3.4786	1.05E+01	28	839	cell communication
22	GO:0044249	6.25E-07	3.4737	1.05E+01	28	840	cellular biosynthetic process
23	GO:0071704	6.69E-07	2.8275	3.37E+01	57	2699	organic substance metabolic process
24	GO:0046907	9.32E-07	6.8865	2.08E+00	12	167	intracellular transport
25	GO:0051234	1.05E-06	3.5137	9.46E+00	26	758	establishment of localization
26	GO:0065008	1.06E-06	3.8978	7.04E+00	22	564	regulation of biological quality
27	GO:0055080	1.16E-06	15.6836	5.62E-01	7	45	cation homeostasis
28	GO:0010033	1.17E-06	4.7202	4.37E+00	17	350	response to organic substance
29	GO:0098771	1.35E-06	15.2794	5.74E-01	7	46	inorganic ion homeostasis
30	GO:0044281	1.47E-06	5.1894	3.47E+00	15	278	small molecule metabolic process
31	GO:0007204	1.49E-06	34.7917	2.12E-01	5	17	positive regulation of cytosolic calcium ion concentration
32	GO:0044271	1.55E-06	3.5090	9.02E+00	25	723	cellular nitrogen compound biosynthetic process
33	GO:0044283	1.74E-06	11.4510	8.49E-01	8	68	small molecule biosynthetic process
34	GO:0006796	1.76E-06	4.0255	6.10E+00	20	489	phosphate-containing compound metabolic process
35	GO:0023052	1.81E-06	3.3312	1.04E+01	27	832	signaling
36	GO:0044238	1.83E-06	2.7015	3.18E+01	54	2548	primary metabolic process
37	GO:0006875	1.84E-06	20.2345	3.87E-01	6	31	cellular metal ion homeostasis
38	GO:0071702	2.39E-06	4.9733	3.61E+00	15	289	organic substance transport
39	GO:0015833	2.40E-06	6.9181	1.88E+00	11	151	peptide transport
40	GO:0042886	2.56E-06	6.8681	1.90E+00	11	152	amide transport
41	GO:0035556	2.69E-06	4.6473	4.13E+00	16	331	intracellular signal transduction
42	GO:0006464	2.91E-06	3.6426	7.47E+00	22	599	cellular protein modification process
43	GO:0036211	2.99E-06	3.6358	7.49E+00	22	600	protein modification process
44	GO:0044237	3.26E-06	2.6341	3.33E+01	55	2665	cellular metabolic process
45	GO:0044260	4.49E-06	2.6623	2.30E+01	43	1840	cellular macromolecule metabolic process
46	GO:0055082	4.67E-06	12.3992	6.86E-01	7	55	cellular chemical homeostasis
47	GO:0043604	4.75E-06	8.2819	1.29E+00	9	103	amide biosynthetic process
48	GO:0009605	4.78E-06	4.9781	3.33E+00	14	267	response to external stimulus
49	GO:0043603	5.01E-06	7.1479	1.65E+00	10	132	cellular amide metabolic process
50	GO:0051179	5.18E-06	2.9768	1.30E+01	30	1045	localization
51	GO:0002376	5.33E-06	4.3849	4.35E+00	16	349	immune system process
52	GO:0050801	5.97E-06	11.9000	7.11E-01	7	57	ion homeostasis
53	GO:0051480	7.65E-06	23.1755	2.87E-01	5	23	regulation of cytosolic calcium ion concentration
54	GO:0046903	9.04E-06	6.6486	1.76E+00	10	141	secretion
55	GO:0015031	9.04E-06	6.6486	1.76E+00	10	141	protein transport
56	GO:0043412	9.47E-06	3.3566	8.04E+00	22	644	macromolecule modification
57	GO:0055065	1.01E-05	14.4335	5.12E-01	6	41	metal ion homeostasis
58	GO:0006812	1.02E-05	6.5468	1.78E+00	10	143	cation transport
59	GO:0016192	1.13E-05	5.3037	2.65E+00	12	212	vesicle-mediated transport
60	GO:0045184	1.39E-05	6.3052	1.85E+00	10	148	establishment of protein localization

Note: Only the first raws are displayed. The complete data set is available in:
http://www.ub.edu/molevol/EGB/Escuer_Supplementary_Tables.zip

Supplementary Table 12. GO enrichment of Molecular Function from RAiSD analyses

Number	GOMFID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
1	GO:0005515	1.60E-14	7.184	8.60E+00	35	1635	protein binding
2	GO:0016747	4.82E-11	26.346	4.89E-01	10	93	transferase activity, transferring acyl groups other than amino-acyl groups
3	GO:0016746	2.17E-10	22.288	5.68E-01	10	108	transferase activity, transferring acyl groups
4	GO:0016740	6.15E-09	4.922	7.56E+00	26	1436	transferase activity
5	GO:0004803	1.40E-07	Inf	1.58E-02	3	3	transposase activity
6	GO:0003824	1.53E-07	3.524	2.00E+01	41	3792	catalytic activity
7	GO:0043167	1.59E-07	4.560	6.45E+00	22	1225	ion binding
8	GO:0050254	5.56E-07	592.103	2.10E-02	3	4	rhodopsin kinase activity
9	GO:0016491	3.15E-06	5.617	2.78E+00	13	528	oxidoreductase activity
10	GO:0016853	3.63E-06	25.345	2.37E-01	5	45	isomerase activity
11	GO:0046914	5.35E-06	8.076	1.29E+00	9	246	transition metal ion binding
12	GO:0016301	5.36E-06	5.770	2.47E+00	12	469	kinase activity
13	GO:0016773	1.04E-05	5.878	2.19E+00	11	417	phosphotransferase activity, alcohol group as acceptor
14	GO:0016667	1.15E-05	98.647	4.74E-02	3	9	oxidoreductase activity, acting on a sulfur group of donors
15	GO:0004672	1.37E-05	6.308	1.84E+00	10	350	protein kinase activity
16	GO:0036094	1.37E-05	4.858	3.18E+00	13	605	small molecule binding
17	GO:0016616	1.79E-05	17.763	3.26E-01	5	62	oxidoreductase activity, acting on the CH-OH group of donors, NAD or NADP as accep
18	GO:0016628	2.24E-05	73.974	5.79E-02	3	11	oxidoreductase activity, acting on the CH-CH group of donors, NAD or NADP as accept
19	GO:0019003	2.73E-05	Inf	1.05E-02	2	2	GDP binding
20	GO:0043168	2.92E-05	5.225	2.45E+00	11	466	anion binding
21	GO:0016627	3.65E-05	24.981	1.89E-01	4	36	oxidoreductase activity, acting on the CH-CH group of donors
22	GO:0046872	3.75E-05	4.138	4.03E+00	14	766	metal ion binding
23	GO:0043169	3.92E-05	4.121	4.05E+00	14	769	cation binding
24	GO:0016614	4.54E-05	14.450	3.95E-01	5	75	oxidoreductase activity, acting on CH-OH group of donors
25	GO:0019899	7.46E-05	6.479	1.40E+00	8	266	enzyme binding
26	GO:0008270	8.52E-05	9.294	7.26E-01	6	138	zinc ion binding
27	GO:1901265	1.22E-04	4.416	2.87E+00	11	546	nucleoside phosphate binding
28	GO:0000166	1.22E-04	4.416	2.87E+00	11	546	nucleotide binding
29	GO:0048037	1.57E-04	16.634	2.74E-01	4	52	cofactor binding
30	GO:0030234	2.06E-04	6.444	1.22E+00	7	231	enzyme regulator activity
31	GO:0016772	2.55E-04	3.405	4.83E+00	14	917	transferase activity, transferring phosphorus-containing groups
32	GO:0050661	2.70E-04	129.652	2.63E-02	2	5	NADP binding
33	GO:0032440	4.04E-04	97.232	3.16E-02	2	6	2-alkenal reductase [NAD(P)] activity
34	GO:0051540	4.04E-04	97.232	3.16E-02	2	6	metal cluster binding
35	GO:0051536	4.04E-04	97.232	3.16E-02	2	6	iron-sulfur cluster binding
36	GO:0015036	4.04E-04	97.232	3.16E-02	2	6	disulfide oxidoreductase activity
37	GO:0097110	4.04E-04	97.232	3.16E-02	2	6	scaffold protein binding
38	GO:0016787	4.22E-04	2.754	8.37E+00	19	1591	hydrolase activity
39	GO:0017046	5.64E-04	77.780	3.68E-02	2	7	peptide hormone binding
40	GO:0005488	6.44E-04	2.405	4.05E+01	54	7693	binding
41	GO:0016903	6.80E-04	19.694	1.74E-01	3	33	oxidoreductase activity, acting on the aldehyde or oxo group of donors
42	GO:0042277	6.80E-04	19.694	1.74E-01	3	33	peptide binding
43	GO:0019829	9.60E-04	55.549	4.74E-02	2	9	cation-transporting ATPase activity
44	GO:0042625	1.20E-03	48.601	5.26E-02	2	10	ATPase coupled ion transmembrane transporter activity
45	GO:0022853	1.20E-03	48.601	5.26E-02	2	10	active ion transmembrane transporter activity
46	GO:0042802	1.44E-03	8.944	4.89E-01	4	93	identical protein binding
47	GO:0031072	1.46E-03	43.198	5.79E-02	2	11	heat shock protein binding
48	GO:0045296	1.74E-03	38.875	6.31E-02	2	12	cadherin binding
49	GO:0044877	1.75E-03	8.465	5.16E-01	4	98	protein-containing complex binding
50	GO:0042626	2.05E-03	35.339	6.84E-02	2	13	ATPase activity, coupled to transmembrane movement of substances
51	GO:0004683	2.05E-03	35.339	6.84E-02	2	13	calmodulin-dependent protein kinase activity
52	GO:0015399	2.05E-03	35.339	6.84E-02	2	13	primary active transmembrane transporter activity
53	GO:0015405	2.05E-03	35.339	6.84E-02	2	13	P-P-bond-hydrolysis-driven transmembrane transporter activity
54	GO:0005543	2.29E-03	12.555	2.63E-01	3	50	phospholipid binding
55	GO:0042562	2.74E-03	29.897	7.89E-02	2	15	hormone binding
56	GO:0140096	2.75E-03	2.609	6.15E+00	14	1169	catalytic activity, acting on a protein
57	GO:0008134	3.17E-03	11.128	2.95E-01	3	56	transcription factor binding
58	GO:0046983	3.23E-03	5.421	1.00E+00	5	190	protein dimerization activity
59	GO:0008289	3.24E-03	7.095	6.10E-01	4	116	lipid binding
60	GO:0098772	3.44E-03	3.891	1.97E+00	7	374	molecular function regulator

Note: Only the first raws are displayed. The complete data set is available in:
http://www.ub.edu/molevol/EGB/Escuer_Supplementary_Tables.zip

Supplementary Table 13. Genes under selection identified by the MK test

Gene	Scaffold	pN	pS	dN	dS	pN.pS	dN.dS	fisher.test.P
g29036	Scaffold_15362_HRSCAF_19823	5	51	5	1	0.098	5.000	0.0002
g15131	Scaffold_15272_HRSCAF_19649	29	13	1	9	2.231	0.111	0.0010
g21396	Scaffold_14178_HRSCAF_16784	4	20	5	0	0.200	Inf	0.0011
g1720	Scaffold_15005_HRSCAF_18897	7	19	6	0	0.368	Inf	0.0019
g3311	Scaffold_15005_HRSCAF_18897	14	6	0	8	2.333	0.000	0.0019
g21043	Scaffold_14178_HRSCAF_16784	1	53	2	0	0.019	Inf	0.0019
g14584	Scaffold_15272_HRSCAF_19649	0	26	2	0	0.000	Inf	0.0026
g2002	Scaffold_15005_HRSCAF_18897	15	165	3	1	0.091	3.000	0.0030
g21232	Scaffold_14178_HRSCAF_16784	2	53	2	0	0.038	Inf	0.0038
g21859	Scaffold_14178_HRSCAF_16784	12	9	0	10	1.333	0.000	0.0041
g32119	Scaffold_15361_HRSCAF_19822	6	1	0	6	6.000	0.000	0.0047
g29317	Scaffold_15362_HRSCAF_19823	9	4	0	7	2.250	0.000	0.0047
g25300	Scaffold_15362_HRSCAF_19823	1	9	4	0	0.111	Inf	0.0050
g27457	Scaffold_15362_HRSCAF_19823	24	10	0	5	2.400	0.000	0.0052
g32394	Scaffold_15361_HRSCAF_19822	9	2	1	7	4.500	0.143	0.0055
g12255	Scaffold_14804_HRSCAF_18385	27	83	5	1	0.325	5.000	0.0060
g22600	Scaffold_14178_HRSCAF_16784	38	164	4	1	0.232	4.000	0.0064
g3221	Scaffold_15005_HRSCAF_18897	16	4	0	4	4.000	0.000	0.0066
g12632	Scaffold_14804_HRSCAF_18385	8	2	0	5	4.000	0.000	0.0070
g24007	Scaffold_14178_HRSCAF_16784	1	22	2	0	0.045	Inf	0.0100
g4239	Scaffold_15005_HRSCAF_18897	10	4	0	6	2.500	0.000	0.0108
g16015	Scaffold_15272_HRSCAF_19649	16	62	3	0	0.258	Inf	0.0114
g3048	Scaffold_15005_HRSCAF_18897	2	29	2	0	0.069	Inf	0.0114
g12704	Scaffold_14804_HRSCAF_18385	0	20	2	1	0.000	2.000	0.0119
g20665	Scaffold_14178_HRSCAF_16784	1	20	2	0	0.050	Inf	0.0119
g25937	Scaffold_15362_HRSCAF_19823	6	0	0	3	Inf	0.000	0.0119
g19749	Scaffold_3225_HRSCAF_3759	4	41	2	0	0.098	Inf	0.0139
g18912	Scaffold_3225_HRSCAF_3759	0	71	1	0	0.000	Inf	0.0139
g1605	Scaffold_15005_HRSCAF_18897	9	1	0	3	9.000	0.000	0.0140
g16155	Scaffold_15272_HRSCAF_19649	46	13	0	3	3.538	0.000	0.0148
g19171	Scaffold_3225_HRSCAF_3759	15	3	0	3	5.000	0.000	0.0150
g3360	Scaffold_15005_HRSCAF_18897	4	18	3	0	0.222	Inf	0.0152
g12386	Scaffold_14804_HRSCAF_18385	27	48	6	1	0.563	6.000	0.0154
g16537	Scaffold_15272_HRSCAF_19649	1	17	2	0	0.059	Inf	0.0158
g14703	Scaffold_15272_HRSCAF_19649	5	20	3	0	0.250	Inf	0.0171
g32922	Scaffold_15361_HRSCAF_19822	2	23	2	0	0.087	Inf	0.0171
g3992	Scaffold_15005_HRSCAF_18897	5	0	0	3	Inf	0.000	0.0179
g25086	Scaffold_15362_HRSCAF_19823	5	0	0	3	Inf	0.000	0.0179
g14769	Scaffold_15272_HRSCAF_19649	9	0	0	2	Inf	0.000	0.0182
g20444	Scaffold_14178_HRSCAF_16784	0	9	2	0	0.000	Inf	0.0182
g22248	Scaffold_14178_HRSCAF_16784	8	104	2	1	0.077	2.000	0.0196
g16619	Scaffold_15272_HRSCAF_19649	0	47	1	0	0.000	Inf	0.0208
g10355	Scaffold_14804_HRSCAF_18385	18	5	0	3	3.600	0.000	0.0215
g21825	Scaffold_14178_HRSCAF_16784	10	2	0	3	5.000	0.000	0.0220
g25632	Scaffold_15362_HRSCAF_19823	8	0	0	2	Inf	0.000	0.0222
g12076	Scaffold_14804_HRSCAF_18385	18	9	0	4	2.000	0.000	0.0227
g31880	Scaffold_15361_HRSCAF_19822	1	24	2	1	0.042	2.000	0.0232
g5146	Scaffold_15005_HRSCAF_18897	33	25	12	1	1.320	12.000	0.0236
g30005	Scaffold_15362_HRSCAF_19823	2	19	2	0	0.105	Inf	0.0237
g29581	Scaffold_15362_HRSCAF_19823	11	9	0	6	1.222	0.000	0.0237
g2927	Scaffold_15005_HRSCAF_18897	3	24	2	0	0.125	Inf	0.0246
g5022	Scaffold_15005_HRSCAF_18897	16	7	1	6	2.286	0.167	0.0247
g25823	Scaffold_15362_HRSCAF_19823	28	16	0	4	1.750	0.000	0.0249
g22752	Scaffold_14178_HRSCAF_16784	13	1	0	2	13.000	0.000	0.0250
g28624	Scaffold_15362_HRSCAF_19823	7	3	0	5	2.333	0.000	0.0256
g13967	Scaffold_15272_HRSCAF_19649	5	28	3	1	0.179	3.000	0.0256
g12362	Scaffold_14804_HRSCAF_18385	7	0	0	2	Inf	0.000	0.0278
g15638	Scaffold_15272_HRSCAF_19649	0	7	2	0	0.000	Inf	0.0278
g30054	Scaffold_15362_HRSCAF_19823	18	10	0	4	1.800	0.000	0.0278
g27242	Scaffold_15362_HRSCAF_19823	4	0	0	3	Inf	0.000	0.0286
g14948	Scaffold_15272_HRSCAF_19649	3	0	0	4	Inf	0.000	0.0286
g20361	Scaffold_14178_HRSCAF_16784	4	0	0	3	Inf	0.000	0.0286
g23113	Scaffold_14178_HRSCAF_16784	1	21	2	1	0.048	2.000	0.0291
g31627	Scaffold_15361_HRSCAF_19822	5	4	0	8	1.250	0.000	0.0294
g13910	Scaffold_15272_HRSCAF_19649	3	6	5	0	0.500	Inf	0.0310
g18520	Scaffold_3225_HRSCAF_3759	5	53	2	1	0.094	2.000	0.0325
g10521	Scaffold_14804_HRSCAF_18385	15	9	0	4	1.667	0.000	0.0349
g21712	Scaffold_14178_HRSCAF_16784	9	15	4	0	0.600	Inf	0.0349

g14537	Scaffold_15272_HRSCAF_19649	21	47	3	0	0.447	Inf	0.0354
g17340	Scaffold_3225_HRSCAF_3759	1	9	4	2	0.111	2.000	0.0357
g21230	Scaffold_14178_HRSCAF_16784	0	6	2	0	0.000	Inf	0.0357
g23187	Scaffold_14178_HRSCAF_16784	18	7	0	3	2.571	0.000	0.0366
g12239	Scaffold_14804_HRSCAF_18385	36	31	7	0	1.161	Inf	0.0370
g11548	Scaffold_14804_HRSCAF_18385	0	26	1	0	0.000	Inf	0.0370
g17787	Scaffold_3225_HRSCAF_3759	0	26	1	0	0.000	Inf	0.0370
g3207	Scaffold_15005_HRSCAF_18897	0	26	1	0	0.000	Inf	0.0370
g26722	Scaffold_15362_HRSCAF_19823	3	11	4	1	0.273	4.000	0.0379
g21663	Scaffold_14178_HRSCAF_16784	0	25	1	0	0.000	Inf	0.0385
g24550	Scaffold_15362_HRSCAF_19823	1	10	2	0	0.100	Inf	0.0385
g4827	Scaffold_15005_HRSCAF_18897	21	32	6	1	0.656	6.000	0.0387
g16450	Scaffold_15272_HRSCAF_19649	2	14	2	0	0.143	Inf	0.0392
g19068	Scaffold_3225_HRSCAF_3759	33	24	0	4	1.375	0.000	0.0392
g16144	Scaffold_15272_HRSCAF_19649	7	10	5	0	0.700	Inf	0.0396
g5037	Scaffold_15005_HRSCAF_18897	7	2	1	5	3.500	0.200	0.0406
g30746	Scaffold_15362_HRSCAF_19823	19	8	0	3	2.375	0.000	0.0406
g28886	Scaffold_15362_HRSCAF_19823	17	7	0	3	2.429	0.000	0.0410
g10184	Scaffold_14804_HRSCAF_18385	10	15	4	0	0.667	Inf	0.0421
g14864	Scaffold_15272_HRSCAF_19649	6	5	0	6	1.200	0.000	0.0427
g4456	Scaffold_15005_HRSCAF_18897	9	8	11	1	1.125	11.000	0.0432
g26311	Scaffold_15362_HRSCAF_19823	3	17	2	0	0.176	Inf	0.0433
g29682	Scaffold_15362_HRSCAF_19823	9	3	0	3	3.000	0.000	0.0440
g12163	Scaffold_14804_HRSCAF_18385	9	12	7	1	0.750	7.000	0.0443
g10464	Scaffold_14804_HRSCAF_18385	0	21	1	0	0.000	Inf	0.0455
g11583	Scaffold_14804_HRSCAF_18385	7	2	0	3	3.500	0.000	0.0455
g15352	Scaffold_15272_HRSCAF_19649	7	31	2	0	0.226	Inf	0.0462
g11814	Scaffold_14804_HRSCAF_18385	12	10	0	5	1.200	0.000	0.0470
g15730	Scaffold_15272_HRSCAF_19649	8	34	2	0	0.235	Inf	0.0476
g20185	Scaffold_14178_HRSCAF_16784	4	1	0	4	4.000	0.000	0.0476
g2140	Scaffold_15005_HRSCAF_18897	1	40	1	0	0.025	Inf	0.0476
g4021	Scaffold_15005_HRSCAF_18897	4	1	0	4	4.000	0.000	0.0476
g9901	Scaffold_14804_HRSCAF_18385	3	1	0	5	3.000	0.000	0.0476
g24635	Scaffold_15362_HRSCAF_19823	5	1	0	3	5.000	0.000	0.0476
g23240	Scaffold_14178_HRSCAF_16784	10	21	3	0	0.476	Inf	0.0478
g14330	Scaffold_15272_HRSCAF_19649	2	12	2	0	0.167	Inf	0.0500

In gray, candiate gene to be under negative selection (n = 49; P-value < 0.05);
In red, candiate gene to be under positive selection (n = 41; P-value < 0.05);
in blue, strong candidate genes to be under positive selection due to a high number of non-synonymous fixations (n = 14; P-value < 0.05).

Supplementary Table 14. Positively selected genes in the MK test and their respective GO. In bold, genes that are the strong candidates

	ID gene	GO																					
	g29036	GO:0004674	GO:0005516	GO:0005524	GO:0005865	GO:0006936	GO:0008307	GO:0008344	GO:0018105	GO:0018107	GO:0019901	GO:0030017	GO:0030018	GO:0031430	GO:0031672	GO:0035095	GO:0040017	GO:0045214	GO:0045989	GO:0046872	GO:0051015	GO:0051371	
	g29317	GO:0000166	GO:0003676	GO:0008270																			
	g25300	GO:0004674	GO:0004697	GO:0005524	GO:0005886	GO:0006468	GO:0018105	GO:0035556	GO:0046872	GO:0002225	GO:0009611	GO:0031435	GO:0043410	GO:0048680	GO:0050832	GO:1900150	GO:1901244	GO:0004672	GO:0004699	GO:0004715	GO:0005634	GO:0005737	
	g27457																						
	g25937																						
	g25086																						
	g25632	GO:0003676	GO:0003964	GO:0004190	GO:0004519	GO:0015074	GO:0002870	GO:0003677	GO:0003723	GO:0003887	GO:0005634	GO:0006310	GO:0046872										
	g30005	GO:0005201	GO:0005581	GO:0005604	GO:0005198	GO:0005587	GO:0005615	GO:0008104	GO:0009792	GO:0016043	GO:0030020	GO:0030198	GO:0031012	GO:0032562	GO:1903354	GO:0007391	GO:0030848	GO:0048621	GO:0055013	GO:0060729	GO:0071711	GO:0005576	
	g29581	GO:0000165	GO:0000226	GO:0005125	GO:0005168	GO:0005170	GO:0001725	GO:0001764	GO:0001932	GO:0001934	GO:0003779	GO:0004672	GO:0004713	GO:0004715	GO:0005178	GO:0005524	GO:0005634	GO:0005730	GO:0005737	GO:0005815	GO:0005829	GO:0005886	
ChrX	g25823	GO:0005654	GO:0016604	GO:0043565	GO:0002230	GO:0005634	GO:0010628	GO:0033141	GO:0045739	GO:0003700	GO:0005044	GO:0016020	GO:0042802										
	g28624																						
	g30054	GO:0003676																					
	g27242	GO:0003676																					
	g26722	GO:0004252	GO:0005615	GO:0006954	GO:0007597	GO:0031639	GO:0042730	GO:0042803	GO:0051919	GO:0000122	GO:0005886	GO:0006508	GO:0006879	GO:0016021	GO:0030514	GO:0033619	GO:0045892	GO:0045944	GO:0055072	GO:0097264	GO:0004175	GO:0005576		
	g24550	GO:0005451	GO:0005743	GO:0005886	GO:0006814	GO:0010008	GO:0010348	GO:0015385	GO:0016021	GO:0016323	GO:0016324	GO:0030054	GO:0030317	GO:0030672	GO:0031966	GO:0034220	GO:0042802	GO:0061178	GO:0072583	GO:0097228	GO:2001206	GO:0097708	
	g30746																						
	g28886	GO:0004190	GO:0006508	GO:0003676	GO:0008270																		
	g26311	GO:0005524	GO:0005741	GO:0006457	GO:0007275	GO:0009408	GO:0031072	GO:0046872	GO:0051082	GO:0005113	GO:0005739	GO:0005829	GO:0007005	GO:0045879	GO:0000122	GO:0005133	GO:0005634	GO:0005737	GO:0005759	GO:0006264	GO:0006469	GO:0006919		
	g29682	GO:0003676	GO:0004523	GO:0005085	GO:0005622	GO:0007264																	
	g24635																						
	g12255																						
	g12386	GO:0004089	GO:0006281	GO:0008270																			
Chr1	**g12239**	GO:0002027	GO:0003283	GO:0005200	GO:0005739	GO:0005764	GO:0005769	GO:0005829	GO:0005856	GO:0005886	GO:0006874	GO:0006888	GO:0006897	GO:0010628	GO:0010881	GO:0010882	GO:0014704	GO:0015031	GO:0016020	GO:0016323	GO:0016324	GO:0019899		
	g11548	GO:0000176	GO:0000460	GO:0003723	GO:0004004	GO:0005524	GO:0005634	GO:0005654	GO:0005730	GO:0006364	GO:0006397	GO:0006401	GO:0008380	GO:0031499	GO:0071013	GO:0000398	GO:0016078	GO:0043630	GO:0071049	GO:0000467	GO:0003729	GO:0008143		
	g10184	GO:0000056	GO:0005634	GO:0005654	GO:0005829	GO:0030688	GO:0031902	GO:0034448	GO:0042274	GO:0000278	GO:0004674	GO:0005524	GO:0005694	GO:0005737	GO:0005813	GO:0005819	GO:0007064	GO:0035556	GO:0071459	GO:0072354	GO:2000751	GO:0016772		
	g12163	GO:0016021	GO:0016486	GO:0016508	GO:0008237	GO:0046872	GO:0000139	GO:0005802	GO:0030658	GO:0008188	GO:0007420	GO:0007507	GO:0010002	GO:0030659	GO:0042445	GO:0016020								
	g21196	GO:0000793	GO:0000796	GO:0003682	GO:0005634	GO:0010032	GO:0051276	GO:0051306	GO:0005654	GO:0016020	GO:0030054	GO:0033077	GO:0045171	GO:0051309	GO:0005694	GO:0007143	GO:0032261;							
	g21043	GO:0003950	GO:0005634	GO:0005694	GO:0006281	GO:0006471	GO:0006974	GO:0008630	GO:0010332	GO:0043069	GO:0003677	GO:0005730	GO:0006302	GO:0070212	GO:1990404	GO:0006303	GO:0005654	GO:0006284	GO:0018312	GO:0030592	GO:0061051	GO:0097191		
	g21232	GO:0001843	GO:0001947	GO:0005929	GO:0005930	GO:0007219	GO:0007224	GO:0007420	GO:0008544	GO:0009953	GO:0016485	GO:0021522	GO:0030992	GO:0031122	GO:0035735	GO:0036064	GO:0045880	GO:0050680	GO:0060021	GO:0060173	GO:0060271	GO:0060348		
	g22600	GO:0000226	GO:0003779	GO:0005178	GO:0005198	GO:0005509	GO:0005604	GO:0005634	GO:0005635	GO:0005737	GO:0005789	GO:0005829	GO:0005882	GO:0005925	GO:0005938	GO:0007010	GO:0007155	GO:0007229	GO:0008017	GO:0008022	GO:0008090	GO:0009611		
	g24007	GO:0007283	GO:0008589	GO:0015631	GO:0030154	GO:0030992	GO:0035735	GO:0036064	GO:0042073	GO:0060271	GO:0097225	GO:0097228	GO:0031514	GO:0097542	GO:0045177	GO:0042384								
	g20665	GO:0005245	GO:0005891	GO:0006816	GO:0034765	GO:0046872	GO:0005886	GO:0005246	GO:1990314	GO:0005908	GO:0070588													
Chr2	g20444	GO:0016021	GO:0003735	GO:0005622	GO:0005840	GO:0006412	GO:0046983																	
	g22248	GO:0001658	GO:0003007	GO:0005509	GO:0005886	GO:0007009	GO:0007156	GO:0007157	GO:0007219	GO:0008543	GO:0016021	GO:0021987	GO:0022008	GO:0033329	GO:0043931	GO:0045177	GO:0048565	GO:0050122	GO:0070062	GO:0072137	GO:0072307	GO:0001736		
	g23113	GO:0001656	GO:0001657	GO:0003148	GO:0003180	GO:0003184	GO:0003272	GO:0005886	GO:0007156	GO:0007411	GO:0007420	GO:0007426	GO:0008046	GO:0009986	GO:0016021	GO:0016199	GO:0021510	GO:0021891	GO:0030673	GO:0032990	GO:0032870	GO:0035385		
	g21712	GO:0004190	GO:0005764	GO:0006508	GO:0008219	GO:0012501	GO:0030163	GO:0004175	GO:0005615	GO:0031667	GO:0042277	GO:0042470	GO:0008233	GO:0016787	GO:0005575;									
	g21230																						
	g21663	GO:0007155	GO:0031012	GO:0046872	GO:0005201	GO:0005615	GO:0005788	GO:0010954	GO:0032092	GO:0050693	GO:1902430	GO:1902993	GO:0062023	GO:0004867									
	g23240	GO:0005737	GO:0007423	GO:0031012	GO:0040011	GO:0050896	GO:0060271	GO:0005509	GO:0005576	GO:0007420	GO:0016020	GO:0021510	GO:0030154	GO:0003180	GO:0003181	GO:0005615	GO:0005739	GO:0007411	GO:0008285	GO:0010629	GO:0030308	GO:0032870		
	g1720	GO:0005261	GO:0005272	GO:0005886	GO:0006812	GO:0016021	GO:0040011	GO:0043025	GO:0016020	GO:0006814														
	g2002	GO:0004674	GO:0005516	GO:0005524	GO:0005865	GO:0006936	GO:0008307	GO:0008344	GO:0018105	GO:0018107	GO:0019901	GO:0030017	GO:0030018	GO:0031430	GO:0031672	GO:0035095	GO:0040017	GO:0045214	GO:0045989	GO:0046872	GO:0051015	GO:0051371		
	g3048	GO:0006814	GO:0015293	GO:0016021	GO:0005886	GO:0006811	GO:0006915	GO:0008028	GO:0015129	GO:0015552	GO:0015913	GO:0016324	GO:0022803	GO:0034356	GO:0070062	GO:0140161	GO:0005343	GO:0015636	GO:0015718	GO:0015730	GO:0035526	GO:0015355		
	g3360	GO:0001666	GO:0001953	GO:0003073	GO:0005201	GO:0005615	GO:0005802	GO:0007155	GO:0008201	GO:0008593	GO:0009612	GO:0014850	GO:0014911	GO:0030198	GO:0031012	GO:0031594	GO:0032355	GO:0044344	GO:0046872	GO:0050839	GO:0062023	GO:0071307		
	g5146																						
Chr3	g2927	GO:0001786	GO:0005044	GO:0005548	GO:0005765	GO:0005886	GO:0006622	GO:0006898	GO:0008565	GO:0010976	GO:0015485	GO:0015917	GO:0016021	GO:0019899	GO:0030666	GO:0031210	GO:0042803	GO:0043202	GO:0043471	GO:0051087	GO:1905123	GO:0000139		
	g3207	GO:0000002	GO:0001018	GO:0001666	GO:0003697	GO:0003727	GO:0004176	GO:0004252	GO:0005524	GO:0005634	GO:0005654	GO:0005739	GO:0005759	GO:0005829	GO:0006515	GO:0007005	GO:0007568	GO:0009725	GO:0010044	GO:0016020	GO:0032042	GO:0034599		
	g4827	GO:0000775	GO:0003700	GO:0003723	GO:0005634	GO:0008270	GO:0030702	GO:0031048	GO:0031380	GO:0000981	GO:0004386	GO:0005524	GO:0005829	GO:0007049	GO:0007059	GO:0031379	GO:0006355							
	g4456	GO:0000724	GO:0000239	GO:0004842	GO:0005634	GO:0005737	GO:0006974	GO:0010212	GO:0016567	GO:0016605	GO:0031297	GO:0031571	GO:0036297	GO:0046872	GO:0090734	GO:0097371	GO:0005654	GO:0031052	GO:0035861	GO:2000001	GO:0005515	GO:0008270		
	g2140	GO:0005096	GO:0051056	GO:0005737	GO:0005856	GO:0014069	GO:0030054	GO:0031532	GO:0043087	GO:0043197	GO:0045211	GO:0046875	GO:0048013	GO:0048167	GO:0050770	GO:0061001	GO:0090630	GO:0019901	GO:0031625	GO:0032991	GO:0043025	GO:0044877		
	g14584	GO:0000226	GO:0005085	GO:0005925	GO:0007264	GO:0007409	GO:0022027	GO:0030424	GO:0030426	GO:0031175	GO:0031138	GO:0043005	GO:0045178	GO:0045200	GO:0048365	GO:0050767	GO:0090630	GO:0120163	GO:1904754	GO:0002244	GO:0043473	GO:0005829		
	g16015	GO:0003697	GO:0004312	GO:0004313	GO:0004314	GO:0004315	GO:0004316	GO:0004320	GO:0005623	GO:0006089	GO:0006633	GO:0008659	GO:0016295	GO:0016296	GO:0031177	GO:0032100	GO:0047117	GO:0047451	GO:0102131	GO:0102132	GO:0005737	GO:0005739		
	g16537	GO:0005509	GO:0005886	GO:0005887	GO:0005918	GO:0005919	GO:0005923	GO:0007158	GO:0007409	GO:0007560	GO:0008045	GO:0008049	GO:0008050	GO:0008366	GO:0016319	GO:0019991	GO:0021682	GO:0030175	GO:0035011	GO:0035151	GO:0045924	GO:0048036		
	g14703	GO:0005886	GO:0005887	GO:0006814	GO:0009408	GO:0010447	GO:0015280	GO:0034220	GO:0042931	GO:0044871	GO:0050915	GO:0050965	GO:0050968	GO:0050982	GO:0005261	GO:0006812	GO:0006912	GO:0050907	GO:0050961	GO:0050974	GO:0016324	GO:0030659		
	g16619	GO:0000165	GO:0001649	GO:0003756	GO:0001666	GO:0001889	GO:0001952	GO:0005096	GO:0005634	GO:0005730	GO:0005737	GO:0006469	GO:0007154	GO:0007265	GO:0007406	GO:0007420	GO:0007422	GO:0007507	GO:0008017	GO:0008429	GO:0008542	GO:0010976		
	g13967	GO:0016021	GO:0016747	GO:0006869	GO:0007275	GO:0008289																		
	g15638	GO:0002244	GO:0005432	GO:0005516	GO:0005741	GO:0005789	GO:0005874	GO:0005886	GO:0005887	GO:0006811	GO:0006851	GO:0006874	GO:0007154	GO:0007612	GO:0007613	GO:0014819	GO:0016528	GO:0021537	GO:0030054	GO:0032126	GO:0035194	GO:0035725		
Chr4	**g13910**	GO:0005737	GO:0005874	GO:0007019	GO:0015631	GO:0031110	GO:0031175	GO:0043005	GO:0051493	GO:0000281	GO:0005622	GO:0005829	GO:0007052	GO:0007165	GO:0007409	GO:0007420	GO:0009615	GO:0016020	GO:0031115	GO:0035024	GO:0035556	GO:0048012		
	g14537	GO:0005886	GO:0015137	GO:0015141	GO:0015746	GO:0016021	GO:0017153	GO:0051362	GO:0070062	GO:0006814	GO:0015293	GO:0005887	GO:0006848	GO:0008340	GO:0010389	GO:0015744	GO:0016323	GO:0050833	GO:0005343	GO:0006842	GO:0015142	GO:0005215		
	g16450	GO:0005102	GO:0005615	GO:0005737	GO:0005886	GO:0007605	GO:0010640	GO:0010977	GO:0016020	GO:0016021	GO:0030426	GO:0031012	GO:0043005	GO:0043025	GO:0048679	GO:0048681	GO:0051045	GO:0060384	GO:0097708	GO:2000010	GO:2001222	GO:0002752		
	g16144	GO:0005506	GO:0005737	GO:0005789	GO:0006082	GO:0006805	GO:0008395	GO:0010164	GO:0010212	GO:0016712	GO:0020037	GO:0033543	GO:0031090	GO:0042359	GO:0042738	GO:0043231	GO:0055114	GO:0060766	GO:0001523	GO:0030858	GO:0005783	GO:0007565		
	g15352	GO:0005576	GO:0030414	GO:0005515	GO:0004869																		
	g15730	GO:0003743	GO:0003924	GO:0005525	GO:0005654	GO:0005739	GO:0008135	GO:0032790	GO:0043024	GO:0070124	GO:0005759	GO:0003723	GO:0006446	GO:0005737	GO:0006412	GO:0006413	GO:0005622							
	g14330																						
	g19749	GO:0006623	GO:0019538	GO:0019898	GO:0031902	GO:0045053	GO:0005741	GO:0005829	GO:0007005	GO:1905090	GO:0005737	GO:0006895	GO:0070062	GO:0005622	GO:0006914	GO:0007399	GO:0007626	GO:0008104	GO:0031045	GO:0035176				
Chr5	g18912	GO:0000278	GO:0000776	GO:0003777	GO:0005524	GO:0005737	GO:0005794	GO:0005868	GO:0005874	GO:0005875	GO:0005938	GO:0006886	GO:0007018	GO:0007051	GO:0007052	GO:0007098	GO:0007279	GO:0007282	GO:0007294	GO:0007298	GO:0007312	GO:0007349		
	g18520	GO:0004725	GO:0005001	GO:0005925	GO:0007155	GO:0007283	GO:0007399	GO:0007411	GO:0007412	GO:0008045	GO:0008201	GO:0008594	GO:0009925	GO:0016021	GO:0030424	GO:0031252	GO:0031290	GO:0032093	GO:0035335	GO:0045467	GO:0048477	GO:0048675		
	g17787	GO:0003677	GO:0005634	GO:0046872	GO:0006310	GO:0015074	GO:0046983	GO:0000981	GO:0000978	GO:0001228	GO:0005654	GO:0019789	GO:0042802	GO:0043565	GO:0045944	GO:0003676								
	g32922	GO:0005509	GO:0005743	GO:0006851	GO:0015369	GO:0016021	GO:0034214	GO:0043022	GO:0051260	GO:0055085	GO:0090093	GO:0005432	GO:0031090	GO:0007269	GO:0015386	GO:0031966	GO:0070584	GO:0071456	GO:0071805	GO:0098793	GO:1902600			
Chr6	g31880	GO:0005634	GO:0006355	GO:0016922	GO:0030099	GO:0030374	GO:0030375	GO:0030851	GO:0032870	GO:0035257	GO:0043009	GO:0045944	GO:0046966	GO:0046983	GO:0048511	GO:0060215	GO:0004871	GO:0007165						

In bold, genes with strong evidence of positive selection
#GO terms: 2262

Note: Only the first columns are displayed. The complete data set is available in:
http://www.ub.edu/molevol/EGB/Escuer_Supplementary_Tables.zip

Supplementary Table 15. GO enrichment of Biological Process from MK analyses

Num	GOBPID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
1	GO:0051179	1.13E-36	5.617	37.437	121	1045	localization
2	GO:0006810	1.93E-30	5.552	25.937	93	724	transport
3	GO:0065008	2.46E-30	6.165	20.205	82	564	regulation of biological quality
4	GO:0051234	2.65E-30	5.434	27.155	95	758	establishment of localization
5	GO:0051641	2.95E-25	6.868	12.324	59	344	cellular localization
6	GO:0065007	3.85E-22	3.431	103.282	180	2883	biological regulation
7	GO:0051649	8.02E-22	7.542	8.562	46	239	establishment of localization in cell
8	GO:0023052	5.89E-20	3.951	29.806	84	832	signaling
9	GO:0007154	1.02E-19	3.910	30.057	84	839	cell communication
10	GO:0016192	1.42E-19	7.442	7.595	41	212	vesicle-mediated transport
11	GO:0046907	6.75E-19	8.393	5.983	36	167	intracellular transport
12	GO:0048869	2.68E-18	3.778	29.054	80	811	cellular developmental process
13	GO:0030154	3.08E-18	3.889	26.618	76	743	cell differentiation
14	GO:0050789	4.72E-18	2.968	93.860	162	2620	regulation of biological process
15	GO:0033036	5.13E-18	5.463	11.930	49	333	macromolecule localization
16	GO:0051049	1.66E-16	6.844	7.022	36	196	regulation of transport
17	GO:0032501	2.14E-16	2.823	77.309	140	2158	multicellular organismal process
18	GO:0042221	3.76E-16	4.001	20.312	62	567	response to chemical
19	GO:0006811	4.07E-16	6.037	8.490	39	237	ion transport
20	GO:0043269	6.45E-16	17.033	1.935	20	54	regulation of ion transport
21	GO:0032879	1.00E-15	5.173	11.034	44	308	regulation of localization
22	GO:0008104	1.80E-15	5.579	9.314	40	260	protein localization
23	GO:0003008	1.96E-15	5.176	10.747	43	300	system process
24	GO:0048468	5.56E-15	4.457	14.115	49	394	cell development
25	GO:0050896	1.15E-14	2.759	58.824	114	1642	response to stimulus
26	GO:0007267	2.19E-14	6.059	7.272	34	203	cell-cell signaling
27	GO:0007399	2.31E-14	3.863	18.557	56	518	nervous system development
28	GO:0023051	8.73E-14	4.168	14.581	48	407	regulation of signaling
29	GO:0006812	8.97E-14	7.203	5.123	28	143	cation transport
30	GO:0010646	1.73E-13	4.148	14.294	47	399	regulation of cell communication
31	GO:0010033	5.06E-13	4.301	12.539	43	350	response to organic substance
32	GO:0042592	6.98E-13	5.805	6.807	31	190	homeostatic process
33	GO:0055080	1.20E-12	15.732	1.612	16	45	cation homeostasis
34	GO:0032502	1.42E-12	2.470	72.115	125	2013	developmental process
35	GO:0098771	1.78E-12	15.205	1.648	16	46	inorganic ion homeostasis
36	GO:0007167	3.07E-12	9.617	2.866	20	80	enzyme linked receptor protein signaling pathway
37	GO:0007165	3.53E-12	3.158	24.719	62	690	signal transduction
38	GO:0055065	3.83E-12	16.392	1.469	15	41	metal ion homeostasis
39	GO:0055085	3.89E-12	7.692	3.941	23	110	transmembrane transport
40	GO:0065009	4.87E-12	5.003	8.240	33	230	regulation of molecular function
41	GO:0050801	6.13E-12	12.148	2.042	17	57	ion homeostasis
42	GO:0071702	7.48E-12	4.426	10.353	37	289	organic substance transport
43	GO:0048856	8.62E-12	2.417	66.992	117	1870	anatomical structure development
44	GO:0030030	8.63E-12	5.564	6.556	29	183	cell projection organization
45	GO:0006928	1.18E-11	4.453	9.995	36	279	movement of cell or subcellular component
46	GO:0046903	2.06E-11	6.296	5.051	25	141	secretion
47	GO:0006936	2.76E-11	15.850	1.397	14	39	muscle contraction
48	GO:0051716	2.88E-11	2.693	36.434	77	1017	cellular response to stimulus
49	GO:0120036	3.20E-11	5.421	6.448	28	180	plasma membrane bounded cell projection organization
50	GO:0003012	3.74E-11	13.307	1.684	15	47	muscle system process
51	GO:1901564	5.45E-11	2.727	33.281	72	929	organonitrogen compound metabolic process
52	GO:0034762	5.67E-11	17.463	1.218	13	34	regulation of transmembrane transport
53	GO:0022008	5.97E-11	3.913	12.180	39	340	neurogenesis
54	GO:0016043	7.99E-11	2.562	40.912	82	1142	cellular component organization
55	GO:0031175	8.12E-11	6.127	4.944	24	138	neuron projection development
56	GO:0019725	8.26E-11	9.904	2.364	17	66	cellular homeostasis
57	GO:0019538	8.49E-11	2.824	28.660	65	800	protein metabolic process
58	GO:0040011	8.53E-11	4.546	8.634	32	241	locomotion
59	GO:0034613	8.94E-11	5.350	6.269	27	175	cellular protein localization
60	GO:0048666	1.02E-10	5.313	6.305	27	176	neuron development

Note: Only the first raws are displayed. The complete data set is available in:
http://www.ub.edu/molevol/EGB/Escuer_Supplementary_Tables.zip

Supplementary Table 16. GO enrichment of Molecular Function from MK analyses

Number	GOMFID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
1	GO:0005515	1.83E-62	15.105	20.721	113	1635	protein binding
2	GO:0019899	1.38E-23	13.433	3.371	33	266	enzyme binding
3	GO:0042802	4.18E-18	22.378	1.179	19	93	identical protein binding
4	GO:0008092	1.08E-16	13.705	2.078	22	164	cytoskeletal protein binding
5	GO:0019904	5.07E-15	35.276	0.558	13	44	protein domain specific binding
6	GO:0015631	1.02E-12	67.227	0.253	9	20	tubulin binding
7	GO:0008134	4.01E-12	22.775	0.710	12	56	transcription factor binding
8	GO:0016491	2.08E-11	5.248	6.691	29	528	oxidoreductase activity
9	GO:0005488	3.92E-11	3.192	97.495	138	7693	binding
10	GO:0019900	5.86E-11	21.231	0.684	11	54	kinase binding
11	GO:0019901	5.45E-10	20.624	0.634	10	50	protein kinase binding
12	GO:0042803	6.26E-10	26.377	0.469	9	37	protein homodimerization activity
13	GO:0046982	1.31E-09	51.650	0.228	7	18	protein heterodimerization activity
14	GO:0045296	3.29E-09	80.703	0.152	6	12	cadherin binding
15	GO:0046872	7.40E-09	3.792	9.708	31	766	metal ion binding
16	GO:0043169	8.12E-09	3.776	9.746	31	769	cation binding
17	GO:0003824	9.83E-09	2.445	48.057	83	3792	catalytic activity
18	GO:0046983	1.82E-08	7.224	2.408	15	190	protein dimerization activity
19	GO:0016301	2.60E-08	4.487	5.944	23	469	kinase activity
20	GO:0044877	4.24E-08	10.459	1.242	11	98	protein-containing complex binding
21	GO:0043167	5.08E-08	3.023	15.525	39	1225	ion binding
22	GO:0030165	7.38E-08	80.223	0.127	5	10	PDZ domain binding
23	GO:0048037	2.64E-07	14.811	0.659	8	52	cofactor binding
24	GO:0016740	4.19E-07	2.696	18.199	41	1436	transferase activity
25	GO:0016705	4.70E-07	10.991	0.963	9	76	oxidoreductase activity, acting on paired donors, with incorporation or reduction of
26	GO:0051015	1.68E-06	79.749	0.101	4	8	actin filament binding
27	GO:0008391	2.00E-06	Inf	0.038	3	3	arachidonic acid monooxygenase activity
28	GO:0008392	2.00E-06	Inf	0.038	3	3	arachidonic acid epoxygenase activity
29	GO:0030506	2.00E-06	Inf	0.038	3	3	ankyrin binding
30	GO:0017134	2.00E-06	Inf	0.038	3	3	fibroblast growth factor binding
31	GO:0036094	2.35E-06	3.402	7.667	23	605	small molecule binding
32	GO:0031625	3.10E-06	28.632	0.241	5	19	ubiquitin protein ligase binding
33	GO:0050839	3.16E-06	17.906	0.418	6	33	cell adhesion molecule binding
34	GO:0016772	4.12E-06	2.860	11.621	29	917	transferase activity, transferring phosphorus-containing groups
35	GO:0016773	5.78E-06	3.810	5.285	18	417	phosphotransferase activity, alcohol group as acceptor
36	GO:0044389	6.82E-06	23.574	0.279	5	22	ubiquitin-like protein ligase binding
37	GO:0008022	7.67E-06	45.560	0.139	4	11	protein C-terminus binding
38	GO:0008270	1.01E-05	6.402	1.749	10	138	zinc ion binding
39	GO:0008017	1.14E-05	39.862	0.152	4	12	microtubule binding
40	GO:0046914	1.50E-05	4.622	3.118	13	246	transition metal ion binding
41	GO:0019903	1.63E-05	35.430	0.165	4	13	protein phosphatase binding
42	GO:0031690	1.96E-05	118.929	0.063	3	5	adrenergic receptor binding
43	GO:0050661	1.96E-05	118.929	0.063	3	5	NADP binding
44	GO:0019838	2.26E-05	31.885	0.177	4	14	growth factor binding
45	GO:0019902	3.05E-05	28.984	0.190	4	15	phosphatase binding
46	GO:0061659	3.89E-05	79.280	0.076	3	6	ubiquitin-like protein ligase activity
47	GO:0061630	3.89E-05	79.280	0.076	3	6	ubiquitin protein ligase activity
48	GO:0030674	4.02E-05	26.567	0.203	4	16	protein binding, bridging
49	GO:0004713	5.32E-05	10.271	0.672	6	53	protein tyrosine kinase activity
50	GO:0016903	5.49E-05	14.301	0.418	5	33	oxidoreductase activity, acting on the aldehyde or oxo group of donors
51	GO:0060090	8.32E-05	21.249	0.241	4	19	molecular adaptor activity
52	GO:0004497	1.27E-04	5.895	1.495	8	118	monooxygenase activity
53	GO:0140096	1.46E-04	2.276	14.815	30	1169	catalytic activity, acting on a protein
54	GO:0004672	1.48E-04	3.446	4.436	14	350	protein kinase activity
55	GO:0046906	1.59E-04	39.631	0.114	3	9	tetrapyrrole binding
56	GO:0016667	1.59E-04	39.631	0.114	3	9	oxidoreductase activity, acting on a sulfur group of donors
57	GO:0016273	1.60E-04	Inf	0.025	2	2	arginine N-methyltransferase activity
58	GO:0016274	1.60E-04	Inf	0.025	2	2	protein-arginine N-methyltransferase activity
59	GO:0034875	1.60E-04	Inf	0.025	2	2	caffeine oxidase activity
60	GO:0070402	1.60E-04	Inf	0.025	2	2	NADPH binding

Note: Only the first raws are displayed. The complete data set is available in:
http://www.ub.edu/molevol/EGB/Escuer_Supplementary_Tables.zip

Fig. S1

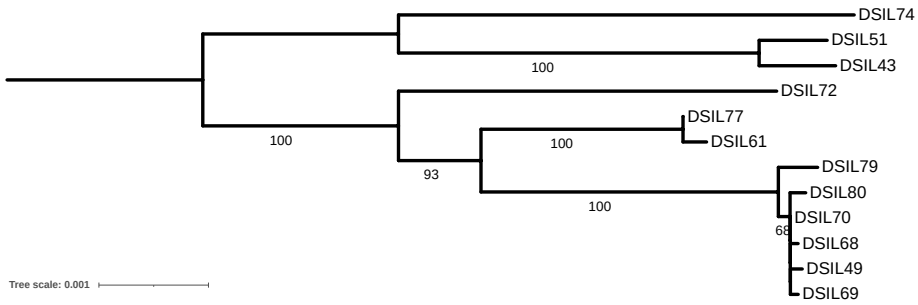
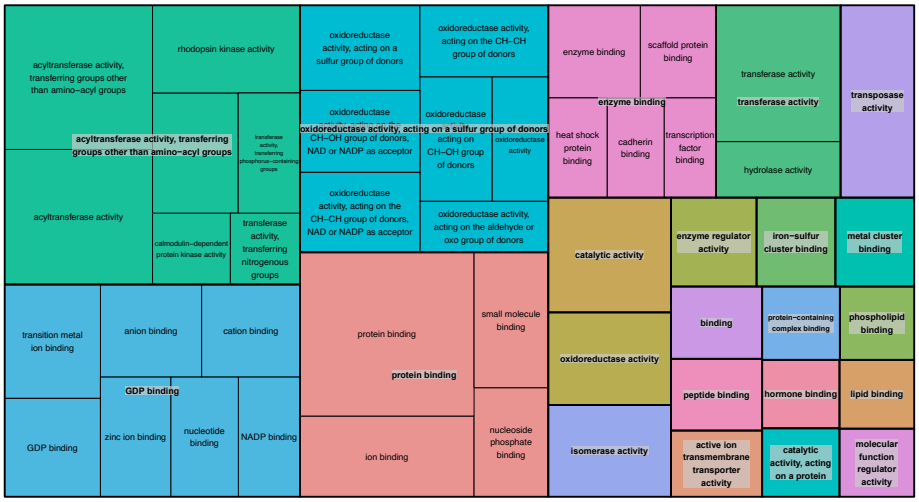

DSIL74
DSIL51
DSIL43
DSIL72
DSIL77
DSIL61
DSIL79
DSIL80
DSIL70
DSIL68
DSIL49
DSIL69
100
100
100
93
100
68
Tree scale: 0.001

Fig. S2

Molecular Function



Biological process

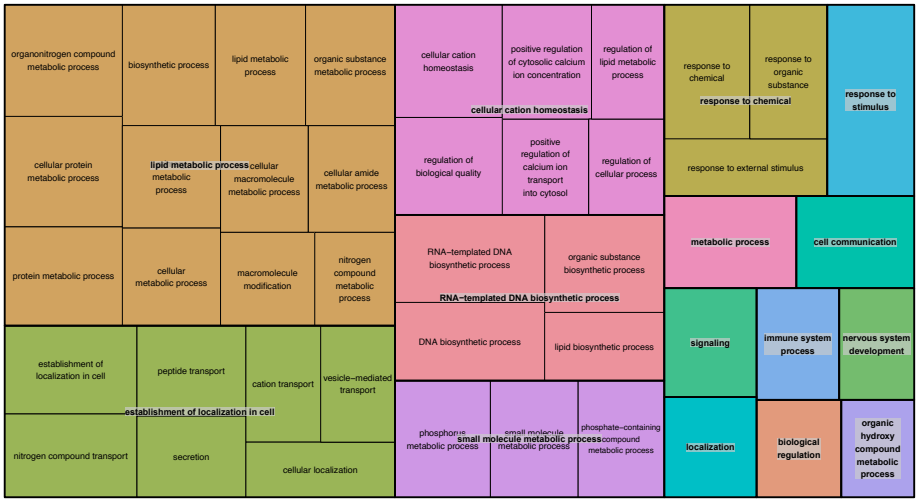
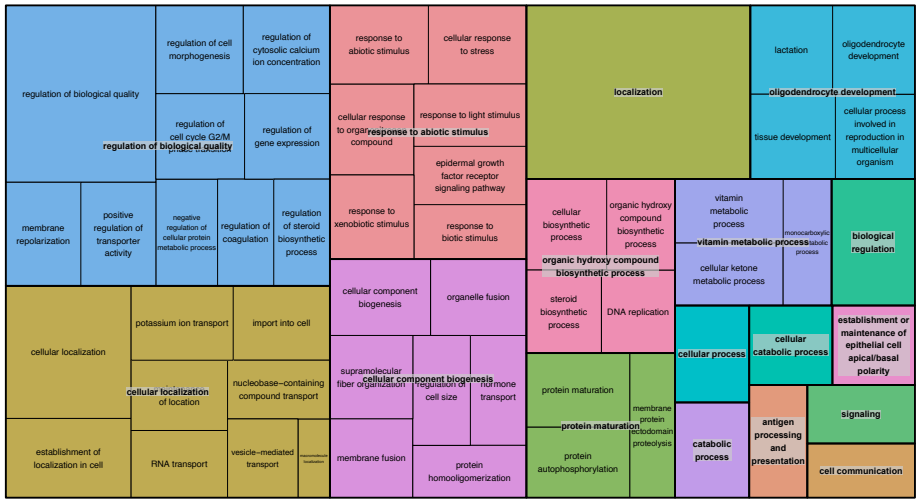


Fig. S3

Molecular function

Biological process



Discussion

Understanding how past natural events have shaped extant phenotypes and determined their current patterns in natural populations is crucial, not only to resolve relevant evolutionary and biological questions, but also to many other aspects such as the maintenance of biodiversity, or their use for the benefit of society. To address these questions, the study of molecular adaptation, i.e., the detection and characterization of the characteristic footprint of positive (and negative) selection at the molecular level, becomes essential.

Combining studies on the action and consequences of molecular adaptation at different evolutionary timescales with genomic re-sequencing studies of extant populations is a very powerful approach to improve our understanding of the genomic basis underlying molecular adaptation in response to environmental variation and that give rise to biological diversification. Comparing genomic and transcriptomic data from species of taxonomic groups differing in one or several phenotypes, such as different lineages of Panarthropoda, is a very powerful tool to identify candidate targets driving phenotypic adaptation.

In this context, the evolution of gene content is especially interesting, since the gain and loss of genetic elements is recognized as one of the most important mechanisms for the generation of new phenotypic diversity. Indeed, the functional relevance of the long-term turnover rate of multigenic families related to key biological processes, such as chemosensory-related gene families, is a very attractive case study[54,145,165–167]. On the other hand, the study of recent, or even

ongoing, adaptive processes driven by mutations in one or a few genes (see Hof et al.[168]) provides new knowledge about the contribution of population demography and non-deterministic forces in the adaptive evolution of traits during very recent rapid diversifications. One very illustrative example of these scenarios is the extraordinary insular adaptive radiation of the *Dysdera* genus[155,169–171].

All these studies are now possible thanks to the development of high throughput sequencing (HTS) technologies. Advances in these technologies are continuously improving the amount and quality of the genomic data available, which has had a direct impact on evolutionary genomics, a field that has undergone decisive changes during the last decades. Our field has moved from the comparative and population genetics of one or a few specific loci to the recent comprehensive analyses of large multigene families and entire genomes using chromosome-level assemblies or re-sequencing studies (population genomics) including many individuals from the same species. In this sense, whole genome sequencing particularly experienced a major change in a very short time interval. Just a few years ago, the reconstruction of the complete sequence of a genome was an expensive and time-consuming process; indeed, they require to generate several libraries with different fragment lengths (e.g. short paired-end or mate pair reads and long reads from PacBio or Nanopore platforms) to obtain only moderately continuous assemblies. Moreover, these tasks needed the use of specific, non-user-friendly software to perform the assembling process, often a difficult task in studies on non-model organisms. Currently, nevertheless, the proliferation of standardized user-friendly assembling pipelines, with the development of promising scaffolding methodologies based on proximity ligation, makes it possible to obtain highly continuous, chromosome-level assemblies for almost any organism (including non-model ones) at reasonable sequencing costs and in a realistic time period. Thanks to these developments, we now have access to thousands of high-quality genomes and transcriptomes with an increasingly better taxonomic distribution across the Tree of Life[172].

In this thesis, we have benefited from recent methodological and technical advances to generate new sequencing data, and high-quality genome assemblies (high continuity) from a non-model organism. We have used this high-quality data to gain insights into the relative role of molecular adaptation in species diversification at different genomic and time scales in Panarthropoda, but also in the study of the impressive island radiation undergone by *Dysdera*, a spider genus from one of the major Arthropoda phylum, the Chelicerata.

1 Improving the genome draft of a non-model organism and expanding the study of chemosensory gene families to Panarthropoda

The chemosensory system (CS) of animals has a critical role, as it is necessary to detect food, prey and predators and perform a key function in the recognition of mates[173,174]. To detect and process the small molecules that provide information about the external environment, the organisms have a series of proteins with specialized functions. Interestingly, in arthropods these proteins are encoded by medium to large sized multigene families including hundreds or even thousands of members. The early diversification of major Panarthropoda lineages occurred long before the colonization of land (or terrestralization) of any descendant lineage[135] (Figure 6); accordingly, we could expect that the CS undergone some important adaptive changes during the out of water period; moreover, since the out of water occur independently in four major phyla of Panarthropoda, we could also expect that these changes occurred independently in each major lineage. Previous studies to that thesis have shown that most chemosensory-related families were already present in the aquatic ancestor (all except the insect olfactory receptor family) and that they are highly dynamic in terms of repertory sizes across Arthropoda[48,54,56]. However, there are many questions to be answered.

The recent increase in the number of high-quality genomes and transcriptomes available to non-model organisms, jointly with the affordable costs of HTS and our experience with the bioinformatics tools needed to perform the involved tasks, made it possible here to extend to the Panarthropoda clade the previous work on Arthropoda chemosensory gene families. We have annotated and performed a comparative genomic analysis of the chemosensory repertoire in several water bears and velvet worm species. Furthermore, we finished the first version (draft genome) of the spider *Dysdera silvatica*[175], to further upgrade to a chromosome-level assembly[145]. This later part is very important since assembly continuity is fundamental for the correct structural annotation of genomic features. This premise is especially true for gene families or repetitive elements, whose identification and complete annotation is seriously compromised in highly fragmented genomes. To study the origin and evolution of chemosensory gene family copies, it was essential to fully identify, correctly annotate and know the genomic distribution of all copies of the family, something only affordable thanks to the new upgraded chromosome-level assembly generated as a part of this thesis.

1.1. A new, chromosome-scale genome draft assembly for *D. silvatica*

To make robust evolutionary inferences based on genomic data it is imperative to have a good-quality reference assembly of our studied organism or, otherwise, from a closely related species. The availability of such reference genomes has greatly increased during the last decade due to the development of new sequencing and assembling/scaffolding technologies, as well as an overall reduction in costs[176]. Indeed, currently, there are several international collaborative projects with the goal of encouraging genome sequencing in different groups of organisms. The Earth BioGenome Project (EBP), for example, aims to sequence, catalog and characterize all the eukaryotic genomes on the Earth in a period of ten years from 2018[177]. Other affiliated or related projects (some of them more specific and started long before the of the EBP), such as the ik5 (Sequencing Five Thousand Arthropod Genome)[178] or the B10K (The Bird 10,000 Genomes)[179] pursue to collect the genome biodiversity in more restricted regions, such as the Catalan Biogenome Project (CBP)[180].

In this thesis, we have contributed to the first draft genome of *D. silvatica* (v1.2), which was posteriorly upgraded to a chromosome-level assembly (v2.3); this is a valuable resource to study the genome structure and its functional relevance of spiders, as well as to conduct further comparative and population genomic analyses in this group. We made a comprehensive structural and functional annotation, including some multigene families and repetitive elements.

The initial draft assembly (v1.2) was obtained after a hybrid assembly of short (Illumina) and long (PacBio and Nanopore) sequencing reads. The inclusion of the libraries based on long reads improved the first short-read-based assembly, especially the Nanopore reads, although the N50 of the assembly continued to be clearly insufficient ~38kb (Table 1). BUSCO statistics were also poor, confirming that we had a highly fragmented genome. The final assembly size was 1.36 Gb, representing 80% of the estimated genome size calculated using flow cytometry (1.7 Gb), and half of the assembly is composed of repetitive elements (53.8%).

The upgraded assembly (v2.3) based on the scaffolding step generating Chicago and Hi-C libraries (and using the Hi-Rise bioinformatic assembly pipeline), increased the genome continuity more than 4,500 times with respect to the previous version. This improved version is reflected in the N50 (174.2 Mb) statistic, and BUSCO results (>80% of genes were complete) (Table 1). Noticeably, the seven largest scaffolds in this assembly show an almost exact correspondence with the seven chromosomes described in the karyotype of this species (M. A. Arnedo, unpublished results), being the largest of them, with 318 Mb. This

result represents a remarkable advance for spider genomics, and a very relevant resource for studies requiring the information of the physical location of loci along chromosomes (see below).

Another important difference between the two versions of the genome is the use of BRAKER pipeline for the structural annotation in v2.3 (33,275 protein coding sequences) instead of MAKER (used in the v1.2 jointly with some in-house scripts). This pipeline searches for hints (putative gene regions), assigns these hints with a confidence score and makes several training rounds to refine the final annotation set. The main advantage of this workflow is the integration of all the steps in just one, being much more efficient than the previous approach. The lower number of gene evidence features found in v2.3 clearly reflects the lower fragmentation of the new assembly, which contains less but more complete gene models, as we can observe from the BUSCO results. About 87% of gene annotations had some functional evidence from Swiss-Prot, ArthropodDB or GO ontology codes. Finally, the functional annotation of v2.3 was more curated, by adding more databases to searches.

Table 1. Comparative of the main assembly quality statistics between the first draft (v1.2) and the chromosome-level (v2.3) assemblies of *D. silvatica*.

	v1.2	v2.3
Genome structure		
Number of scaffolds	65,205	15,360
Longest scaffold (Mb)	0.340	317.95
Scaffold N50 (Mb)	0.038	174.19
Genome annotation		
Protein-coding genes	48,619	33,275
Functionally annotated	36,398 (74.9%)	28,904 (86.9%)
BUSCO analysis		
Arachnida set. Complete genes n = 2,934	1,930 (65.8%)	2,532 (86.3%)
Arthropoda set. Complete genes n = 1,013	689 (70.5%)	818 (80.8%)

The contribution of this resource is still more relevant if we consider that very few genomes of Arachnida have been sequenced at the chromosome-level. The first assemblies of a similar quality were published in 2021[143,144], just a few months earlier than that of *D. silvatica*[145]. The rest of chromosome-level assemblies available today have been published very recently, during this year[146,148,181], demonstrating that the subject of this thesis is a cutting edge within spider genomics (Table 2). Although the total number of scaffolds was comparatively high in the first version of the *D. silvatica* genome, it was clearly improved in the second one. In fact, 87% of the total genome size is included in the seven largest scaffolds (i.e., chromosomes), which makes the *D. silvatica* v2.3 assembly possibly one of the best spider genomes published to date.

Table 2. Statistics of the chromosome-level assemblies currently available for the class Arachnida. Data obtained in September 2022.

Species	Assembly size (Gb)	Scaffold N50 (Mb)	Contig N50 (kb)	Number of scaffolds
Argiope bruennichi	1.67	124.24	284.77	2,231
Trichonephila antipodiana	2.29	173	1,138	377
Dysdera silvatica	1.36	174.19	21.95	15,360
Hylyphantes graminicola	0.93	77,07	889.59	103
Latrodectus elegans	1.57	114.31	4,340	164
Uloborus diversus	2.15	185.51	452.78	1,586

Furthermore, within the scope of this thesis, the generation of a chromosome-level assembly of *D. silvatica* had even more importance since i) it has been the reference sequence in the first comprehensive population genomics study of a spider in the context of the adaptive radiation of the genus *Dysdera* in the Canary Islands, and ii) it has been the backbone to characterize with an unprecedented detail, localizing the genes in the chromosomes, the genes encoding the chemosensory receptors in the genome of an arachnid.

1.2. Evolution of chemosensory-related gene families in Panarthropoda

We performed an evolutionary genomics study at a deep evolutionary scale performing some comparative genomics and transcriptomics analysis of the chemosensory-related gene families in several Panarthropoda species, onychophorans and tardigrades. We generated the first antennal and head specific transcriptomes of an onychophoran species, *E. rowelli* (these tissues include the chemosensory structures in this species[95]). We used the available whole transcriptomes of this and other velvet worms and the genomes of the tardigrades *H. exemplaris* and *R. varieornatus* to identify and characterize the genes and the transcripts encoding the putative chemoreceptors and soluble proteins associated with chemosensory organs in these species. We also searched for these genes in the reported genome draft of *E. rowelli*; this assembly, however, is very incomplete and of low quality. Therefore, our organ-specific transcriptome newly generated in this thesis becomes a valuable resource to study chemosensory-related gene families in onychophorans, representing the first, mostly complete onychophoran reference chemosensory-related gene set to date.

We identified transcripts for the two main arthropod chemoreceptor families IRs/iGluRs and GRs, as well as for other receptors and proteins involved in the chemoreception, such as DEG-ENaC, CD36-SNMP and NPC2 families, expressed in the antenna of *E. rowelli*. Interestingly, some of these transcripts were specifically or differentially expressed in the head or the antenna, suggesting their chemosensory role. The most relevant findings are that the number of GRs transcripts found in this species is by far the lowest found in Panarthropoda, and that, surprisingly, we fail to find an homolog of the highly conserved co-receptor IR25a, which casts doubt on the olfactory role of the IR expressed in the antenna of this group (IR25a is present and highly conserved in all protostomes surveyed so far[182]). Although we cannot rule out the possibility that another member of the IR family is acting as a co-receptor in these organisms, these results open the possibility that they adapted to the new terrestrial environment by the co-option of other receptors to smell (e.g., DEG-ENaC, GPCR family 3, TRP channels, among others; all of them with members expressed in the antenna of *E. rowelli*). (Figure 10a). Conversely, in tardigrades, we didn't find any genomic region candidate to encode neither complete nor partial DEG-ENaC and CD36/SNMPs sequences. We also performed searches of DEG-ENaC in six transcriptomes of tardigrades from Heterotardigrada class, and found it was missing only in Eutardigrada (Figure 10b). In addition to the possible chemosensory function of some members of these families, these proteins have also been implicated in processes related to the maintenance of osmotic and intestinal stem cell homeostasis and resistance to

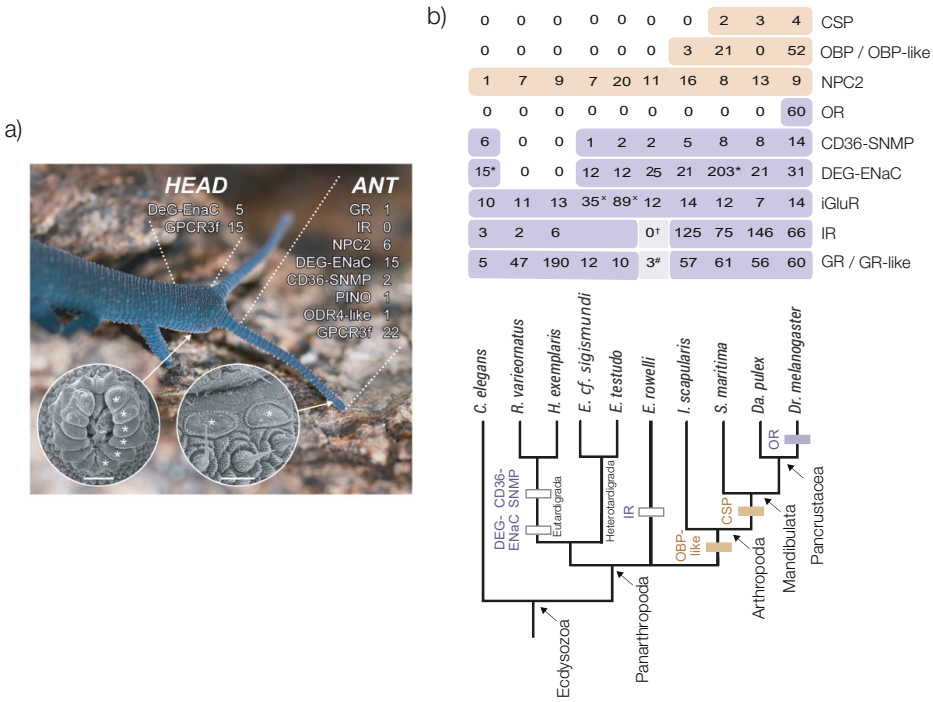


Figure 10. Chemosensory structures in the onychophoran *E. rowelli* and summary of CS repertoire in each tissue and across Panarthropoda. a) Anterior end of an individual with anatomical compartments indicated by dotted lines. Numbers refer to those genes specifically or differentially expressed in antenna (ANT) and head (HEAD). b) Minimum estimates of gene family sizes in the genomes from nine major Ecdysozoan lineages. Colored boxes indicate gains and losses of gene families, respectively. Purple and light-brown shadings means membrane receptors and soluble proteins, respectively. †Three very short sequences encoding parts of the iGluR/IR LCD, although they are phylogenetically related to IRs, could not be unambiguously assigned to this subfamily. #One complete GR receptor and two sequences resulting from partial BLAST hits. xRepertoire of iGluRs and IR are not separated. Modified from Vizueta et al.[95]

desiccation[95,183,184], suggesting that the loss of DEG-ENaC family could be part of a lineage-specific adaptation of eutardigrades to survive in extreme environments.

Overall, our deep scale evolutionary analysis of the evolution of chemosensory-related gene families in Panarthropoda showed a very high gene turnover (birth and death) rates in the GR-like proteins in tardigrades and arachnids, while onychophorans appear to have lost most of their members. Our analyses also confirm that the GR family is the oldest receptor family known in Panarthropoda and was probably co-opted from an ancient receptor with a non-chemosensory function already present in the ancestors of animals[185]. On the other hand, the IR family is also a conserved family of invertebrate receptors, probably originating from ionotropic glutamate receptors present in the ancestor of

Protostomes[182,186]. Although it is still under debate, the results of this thesis (the presence of GR in the antennae of onychophorans, and the large repertory size of this family found in tardigrades with respect to Nematodes), would point to the co-option of the chemosensory function of these receptors in the ancestor of Panarthropoda. Conversely, the phylogenetic analysis of the members of the IR and iGluR families obtained in this tesis[95,145], support a monophyletic origin of divergent (chemosensory) IRs, i.e, they originated from ionotropic glutamate receptors (iGluR) in a single evolutionary event, at some time very early in the diversification of Protostomes. The origin of coreceptor IR25a, on the other hand, is much more doubtful, since its phylogenetic position, perhaps influenced by its different rate of evolution, seems to point to an independent (earlier) origin than the divergent IR family.

Another important contribution of this thesis is the finding that *D. silvatica* has the largest repertoire of IR found in Arthropoda (see also below), except for the cockroach *Blattella germanica*[187]. By contrast, the GR genome of this spider is relatively low in comparison to other arachnids[54]. As we have seen in previous studies in our research group, the repertoire of these two receptor families displays a huge variation between lineages across Arthropoda, despite showing very similar birth and death rates in the different groups[54,56,95,165,182]. This large variation in the number of copies likely reflects the long-term effect of such high turnover rates, which even being similar and constant across lineages, are still mainly stochastic. We have detected, however, some lineage-specific expansions and contractions that would be associated with particular (and complex) ecological aspects related to important diet or habitat changes[187,188].

Regarding soluble proteins, the only candidates found in onychophorans and tardigrades are the NPC2, with some of its members specifically or differentially expressed in the *E. rowelli* antenna. It is possible then that some NPC2 copies are performing a similar role to the Arthropod OBP/OBP-like proteins or CSP families. We found genes encoding the OBP-like family in *D. silvatica* and other chelicerates, so the origin of this family can be traced back to the ancestor of arthropods. By contrast, the absence of the CSP family in the genomes of chelicerates would suggest that these soluble proteins are specific to Mandibulata (Hexapoda and Myriapoda)[166].

This thesis sheds light in our understanding of the origin and evolution of the CS in Panarthropoda, however, it is necessary to consider that the phylogeny of this group is still in discussion. Indeed, if the true topology was different than that considered here (although we used the most supported to date), our evolutionary

inference could be inaccurate. Resolving the phylogenetic relationships among Panarthropoda lineages is crucial for correctly interpreting our results. Furthermore, having a better quality reference genome of *E. rowelli* would help to identify, annotate and curate the members of the CS families in onychophorans, which is crucial to confirming the loss of the IR family and the reduction of the GR repertory size in this species.

1.3. The chemosensory system in *D. silvatica* genome

In this thesis, we have also identified the main chemosensory-related genes encoded in the genome of *D. silvatica*. We characterized the members of the arthropod GR and IR receptor families, using sequence similarity-based searches as in Vizueta et al.[56]. As expected, our annotations based on the chromosome-scale genome assembly of *D. silvatica* largely improves those published in this previous transcriptome-based work. Despite being frequently clustered in tandem in the same genomic region, which complicates the identification of many copies and their structural annotation, most chemoreceptor genes seemed correctly annotated and complete in our analysis. This result not only upgrades previous annotations in this species (see also Cerca et al.[165]) but also denotes a remarkable improvement with respect to the annotations of the members of these families in other published arthropod genomes, where fragmentation precluded the accurate reconstruction of these puzzling clusters.

We found that the X chromosome of *D. silvatica*, which is far the longest chromosome (representing about ~23% of the assembled genome) encodes the fewest number of chemoreceptors relative to its size (only ~5%). Chemoreceptor genes are distributed uniformly across all autosomes with the highest abundance being found in smaller scaffolds. This last observation could be explained by the presence of many chemosensory receptors in dense genome clusters (and likely recent); this tandem repeating structure causes many problems when connecting contigs in scaffolds and tends to generate non-jointly scaffolds that include a complete or partially assembled cluster of genes of the same family. Since our chromosome-scale scaffolding of the *D. silvatica* genome was assembled using essentially short reads, despite the high continuity, we already have many chemoreceptor genes located in the minor scaffolds. Obviously, performing the initial assembly using long reads (and high coverage) will be of great interest to perform a very good analysis of the distribution of these genes in the future. The use of long reads could help the correct scaffolding of gene family clusters but also of other repetitive regions across the genome.

In our analysis we have detected a large number of genomic clusters (83 clusters, 17 and 66 of them including Gr and IR genes, respectively), which can harbor up to 10 copies of the same family. As we would expect if the origin of new copies is mainly caused by unequal crossing over, the evolutionary (genetic) distances between chemoreceptors in the chromosome of *D. silvatica* will increase with their physical distance. Gene clusters usually end up being broken and gene copies are dispersed across the genome, mainly by translocations[189], or by inversions. This could explain the fact of finding pairs of genes that have diverged little despite being very distant in the genome. Also, it is possible the existence of gene copies originated in an old duplication event but remaining clustered, maintained by some functional or gene regulation constraint (e.g., Hox genes[190]; see also Librado and Rozas[191]). After gene duplication, new segregating copies can be lost by genetic drift or purifying selection or become fixed by change or by positive selection. Once fixed in the population, they can be also eliminated due to the accumulation of negative or neutral mutations (purifying selection or pseudogenization). Conversely, these new copies can also be retained due to other mechanisms, such as subfunctionalization, dose compensation or neofunctionalization, processes that could also include post-duplication positive selected mutations[192]. Some of these mechanisms may also result in deviations from the expected relationship between physical distances and evolutionary distances, within and between clusters, and thus be responsible for the exceptions discussed above. More analyses are required to uncover the relative role of these mechanisms in the chromosomal organization and evolution of chemoreceptors in spiders, and by extension, in arthropods.

2 The adaptive radiation of *Dysdera* in the Canary Islands. Analysis of the genomic variation patterns of *D. silvatica*

Dysdera spiders suffered one of the most interesting adaptive radiation in arachnids[153,170,171,193], generating ~60 species endemic in the Canary Islands[171]. One of the main drivers of this radiation is oniscophagy, with the rapid and, even parallel, diversification of diet preferences to feed on woodlice in different islands. The spiders that feed preferably on these isopods often present similar modifications of chelicerae and capture behaviors[193]. The patterns of phenotypic convergence observed in the Canary Islands makes *Dysdera* a suitable model to study the genomics basis of adaptive radiations.

The availability of the chromosome-level reference genome of *D. silvatica*, also generated in this thesis, has allowed us to perform the first population genomics

analysis at a chromosome-scale in a spider. In this thesis, we have sequenced the genome of 12 individuals of a natural population of *D. silvatica* from La Gomera. We have quantified the levels of polymorphism, divergence and linkage disequilibrium across chromosomes. We also applied a set of neutrality tests and a genome scan for positive selection and inferred the demographic history of the population, using a coalescent-based approach. We found that *D. silvatica* presents an exceptionally high levels of nucleotide polymorphism ($\pi_{Autosomes}$ = 0.015, π_X = 0.008), a very unexpected result if we consider that this species is endemic and restricted to three very recent and small islands (Gomera, El Hierro and La Palma), and that the sample was collected in just a single locality. These levels of polymorphism have been only observed in broadly distributed species with very large effective population sizes (Ne)[194]. Based on our estimated neutral mutation rate ($\mu = 10^{-8}$), and assuming a constant size population, we obtain a Ne of 375,000.

It is known from mitochondrial phylogenies that for a particular *Dysdera* species, populations from different islands are genetically differentiated (M. A. Arnedo, unpublished results). Moreover, there is also evidence of intra-island local population structure within species (e.g., in populations of *D. verneaui*) likely due to geographical isolation in the past[195]. *D. silvatica* is present in La Palma and El Hierro in addition to La Gomera, so the high levels of nucleotide polymorphism observed in this species could be explained by a recent the admixture of genetic information from already structured populations from different islands, due to secondary contacts or colonization events. In fact, our demographic inference analysis suggests a very large historical population size in *D. silvatica* maintained for more than ~200 Kya until a few thousands of years ago. Further analyses adding data from other populations of this species, including samples from the same and different islands, are required to confirm this hypothesis.

We searched for the signals of recent adaptation in the population of *D. silvatica* or in the divergence of this species and *D. bandamae*. To study the recent (or ongoing) processes we performed a genome scan of selection to search for the molecular hallmark of putative selected sweeps. To assess older selective events, we applied the MK test. On one hand, we found no support for the "faster X" hypothesis[196] as the proportion of protein-coding genes with evidence of positive selection (by the MK test) in the X chromosome is not substantially higher than that in autosomes. Nevertheless, this result is not conclusive since the number of genes and substitutions used for the MK test was only a very small representation of the set of genes annotated for this species. To increase the power of the analysis we would need to perform the MK test using a more distant outgroup to compute

synonymous and nonsynonymous substitutions. On the other hand, the results of the RAiSD analysis identified 19 regions candidates to have suffered a recent selective sweep. Remarkably, chromosome 1 harbors some of the regions with stronger hallmarks, with genes associated to lipid biosynthetic processes, activation of appetite and protein binding. Positively selected genes predicted by the MK test are enriched in functions such as "protein binding" (a function also enriched in RAiSD candidates) and "cellular localization". Noticeably, GO enrichment results from both RAiDS and MKT demonstrate that in this species, recent molecular adaptation is partially associated with dietary diversification; indeed, some of the target functions identified in our thesis were also overrepresented in the comparative transcriptomic analysis of Vizueta et al.[197].

These results indicate that oniscophagy has been an important driver of adaptation in the species *D. silvatica*, and therefore it is still an active driver of diversification in the genus *Dysdera*. Moreover, one of the most enriched GO among the "selective sweep" candidates is related to rhodopsin activity. Rhodopsin is very sensitive to light and allows vision in dark conditions[198]. The spiders from the genus *Dysdera* are nocturnal hunters with a very limited vision (without the anterior median eyes)[199]. It is conceivable, therefore, that positive selection could shape genes involved in rhodopsin biosynthesis to facilitate dark adaptation[200]. It is also worth noting that "response to chemical stimulus" is also enriched among candidates, pointing to the chemosensory system as another important target of recent selection.

Conclusions

1 We generated the first assembly of the Canary Island endemic spider *D. silvatica*, the first available genome of a species from the Dysderidae family.

2 We upgraded the genome of *D. silvatica* to a chromosome-level assembly using Chicago and Hi-C libraries. We obtained a highly continuous genome that constitutes one of the most quality genomic resources available for spiders.

3 Our comparative transcriptomics analysis across tissues allows identifying the members of arthropod chemosensory gene families expressed in the antenna of *E. rowelli*. Remarkably, the expression of the IR25a receptor, a co-receptor that is highly conserved across protostomes, was not detected in velvet worms.

4 The results of the RNAseq study in *E. rowelli* point to NPC2 as the only candidate family encoding chemosensory-like soluble proteins in the Panarthropoda ancestor.

5 The Eutardigrada genomes surveyed in this work do not encode the DEG-ENaC and CD36-sensory neuron membrane proteins, two families with members involved in chemosensation in other panarthropods.

6 We found that *D. silvatica* encodes 545 chemosensory receptors, 411 IRs and 134 GRs. The IR repertoire family size is one of the largest found in arthropods. Conversely, the size of the GR family is smaller than in other arachnids.

7 Chemoreceptor genes are highly clustered in the genome of *D. silvatica*, and their physical and evolutionary distances correlate across the genome. Our analysis confirms the unequal crossing over as the main mechanism for gene duplication in these families and uncovers recent bursts in gene duplication of chemoreceptors in spiders.

8 The X chromosome of *D. silvatica* encodes proportionally a smaller number of chemoreceptors than the autosomes; nevertheless, the fact that many members of these families are in small scaffolds precludes making a firm conclusion about this uneven distribution.

9 The surveyed population of *D. silvatica* presents exceptional high levels of nucleotide polymorphism, and a rapid decay of linkage disequilibrium with distance.

10 The high levels of variation found in *D. silvatica* would have been caused by a long period in the past with a large effective population size, likely due to some ancestral population structure.

Bibliography

1. Cutter, A. D. *A Primer of Molecular Population Genetics*. (Oxford University Press, 2019).

2. Nicholas H. Barton, Derek E. G. Briggs, Jonathan A. Eisen, David B. Goldstein & Nipam H. Patel. *Evolution*. (Cold Spring Harbor Laboratory Press, 2007).

3. Slatko, B. E., Gardner, A. F. & Ausubel, F. M. Overview of Next-Generation Sequencing Technologies. *Curr. Protoc. Mol. Biol.* **122**, e59 (2018).

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

5. Calvo Martín, J. M. *Evolución molecular de los genes del grupo Polycomb en el género Drosophila*. (Universitat de Barcelona, 2017).

6. Ellegren, H. Genome sequencing and population genomics in non-model organisms. *Trends in Ecology and Evolution* vol. 29 51–63 (2014).

7. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10**, 57 (2009).

8. Begun, D. J. *et al.* Population Genomics: Whole-Genome Analysis of Polymorphism and Divergence in Drosophila simulans. *PLOS Biol.* **5**, e310 (2007).

9. Rundell, R. J. & Price, T. D. Adaptive radiation, nonadaptive radiation, ecological speciation and nonecological speciation. *Trends Ecol. Evol.* **24**, 394–399 (2009).

10. Hodges, S. A. & Derieg, N. J. Adaptive radiations: From field to genomic studies. in *In the Light of Evolution* vol. 3 27–45 (National Academies Press (US), 2009).

11. Wilson, E. O. . M. & H., R. The Theory of Island Biogeography. *Nature* **413**, 1 (1988).

12. Whittaker, R. J. & Martínez, J. M. F.-P. Island Biogeography: Ecology, Evolution, and Conservation. (Oxford University Press, 2007).

13. Wolf, J. B. W. & Ellegren, H. Making sense of genomic islands of differentiation in light of speciation. *Nat. Rev. Genet.* **18**, 87–100 (2017).

14. Grant, P. R. & Grant, B. R. *How and Why Species Multiply*. *How and Why Species Multiply* (Princeton University Press, 2020).

15. Losos, J. B. *Lizards in an Evolutionary Tree: Ecology and Adaptive Radiation of Anoles*. Vol. 10. (University of California Press, 2011).

16. Almén, M. S. *et al.* Adaptive radiation of Darwin's finches revisited using whole genome sequencing. *BioEssays* **38**, 14–20 (2016).

17. Muschick, M., Indermaur, A. & Salzburger, W. Convergent Evolution within an Adaptive Radiation of Cichlid Fishes. *Curr. Biol.* **22**, 2362–2368 (2012).

18. Gillespie, R. G. & Roderick, G. K. Arthropods on islands: Colonization, speciation, and conservation. *Annual Review of Entomology* vol. 47 595–632 (2002).

19. Zallen, D. T. Despite Franklin's work, Wilkins earned his Nobel. *Nature* vol. 425 15 (2003).

20. Watson, J. D. & Crick, F. H. C. Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. *Nature* **171**, 737–738 (1953).

21. Heather, J. M. & Chain, B. The sequence of sequencers: The history of sequencing DNA. *Genomics* vol. 107 1–8 (2016).

22. Sanger, F. & Coulson, A. R. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J. Mol. Biol.* **94**, 441–448 (1975).

23. Maxam, A. M. & Gilbert, W. A new method for sequencing DNA. *Proc. Natl. Acad. Sci. U. S. A.* **74**, 560–564 (1977).

24. Cohen, S. N., Chang, A. C. Y., Boyer, H. W. & Helling, R. B. Construction of biologically functional bacterial plasmids in vitro. *Proc. Natl. Acad. Sci. U. S. A.* **70**, 3240–3244 (1973).

25. Saiki, R. K. *et al.* Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science.* **239**, 487–491 (1988).

26. Smith, L. M. *et al.* Fluorescence detection in automated DNA sequence analysis. *Nature* **321**, 674–679 (1986).

27. Nyrén, P. Enzymatic method for continuous monitoring of DNA polymerase activity. *Anal. Biochem.* **167**, 235–238 (1987).

28. Ronaghi, M., Karamohamed, S., Pettersson, B., Uhlén, M. & Nyrén, P. Real-time DNA sequencing using detection of pyrophosphate release. *Anal. Biochem.* **242**, 84–89 (1996).

29. Ronaghi, M., Uhlén, M. & Nyrén, P. A sequencing method based on real-time pyrophosphate. *Science* vol. 281 363–365 (1998).

30. Gužvić, M. The history of DNA sequencing. *J. Med. Biochem.* **32**, 301–312 (2013).

31. Schadt, E. E., Turner, S. & Kasarskis, A. A window into third-generation sequencing. *Hum. Mol. Genet.* **19**, (2010).

32. Branton, D. *et al.* The potential and challenges of nanopore sequencing. *Nat. Biotechnol. 2008 2610* **26**, 1146–1153 (2008).

33. Bleidorn, C. Third generation sequencing: technology and its potential impact on evolutionary biodiversity research. *Syst. Biodivers.* **14**, 1–8 (2016).

34. Eisenstein, M. Oxford Nanopore announcement sets sequencing sector abuzz. *Nature Biotechnology* vol. 30 295–296 (2012).

35. Clarke, J. *et al.* Continuous base identification for single-molecule nanopore DNA sequencing. *Nat. Nanotechnol.* **4**, 265–270 (2009).

36. Mao, Y. & Zhang, G. A complete, telomere-to-telomere human genome sequence presents new opportunities for evolutionary genomics. *Nature Methods* vol. 19 635–638 (2022).

37. Miga, K. H. *et al.* Telomere-to-telomere assembly of a complete human X chromosome. *Nat. 2020 5857823* **585**, 79–84 (2020).

38. Hakim, O. & Misteli, T. SnapShot: Chromosome conformation capture. *Cell* vol. 148 1068.e1 (2012).

39. Belton, J. M. *et al.* Hi-C: A comprehensive technique to capture the conformation of genomes. *Methods* **58**, 268–276 (2012).

40. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science (80).* **326**, 289–293 (2009).

41. Eagen, K. P. Principles of Chromosome Architecture Revealed by Hi-C. *Trends in Biochemical Sciences* vol. 43 469–478 (2018).

42. Putnam, N. H. *et al.* Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. *Genome Res.* **26**, 342–350 (2016).

43. Allendorf, F. W., Hohenlohe, P. A. & Luikart, G. Genomics and the future of conservation genetics. *Nature Reviews Genetics* vol. 11 697–709 (2010).

44. Holliday, J. A., Hallerman, E. M. & Haak, D. C. Genotyping and Sequencing Technologies in Population Genetics and Genomics. in 83–125 (Springer, Cham, 2018).

45. Lou, R. N., Jacobs, A., Wilder, A. P. & Therkildsen, N. O. A beginner's guide to low-coverage whole genome sequencing for population genomics. in *Molecular Ecology* vol. 30 5966–5993 (John Wiley & Sons, Ltd, 2021).

46. Korneliussen, T. S., Albrechtsen, A. & Nielsen, R. ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinformatics* **15**, 1–13 (2014).

47. Kaupp, U. B. Olfactory signalling in vertebrates and insects: differences and commonalities. *Nat. Rev. Neurosci.* **11**, 188–200 (2010).

48. Sánchez-Gracia, A., Vieira, F. G., Almeida, F. C. & Rozas, J. Comparative Genomics of the Major Chemosensory Gene Families in Arthropods. *Encycl. Life Sci.* (2011).

49. Vieira, F. G. & Rozas, J. 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 (2011).

50. Vieira, F. G., Sánchez-Gracia, A. & Rozas, J. Comparative genomic analysis of the odorant-binding protein family in 12 *Drosophila* genomes: Purifying selection and birth-and-death evolution. *Genome Biol.* **8**, R235 (2007).

51. Nei, M. & Rooney, A. P. Concerted and birth-and-death evolution of multigene families. *Annu. Rev. Genet.* **39**, 121–52 (2005).

52. Librado, P., Vieira, F. G. & Rozas, J. BadiRate: Estimating family turnover rates by likelihood-based methods. *Bioinformatics* **28**, 279–281 (2012).

53. Sánchez-Gracia, A., Vieira, F. G. & Rozas, J. Molecular evolution of the major chemosensory gene families in insects. *Heredity (Edinb).* (2009)

54. 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.* **10**, 1221–1236 (2018).

55. Croset, V. *et al.* Ancient Protostome Origin of Chemosensory Ionotropic Glutamate Receptors and the Evolution of Insect Taste and Olfaction. *PLOS Genet.* **6**, e1001064 (2010).

56. Vizueta, J. *et al.* Evolution of Chemosensory Gene Families in Arthropods: Insight from the First Inclusive Comparative Transcriptome Analysis across Spider Appendages. *Genome Biol. Evol.* **9**, 178 (2017).

57. Nei, M. & Rooney, A. P. Concerted and Birth-and-Death Evolution of Multigene Families. *Annu. Rev. Genet.* **39**, 121 (2005).

58. Conrad, B. & Antonarakis, S. E. Gene duplication: a drive for phenotypic diversity and cause of human disease. *Annu. Rev. Genomics Hum. Genet.* **8**, 17–35 (2007).

59. Vogt, R. G. & Riddiford, L. M. Pheromone binding and inactivation by moth antennae. *Nature* vol. 293 161–163 (1981).

60. Leal, W. S., Nikonova, L. & Peng, G. Disulfide structure of the pheromone binding protein from the silkworm moth, *Bombyx mori. FEBS Lett.* **464**, 85–90 (1999).

61. Scaloni, A., Monti, M., Angeli, S. & Pelosi, P. Structural analysis and disulfide-bridge pairing of two odorant-binding proteins from Bombyx mori. *Biochem. Biophys. Res. Commun.* **266**, 386–391 (1999).

62. Kaissling, K. E. Olfactory perireceptor and receptor events in moths: A kinetic model revised. *J. Comp. Physiol. A Neuroethol. Sensory, Neural, Behav. Physiol.* **195**, 895–922 (2009).

63. Leal, W. S. *et al.* Kinetics and molecular properties of pheromone binding and release. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 5386–5391 (2005).

64. Pelosi, P. & Maida, R. Odorant-binding proteins in vertebrates and insects: Similarities and possible common function. *Chem. Senses* **15**, 205–215 (1990).

65. Nomura Kitabayashi, A., Arai, T., Kubo, T. & Natori, S. Molecular cloning of cDNA for p10, a novel protein that increases in the regenerating legs of *Periplaneta americana* (American cockroach). *Insect Biochem. Mol. Biol.* **28**, 785–790 (1998).

66. Li, S. *et al.* Multiple functions of an odorant-binding protein in the mosquito *Aedes aegypti*. *Biochem. Biophys. Res. Commun.* **372**, 464–468 (2008).

67. Pelosi, P., Calvello, M. & Ban, L. Diversity of odorant-binding proteins and chemosensory proteins in insects. *Chem. Senses* **30**, i291–i292 (2005).

68. Zhou, J. J., Kan, Y., Antoniw, J., Pickett, J. A. & Field, L. M. Genome and EST analyses and expression of a gene family with putative functions in insect chemoreception. *Chem. Senses* **31**, 453–465 (2006).

69. Vieira, F. G. & Rozas, J. 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 (2011).

70. Renthal, R. *et al.* The chemosensory appendage proteome of *Amblyomma americanum* (Acari: Ixodidae) reveals putative odorant-binding and other chemoreception-related proteins. *Insect Sci.* **24**, 730–742 (2017).

71. Vizueta, J., Macías-Hernández, N., Arnedo, M. A., Rozas, J. & Sánchez-Gracia, A. Chance and predictability in evolution: The genomic basis of convergent dietary specializations in an adaptive radiation. *Mol. Ecol.* **28**, 4028–4045 (2019).

72. Pelosi, P., Iovinella, I., Felicioli, A. & Dani, F. R. Soluble proteins of chemical communication: An overview across arthropods. *Front. Physiol.* **5**, 320 (2014).

73. Storch, J. & Xu, Z. Niemann-Pick C2 (NPC2) and intracellular cholesterol trafficking. *Biochim. Biophys. Acta - Mol. Cell Biol. Lipids* **1791**, 671–678 (2009).

74. Vizueta Moraga, J. Genómica de la adaptación en artrópodos: estudio del sistema quimiosensorial y de la radiación del género *Dysdera* (Araneae) en Canarias. (2020)

.75. Benton, R., Sachse, S., Michnick, S. W. & Vosshall, L. B. Atypical membrane topology and heteromeric function of *Drosophila* odorant receptors in vivo. *PLoS Biol.* **4**, 240–257 (2006).

76. Sato, K. *et al.* Insect olfactory receptors are heteromeric ligand-gated ion channels. *Nature* **452**, 1002–1006 (2008).

77. Wicher, D. *et al. Drosophila* odorant receptors are both ligand-gated and cyclic-nucleotide-activated cation channels. *Nature* **452**, 1007-U10 (2008).

78. Fishilevich, E. & Vosshall, L. B. Genetic and functional subdivision of the *Drosophila* antennal lobe. *Curr. Biol.* **15**, 1548–1553 (2005).

79. Jones, W. D., Cayirlioglu, P., Kadow, I. G. & Vosshall, L. B. Two chemosensory receptors together mediate carbon dioxide detection in *Drosophila*. *Nature* **445**, 86–90 (2007).

80. Amrein, H. & Bray, S. Bitter-sweet solution in taste transduction. *Cell* vol. 112 283–284 (2003).

81. Larsson, M. C. *et al.* Or83b encodes a broadly expressed odorant receptor essential for *Drosophila* olfaction. *Neuron* **43**, 703–714 (2004).

82. Kwon, J. Y., Dahanukar, A., Weiss, L. A. & Carlson, J. R. The molecular basis of CO2 reception in *Drosophila*. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 3574–3578 (2007).

83. Robertson, H. M. Molecular evolution of the major arthropod chemoreceptor gene families. *Annu. Rev. Entomol.* **64**, 227–242 (2019).

84. Eyun, S. Il *et al.* Evolutionary History of Chemosensory-Related Gene Families across the Arthropoda. *Mol. Biol. Evol.* **34**, 1838–1862 (2017).

85. Brand, P. *et al.* The origin of the odorant receptor gene family in insects. *Elife* **7**, (2018).

86. Robertson, H. M. The Insect Chemoreceptor Superfamily Is Ancient in Animals. *Chem. Senses* **40**, 609–614 (2015).

87. Saina, M. *et al.* A cnidarian homologue of an insect gustatory receptor functions in developmental body patterning. *Nat. Commun.* **6**, (2015).

88. Wicher, D. & Miazzi, F. Functional properties of insect olfactory receptors: ionotropic receptors and odorant receptors. *Cell Tissue Res.* **383**, 7–19 (2021).

89. Kuner, T., Seeburg, P. H. & Guy, H. R. A common architecture for K+ channels and ionotropic glutamate receptors? *Trends Neurosci.* **26**, 27–32 (2003).

90. Benton, R., Vannice, K. S., Gomez-Diaz, C. & Vosshall, L. B. Variant Ionotropic Glutamate Receptors as Chemosensory Receptors in *Drosophila*. *Cell* **136**, 149–162 (2009).

91. Chen, Z., Wang, Q. & Wang, Z. The amiloride-sensitive epithelial Na+ channel PPK28 is essential for *Drosophila* gustatory water reception. *J. Neurosci.* **30**, 6247–6252 (2010).

92. Lu, B., LaMora, A., Sun, Y., Welsh, M. J. & Ben-Shahar, Y. Ppk23-dependent chemosensory functions contribute to courtship behavior in *Drosophila melanogaster*. *PLoS Genet.* **8**, (2012).

93. Rogers, M. E., Sun, M., Lerner, M. R. & Vogt, R. G. Snmp-1, a novel membrane protein of olfactory neurons of the silk moth *Antheraea polyphemus* with homology to the CD36 family of membrane proteins. *J. Biol. Chem.* **272**, 14792–14799 (1997).

94. Nichols, Z. & Vogt, R. G. 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 (2008).

95. Vizueta, J. *et al.* Evolutionary history of major chemosensory gene families across Panarthropoda. *Mol. Biol. Evol.* **37**, 3601–3615 (2020).

96. Benton, R. Sensitivity and specificity in *Drosophila* pheromone perception. *Trends Neurosci.* **30**, 512–519 (2007).

97. Benton, R., Vannice, K. S. & Vosshall, L. B. An essential role for a CD36-related receptor in pheromone detection in *Drosophila*. *Nature* **450**, 289–293 (2007).

98. Forstner, M. *et al.* Differential expression of SNMP-1 and SNMP-2 proteins in pheromone-sensitive hairs of moths. *Chem. Senses* **33**, 291–299 (2008).

99. Gomez-Diaz, C. *et al.* A CD36 ectodomain mediates insect pheromone detection via a putative tunnelling mechanism. *Nat. Commun.* **7**, 11866 (2016).

100. Møbjerg, N. *et al.* Survival in extreme environments - on the current knowledge of adaptations in tardigrades. *Acta Physiol. (Oxf).* **202**, 409–420 (2011).

101. Gabriel, W. N. *et al.* The tardigrade *Hypsibius dujardini*, a new model for studying the evolution of development. *Dev. Biol.* **312**, 545–559 (2007).

102. Müller, K. J., Walossek, D. & Zakharov, A. 'Orsten' type phosphatized soft-integument preservation and a new record from the Middle Cambrian Kuonamka Formation in Siberia. *Neues Jahrb. für Geol. und Paläontologie - Abhandlungen* **197**, 101–118 (1995).

103. Degma, P., Bertolani, R. & Guidetti, R. Actual checklist of Tardigrada species. 1–44 (2020)

104. Hashimoto, T. *et al.* Extremotolerant tardigrade genome and improved radiotolerance of human cultured cells by tardigrade-unique protein. *Nat. Commun.* **7**, 1–14 (2016).

105. Becquerel, P. La suspension de la vie au-dessous de 1/20K absolu par demagnetisation adiabatique de l'alun de fer dans le vide le plus eleve. *Comptes rendus Hebd. des seances l'Aca demie des Sci.* **231**, 261–263 (1950).

106. Hengherr, S., Worland, M. R., Reuner, A., Brümmer, F. & Schill, R. O. High-temperature tolerance in anhydrobiotic tardigrades is limited by glass transition. *Physiol. Biochem. Zool.* **82**, 749–755 (2009).

107. Horikawa, D. D. *et al.* Establishment of a rearing system of the extremotolerant tardigrade *Ramazzottius varieornatus*: A new model animal for astrobiology. *Astrobiology* **8**, 549–556 (2008).

108. Jönsson, K. I., Harms-Ringdahl, M. & Torudd, J. Radiation tolerance in the eutardigrade *Richtersius coronifer*. *Int. J. Radiat. Biol.* **81**, 649–656 (2005).

109. Horikawa, D. D. *et al.* Radiation tolerance in the tardigrade *Milnesium tardigradum*. *Int. J. Radiat. Biol.* **82**, 843–848 (2006).

110. Ono, F. *et al.* Effect of high hydrostatic pressure on to life of the tiny animal tardigrade. *J. Phys. Chem. Solids* **69**, 2297–2300 (2008).

111. Jönsson, K. I., Rabbow, E., Schill, R. O., Harms-Ringdahl, M. & Rettberg, P. Tardigrades survive exposure to space in low Earth orbit. *Curr. Biol.* **18**, (2008).

112. Mapalo, M. A. *et al.* The Unique Antimicrobial Recognition and Signaling Pathways in Tardigrades with a Comparison Across Ecdysozoa. *G3 Genes|Genomes|Genetics* **10**, 1137–1148 (2020).

113. Hering, L. & Mayer, G. Analysis of the opsin repertoire in the tardigrade *Hypsibius dujardini* provides insights into the evolution of opsin genes in Panarthropoda. *Genome Biol. Evol.* **6**, 2380–2391 (2014).

114. Delmont, T. O. & Eren, A. M. Identifying contamination with advanced visualization and analysis practices: Metagenomic approaches for eukaryotic genome assemblies. *PeerJ* **2016**, e1839 (2016).

115. Hara, Y., Shibahara, R., Kondo, K., Abe, W. & Kunieda, T. Parallel evolution of trehalose production machinery in anhydrobiotic animals via recurrent gene loss and horizontal transfer. *Open Biol.* **11**, (2021).

116. Yoshida, Y. *et al.* Comparative genomics of the tardigrades *Hypsibius dujardini* and *Ramazzottius varieornatus*. *PLOS Biol.* **15**, e2002266 (2017).

117. Boothby, T. C. *et al.* Evidence for extensive horizontal gene transfer from the draft genome of a tardigrade. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 15976–15981 (2015).

118. Koutsovoulos, G. *et al.* No evidence for extensive horizontal gene transfer in the genome of the tardigrade *Hypsibius dujardini*. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 5053–5058 (2016).

119. Jørgensen, A., Kristensen, R. M. & Møbjerg, N. Phylogeny and Integrative Taxonomy of Tardigrada. 95–114 (2018).

120. Giribet, G. & Edgecombe, G. D. *The Invertebrate Tree of Life*. *The Invertebrate Tree of Life* (2020).

121. Concha, A. *et al.* Oscillation of the velvet worm slime jet by passive hydrodynamic instability. *Nat. Commun. 2015 61* **6**, 1–6 (2015).

122. Murienne, J., Daniels, S. R., Buckley, T. R., Mayer, G. & Giribet, G. A living fossil tale of Pangaean biogeography. *Proc. R. Soc. B Biol. Sci.* **281**, (2013).

123. Ghiselin, M. T. Peripatus as a Living Fossil. in 214–217 (Springer New York, 1984).

124. Ou, Q., Shu, D. & Mayer, G. Cambrian lobopodians and extant onychophorans provide new insights into early cephalization in Panarthropoda. *Nat. Commun.* **3**, 1–7 (2012).

125. Pauli, T. *et al.* Transcriptomic data from panarthropods shed new light on the evolution of insulator binding proteins in insects. *BMC Genomics* **17**, 1–10 (2016).

126. Ortega-Hernández, J., Janssen, R. & Budd, G. E. Origin and evolution of the panarthropod head – A palaeobiological and developmental perspective. *Arthropod Structure and Development* vol. 46 354–379 (2017).

127. Petersen, M. *et al.* Diversity and evolution of the transposable element repertoire in arthropods with particular reference to insects. *BMC Evol. Biol.* **19**, 1–15 (2019).

128. Allwood, J. *et al.* Support for vicariant origins of the New Zealand Onychophora. *J. Biogeogr.* **37**, 669–681 (2010).

129. Kristensen, N. P. Phylogeny of Insect Orders. *Annu. Rev. Entomol.* **26**, 135–157 (1981).

130. Storch, V. & Ruhberg, H. Fine structure of the sensilla of Peripatopsis moseleyi (Onychophora). *Cell Tissue Res. 1977 1774* **177**, 539–553 (1977).

131. Storch, V., Ruhberg, H., Harrison, F. W., & Rice, M. E. *Microscopic anatomy of invertebrates. Onychophora, Chilopoda and Lesser Protostomata*. (1993).

132. Nielsen, C. *Animal Evolution*. (Oxford University Press, 2013).

133. Giribet, G. & Edgecombe, G. D. Current understanding of Ecdysozoa and its internal phylogenetic relationships. *Integr. Comp. Biol.* **57**, 455–466 (2017).

134. Hejnol, A. *et al.* Assessing the root of bilaterian animals with scalable phylogenomic methods. *Proc. R. Soc. B Biol. Sci.* **276**, 4261–4270 (2009).

135. Rota-Stabelli, O., Daley, A. C. & Pisani, D. Molecular timetrees reveal a cambrian colonization of land and a new scenario for ecdysozoan evolution. *Curr. Biol.* **23**, 392–398 (2013).

136. Borner, J., Rehm, P., Schill, R. O., Ebersberger, I. & Burmester, T. A transcriptome approach to ecdysozoan phylogeny. *Mol. Phylogenet. Evol.* **80**, 79–87 (2014).

137. Laumer, C. E. *et al.* Revisiting metazoan phylogeny with genomic sampling of all phyla. *Proc. R. Soc. B* **286**, (2019).

138. Zhang, Z.-Q. Animal biodiversity: An outline of higher-level classification and survey of taxonomic richness. *Zootaxa* **3148**, 3 (2011).

139. Coddington, J. A., Harvey, M. S., Prendini, L. & Walter, D. E. *Arachnida* in Assembling Tree of Life. Oxford University Press. (2014).

140. Garb, J. E., Sharma, P. P. & Ayoub, N. A. Recent progress and prospects for advancing arachnid genomics. *Curr. Opin. Insect Sci.* **25**, 51–57 (2018).

141. Sharma, P. P. *et al.* Phylogenomic interrogation of arachnida reveals systemic conflicts in phylogenetic signal. *Mol. Biol. Evol.* **31**, 2963–2984 (2014).

142. Thomas, G. W. C. *et al.* Gene Content Evolution in the Arthropods. *Genome Biol.* **21**, 1–14 (2020).

143. Fan, Z. *et al.* A chromosome-level genome of the spider *Trichonephila antipodiana* reveals the genetic basis of its polyphagy and evidence of an ancient whole-genome duplication event. *Gigascience* **10**, 1–15 (2021).

144. Sheffer, M. M. *et al.* Chromosome-level reference genome of the European wasp spider *Argiope bruennichi*: A resource for studies on range expansion and evolutionary adaptation. *Gigascience* **10**, 1–12 (2021).

145. Escuer, P. *et al.* The chromosome-scale assembly of the Canary Islands endemic spider *Dysdera silvatica* (Arachnida, Araneae) sheds light on the origin and genome structure of chemoreceptor gene families in chelicerates. *Mol. Ecol. Resour.* **22**, 375–390 (2022).

146. Miller, J., Zimin, A. V & Gordus, A. Chromosome-level genome and the identification of sex chromosomes in Uloborus diversus. *bioRxiv* 2022.06.14.495972 (2022)

147. Wang, Z. *et al.* Chromosome-level genome assembly of the black widow spider Latrodectus elegans illuminates composition and evolution of venom and silk proteins. *Gigascience* **11**, (2022).

148. Zhu, B. *et al.* Chromosomal-level genome of a sheet-web spider provides insight into the composition and evolution of venom. *Mol. Ecol. Resour.* **22**, 2333–2348 (2022).

149. Lozano-Fernandez, J. *et al.* Increasing species sampling in chelicerate genomic-scale datasets provides support for monophyly of Acari and Arachnida. *Nat. Commun. 2019 101* **10**, 1–8 (2019).

150. Ballesteros, J. A. *et al.* Comprehensive Species Sampling and Sophisticated Algorithmic Approaches Refute the Monophyly of Arachnida. *Mol. Biol. Evol.* **39**, (2022).

151. World Spider Catalog. *Online at http://wsc.nmbe.ch. Version 23.5,* Accessed on September 2022.

152. Arnedo, M. A. *et al.* The dark side of an island radiation: systematics and evolution of troglobitic spiders of the genus *Dysdera* Latreille (Araneae : Dysderidae) in the Canary Islands. *Invertebr. Syst.* **21**, 623–660 (2007).

153. Arnedo, M. A., Oromí, P. & Ribera, C. Radiation of the spider genus *Dysdera* (Araneae, Dysderidae) in the Canary Islands: Cladistic assessment based on multiple data sets. *Cladistics* **17**, 313–353 (2001).

154. Bidegaray-Batista, L., Macías-Hernández, N., Oromí, P. & Arnedo, M. A. Living on the edge: demographic and phylogeographical patterns in the woodlouse-hunter spider *Dysdera lancerotensis* Simon, 1907 on the eastern volcanic ridge of the Canary Islands. *Mol. Ecol.* **16**, 3198–3214 (2007).

155. Crespo, L. C., Silva, I., Enguídanos, A., Cardoso, P. & Arnedo, M. A. Integrative taxonomic revision of the woodlouse-hunter spider genus *Dysdera* (Araneae: Dysderidae) in the Madeira archipelago with notes on its conservation status. *Zool. J. Linn. Soc.* **192**, 356–415 (2021).

156. Macías-Hernández, N., Oromí, P. & Arnedo, M. A. Patterns of diversification on old volcanic islands as revealed by the woodlouse-hunter spider genus *Dysdera* (Araneae, Dysderidae) in the eastern Canary Islands. *Biol. J. Linn. Soc.* **94**, 589–615 (2008).

157. Macías-Hernández, N., Oromí, P. & Arnedo, M. A. Integrative taxonomy uncovers hidden species diversity in woodlouse hunter spiders (Araneae, Dysderidae) endemic to the Macaronesian archipelagos. *Syst. Biodivers.* **8**, 531–553 (2010).

158. Macías-Hernández, N., Bidegaray-Batista, L., Emerson, B. C., Oromí, P. & Arnedo, M. The Imprint of Geologic History on Within-Island Diversification of Woodlouse-Hunter Spiders (Araneae, Dysderidae) in the Canary Islands. *J. Hered.* **104**, 341–356 (2013).

159. Macías-Hernández, N., López, S. de la C., Roca-Cusachs, M., Oromí, P. & Arnedo, M. A. A geographical distribution database of the genus *Dysdera* in the Canary Islands (Araneae, Dysderidae). *Zookeys* **2016**, 11–23 (2016).

160. Pekár, S., Líznarová, E. & Řezáč, M. Suitability of woodlice prey for generalist and specialist spider predators: A comparative study. *Ecol. Entomol.* **41**, 123–130 (2016).

161. Gorvett, H. *Tegumental glands and terrestrial life in woodlice*. *Proc. Zool. Soc. London* **126**, 291–314 (1956).

162. Sutton, S. L. *Woodlice*. (Pergamon Press., 1980).

163. Řezáč, M., Pekár, S., Arnedo, M., Macías-Hernández, N. & Řezáčová, V. Evolutionary insights into the eco-phenotypic diversification of *Dysdera* spiders in the Canary Islands. *Org. Divers. Evol.* **21**, 79–92 (2021).

164. Vizueta, J., Macías-Hernández, N., Arnedo, M. A., Rozas, J. & Sánchez-Gracia, A. Chance and predictability in evolution: The genomic basis of convergent dietary specializations in an adaptive radiation. *Mol. Ecol.* **28**, 4028–4045 (2019).

165. Cerca, J. *et al.* The Tetragnatha kauaiensis Genome Sheds Light on the Origins of Genomic Novelty in Spiders. *Genome Biol. Evol.* **13**, (2021).

166. Vieira, F. G. & Rozas, J. 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 (2011).

167. Croset, V. *et al.* Ancient protostome origin of chemosensory ionotropic glutamate receptors and the evolution of insect taste and olfaction. *PLoS Genet.* **6**, (2010).

168. Hof, A. E. V. t. *et al.* The industrial melanism mutation in British peppered moths is a transposable element. *Nat. 5347605* **534**, 102–105 (2016).

169. Arnedo, M. A., Oromí, P. & Ribera, C. Radiation of the Spider Genus *Dysdera* (Araneae, Dysderidae) in the Canary Islands: Cladistic Assessment Based on Multiple Data Sets. *Cladistics* **17**, 313–353 (2001).

170. Arnedo, M. A. *et al.* The dark side of an island radiation: systematics and evolution of troglobitic spiders of the genus *Dysdera* Latreille (Araneae : Dysderidae) in the Canary Islands. *Invertebr. Syst.* **21**, 623–660 (2007).

171. Macías-Hernández, N., López, S. de la C., Roca-Cusachs, M., Oromí, P. & Arnedo, M. A. A geographical distribution database of the genus *Dysdera* in the Canary Islands (Araneae, Dysderidae). *Zookeys* **2016**, 11 (2016).

172. Vargas, P.; Zardoya, R. *The Tree of Life*. (Sinauer Associates. Sunderland (EE.UU.), 2014).

173. Smadja, C. & Butlin, R. K. On the scent of speciation: the chemosensory system and its role in premating isolation. *Hered. 2009 1021* **102**, 77–97 (2008).

174. Asahina, K., Pavlenkovich, V. & Vosshall, L. B. The Survival Advantage of Olfaction in a Competitive Environment. *Curr. Biol.* **18**, 1153–1155 (2008).

175. Sánchez-Herrero, J. F. *et al.* The draft genome sequence of the spider *Dysdera silvatica* (Araneae, Dysderidae): A valuable resource for functional and evolutionary genomic studies in chelicerates. *Gigascience* **8**, 1–9 (2019).

176. Green, E. D. *et al.* Strategic vision for improving human health at The Forefront of Genomics. *Nat. 2020 5867831* **586**, 683–692 (2020).

177. Lewin, H. A. *et al.* Earth BioGenome Project: Sequencing life for the future of life. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 4325–4333 (2018).

178. Evans, J. D. *et al.* The i5K Initiative: Advancing Arthropod Genomics for Knowledge, Human Health, Agriculture, and the Environment. *J. Hered.* **104**, 595–600 (2013).

179. Zhang, G. Bird sequencing project takes off. *Nat. 2015 5227554* **522**, 34–34 (2015).

180. Societat Catalana de Biologia. Catalan Biogenome Project. Available online at: *https://www.scb.cat/biogenoma* (2020).

181. Wang, Z. *et al.* Chromosome-level genome assembly of the black widow spider *Latrodectus elegans* illuminates composition and evolution of venom and silk proteins. *Gigascience* **11**, (2022).

182. Croset, V. *et al.* Ancient protostome origin of chemosensory ionotropic glutamate receptors and the evolution of insect taste and olfaction. *PLoS Genet.* **6**, e1001064 (2010).

183. Zheng, X., Valakh, V., DiAntonio, A. & Ben-Shahar, Y. Natural antisense transcripts regulate the neuronal stress response and excitability. *Elife* **2014**, (2014).

184. Waterson, M. J. *et al.* Water sensor ppk28 modulates *Drosophila* lifespan and physiology through AKH signaling. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 8137–8142 (2014).

185. Saina, M. *et al.* A cnidarian homologue of an insect gustatory receptor functions in developmental body patterning. *Nat. Commun.* **6**, 6243 (2015).

186. Eyun, S. Il *et al.* Evolutionary History of Chemosensory-Related Gene Families across the Arthropoda. *Mol. Biol. Evol.* **34**, 1838–1862 (2017).

187. Robertson, H. M., Baits, R. L., Walden, K. K. O., Wada-Katsumata, A. & Schal, C. Enormous expansion of the chemosensory gene repertoire in the omnivorous German cockroach *Blattella germanica*. *J. Exp. Zool. Part B Mol. Dev. Evol.* **330**, 265–278 (2018).

188. Cao Thi Ngoc, P. *et al.* Complex evolutionary dynamics of massively expanded chemosensory receptor families in an extreme generalist chelicerate herbivore. *Genome Biol. Evol.* **8**, 3323–3339 (2016).

189. Holland, P. W. H. Gene duplication: Past, present and future. *Semin. Cell Dev. Biol.* **10**, 541–547 (1999).

190. Lemons, D. & McGinnis, W. Genomic Evolution of Hox Gene Clusters. *Science (80).* **313**, 1918–1922 (2006).

191. Librado, P. & Rozas, J. Uncovering the Functional Constraints Underlying the Genomic Organization of the Odorant-Binding Protein Genes. *Genome Biol. Evol.* **5**, 2096–2108 (2013).

192. Hahn, M. W. Distinguishing Among Evolutionary Models for the Maintenance of Gene Duplicates. *J. Hered.* **100**, 605–617 (2009).

193. Řezáč, M., Pekár, S., Arnedo, M., Macías-Hernández, N. & Řezáčová, V. Evolutionary insights into the eco-phenotypic diversification of *Dysdera* spiders in the Canary Islands. *Org. Divers. Evol.* 79–92 (2021)

194. Signor, S. A., New, F. N. & Nuzhdin, S. A large panel of drosophila simulans reveals an abundance of common variants. *Genome Biol. Evol.* **10**, 189–206 (2018).

195. Macías-Hernández, N., Bidegaray-Batista, L., Emerson, B. C., Oromí, P. & Arnedo, M. The imprint of geologic history on within-island diversification of woodlouse-hunter spiders (Araneae, Dysderidae) in the canary islands. *Journal of Heredity* vol. 104 341–356 (2013).

196. Charlesworth, B., Campos, J. L. & Jackson, B. C. Faster-X evolution: Theory and evidence from Drosophila. *Mol. Ecol.* **27**, 3753–3771 (2018).

197. Vizueta, J., Macías-Hernández, N., Arnedo, M. A., Rozas, J. & Sánchez-Gracia, A. Chance and predictability in evolution: The genomic basis of convergent dietary specializations in an adaptive radiation. *Mol. Ecol.* **28**, 4028–4045 (2019).

198. J. Litman, B. & C. Mitchell, D. Rhodopsin structure and function. *Biomembr. A Multi-Volume Treatise* **2**, 1–32 (1996).

199. Morehouse, N. I., Buschbeck, E. K., Zurek, D. B., Steck, M. & Porter, M. L. Molecular evolution of spider vision: New opportunities, familiar players. *Biol. Bull.* **233**, 21–38 (2017).

200. Leibrock, C. S., Reuter, T. & Lamb, T. D. Molecular basis of dark adaptation in rod photoreceptors. *Eye* **12**, 511–520 (1998).

Appendix

A

The genome sequence of the grape phylloxera provides insights into the evolution, adaptation, and invasion routes of an iconic pest

Rispe *et al. BMC Biology* (2020) 18:90
https://doi.org/10.1186/s12915-020-00820-5

BMC Biology

RESEARCH ARTICLE Open Access

The genome sequence of the grape phylloxera provides insights into the evolution, adaptation, and invasion routes of an iconic pest

Claude Rispe[1*†], Fabrice Legeai[2*†], Paul D. Nabity[3], Rosa Fernández[4,5], Arinder K. Arora[6], Patrice Baa-Puyoulet[7], Celeste R. Banfill[8], Leticia Bao[9], Miquel Barberà[10], Maryem Bouallègue[11], Anthony Bretaudeau[2], Jennifer A. Brisson[12], Federica Calevro[7], Pierre Capy[13], Olivier Catrice[14], Thomas Chertemps[15], Carole Couture[16], Laurent Delière[16], Angela E. Douglas[6,17], Keith Dufault-Thompson[18], Paula Escuer[19], Honglin Feng[20,21], Astrid Forneck[22], Toni Gabaldón[4,23,24], Roderic Guigó[25,26], Frédérique Hilliou[27], Silvia Hinojosa-Alvarez[19], Yi-min Hsiao[28,29], Sylvie Hudaverdian[30], Emmanuelle Jacquin-Joly[31], Edward B. James[8], Spencer Johnston[32], Benjamin Joubard[16], Gaëlle Le Goff[33], Gaël Le Trionnaire[30], Pablo Librado[34], Shanlin Liu[35,36,37], Eric Lombaert[38], Hsiao-ling Lu[39], Martine Maïbèche[15], Mohamed Makni[11], Marina Marcet-Houben[4], David Martínez-Torres[40], Camille Meslin[31], Nicolas Montagné[41], Nancy A. Moran[42], Daciana Papura[16], Nicolas Parisot[7], Yvan Rahbé[43], Mélanie Ribeiro Lopes[7], Aida Ripoll-Cladellas[25], Stéphanie Robin[44], Céline Roques[45], Pascale Roux[16], Julio Rozas[19], Alejandro Sánchez-Gracia[19], Jose F. Sánchez-Herrero[19], Didac Santesmasses[25,46], Iris Scatoni[47], Rémy-Félix Serre[45], Ming Tang[37], Wenhua Tian[3], Paul A. Umina[48], Manuella van Munster[49], Carole Vincent-Monégat[7], Joshua Wemmer[3], Alex C. C. Wilson[8], Ying Zhang[18], Chaoyang Zhao[3], Jing Zhao[35,36], Serena Zhao[42], Xin Zhou[37], François Delmotte[16*†] and Denis Tagu[30*†]

Abstract

Background: Although native to North America, the invasion of the aphid-like grape phylloxera *Daktulosphaira vitifoliae* across the globe altered the course of grape cultivation. For the past 150 years, viticulture relied on grafting-resistant North American *Vitis* species as rootstocks, thereby limiting genetic stocks tolerant to other
(Continued on next page)

* Correspondence: claude.rispe@inrae.fr; fabrice.legeai@inrae.fr; francois.delmotte@inrae.fr; denis.tagu@inrae.fr
†Claude Rispe and Fabrice Legeai are co-first authors.
†François Delmotte and Denis Tagu are co-last authors.
[1]BIOEPAR, INRAE, Oniris, Nantes, France
[2]BIPAA, IGEPP, Agrocampus Ouest, INRAE, Université de Rennes 1, 35650 Le Rheu, France
[16]SAVE, INRAE, Bordeaux Sciences Agro, Villenave d'Ornon, France
[30]IGEPP, Agrocampus Ouest, INRAE, Université de Rennes 1, 35650 Le Rheu, France
Full list of author information is available at the end of the article

(Continued from previous page)
stressors such as pathogens and climate change. Limited understanding of the insect genetics resulted in successive outbreaks across the globe when rootstocks failed. Here we report the 294-Mb genome of *D. vitifoliae* as a basic tool to understand host plant manipulation, nutritional endosymbiosis, and enhance global viticulture.

Results: Using a combination of genome, RNA, and population resequencing, we found grape phylloxera showed high duplication rates since its common ancestor with aphids, but similarity in most metabolic genes, despite lacking obligate nutritional symbioses and feeding from parenchyma. Similarly, no enrichment occurred in development genes in relation to viviparity. However, phylloxera evolved > 2700 unique genes that resemble putative effectors and are active during feeding. Population sequencing revealed the global invasion began from the upper Mississippi River in North America, spread to Europe and from there to the rest of the world.

Conclusions: The grape phylloxera genome reveals genetic architecture relative to the evolution of nutritional endosymbiosis, viviparity, and herbivory. The extraordinary expansion in effector genes also suggests novel adaptations to plant feeding and how insects induce complex plant phenotypes, for instance galls. Finally, our understanding of the origin of this invasive species and its genome provide genetics resources to alleviate rootstock bottlenecks restricting the advancement of viticulture.

Keywords: Arthropod genomes, *Daktulosphaira vitifoliae*, Gene duplications, Host plant interactions, Effectors, Biological invasions

Introduction

Biological invasions can affect ecosystems and severely impact human societies and economies by threatening global food production when the invader is a pest or pathogen [1]. How invading species become so successful in their new environments remains enigmatic, and although numerous hypotheses are supported by various organisms [2], deciphering the genetics underlying invaders provides deep insight into population or genotype-specific success [3]. Few biological invasions have wreaked as much havoc on a cultivated plant species as the grape phylloxera, *Daktulosphaira vitifoliae* (Fitch), did on the European grape, *Vitis vinifera* [4, 5]. The accidental introduction of *D. vitifoliae* in the 1860s from its native range in North America to France precipitated the start of a "phylloxeric plague" that rapidly spreads across Europe and later to other grape-growing regions of the world [6, 7], wiping out many vineyards. But it took several years to identify *D. vitifoliae* as the causative agent, largely through a fruitful collaboration between C. V. Riley (USA) and J.-E. Planchon (France) [8]. Yet, in the 150 years since the invasion began, little is known about how *D. vitifoliae* spread or what enables its success across *Vitis* species.

D. vitifoliae is a minuscule cyclically parthenogenetic insect, alternating sexual and asexual reproduction, like aphids, a related group (Fig. 1). But unlike aphids, which are viviparous in asexual stages, feed on phloem sap, and are associated with the endosymbiont *Buchnera* [10], phylloxera is oviparous at all stages, feeds on parenchymatous cells, and does not have a known obligatory bacterial endosymbiont. A further peculiarity of grape phylloxera compared to other species of its group, Phylloxeroidea, is that this insect feeds either underground on roots or on leaves (Fig. 1). Leaf-galling forms are predominant on native American *Vitis* species whereas root galling is the predominant feeding mode in cultivated varieties of *V. vinifera* worldwide. Indeed, symptoms on leaves of cultivated vines are barely observed, suggesting rarity of sexuality [11]. Root feeding is lethal on cultivated grapevine as it creates wounds that are vulnerable to entry of soil-borne fungal and bacterial pathogens [12].

Viticulture in Europe was rescued by the discovery that many *Vitis* species of American origin exhibit tolerance or resistance to *D. vitifoliae* and could be used as rootstock for grafting *V. vinifera* cultivars, thereby retaining desirable characteristics of the fruit [8]. This grafting solution exploits the coevolutionary relationship between parasite and host in native populations that resulted in the coexistence of these species. This approach has proven a successful management strategy worldwide. However, past rootstock failures [12] and the use of non-grafted vines in some regions of the world (Australia, Chile, China, and occasionally in the USA) demand constant surveillance for phylloxera infestation to prevent invasions. Ultimately, the overall success of grafting as a control strategy precipitated decreased research on phylloxera biology. Thus, many aspects of *D. vitifoliae* ecology, evolution, and population genetics, including knowledge of how its genetic architecture enables or is constrained by interactions with its host plants, remained unknown.

Genome sequencing of the grape phylloxera—with annotation performed with the help of the International Aphid Genomics Consortium—has allowed us to address evolutionary processes shaping the biology of this organism at different time-scales. First, our comparative analyses allowed us to evaluate

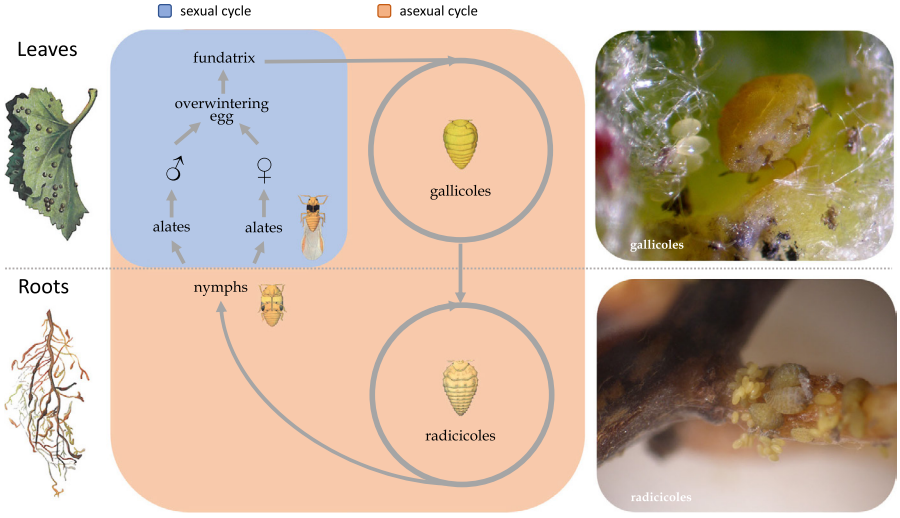


Fig. 1 The life cycle of grape phylloxera, alternating between a sexual phase and an asexual phase. Asexual females can feed either on leaves where they form galls (gallicoles) or on roots (radicicoles) of *Vitis* species. Eggs resulting from sexual reproduction hatch in the spring to produce a mobile stage (fundatrix) that initiates a colony on leaves. Gallicoles and radicicoles then undergo several asexual generations during a season. First instars of gallicoles and radicicoles are also mobile forms (crawlers) that allow the establishment of new feeding sites on active growing shoot tips or roots. Gallicoles first instars can migrate to roots, giving rise to radicicoles. Under inducing conditions, radicicoles can give rise to nymphal instars which emerge from the ground and molt to the alate adult stage achieving long-range dispersal. Alates, although morphologically identical, either engender exclusively sexual females or males. After mating, each sexual female lays a unique overwintering egg. Radicicoles can alternatively overwinter as first instar hibernants, implying a possible permanent looping of asexual cycles on roots. Gallicoles are the typical form found on native *Vitis* sp. in North America while radicicoles are most common on the cultivated varieties of *Vitis vinifera* throughout the world. Drawings included in the figure were taken from [9]

ancient evolutionary events dating back to the common ancestor between phylloxera and aphids or earlier. Grape phylloxera is related to aphids, a group with which it shares important evolutionary innovations (such as cyclical parthenogenesis, the alternation of sexual and asexual reproduction) but differs in other traits (strict oviparity, lack of obligate endosymbionts). We expected the genome sequence to exhibit evidence of these differences, in terms of gene repertoires and metabolic pathways. Also, given that aphids retain an exceptional level of gene duplication [13, 14], we examined if this observation extended to phylloxera, or even to a common ancestor of the Sternorrhyncha, the group of plant-feeding insects that includes phylloxera and aphid. We then analyzed patterns of gene expansion along the phylogenetic tree to better understand how plant feeding alters the evolution of herbivore genomes. Second, we addressed more recent evolutionary processes that influenced the genomes of each lineage (e.g., aphids versus phylloxera). Nutritional constraints, resulting from a strict diet of plant sap, are expected to have affected the genomes of both aphids and phylloxera, with expected common points and differences linked to the shared or unique traits between these groups. To address this question, we compared genome repertoires, which pointed to rapid changes possibly shaped by intense evolutionary pressure in the context of host plant specialization and manipulation. Third, our work addresses a very recent biological event, the invasion of Europe, and other grape-producing regions by phylloxera. With the intention to trace back the geographical routes of this invasion, we performed genome-wide sequencing of phylloxera samples from both the native (North America) and introduced populations (Europe and New World vineyards).

Our study, supported by a highly complete genome and an uncommon community effort on curated annotation, revealed that:

- Phylloxera (like aphids) has a high number of coding genes compared to other arthropods, with both an

increased level of gene duplication mapped to the common ancestor of aphids and phylloxera and high rates of recent duplications
- An extraordinarily large expansion of a novel gene family is comprised of putative effectors; we expect that they represent a key component of the interactions and adaptation between this insect species and its host plants
- Phylloxera populations of the upper Mississippi River basin, feeding on the wild species *Vitis riparia*, are likely to be the principal source of the invasion to Europe. Subsequent invasions of South America and western Australia were the result of secondary introductions, from European sources

Results and discussion

Genome features

The haploid genome size of the *D. vitifoliae* Pcf7 strain was estimated by flow cytometry at 294 Mb by two independent measures (± SD = 1 Mb with *Drosophila melanogaster* and ± SD = 5 Mb with alfalfa as references, respectively). The final assembly (v3.1) summed to 282.7 Mb, a total close to that estimated by flow cytometry. The genome assembly comprised 10,492 scaffolds with a median size of 1077 bp and an N50 of 342 kb. A BUSCO analysis based on insect conserved genes indicated the presence of 94.2% of these as complete genes (Table 1). A total of 24,581 genes (OGS 3.0) were automatically predicted. Extensive manual annotation (see below) led to gene corrections of more than 15% of the inspected genes as well as new gene detection (see the "Effectors" section), such that the final gene catalog contained 25,814 predicted genes and 25,825 transcripts (OGS 3.2). The genomic GC content was low for an arthropod (27.2%) but comparable to that of other aphid genomes (e.g., 27.8% for *A. pisum*, 30.1% for *M. persicae* [13, 14]). The recovered mitochondrial genome had gene content and order typical of insect and aphid mitochondrial genomes with 13 protein-coding genes, 22 tRNA genes, and 2 rRNA genes (Additional File 1: Fig.S1): the *D. vitifoliae* mitochondrial scaffold was smaller (15,568 bp) than the mitochondrial genomes from the pea aphid, *Acyrthosiphon pisum* (16,971 bp), and the fruit fly, *Drosophila melanogaster* (19,517 bp), and had similar GC content to both other species (15.5% vs 15.2% and 17.8%, respectively) [13].

Table 1 Assembly parameters and genome features of the grape phylloxera genome, version V3.1

Parameters	Numbers
Assembly	Version 3.1
Contigs	
Total assembly size	282,671,353
Number of contigs	17,162
Contig N50 length (bp)	74,750
Longest contig (bp)	718,286
Shortest contig (bp)	83
Number of contigs > 10 kb	4914
Mean (median) contig size, in bp	16,107 (1635)
Scaffolds	
Number of scaffolds	10,492
Longest scaffold (bp)	2,080,308
Shortest scaffold (bp)	141
Number of scaffolds > 1 Mb	19
Mean (median) scaffold size, in bp	26,942 (1077)
N50 scaffold length (bb)	341,590
Genomic features (OGS 3.2)	
Mean transcripts length (bp)	4653
Mean CDS length (bp)	1053
Mean exon length (bp)	244
Mean exon number per gene	5.4
Gene count	25,825
BUSCO analysis (genome v3.1)	
Complete BUSCO	1563/1658 (94.2%)
Complete and single-copy BUSCOs	1531/1658 (92.3%)
Complete and duplicated BUSCOs	32/1658 (1.9%)
Fragmented BUSCOs	26/1658 (1.6%)
Missing BUSCOs	69/1658 (4.2%)

Horizontal gene transfer from bacteria and fungi into the phylloxera genome

Genomes of Aphididae and Adelgidae species were previously shown to contain genes underlying carotenoid biosynthesis as the result of a horizontal transfer event from a fungus [15]. Homologs of these genes were recently found to be present in nine Phylloxeridae species [16], including the grape phylloxera. Confirming these results, our searches of the phylloxera genome revealed that the carotenoid biosynthetic gene cluster is present, as a single copy, and containing the fused phytoene synthase/lycopene cyclase that is characteristic of aphids and of some fungi (Additional File 1: Table S1) [17, 18]. Based on BLASTp searches (*e* value cutoff = 0.01) of published genomes using query sequences from *A. pisum*, homologs of these genes appear to be absent from sequenced genomes of the Psyllidae and Aleyrodidae. Phytoene desaturase is present in adelgids based on PCR amplification and Sanger sequencing [15], but genome sequences of adelgids are not available for further screening. Based on this distribution, it is likely that these fungal genes were transferred to an ancestor of all Aphidomorpha (Aphididae, Adelgidae, Phylloxeridae) in one event and underwent subsequent duplications in lineages of Aphididae.

The *A. pisum* genome also contains genes of bacterial origin (*ldcA*, *rlpA*, and *amiD*) that are highly expressed in the bacteriocytes housing the obligate bacterial endosymbiont *Buchnera aphidicola*, but that were acquired from bacterial sources other than the symbionts [17, 18]. None of these genes could be found in the phylloxera genome. Because Phylloxeridae lack obligate bacterial symbionts, the absence of these genes is consistent with the hypothesis that they were acquired by ancestral aphids in the context of adaptation for the obligate symbiosis. The absence of these genes could reflect loss in Phylloxeridae or acquisition in Aphididae after divergence from Phylloxeridae (Additional File 1: Table S1).

Repetitive DNA

In addition, 317,612 TE copies were identified; these constitute 119 Mb, or 42.2% of the draft sequence (Additional File 1: Fig.S2) [19], slightly above the 38% found in *A. pisum* [13] and the maximum for known hemipteran genomes. These sequences were classified according to their structural and coding features into 1996 TE families. LINE elements (26.5%) and Class Terminal Inverted Repeats (TIR, 13%) were the most prevalent in class I and II, respectively. LTR and TIR orders were dominated by *Gypsy* and *hAT* elements, respectively (Additional File 1: Table S2), as also found in *A. glycines* and *B. tabaci* [14, 20]. Comparisons of these copies within each order of TE and within the clusters defined by REPET show that average identities were generally below 95% (Additional File 1: Fig.S2), suggesting that most superfamilies correspond to ancient invasions. However, a few clusters, corresponding to *Gypsy*, *Bel/pao*, *Tc1-mariner*, and *hAT* elements, showed high degrees of identity (> 95%), suggesting recent expansions of these elements.

Annotation of protein-coding genes

To improve the quality of gene prediction and to elucidate key biological processes in grape phylloxera, the IAGC fostered a community effort of manual curation, leading to the expert annotation of 4815 genes, or approximately 18.6% of the final gene set (OGS 3.2). All annotation steps and transcription data are stored in AphidBase [21]. This allowed us to perform a phylogenomic study of the phylloxera gene content and specific analyses of functional groups as detailed below.

Evolution of gene content and duplication rates

A comparison of gene content across 14 taxa, including phylloxera, other hemipteran species and several outgroup insect species revealed many widespread genes (red bars in Fig. 2). Lineage-specific and/or orphan genes also were often abundant, particularly in phylloxera and some aphid species (*A. pisum* and *R. padi*) but not all. Furthermore, a relatively large number of genes were specific to the Phylloxeridae + Aphididae clade (purple bars, Fig. 2). A total of 6623 genes from the phylloxera genome (25.9% of the total) were phylloxera-specific (i.e., did not have any homologs in the phylogenetic context of our study). These were enriched in GO terms related to sensory perception of taste, protein metabolism, microtubule-based processes, ribosome biogenesis, and G-protein coupled receptor signaling pathway, among others (Additional File 1: Fig.S3). Enriched GO terms in the phylloxera genome, excluding TEs, related to host cell surface receptor binding, hydrogen ion

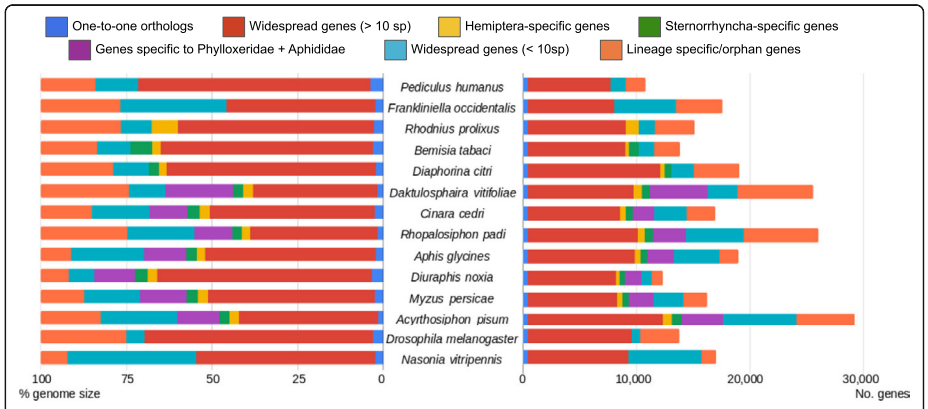


Fig. 2 Comparative gene content across insects, with emphasis on Hemiptera. Total number of genes (right) or percentage of genome (left) are indicated

A

transmembrane transporter activity, odorant binding, and olfactory receptor activity, which suggests that some of the phylloxera-specific gene expansions are involved in sensory perception (Additional File 1: Fig.S4). Among phylloxera-specific genes, 1115 had hits with InterProScan databases, indicating that they may have homologs outside the phylogenetic context of this study. This still leaves 5508 genes in the phylloxera genome with strictly no hit. These results are in line with those found for other aphids. For instance, 4530 genes were inferred as species-specific and/or orphan in *Aphis glycines*, which represents a 23.3% of its genes [22]. We analyzed gene gain and loss patterns across Sternorrhyncha, the hemipteran suborder containing phylloxera; the Sternorrhyncha is defined by its characteristic mouthpart position, adapted for plant sap feeding. Rates and patterns of gene gain and loss varied widely among taxa. The highest level of net gene gain and loss was found for *Diaphorina citri*, with ca. 6500 genes lost in comparison with phylloxera (4442 excluding TEs) (Additional File 1: Fig.S5). The lowest values were obtained for the aphid species *A. pisum*, *M. persicae*, *A. glycines*, and *R. padi*. Interestingly, gene gain and loss were lower at more basal nodes (N1 to N8) than at the tips of the phylogeny (Fig. 3). Our phylome approach for Sternorrhyncha species and outgroups showed a high duplication rate at the base of Phylloxeridae, Adelgidae, and Aphididae (i.e., at node C, where this metric ranged between 0.49 and 1.59 depending on the inclusion of TEs and gene expansions) (Additional File 1: Fig.S6). This, along with our analysis of duplication ages (Fig. 4), indicates an excess of old duplicates predating the diversification of Aphidomorpha. In addition, for relatively recent duplications (nodes for which $dS < 1$), we found many more duplication events in phylloxera ($n = 6005$ nodes) than in *D. melanogaster* ($n = 440$) (Fig. 3). We found in particular 2717 pairs of paralogs with $dS < 0.1$, which is 13 times the number found in the *D. melanogaster* genome. An even stronger burst of recent duplications was found for *A. pisum* (10,399 nodes with $dS < 0.1$, a 51-fold increase compared to *D. melanogaster*) (see [13]). For *A. pisum*, a recent study based on a chromosomal-level assembly showed that duplications in this lineage were dominated by small-scale events, with no signs of larger-scale events [23]. With the goal of understanding the putative role of gene duplicates in the generation of new adaptations in pest species, we explored GO enrichment in the genes duplicated at nodes

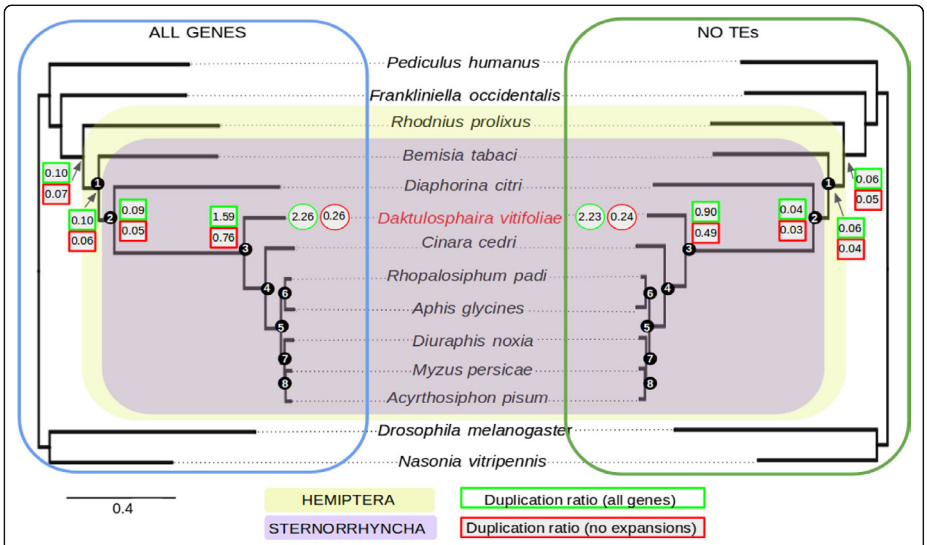


Fig. 3 Species tree based on one-to-one orthologs inferred in our data set. The grape phylloxera is indicated in red. All nodes were highly supported in all analyses (> 0.95 SH-like support). Duplication ratios considering all genes and excluding proteins encoded by transposable elements (TE) are plotted in the three most basal nodes of Sternorrhyncha and Hemiptera. The two ratios provided per node, as well as for the phylloxera terminal, correspond to the values resulting from the inclusion (green) or exclusion (red) of gene expansions. Some nodes mentioned in the "Results and discussion" section (marked as 1, 2, and 3) are highlighted

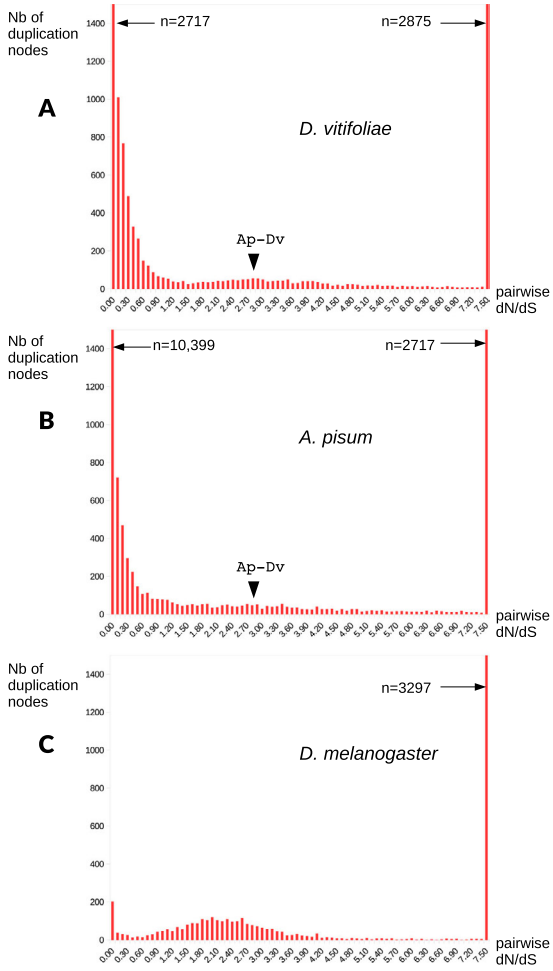


Fig. 4 Distribution of synonymous distances among paralogs for grape vine phylloxera (panel **a**, *D. vitifoliae),* pea aphid (panel **b**, *A. pisum*), and fruit fly (panel **c**, *D. melanogaster*). Paralogs were identified as RBH pairs, with an iterative approach allowing to cover both recent duplications (terminal nodes in gene families) and more ancient duplications (internal nodes). For readability, the *y*-axis (number of dS classes) is truncated to 1500 (numbers above that threshold are indicated on the figures). For both *A. pisum* and *D. vitifoliae*, an arrowhead indicates the median dS between orthologs (RBH genes between the two species), dS = 2.83: this metric, a proxy of the age of separation between the two species allows to distinguish duplications that are more recent (left of the arrow, lower dS values) *or more ancient (right of the arrowhead, higher dS) than the speciation event

preceding the diversification of phylloxera (Sternorrhyncha, Psyllidae + Aphidomorpha, and Aphidomorpha). While almost no enrichment was detected in genes duplicated at the nodes respectively preceding Sternorrhyncha and the clade comprising Psyllidae plus Aphidomorpha, genes duplicated at Aphidomorpha were enriched in several functions, including regulation of transcription, protein modification (phosphorylation,

protein binding, etc.), neurogenesis, oogenesis, and sensory perception (Additional File 1: Table S3). On top of this, an important part of the recent phylloxera expansions was constituted by lineage-specific genes (most of them, with no GO assigned), which we characterized as putative effectors, as developed below. Altogether, these results suggest that a burst of duplication at the origin of Aphidomorpha, but also more recent species-specific bursts of duplicates, both affecting diverse biological functions, could have contributed to feeding-related adaptations in these lineages.

Metabolism and immunity genes

Metabolic pathways were reconstructed combining the CycADS database and a so-called gap-filling procedure (see the "Material and method" section). Gap filling allowed improving annotation by detecting 29 putative additional functions, associated with 39 genes (Additional File 1: Table S4). For example, this includes a candidate gene for phosphopantothenoylcysteine synthetase (DV3025962.1, EC: 6.3.2.5), an enzyme of the coenzyme A biosynthesis pathway, and a candidate gene for nicotinamidase (DV3000063, EC: 3.5.1.19), an enzyme involved in nicotinamide metabolism (Additional File 1: Fig.S7). Thus, the DakviCyc database contains a metabolic network reconstruction of the phylloxera genome. Metabolism was found to be largely conserved between grape phylloxera and the aphids *M. persicae* and *A. pisum* (Fig. 5), as 335 pathways were present in all three species, while we found 11, 9, and 5 unique pathways for *D. vitifoliae*, *M. persicae*, and *A. pisum*, respectively. But 22 pathways were missing in phylloxera compared to the two aphid species (Fig. 5, Additional File 1: Table S5). Finally, the urea cycle (Additional File 1: Fig.S8) was absent from all three species [13, 24]. We identified 1097 different EC numbers with at least one protein in the phylloxera genome (Fig. 5, Additional File 1: Table S5). Of these, 961 appear to be core enzyme functions shared with both *M. persicae* and *A. pisum*. Only 66 were found to be unique to phylloxera, while 34 were found in *M. persicae* and 116 in *A. pisum*. In addition, 221 enzyme functions were found to be missing in grape phylloxera, including 71 shared by *M. persicae* and *A. pisum*. All genes required for amino acid metabolism and found in phylloxera were present in *M. persicae* and *A. pisum* (Fig. 6). Broken metabolic pathways in the two species of aphids are frequently completed by genes encoded by *Buchnera*, the aphid's primary endosymbiont. However, phylloxera does not have symbionts [25, 26] which would imply that phylloxera cannot synthesize amino acids such as cysteine or arginine (Fig. 6). The bacterium *Pantoea agglomerans* is occasionally found in phylloxera [27], but is not an obligate symbiont, so it probably does not provide missing essential amino acids to this insect. This inability is probably compensated by the specific feeding mode of phylloxerids (modified parenchymal cells which contain essential amino acids) [28–30].

Concerning immunity genes, all genes of the TOLL pathway were found, though some had low similarity to *D. melanogaster* homologs (Additional File 1: Table S6). By contrast, and as previously observed for *A. pisum* and other aphids and the psyllid *D. citri*, several key genes of the IMD pathway present in *D. melanogaster* or other arthropods were missing in phylloxera: *Imd*, *CYLD*, *Fadd*, and *Tab2* (Additional File 1: Table S7, Additional File 1: Fig.S9). Genes encoding *PGRPs* and other antimicrobial peptides

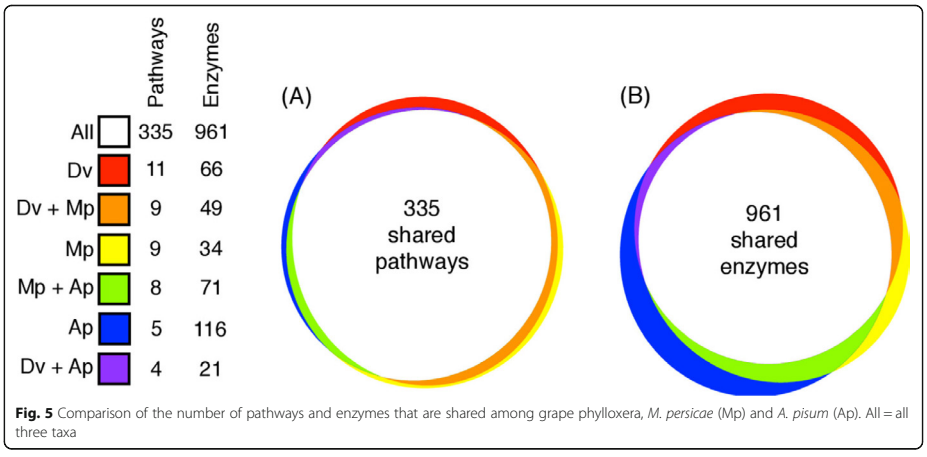


Fig. 5 Comparison of the number of pathways and enzymes that are shared among grape phylloxera, *M. persicae* (Mp) and *A. pisum* (Ap). All = all three taxa

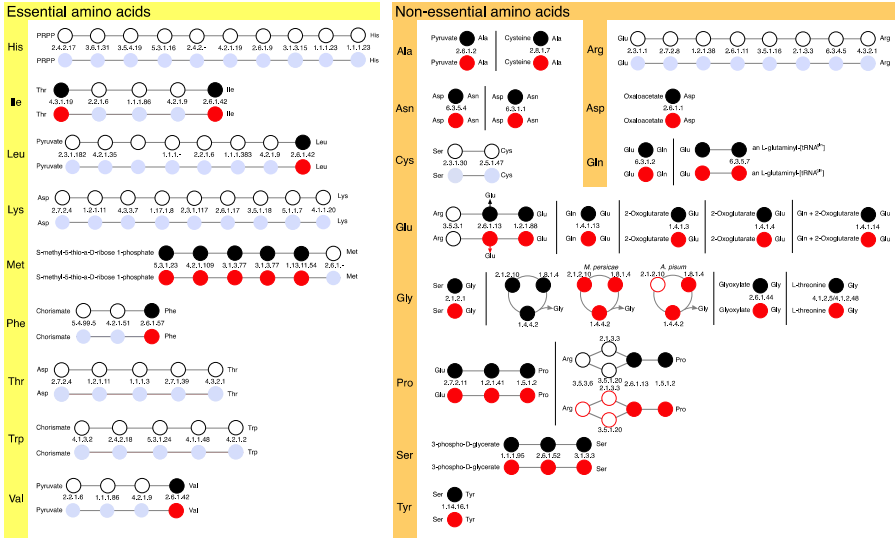


Fig. 6 A comparison of amino acid biosynthesis pathways between *D. vitifoliae* (upper or left pathway for each amino acid, with enzymes depicted by black circles) and the aphids *A. pisum* and *M. persicae* (lower or right pathway for each amino acid, with present enzymes depicted by red circles). The presence of an enzyme in a pathway is shown by a filled circle, and the absence by an empty circle. In aphids, the endosymbiotic bacteria *Buchnera aphidicola* is involved in amino acid metabolism: *Buchnera*-produced enzymes are depicted by a filled blue circle. Pathways read from left to right. Where known, enzymes are identified by their EC number

were also absent, whereas the JNK pathway, which is connected to IMD in *D. melanogaster*, was complete. Some differences, which are difficult to explain, exist between phylloxera and aphids: *Dredd* and *Kay* are present in phylloxera but not in *A. pisum*, while *Tab2*, absent in the phylloxera genome, is present in *A. pisum*. We also found that only one known transcription factor (TF) regulating the IMD pathway was present in the phylloxera genome instead of the three (*Dif*, *Dorsal*, and *Relish*) in other insects. This phylloxera immune TF matches to the three present in other insect species, and it was not possible to establish clear relationships of orthology between these genes. The lack of an intact IMD pathway in psyllids and aphids has been suggested to relate to their obligate symbiotic associations [31, 32]; however, this hypothesis does not explain the apparent lack of intact IMD pathway in phylloxera, which lacks obligate symbionts. We however note the possibility of divergent genes that would represent a functional pathway as recently shown for *R. prolixus* [33].

Functional groups that are similar between phylloxera and aphids

Manual annotation combined with phylogenetic and evolutionary analyses indicates that genes in several functional groups have not changed drastically between Phylloxeridae and Aphididae. This is the case of selenoproteins, which are proteins that include a selenocysteine amino acid residue, this requiring a specific machinery. Although most animals have selenoproteins, several insects including *A. pisum* lack them [13]. We find that grape phylloxera also lacks both selenoproteins and the Sec machinery. Selenoproteins known to be present in other Paraneoptera species were found only as cysteine-containing homologs (MSRB1 and TR) or not found at all (SPS2, GPXx, and SelenoK). Most of the essential factors for selenoprotein synthesis of the Sec machinery, (tRNA-Sec, PSTKpstk, SEPSECSSecS, SECI SBP2SBP2, and EEFSECeEFSec and SEPHS2) were absent. Analyzing additional genomes of Paraneoptera allowed to map the selenoprotein extinction event in the common ancestor of scale insects, phylloxera, and aphids (Fig. 7). The gene set underlying structural components of the cuticle is also highly conserved between phylloxera and aphids. A total of 94 unique cuticular proteins (including 11 RR-1 and 61 RR-2) were found in the phylloxera genome. These numbers were similar in aphids, although *A. pisum* showed a larger expansion of the RR-2 protein subfamily (Additional File 1: Table S8).

A

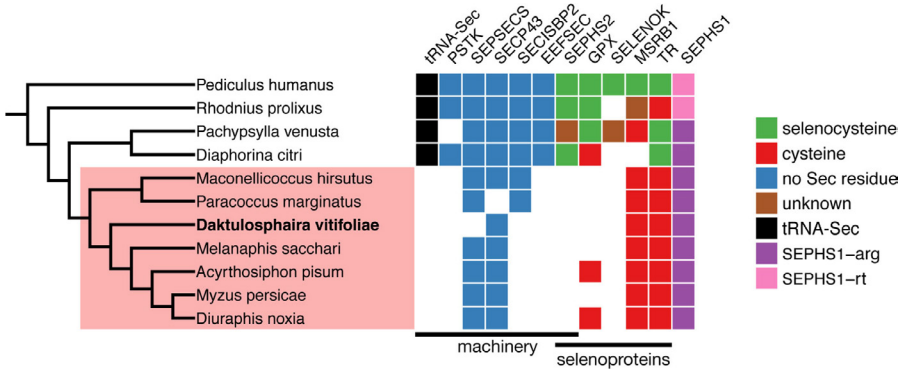


Fig. 7 Selenoproteins in Paraneoptera. Species tree annotated with the prediction of selenoproteins and the Sec machinery. The species highlighted in red (Aphidomorpha and Coccoidea) lack selenoproteins and several genes required for their synthesis (the Sec machinery). Each column corresponds to a gene family. Selenoprotein families are colored based on the codon found at the Sec position: selenocysteine in green; cysteine in red; brown indicates incomplete results where the codon at the Sec position was not known. Sec machinery genes are colored in black (tRNA-Sec) or blue (proteins). SEPHS1 is a paralog of SEPHS2, which was found here to have either an arginine codon (SEPHS-arg) or a non-Sec UGA readthrough codon (SEPHS1-rt) at the Sec position

Most RR-1 proteins from phylloxera seem to display 1-to-1 or 1-to-2 orthology relationships with their *A. pisum* and *M. persicae* homologs (Additional File 1: Fig.S10). This reduced complexity signals the absence of specific duplication trends for this protein subfamily (in contrast with the RR-2 subfamily). Concerning the RR-2 subfamily, the main trend was the presence of three clades of high diversification within aphid species, and therefore absent from the phylloxera clade (labeled Post-Dv diversification clusters A, B, and C in Additional File 1: Fig.S11), while a few cases of RR-2 genes from phylloxera phylogenetically close and localized in tandem suggest recent duplications. We found that phylloxera retains standard sets of chitin-metabolizing genes (chitin synthase, chitinases, chitin-binding, chitin deacetylase genes). A single chitin synthase has been detected in all aphid species, and also in phylloxera, a situation correlated with the absence of peritrophic membrane in aphid guts. Lastly, we did not see major differences in the gene complement of the "development" function, even though phylloxera lacks viviparity, a major developmental difference from aphids [34]. This suggests that viviparity in aphids evolved through sub- or neo-functionalization of genes that existed in the common ancestor of the two groups. The developmental gene catalog of phylloxera is nearly complete, with 97 genes annotated (Additional File 1: Table S9). Most of the missing genes were also absent in Aphididae, e.g., *bicoid*, *gurken*, or *oskar*. We found fewer gene duplications in the phylloxera genome than in the *A. pisum* genome (e.g., for *piwi*). Annotations and analyses on microRNAs (Additional File 1: Fig.S12, Additional File 1: Table S10), DNA methylation genes, aquaporins (Additional File 1: Fig.S13), the circadian clock machinery (Additional File 1: Table S11, Fig.S14, Fig.S15), and odorant and gustatory receptors or ligand proteins and detoxification proteins (Additional File 1: Table S12, Table S13, Table S14, Table S15, Fig.S16, Fig.S17, Fig.S18, Fig.S19, Fig.S20) [17, 18, 35–43] are described in the supplementary information document, along with the corresponding methods and results (Additional File 1: Supplementary Methods and Results) [44–76].

Extraordinary large expansion of candidate effector genes

We identified over 2700 genes with effector attributes, indicating that a large repertoire of genes underlies nutrition, growth, and defense-related processes during interactions with *Vitis* species. Of these, 419 had domains with known function (Additional File 1: Table S16), yet most genes did not show clear homology to genes in any other organisms (> 86% were no-hit). The three most numerous domains belong to the RING-type zinc finger, ankyrin repeat, and EF-hand domains, which function generally to respectively modulate the ubiquitin-proteasome pathway [77], mediate protein-protein interactions [78], or bind to calcium, e.g., calmodulins, to regulate the cellular calcium signaling pathway [79]. Notably, all pathways play important roles in a vast array of cellular processes and impact nearly every aspect of cellular life including stress response, growth, and

development [78, 80]. The largest four groups contained the majority of genes (80% or 2165 of 2741 genes, Additional File 1: Fig.S21), but this was driven by the largest single cluster of 1551 genes (Fig. 8C). This species-specific expansion likely reflects the influence of host specialization as observed for other insect effector genes

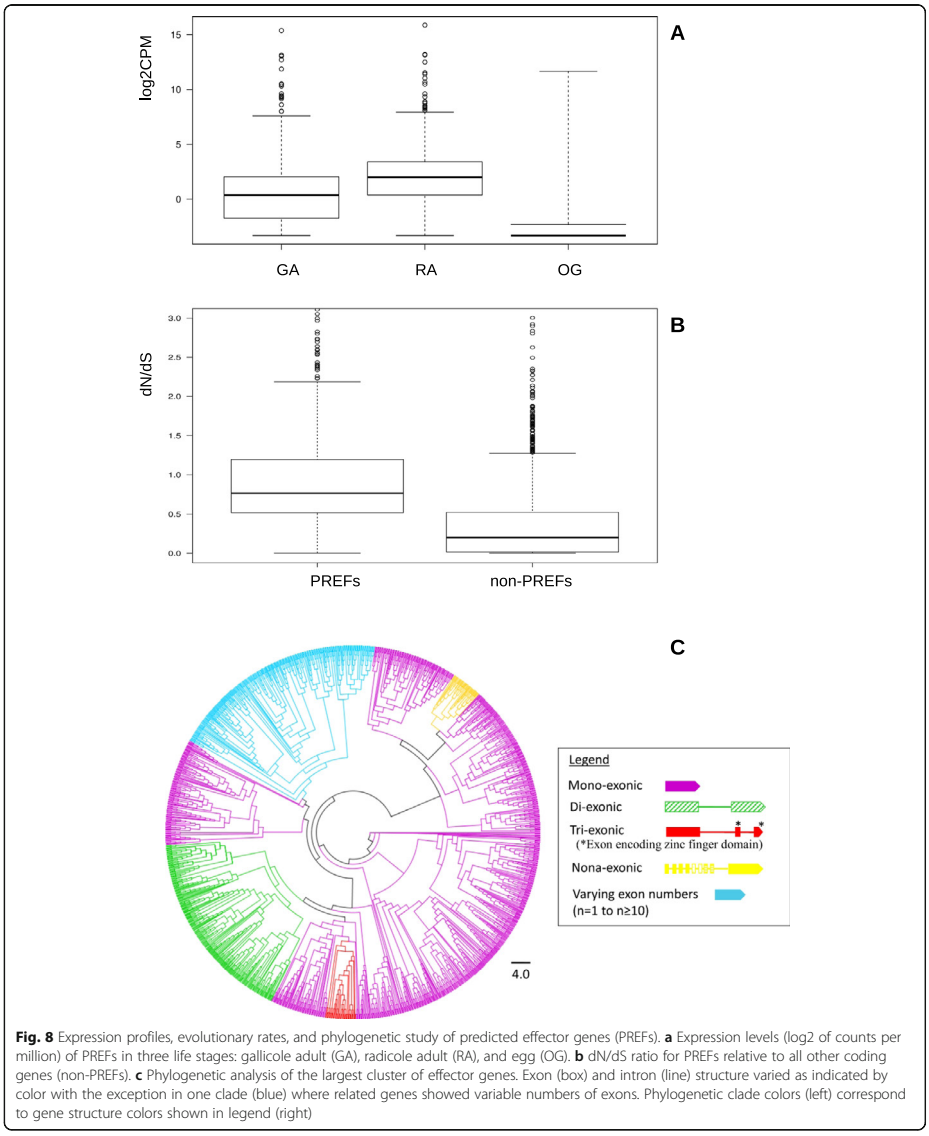


Fig. 8 Expression profiles, evolutionary rates, and phylogenetic study of predicted effector genes (PREFs). **a** Expression levels (log2 of counts per million) of PREFs in three life stages: gallicole adult (GA), radicole adult (RA), and egg (OG). **b** dN/dS ratio for PREFs relative to all other coding genes (non-PREFs). **c** Phylogenetic analysis of the largest cluster of effector genes. Exon (box) and intron (line) structure varied as indicated by color with the exception in one clade (blue) where related genes showed variable numbers of exons. Phylogenetic clade colors (left) correspond to gene structure colors shown in legend (right)

[81, 82]. Phylogenetic study of this cluster combined with the analysis of exon-intron architecture revealed that most genes lack introns, a feature of genes that function in rapid turnover [83]. Interestingly, some subclades (i) evolved additional (up to and ≥ 10) exons specific to gene clades, (ii) duplicated existing exons to form motif repeats, or (iii) lost exons (Fig. 8c). While the gain of novel exons contributes to the development of new gene functions, exon duplications to form motif repeats help establish stable structures that play versatile roles in many biological processes [84]. A subgroup of genes within the largest cluster contained RING domains (this domain was the most frequent among all domains identified). Thus, genes within this large cluster may mediate protein-protein interactions in part via the ubiquitin-proteasome pathway [77]. This is hypothesized to represent an evolutionary innovation to manipulate plant development, perhaps through molecular mimicry [85–88]. In insects, for example, the Hessian fly delivers hundreds of F-box proteins, a component of SCF-type E3 ubiquitin ligase complex, as effectors likely for insect colonization and gall formation [81], and the green peach aphid (*M. persicae*) and the green rice leafhopper (*Nephotettix cincticeps*) inject EF-hand proteins as calcium binding molecules into host cells during feeding [89, 90]. This hypothesis was also supported by recent evidence of interactions between secretory RING proteins of phylloxera and plant proteins and by the finding of strong downregulation of plant genes related to protein synthesis in *Vitis* galls [91]. Our findings thus suggest that *D. vitifoliae* secretes a pool of effectors to mimic host proteins for plant manipulation.

The great expansion of effector genes is accompanied by a specific pattern of expression restricted to feeding forms, especially on roots (Fig. 8a) and fast evolution, as indicated by high dN/dS values, diversity of exon-intron structures within clusters, and tandem duplication (Fig. 8b, Additional File 1: Fig.S22) [92]. Taken together, these effector gene attributes highlight the complexities that underlie construction of an extended phenotype suggesting a role in coevolution with *Vitis* hosts and mirroring patterns observed to a lesser extent in other insects [81, 93].

Invasion routes of phylloxera

Genome sequencing of pools of insects from several populations of both the native and introduced range was used as a tool to infer the most probable routes of the phylloxera invasion(s) from Northern America to the rest of the world (Fig. 9a) and to compare genetic profiles and variability of the different populations. Samples from the introduced range in Europe clustered together, which is broadly consistent with a single origin for the invasion of the different European countries (Fig. 9b). This European cluster also included two populations from the native range, Wisconsin and Illinois. Therefore, native populations of the upper Mississippi River region, which feed on the wild riverbank grape (*Vitis riparia*), could represent the source of the historic invasion of Europe by phylloxera. This result linking European population and *V. riparia* native populations is consistent with preceding studies using mitochondrial [94] or microsatellite markers [95]. This area, known as French Louisiana in the seventeenth and eighteenth centuries, was under strong French influence and had intense commercial exchanges with France and the rest of Europe into the nineteenth century. At that time, exotic plants were fashionable, and botanists and vine growers had established many personal collections of American vine varieties through the importation of seedlings, cuttings, and rooted plants [7]. Several reports indicate Missouri as the source of resistant rootstocks, suggesting an established grape culture in the Mississippi River region. The French sample however had a distinct profile from the rest of the European populations (Germany, Austria, Romania, Armenia) which were all very tightly clustered with Mississippi valley populations (Wisconsin, Illinois) (Fig. 9b). Using ABC methods, we found that the genetic profile of French populations was best explained as the result of admixture between populations from the Middle West (Wisconsin or Illinois) and from the New York region (Fig. 10a). It is tempting to relate the more diverse genetic profile of French phylloxera population with the historical reports of two independent fronts of colonization in that country, respectively, in Pujaux in the Gard department in 1861 and in Floirac near Bordeaux in 1866 [96] (two sites separated by ~ 430 km). While distinct North American localities may have been sources for the two sites of introduction in France, this hypothesis is difficult to validate without historical phylloxera collections. Also, movements of populations might have erased the possible initial genetic structure resulting from this admixture.

Considering the introduction from the Mississippi valley (represented by Illinois) into the rest of Europe, we tested two scenarios: in the first scenario, colonization of European vineyards would have followed the colonization of France (which served as a bridgehead) by the Illinois population, whereas in the second scenario, there would have been two independent introductions (and two different bottlenecks) from Illinois to France and from Illinois to the rest of Europe. ABC scenarios supported in majority the first scenario (Fig. 10b). Historical reports have documented that the invasion of French septentrional vineyards and central European countries occurred through progressive colonization from sources in South France, which is consistent with this hypothesis [7]. Our data also give new insight into

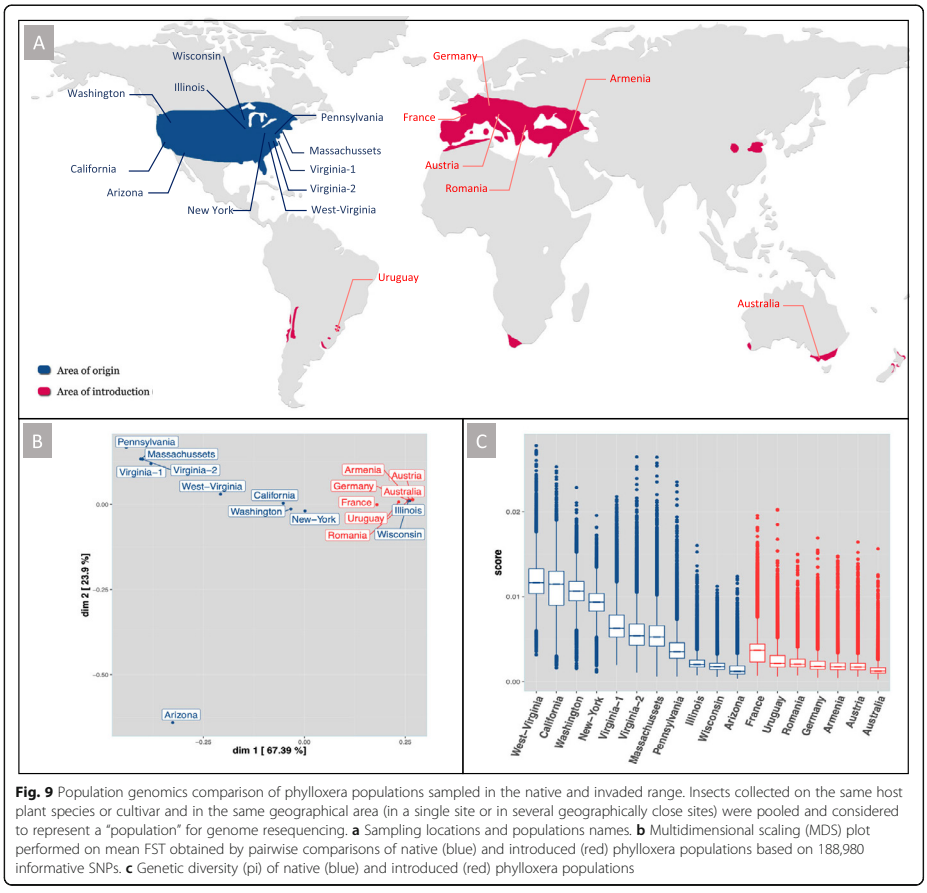


Fig. 9 Population genomics comparison of phylloxera populations sampled in the native and invaded range. Insects collected on the same host plant species or cultivar and in the same geographical area (in a single site or in several geographically close sites) were pooled and considered to represent a "population" for genome resequencing. **a** Sampling locations and populations names. **b** Multidimensional scaling (MDS) plot performed on mean FST obtained by pairwise comparisons of native (blue) and introduced (red) phylloxera populations based on 188,980 informative SNPs. **c** Genetic diversity (pi) of native (blue) and introduced (red) phylloxera populations

the worldwide invasion of phylloxera, as we found that introduced populations from South America (Uruguay) and Australia were extremely close to European populations. This may result either from an introduction from the same North American source of the European invasion or from a secondary introduction from Europe. The second scenario is likely, as Southern hemisphere vineyards were planted with traditional *V. vinifera* varieties imported from Europe.

The native population from Arizona was found to be highly divergent from all other populations, with a very low level of genetic diversity (Fig. 9c). It is a geographically distant population with insects feeding on a locally distributed host, *Vitis arizonica*. Lund et al. [97] also reported that Arizona populations were markedly different from other North American populations, suggesting that these populations represent a different host race within grape phylloxera or even a distinct species. The estimated divergence between the population from Arizona and other native populations for the *co1* mitochondrial gene (~ 1%) represents a relatively high variation for two races, but could still remain below commonly used thresholds for defining different species [98]. However, relatively low levels of divergence can correspond to a recent event of speciation, a scenario that would fit with the reproductive isolation of this population located on sky islands and likely disconnected from all other populations from the rest of the USA. Surprisingly,

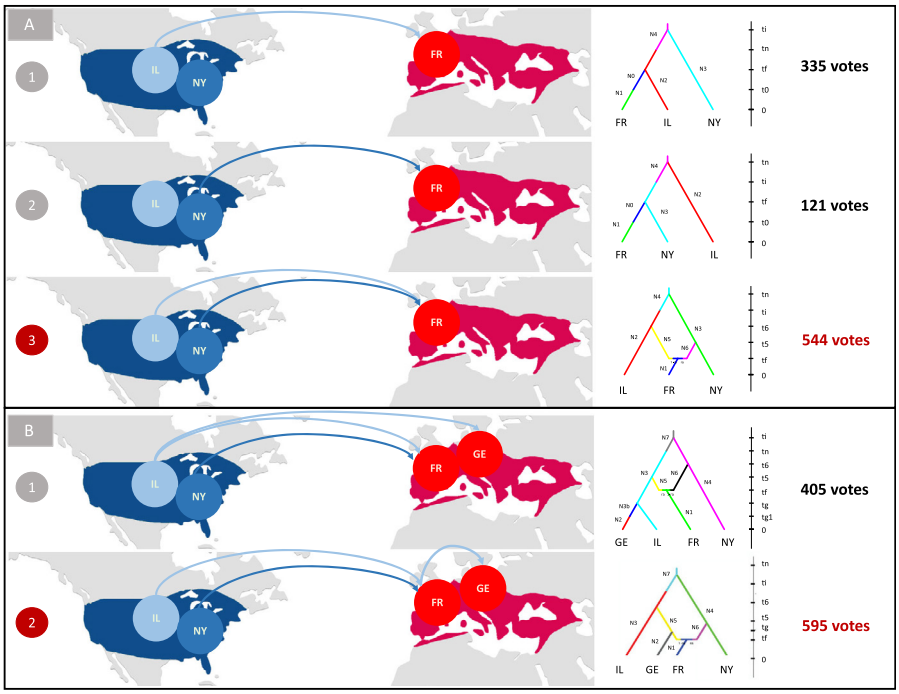


Fig. 10 Routes of the phylloxera invasion of Europe inferred from the analysis of genome-wide resequencing data from native and introduced populations. Most likely scenarios of phylloxera introduction into Europe identified by the approximate Bayesian computation (ABC) approach using native populations collected on *Vitis riparia* (New York = NY, Illinois = IL) and introduced populations (France, Germany). **a** Scenarios for the introduction of France, testing a single origin (NY or IL) and admixture. **b** Scenarios for the secondary introduction in Germany, testing an independent introduction from IL versus a common introduction between France and Germany. Detailed legends of the diyABC trees on the right are given in Additional File 1: Table S20

Illinois and Wisconsin populations also had very low genetic diversity, similar to what is observed in introduced populations. This suggests a complex story in the native range itself, since genetic bottlenecks could likely explain these patterns (possibly after recolonization or long-term isolation). Our data therefore suggest that some native phylloxera populations had low genetic diversity before they served as a source for the invasion into Europe, suggesting that founder effects [94] are not the sole factor of the limited genetic diversity of introduced populations.

Conclusions

The genomic resources presented here provide new insights into genome evolution that change our understanding of grape phylloxera interactions. They also open the door to research lines such as the role of the expanded family of effectors in plant feeding, the adaptation of the metabolism in absence of bacterial symbionts, and the influence of host plant specialization on genome architecture. Our results provide a detailed understanding of the genetics underlying invasion and reveal the potential threat to viticulture and native grapes naïve to phylloxera should naturally occurring populations that vary in gene repertoires invade. Given both genotypic diversity and transcriptional plasticity underlie the adaptation of species to novel hosts [24], the genome of grape phylloxera provides the means to understand how populations or even single genotypes adapted to local climates when existing or new populations of phylloxera expanded from North America or Europe to other parts of the world.

Material and methods

Biological material for genome sequencing

The isofemale INRA-Pcf7 clone was established from grape phylloxera individuals collected in 2010 at Pineuilh (France) on "Cabernet franc" scions grafted on S04 rootstock (*V. berlandieri x V. riparia*). The clone was maintained in rearing collection at INRAE Bordeaux through parthenogenetic reproduction (controlled chamber at 22 °C, L:16/D:8 and 60% of humidity) on American variety "Harmony" leaves, a Dog-Ridge hybrid of *V. champinii* and accession 1613C (*V. labrusca* x *V. riparia* x *V. vinifera*), and on root pieces of *V. vinifera* "Cabernet sauvignon."

Flow cytometry

Two measures were performed independently, using protocols described in [99, 100], respectively. Briefly, measures were performed from the whole body of a female phylloxera INRA Pcf7 clone, using D. *melanogaster* female (1C = 175 Mbp) or alfalfa (*Medicago sativa*) leaf tissue (1C = 206.4 Mbp) as a standard. Nuclei from a mixture of both biological materials (phylloxera vs standard) were prepared and propidium iodide-stained. The relative 2C red fluorescent peak positions of the sample and standard were determined by flow cytometry with the amount of DNA in phylloxera calculated as the ratio of the sample and standard 2C peak means times the 1C amount of DNA of the standard. This was done for $n = 12$ replicates (using *D. melanogaster*) and $n = 9$ replicates (using alfalfa).

DNA extraction and sequencing

For Illumina sequencing, genomic DNA was extracted from six samples of the Pcf7 clone, each corresponding to approximately 200 individuals, with a mix of adults and larvae. The insects were homogenized using three sterilized glass beads (2 mm diameter) for 30 s at 30 Hz (Tissuelyser, Retsch), and DNA was extracted using DNeasy Blood & Tissue kit (Qiagen Inc., Chatsworth, CA). Between 14 and 25 μg of DNA were obtained from each sample after column elution with 100 μl of 10 mM Tris-HCl-1 mM EDTA, pH 7.8. Quantitation of DNA was performed using DeNovix Fluorescence Assays. Four pair-end and two mate-pair libraries were prepared according to the Illumina manufacturer's protocol (Additional File 1: Table S17). For PacBio sequencing, four samples, each with ~ 600 adults of the Pcf7 clone, were extracted with a salting-out protocol [101]. Through this protocol, a total of 120 um of long and ultrapure genomic DNA fragments were obtained. Quality was assessed with a NanoDrop (A260/280 ratio between 1.8 and 2.0 and A260/230 ratio ≥ 2.0). Illumina sequencing was performed at the BGI Shenzhen facilities (Shenzen, China) on a HiSeq2500 machine. PacBio was performed at the Genotoul facilities (Toulouse, France) using the SMRT sequencing technology. Illumina pair-end, Illumina mate-pair, and PacBio reads gave a genome sequencing coverage of 147X, 36X, and 58X respectively (Additional File 1: Table S17).

Reads processing and assembly

We first eliminated adaptors and removed duplicate reads. The remaining sequences were then corrected using the Soap Error Correction (SOAPec_v2.01) tool and assembled using the SOAPdenovo pipeline (version 2.04: released on July 13, 2012) with the options -K 81 (kmer size) and -d 2 (edges cutoff), resulting in 414,258 scaffolds. Scaffolds longer than 500 bp or including a gene annotation (see below) were kept ($n = 16,380$) and scaffolded with PacBio subreads (without correction) using a modified version of SSPACE-LR ver 1.1 [102], with the option "-s 1 -a 250". Finally, the gaps of this last version were filled with Illumina reads using GapFiller [103].

Automatic annotation and manual curation

Gene predictions were generated using MAKER2 [104]. Within MAKER2, a first gene set was predicted by similarity to known proteins, or contigs of RNA-Seq (see below). This gene set was used thereafter for training both Augustus [105] and SNAP [106], in two steps, using results from an initial training to retrain again the software. Transcriptomic evidence came from two previous RNA-Seq projects [107, 108], which included whole bodies of leaf-galling adults (gallicoles), whole bodies of root-feeding adults (radicicoles), and eggs from radicicoles. Proteomic evidence came from SwissProt (release 2016_10) and a protein set from various hemipteran species, including *A. pisum* (NCBI), *M. persicae* Clones G006 and O (AphidBase), *D. noxia* (NCBI), *Cimex lectularius* (NCBI), and *Rhodnius prolixus* (EnsEMBL). An Apollo [59] server was set up to allow manual curation of a set of genes from the automatic annotation. As many as 4815 genes were curated and checked based on guidelines defined by BIPAA [https://bipaa.genouest.org/is/how-to-annotate-a-genome/]. Curated genes were merged with the automatic annotation using a custom script [https://github.com/abretaud/ogs-tools/tree/master/ogs_merge]. Putative functions of predicted proteins by the above pipeline were identified with blastp (v2.6.0) against Genbank NR (non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF, release 09/2017), and interproscan v5.13-52.0 against Interpro. Associated GO terms were collected from blast NR and interproscan results with blast2GO (v2.5). Transmembrane domain signal peptides were identified by tmhmm v2.0c and signalP (euk v4.1), respectively. All genome resources and the Apollo server were made available online on BIPAA, within the AphidBase section [http://

bipaa.genouest.org/is/aphidbase/daktulosphaira_vitifoliae/] [21]. This system was rolled out using different projects from the GMOD tool suite (JBrowse [109], Chado [110], Tripal [111]) and developments from the Galaxy Genome Annotation project [https://galaxy-genome-annotation.github.io/] [112].

Detection of contaminant scaffolds

A screening of contaminants was performed on scaffolds (blastp of the predicted proteins to nr), which allowed to eliminate 3 scaffolds identified as bacterial. We also used BLOB [113] which screens viral, bacterial, and eukaryotic contaminants based on GC content and similarity. We identified a very small number of potential residual contaminants: they totalled only 1.352 Mbases in 92 scaffolds (0.5% of the assembly size), suggesting that contamination would at best be marginal.

Characterization of the mitochondrial genome

The mitochondrial genome sequence from the grape phylloxera was found during genome assembly. The initial mitochondrial scaffold was 15,613 bp in length, and inspection of the predicted gene sequences revealed a frameshift within the *nad5* sequence. Closer inspection showed a possible insertion of 45 T nucleotides within *nad5*. PacBio reads were mapped to this region, revealing that the insertion was likely due to a sequencing or assembly error. This insertion was removed, resulting in an intact *nad5* gene sequence. The final assembled scaffold is 15,568 bp in length and has a GC content of 15.6%. A gene prediction analysis was performed on this scaffold using MITOS [114] and ARWEN v1.2 [115].

Horizontal gene transfers

To determine if genes for carotenoid biosynthesis were present in the phylloxera genome, we used genes of this pathway previously characterized in *A. pisum* as query sequences for blastp searches on the predicted proteins of the phylloxera genome. The *A. pisum* genome also contains genes from bacterial sources [17, 18], and, again using the *A. pisum* sequences as queries, we performed blastp searches on protein databases for the genomes of *D. vitifoliae*, *A. pisum*, *Myzus persicae*, *Diuraphis noxia*, *Aphis glycines*, *Rhopalosiphum padi*, *Diaphorina citri* (Psyllidae), *Pachypsylla venusta* (Psyllidae), and *Bemisia tabaci* (Aleyrodidae). A blastp search was also conducted in the NCBI non-redundant protein sequence database, in order to identify other species where these genes might be present. The alignments were made using MAFFT v7.313 using default parameters [47]. Phylogenetic trees were constructed from sequences retrieved from blastp searches, using RAxML under the PROTCATJTTF model, with 100 bootstrap replicates.

Repetitive DNA

Transposable elements (TEs) were identified and annotated using the REPET package v2.2 [116, 117]. Manual inspection was performed to confirm TE orders, clusters, and families. The level of identity between a fragment and its reference TE/repeat consensus was used to estimate ages of TE expansions.

Annotation of protein-coding genes

Gene expansions

Phylome reconstruction The phylome (i.e., the complete collection of phylogenetic trees for each gene in a genome) of grape phylloxera was reconstructed to obtain a dynamic view of gene family expansion within this genome. We included nine other fully sequenced genomes of Hemiptera based on their phylogenetic position: *A. pisum* (Harris, 1776) (Sternorrhyncha, Aphididae), *M. persicae* (Sulzer, 1776) (Sternorrhyncha, Aphididae), *D. noxia* (Kurdjumov, 1913) (Sternorrhyncha, Aphididae), *C. cedri* (Curtis, 1835) (Sternorrhyncha, Aphididae), *A. glycines* Matsumara, 1917 (Sternorrhyncha, Aphididae), *R. padi* (Stal, Linnaeus, 1758) (Sternorrhyncha, Aphididae), *D. citri* Kuwayama, 1908 (Sternorrhyncha, Psylloidea), *B. tabaci* (Gennadius, 1889) (Sternorrhyncha, Aleyrodoidea), and the true bug *Rhodnius prolixus* (Stål, 1859) (Heteroptera, Reduviidae). As outgroups, we selected four insect taxa: *D. melanogaster* Meigen, 1830 (Diptera, Drosophilidae), *Nasonia vitripennis* (Ashmead, 1904) (Hymenoptera, Pteromalidae), *Frankliniella occidentalis* (Pergande, 1895) (Thysanoptera, Thripidae), and *Pediculus humanus* (Linnaeus, 1758) (Phthiraptera, Pediculidae). Genome versions are indicated in Additional File 1: Table S18. Phylomes were reconstructed using the PhylomeDB pipeline [118]. For each protein encoded in the grape phylloxera genome (25,567 annotated proteins, Official Gene Set version 3.2) (http://bipaa.genouest.org/sp/daktulosphaira_vitifoliae/), we performed a blastp search against the custom proteome database built from the genomes listed above, which included a total of 252,530 proteins. Results were filtered using an *e* value of 1e−05 and a minimum of 50% overlap between the query and the hit sequences. Multiple sequence alignments were reconstructed in forward and in reverse [119] using three different programs: MUSCLE v3.8 [120], MAFFT v6.712b [61], and Kalign v2.04 [121]. The resulting alignments were then combined using M-COFFEE v10.00.r1607 [122]. A trimming step was performed using trimAl v1.3 [48] (consistency-score cutoff 0.1667, gap-score cutoff 0.9). The best fitting model was selected by reconstructing neighbor joining trees as implemented in BioNJ [123] using seven different models (JTT, LG, WAG, Blosum62, MtREV, VT, and Dayhoff). The best

model in terms of likelihood as selected by the Akaike Information Criterion (AIC) [124] was chosen for tree reconstruction. Trees were reconstructed using PhyML v20120412 [62]. Four rate categories were used, and invariant positions were inferred from the data. Branch support was computed using an aLRT (approximate likelihood ratio test) based on a chi-square distribution. Resulting trees and alignments are stored in phylomeDB 4.0 [118] (http://phylomedb.org), with the phylomeID 196. A species-overlap algorithm, as implemented in ETE v3.0 [125], was used to infer orthology and paralogy relationships from the phylogenetic trees reconstructed in the phylome. The algorithm traverses the tree and calls speciation or duplication events at internal nodes based on the presence of common species at both daughter partitions defined by the node. Gene gains and losses were calculated on this basis. Duplication ratios per node were calculated by dividing the number of duplications observed in each node by the total number of gene trees containing that node: theoretically, a value of 0 would indicate no duplication, a value of 1 an average of one duplication per gene in the genome, and > 1 an average of more than 1 duplication per gene and node.

Species tree reconstruction The species tree was built using one-to-one orthologs present in all 14 included species, with a final alignment of 409 genes and 245,463 concatenated amino acid positions. To ensure a congruent phylogenetic hypothesis under different models, a series of approaches were followed to infer the species tree. First, an approximately maximum-likelihood tree was reconstructed with FastTree v. 2.1 [126] under the LG [127] model of amino acid evolution. Second, a supertree was reconstructed using DupTree [128] based on all the trees reconstructed in the phylome. Both phylogenies were congruent.

Removal of proteins from transposable elements In order to disentangle the effect of transposable elements (TEs) and of other factors, we removed all genes annotated as proteins encoded by TEs, prior to the inference of gene expansions, GO term enrichment and gene gains, losses, and duplications.

Detection of expanded protein families For each gene tree, we selected the nodes that contained only phylloxera sequences with ETE v3.0 [125]. Nodes with more than 5 sequences were counted as expansions. Overlapping expansions (i.e., partial gene trees with terminals in common) were fused when they shared more than 20% of their members.

Gene annotation and scrutiny of putative phylloxera-specific genes In addition to the automatic and manual annotation performed on the phylloxera Official Gene Set (OGS) (http://bipaa.genouest.org/sp/daktulosphaira_vitifoliae/), all genes in the phylloxera genome were functionally annotated with InterProScan v.5.19 [129]. Gene Ontology [130] annotations and PFAM [65] motifs were assigned to these genes as well. All genes that did not show any BLAST hits during the all-by-all comparison (see the "Phylome reconstruction" section) were interpreted as putative phylloxera-specific genes. These genes were further scrutinized through functional annotations with InterProScan v.5.19 [129] as well.

GO term enrichment FatiGO [131] was used to check for enrichment in GO terms between the phylloxera genes and the rest of the database (i.e., the sum of the genes belonging to the other species included in the phylome). Sets of enriched GO terms were summarized and visualized in REVIGO [132]. GO enrichment was explored for phylloxera-specific genes, as well as for genes duplicated in each of the nodes to evaluate potential specific adaptation at different time points of the evolution of this species and group.

Synonymous distance-based assessment of duplication ages To remove potentially spurious gene models from the official gene set, we first used a filtering step, eliminating genes which had very weak support: these were defined as genes with no manual annotation, no hit to the nr database of GenBank, and very low RNA-Seq support (< 0.5 CPM for the average of expression counts between radicicoles, gallicoles, and eggs). This left us with $n = 21,863$ genes (a filtering of nearly 4000 genes). To evaluate synonymous distances (dS) among paralogs, we used a Reciprocal Best Hit approach (RBH) by blasting gene collections against themselves, determining pairs of genes that matched the RBH criteria. Doing this in just one step would lead us to focus on terminal branches in expanded gene families, neglecting deeper nodes and thus missing the ancient dynamics in the history of duplications. To account for this, we applied an approach similar to that used in [133]: after a first round of RBH identification, a member of each RBH pair was tagged for elimination (we chose the shortest sequence, or randomly selected one of the genes in case of equal lengths). We then re-started the RBH identification, allowing to gradually reach deeper nodes in gene families. The process was reiterated 10 times, as the number of duplications decreased sharply in the last runs. Each RBH pair of genes in the different runs (representing a node in gene families) was used for a pairwise estimation of synonymous distance. For this, the protein sequences were aligned; this alignment was then reported on the nucleotide sequence and cleaned using GBlocks [134]; this step eliminated poorly aligned regions, giving a

conservative estimate of the distances among copies. Finally, dS was estimated using Codeml (PAML software [135]). For comparison, we applied the same procedure to the *A. pisum* genome (using the NCBI update prediction, n = 27,986 genes) and for *D. melanogaster* (using the r6.21 annotation, and selecting the longest alternative transcript of each gene, n = 13,931 gene sequences).

Metabolism

CycADS annotation and DakviCyc database generation We used the Cyc Annotation Database System (CycADS [136]), an automated annotation management system, to integrate protein annotations from different sources into a Cyc metabolic network. Using the CycADS pipeline, proteins were annotated using Blast2GO [137], InterProScan [129], KAAS [138], PRIAM [139], and PhylomeDB [118] to obtain EC and GO numbers. These data were processed in the CycADS SQL database and automatically extracted to generate appropriate input files to build or update BioCyc databases [140] using the Pathway Tools software [141]. The DakviCyc database, representing the metabolic protein-coding genes of phylloxera, was thus generated and is now included in the ArthropodaCyc database, a collection of arthropod metabolic network databases [142] (http://arthropodacyc.cycadsys.org/).

Metabolic pathway gap filling Metabolic reconstructions from the ArthropodaCyc databases for *D. vitifoliae*, *A. pisum*, and *M. persicae* (clone G006) were exported in the SBML format and imported into the PSAMM software [143]. First, metabolic pathway gaps were identified using the "*gapcheck*" function, which reports a list of all metabolites not produced in the metabolic network. Then, the objective function was defined for each non-producing metabolite, and a gap-filling procedure was performed for each objective function through individual rounds of simulations using the PSAMM implementation of the *fastgapfill* algorithm [144]. In the gap-filling step, results from *A. pisum* and *M. persicae* were used as candidates for identifying potentially missing annotations. Following the gap-filling predictions, candidate missing genes were identified through the identification of homologs to annotated genes in *A. pisum* and *M. persicae*. This was achieved with manual curations using evidence from blast alignments, Pfam protein domain identifications [65], phylomeDB [118], transcriptomic support of gene expression, and literature review. Two rounds of annotation were performed with the above procedure, and predictions in the DakviCyc database were updated through these iterations. External links to resources that include the comprehensive enzyme information system: BRENDA (https://www.brenda-enzymes.org/), InterPro [129], KEGG orthology (https://www.genome.jp/kegg/), PhylomeDB [118], and crosslinks to the AphidBase [21] genome browser were added for all predicted genes.

Immunity genes

Immune genes were annotated using bidirectional blastp analyses. We first used the phylloxera gene set to identify proteins with similarity to genes of the IMD and TOLL pathways. These putative phylloxera proteins were then blasted against *D. melanogaster* reference proteins. This approach was then extended to a complete collection of *D. melanogaster* immune genes. For reciprocal best hits (RBH) between phylloxera and *D. melanogaster*, the *D. melanogaster* annotation was directly transferred to phylloxera. In other cases (non-RBH relationship), a manual curation was performed, using the genomic information for arthropods with well-annotated immune pathways (*Nasonia vitripennis*, *Plautia stali*, *Rhodnius prolixus*, *Tribolium castaneum*) or for other aphid genomes (*A. pisum and M. persicae*) available in Genbank, ArthropodaCyc, and ImmunoDB [145] databases.

Cuticular proteins

To determine the full set of genes coding for cuticular proteins (CPs) (including cuticular proteins with R&R motif defined as CPR proteins [146]), we searched CPs among the initial prediction by using the CutProtFam annotation website [147] (http://aias.biol.uoa.gr/CutProtFam-Pred/), with standard settings. Candidate genes were then fully manually curated on AphidBase through Apollo. Phylogenetic analyses were performed using the updated protein sequences of sets of RR-1 or RR-2 genes of *M. persicae* [24], *A. pisum* [148], and *D. noxia* [149]. RR-1 and RR-2 sub-groups were treated separately. For RR-1 proteins, signal peptides were predicted using ExPASy tools (http://www.expasy.org/tools/) and removed; then phylogenetic analyses were conducted on the mature sequences. For RR-2 proteins, only the extended 69 amino acids RR domain (pfam00379) was used for phylogenetic analysis, the rest of the sequences being too divergent to align. Alignments were made with Clustal Omega [150], and phylogenetic analyses were made using the Phylogeny.fr platform [151] where alignments were cleaned with Gblocks and a maximum likelihood method as implemented in the PhyML program was used to infer a phylogenetic tree.

Selenoproteins

Selenoproteins contain the non-canonical amino acid selenocysteine (Sec), known as the 21st amino acid. Sec is encoded by a UGA codon, normally a stop codon, and is inserted through a recording mechanism that requires

a dedicated set of factors known as the Sec machinery [152]. Selenoproteins exist in different domains of life and are widespread in Metazoa, but appear to be lacking in some insect species [153] including the pea aphid [13], two Astigmata (non-insect arthropods) species [154], and plant parasitic nematodes [155]. To search for selenoproteins and the Sec machinery, the genome of grape phylloxera was analyzed with Selenoprofiles [156] and Secmarker [154].

Effectors

To identify genes underlying effector proteins active when grape phylloxera interacts with *Vitis* host plants, we modified a bioinformatics pipeline from [157]. This pipeline was designed based on four features of effectors: (1) secretory, (2) small-sized (≤ 500 amino acids, and this only applied on the initial screening), (3) herbivore-only, and (4) gene-duplicating [157, 158]. Testing of this pipeline on the genome dataset of the Hessian fly (*Mayetiola destructor*), a plant manipulating herbivore [81], showed that 95% of the predicted effector genes matched (blastp *e* value < 1e−5) the salivary gland-derived Hessian fly effector genes. We therefore screened the 24,585 automated gene models (OGS3.0_20161223_proteins) and predicted a first set of 354 effector genes that classified (using OrthoMCL) into 78 clusters according to sequence similarity. We then performed manual annotation on each of these clusters to (1) correct gene models based on the transcript data from gallicole, radicicole, and egg samples [107] integrated into Apollo and sequence similarity to other members of the same cluster and (2) recover gene models, through tblastn searches, that were not included in the automated annotation and prediction of OGS3.0_20161223_proteins because of mis-prediction. Using this automated gene model-based (AGMB) approach (note that it also identified effector candidates that were absent of automated gene models but shared sequence similarity to the ones predicted from the automated gene model collections), and eliminating our sequence size limit to include proteins > 500AA, we predicted 1766 effector candidates from the genome of *D. vitifoliae*. While conducting manual annotation on the genome, we detected a number of putative genes which had particular characteristics: (1) absence of automatic annotation (i.e., no gene model was predicted), (2) presence of an ORF usually encoding more than 200 amino acids and corresponding to a monoexonic structure, and (3) clear RNA-Seq support, in particular in the radicicole samples. The two former points suggested that these ORFs represented bona fide genes, with a particular intron-less structure. Such pattern is usually penalized in gene prediction tools for Eukaryotes, which could explain the absence from the automatic gene model prediction. Additional traits of these genes suggested that they encoded effectors because of (1) the presence of a secretory signal peptide in the N-terminus; (2) clusters of similar gene copies, indicative of tandem gene duplication; and (3) some sequence similarity to the putative effectors predicted using the AGMB approach described above. To generalize the search of similar genes, we performed tblastn searches to the genome and annotated matching regions which shared the above patterns. Because we usually found different hits in each search, but with a relatively low amino acid identity (as low as 20%), it appeared that the grape phylloxera genome encodes highly expanded gene families characterized by high evolutionary rates. To ensure that we collected the most complete collection of genes, the tblastn searches were performed iteratively, each time using the collection of manually annotated monoexonic effector candidates as a query data set, then annotating the new hits, and repeating this process until no new hits were detected. Some of the effector candidates identified using this non-automated gene model-based (NAGMB) approach overlapped with those identified through the AGMB pipeline and therefore were combined with the latter resulting into a total number of 2741 manually annotated predicted effectors (PREFs) in the phylloxera genome. Genes were clustered using SiLiX. Because numerous PREFs appeared unique, lacking sequence homology to other PREFs, and comparisons were based within species rather than among species, the final clusters were determined through an iterative process in SiLiX. As overlap among sequences increased to 60%, very few new clusters were formed. Similarly, as identity decreased down to 20%, the number of clusters reached a minimum. Thus, 60% overlap and 20% identity were designated as conservative thresholds per parameters defined in SiLiX. As PREF function is validated through further study, these thresholds may change to best organize clusters without breaking up families of known function predicted from sequence motifs. Phylogenetic analyses of the largest orthogroup (cluster3, n = 1551 PREFs) were performed following the protocol described by [30] with modifications. Briefly, the deduced protein sequences were aligned using MAFFT (v7.271) [47] with "auto" setting and the alignments were trimmed using TRIMAL(v1.4) based on a gap threshold of 0.25. One PREF (DV3018723) was removed because its sequence is composed only by gaps after trimming, leaving a total of 1550 PREFs with 375 amino acids each (including gaps) for phylogenetic tree construction. Lastly, these aligned sequences were run on PhyML (v3.0) with the value of approximate Likelihood-Ratio Test (aLRT) for branches set as "-1." To evaluate selective pressures acting on PREFs (comparing the different orthogroups, and comparing PREFs and non-PREFs), we estimated evolutionary rates for the

most recent duplication events in the genome. These events were pointed by determining reciprocal best hits (RBH) and by estimating the pairwise non-synonymous to synonymous ratio (dN/dS) for each pair of sequences found to be RBH. For that, we aligned sequences, trimmed the alignments (with Gblocks), and evaluated rates with codeml (PAML). For RBH detection, we included all manually curated genes (including PREFs); among the other genes, we eliminated gene models with very low support (genes with no hit and a very low RNA-Seq support, i.e., < 0.5 counts per million reads in radicicoles, gallicoles, and eggs data sets). This filtering was intended to remove noise and potentially inflated rate estimates that might occur for spurious gene models (the resulting data set comprised 23,961 genes).

Genome resequencing of phylloxera populations and invasion route inference

Phylloxera individuals were collected from both native and introduced areas. All samples consisted of gall-feeding adult insects except for two American populations (California, Washington) that were sampled as root-feeding insects. Insects collected in the same geographical area (in a single site or in several geographically close sites) and on the same host plant species or cultivar were pooled and considered to represent one population for genome resequencing. In the native area, samples were collected either on cultivated grapevines or on wild *Vitis* species: *Vitis arizonica* (Arizona), *Vitis labrusca* (Massachusetts), *Vitis aestivalis* (West Virginia), *Vitis vulpina* (Pennsylvania, Virginia1), *Vitis riparia* (Wisconsin, Illinois, New York), interspecific hybrid Chambourcin and Concord (Virginia2 and Washington, respectively), rootstocks 1103P (California). Populations from introduced areas (France, Germany, Hungary, Austria, Romania, Armenia, Uruguay, Australia) were collected from galls on leaves of *Vitis vinifera cv.* Details on this sampling are presented in Additional File 1: Table S19. For each pool, which comprised between 30 and 100 individuals (adult insects), a DNA library was prepared with the TruSeq Nano Illumina kit, and sequenced on one lane of an Illumina HiSeq3000 sequencing machine at the Genotoul platform (reaching a genome coverage of ~ 60X for each pool). The reads (paired-end 2 × 150 bp) were mapped on the genome reference with BWA mem v0.7.10, with default parameters. Only primary alignments of properly paired reads were kept using samtools, and PCR duplicates were removed using Picard tools (https://github.com/broadinstitute/picard). Each pileup file was then subsampled with Popoolation2/subsample-pileup.pl [159] in order to reach a coverage of 15 at each site, and individual population genetic statistics (diversity, mutation rate, and Tajima's *D*) were calculated with Popoolation2/Variancesliding.pl. The counts of major alleles for each population and for each position were calculated from the subsamples and used as entry for the PCA (FactomineR). We used popoolation2/FST_sliding to estimate pairwise FST after the synchronization of the pileup files with Popoolation2/mpileup2sync.jar, extraction of polymorphic sites (minimal count of the minor allele over all the samples = 4, and coverage at each site and each sample > 10) and subsampling (as above). The average of FST pairwise were computed and used for generating a distance matrix distance for the MDS plot (done with R/ggplot2).

In order to test various demographic scenarios for the introduction of phylloxera in Europe, we used diyABC, and abcRF [160]. We randomly selected 10,000 polymorphic SNPs and 100 monomorphic SNPs in 5 populations (France, Germany, Illinois, Wisconsin, and New York) and generated individual data (respectively 200, 140, 170, 120, and 200 individuals) based on the observed allelic frequencies at each site. With diyABC, we extracted summary statistics (with respect to the distributions of the diversity, FST, and Nei's distances) for more than 10,000 simulations by scenarios and used abcRF to compare simulations results with summary statistics from our observed genotypes in order to choose the most realistic model (i.e., those with more votes among 1000 trees in the random forest). We first compared the demographic scenarios in the native area, with and without admixture, selecting the best model, then introduced sequentially the French and German populations (representing the two genetic profiles of phylloxera populations found in Europe)—for detailed statistics, see Additional File 1: Table S20.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12915-020-00820-5.

Additional file 1: Figures. S1-S22, Table S1-S20, Methods and Results. Figure S1. Mitochondrial genome view of grape phylloxera. **Figure S2.** Proportion of transposable elements (TE) in the genome. **Figure S3.** GO terms of phylloxera-specific genes. **Figure S4.** Enriched GO terms in the phylloxera genome with and without TEs. **Figure S5.** Gene gain/loss at different nodes or branches. **Figure S6.** Species phylogenetic tree based on insect genomes and the transcriptomes of *Planococcus citri* and *Adelges tsugae*. **Figure S7.** Diagram of the gap-filling and annotation process. **Figure S8.** Urea cycle in *D. vitifoliae* and *A. pisum*. **Figure S9.** IMD immune pathway in *D. vitifoliae*. **Figure S10.** Phylogenetic tree of RR-1 cuticular proteins. **Figure S11.** Phylogenetic tree of RR-2 cuticular proteins. **Figure S12.** Comparison of miRNAs in *D. vitifoliae* and other insect genomes. **Figure S13.** Phylogenetic tree of aquaporin protein sequences. **Figure S14.** Comparison of the phylloxera PER protein with other insects. **Figure S15.** Amino acid alignment of PTTH amino acid sequences. **Figure S16.** Phylogeny of hemipteran ORs. **Figure S17.** Phylogeny of hemipteran GRs. **Figure S18.** Phylogenetic analysis of OBPs. **Figure S19.** Phylogenetic analysis of CSPs. **Figure S20.** Phylogenetic analysis of NPC2s. **Figure S21.** Distribution of cluster sizes of putative effectors. **Figure S22.** Physical distribution of the three largest clusters of effectors. **Table S1.** Genes of bacterial and fungal origin.

Table S2. Statistics on TEs. **Table S3.** GO enrichment of genes duplicated at different ancestral nodes. **Table S4.** Metabolic gaps in the D. vitifoliae reaction network. **Table S5.** Functional annotation of metabolic genes. **Table S6.** Genes of the TOLL pathway. **Table S7.** Genes of the IMD pathway. **Table S8.** Statistics on cuticular proteins. **Table S9.** Developmental genes in *D. vitifoliae* and *A. pisum*. **Table S10.** miRNAs. **Table S11.** Clock-related genes. **Table S12.** List of ORs and GRs. **Table S13.** Number of OBPs, CSPs and NPC2s. **Table S14.** List of Cytochromes P450. **Table S15.** List of genes involved in detoxification. **Table S16.** Effector genes with predicted domains and their corresponding functions. **Table S17.** Statistics on sequence reads and SRA accessions used for the reference genome. **Table S18.** List of species used to study gene expansions. **Table S19.** Sampling sites and SRA used for population genomics analyses. **Table S20.** Prior distribution of parameters used for ABC modeling of invasion routes.

Acknowledgements
The International Aphid Genomics Consortium (IAGC) has been a key player for the coordination of the annotation process. The 1KITE consortium is acknowledged for providing transcriptomic data on the whole insect from the grape phylloxera. The i5k consortium is acknowledged for its support on Apollo training and for expert genome annotation. The bioinformatic platforms GenOuest and BIPAA (Bioinformatics Platform for Agroecosystem Arthropods) in Rennes (France) and Bioinfo Genotoul in Toulouse (France) are acknowledged for their support in bioinformatics. We thank Fabien Gagné (Canada), Mauro Jermini (Switzerland), Ann-Kristin Jung (Germany), Laszlo Kocsis (Hungary), Roberto Lopez (Argentina), and M. Andrew Walker (USA) for collecting phylloxera populations—although not all populations could be included in the present study. We thank Dorith Rotenberg (North Carolina State University, USA) for giving early access to the unpublished genome assembly of the Thysanoptera species *Frankliniella occidentalis* (ik5 project). Three reviewers made helpful comments or suggestions, which helped in improving the manuscript.

Authors' contributions
Authors contributing specific working groups are indicated below. Steering committee: F Delmotte, F Legeai, C Rispe, and D Tagu; genome sequencing: Shanlin Liu, Jing Zhao, M Tang, X Zhou, C Couture, D Papura, B Joubard, P Roux, R-F Serre, and C Roques; flow cytometry: O Catrice and S Johnston; genome assembly and annotation: F Legeai, A Bretaudeau, and S Robin; mitochondrial DNA: K Dufault-Thompson and Y Zhang; horizontal gene transfer: S Zhao and N Moran; repetitive DNA: P Capy, M Bouallègue, M Makni, and F Legeai; miRNAs: H Feng, ACC Wilson, F Legeai, S Hudaverdian, and G Le Trionnaire; duplications: R Fernández, M Marcet-Houben, C Rispe, and T Gabaldón; metabolism and immunity: P Baa-Puyoulet, C R. Banfill, F Calevro, E B James, N Parisot, M Ribeiro Lopes, K Thompson, C Vincent-Monégat, ACC Wilson, and Y Zhang; aquaporins: AK Arora and AE Douglas; DNA methylation: JA Brisson; selenoproteins: A Ripoll-Cladellas, D Santesmasses, and R Guigó; circadian clock and related genes: M Barberà and D Martínez-Torres; cuticular proteins: M van Munster and Y Rahbé; developmental genes: H Lu and Y Hsiao; odorant and gustatory receptors: C Meslin, E Jacquin-Joly, and N Montagné; odorant, gustatory ligand-binding and extracellular-binding families: P Escuer, S Hinojosa-Alvarez, P Librado, J Rozas, A Sánchez-Gracia, and J F. Sánchez-Herrero; detoxification genes: F Hilliou, Gaëlle Le Goff, Thomas Chertemps, and M Maïbèche; effectors: PD Nabity, C Rispe, W Tian, J Wemmer, and C Zhao; invasion routes: F Delmotte, F Legeai, E Lombaert, and C Rispe; sampling: PD Nabity, L Delière, IB Scatoni, A Forneck, and PA Umina. The authors read and approved the final manuscript.

Authors' information
Claude Rispe and Fabrice Legeai are shared first authors and contributed equally to the work. François Delmotte and Denis Tagu are shared last authors and contributed equally to the work. The third author, Paul D Nabity, led the analyses on putative effectors. The fourth author, Rosa Fernández, led the analyses on duplications. All other contributors are listed in alphabetical order.

Funding
This work has been funded by INRAE (France) and by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement no. 764840 for the ITN IGNITE project. Rosa Fernandez was funded by a Juan de la Cierva-Incorporación Fellowship (Government of Spain, IJCI-2015-26627) and a Marie Skłodowska-Curie Fellowship (747607). Angela Douglas was supported by the US National Institute of Food and Agriculture Grant 12216941. Honglin Feng was supported by a University of Miami Maytag Fellowship, William H. Evoy Graduate Research Support Fund, and a Molecular Biosciences Graduate Research Award from the Department of Biology.

Availability of data and materials
All data generated or analyzed during this study are included in this published article and its supplementary information files. Reads produced in this study are available at the NCBI Short Read Archive (SRA) under accession PRJNA588186 [161] for reads used for the reference genome, and under accession PRJNA588387 [162] for reads used for the population genomics study. Other datasets (assembled sequence, official gene sets, microRNAs, mitochondrial genome) are available at the Aphidbase repository [163].

Ethics approval and consent to participate
Not applicable

Competing interests
The authors declare that they have no competing interests.

Author details
[1]BIOEPAR, INRAE, Oniris, Nantes, France. [2]BIPAA, IGEPP, Agrocampus Ouest, INRAE, Université de Rennes 1, 35650 Le Rheu, France. [3]Department of Botany and Plant Sciences, University of California, Riverside, USA. [4]Bioinformatics and Genomics Unit, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology, Dr. Aiguader, 88, 08003 Barcelona, Spain. [5]Present address: Institute of Evolutionary Biology (CSIC-UPF), Passeig marítim de la Barceloneta 37-49, 08003 Barcelona, Spain. [6]Department of Entomology, Cornell University, Ithaca, NY 14853, USA. [7]Univ Lyon, INSA-Lyon, INRAE, BF2I, UMR0203, F-69621, Villeurbanne, France. [8]Department of Biology, University of Miami, Coral Gables, FL 33146, USA. [9]Facultad de Agronomía, Montevideo, Uruguay. [10]Institut de Biologia Integrativa de Sistemes, Parc Científic Universitat de Valencia, C/ Catedrático José Beltrán n° 2, 46980 Paterna, València, Spain. [11]Université de Tunis El Manar, Faculté des Sciences de Tunis, LR01ES05 Biochimie et Biotechnologie, 2092 Tunis, Tunisia. [12]Department Biol, Univ Rochester, Rochester, NY 14627, USA. [13]Laboratoire Evolution, Génomes, Comportement, Ecologie CNRS, Univ. Paris-Sud, IRD, Université Paris-Saclay, Gif-sur-Yvette, France. [14]LIPM, Université de Toulouse, INRAE, CNRS, Castanet-Tolosan, France. [15]Sorbonne Université, UPEC, Université Paris 7, INRAE, CNRS, IRD, Institute of Ecology and Environmental Sciences, Paris, France. [16]SAVE, INRAE, Bordeaux Sciences Agro, Villenave d'Ornon, France. [17]Department of Molecular Biology and Genetics, Cornell University, Ithaca, NY 14853, USA. [18]Department of Cell and Molecular Biology, College of the Environment and Life Sciences, University of Rhode Island, Kingston, RI, USA. [19]Departament de Genètica, Microbiologia i Estadística and Institut de Recerca de la Biodiversitat (IRBio), Universitat de Barcelona, 08028 Barcelona, Spain. [20]Department of Biology, University of Miami, Coral Gables, USA. [21]Current affiliation: Boyce Thompson Institute for Plant Research, Cornell University, Ithaca, USA. [22]Universität für Bodenkultur (BOKU), Vienna, Austria. [23]Universitat Pompeu Fabra, 08003 Barcelona, Spain. [24]Institució Catalana de Recerca i Estudis Avançats (ICREA), Pg. Lluís Companys 23, 08010 Barcelona, Spain. [25]Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Barcelona, Spain. [26]Universitat Pompeu Fabra (UPF), Barcelona, Spain. [27]Université Côte d'Azur, INRAE, CNRS, Institut Sophia Agrobiotech, Sophia-Antipolis, France. [28]Institute of Biotechnology and Department of Entomology, College of Bioresources and Agriculture, National Taiwan University, Taipei, Taiwan. [29]Present affiliation: Bone and Joint Research Center, Chang Gung Memorial Hospital, Taoyuan, Taiwan. [30]IGEPP, Agrocampus Ouest, INRAE, Université de Rennes 1, 35650 Le Rheu, France. [31]INRAE, Institute of Ecology and Environmental Sciences, Versailles, France. [32]Department of Entomology, Texas A&M University, College Station, TX 77843, USA. [33]Université Côte d'Azur, INRAE, CNRS, Institut Sophia Agrobiotech, Sophia-Antipolis, France. [34]Laboratoire d'Anthropobiologie Moléculaire et d'Imagerie de Synthèse, CNRS UMR 5288, Université de Toulouse, Université Paul Sabatier, Toulouse, France. [35]China National GeneBank-Shenzhen, BGI-Shenzhen, Shenzhen 518083, Guangdong Province, People's Republic of China. [36]BGI-Shenzhen, Shenzhen 518083,

A

Guangdong Province, People's Republic of China. [37]Department of Entomology, College of Plant Protection, China Agricultural University, Beijing 100193, People's Republic of China. [38]Université Côte d'Azur, INRAE, CNRS, ISA, Sophia Antipolis, France. [39]Department of Post-Modern Agriculture, MingDao University, Changhua, Taiwan. [40]Institut de Biologia Integrativa de Sistemes, Parc Cientific Universitat de Valencia, C/ Catedrático José Beltrán n° 2, 46980 Paterna, València, Spain. [41]Sorbonne Université, Institute of Ecology and Environmental Sciences, Paris, France. [42]Department of Integrative Biology, University of Texas at Austin, Austin, USA. [43]Univ Lyon, INRAE, INSA-Lyon, CNRS, UCBL, UMR5240 MAP, F-69622 Villeurbanne, France. [44]BIPAA IGEPP, Agrocampus Ouest, INRAE, Université de Rennes 1, 35650 Le Rheu, France. [45]Plateforme Génomique GeT-PlaGe, Centre INRAE de Toulouse Midi-Pyrénées, 24 Chemin de Borde Rouge, Auzeville, CS 52627, 31326 Castanet-Tolosan Cedex, France. [46]Division of Genetics, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115, USA. [47]Facultad de Agronoía, Montevideo, Uruguay. [48]School of BioSciences, The University of Melbourne, Parkville, VIC, Australia. [49]BGPI, Université Montpellier, CIRAD, INRAE, Montpellier SupAgro, Montpellier, France.

Received: 16 December 2019 Accepted: 22 June 2020
Published online: 23 July 2020

References

1. Pimentel D, Zuniga R, Morrison D. Update on the environmental and economic costs associated with alien-invasive species in the United States. Ecol Econ. 2005;52:273–88 https://doi.org/10.1016/j.ecolecon.2004.10.002.
2. Gurevitch J, Fox GA, Wardle GM, Inderjit TD. Emergent insights from the synthesis of conceptual frameworks for biological invasions. Ecol Lett. 2011; 14:407–18 https://doi.org/10.1111/j.1461-0248.2011.01594.x.
3. Kueffer C, Pyšek P, Richardson DM. Integrative invasion science: model systems, multi-site studies, focused meta-analysis and invasion syndromes. New Phytol. 2013;200:615–33 https://doi.org/10.1111/nph.12415.
4. Banerjee A, Duflo E, Postei-Vinay G, Watts T. Long-run health impacts of income shocks: wine and phylloxera in nineteenth-century France. Rev Econ Stat. 2010;92:714–28.
5. Simberloff D. Non-native invasive species and novel ecosystems. F1000Prime Rep 2015;7. https://doi.org/10.12703/P7-47.
6. Galet P. Phylloxera. Mal. Parasites Vigne Tome II, Montpellier: Paysan du Midi; 1982, p. 1059–313.
7. Pouget R. Le Phylloxera et les maladies de la vigne. Edilivre-Aparis; 2015.
8. Carton Y, Sorensen C, Smith J, Smith E. Une coopération exemplaire entre entomologistes français et américains pendant la crise du Phylloxera en France (1868–1895). Ann Société Entomol Fr NS. 2007;43:103–25 https://doi.org/10.1080/00379271.2007.10697500.
9. Marchal P, Feytaud J. Les données nouvelles sur le phylloxéra. Rev Vitic - Tome XL Ed P Viala. 1913.
10. Moran NA, McLaughlin HJ, Sorek R. The dynamics and time scale of ongoing genomic erosion in symbiotic bacteria. Science. 2009;323:379–82 https://doi.org/10.1126/science.1167140.
11. Riaz S, Lund KT, Granett J, Walker MA. Population diversity of grape phylloxera in California and evidence for sexual reproduction. Am J Enol Vitic. 2017;68:218–27 https://doi.org/10.5344/ajev.2016.15114.
12. Granett J, Walker MA, Kocsis L, Omer AD. Biology and management of grape phylloxera. Annu Rev Entomol. 2001;46:387–412 https://doi.org/10.1146/annurev.ento.46.1.387.
13. The International Aphid Genomics Consortium. Genome sequence of the pea aphid Acyrthosiphon pisum. PLoS Biol. 2010;8:e1000313 https://doi.org/10.1371/journal.pbio.1000313.
14. Wenger JA, Cassone BJ, Legeai F, Johnston JS, Bansal R, Yates AD, et al. Whole genome sequence of the soybean aphid, Aphis glycines Insect Biochem Mol Biol 2017. https://doi.org/10.1016/j.ibmb.2017.01.005.
15. Nováková E, Moran NA. Diversification of genes for carotenoid biosynthesis in aphids following an ancient transfer from a fungus. Mol Biol Evol. 2012; 29:313–23 https://doi.org/10.1093/molbev/msr206.
16. Zhao C, Nabity PD. Phylloxerids share ancestral carotenoid biosynthesis genes of fungal origin with aphids and adelgids. PLoS One. 2017;12: e0185484 https://doi.org/10.1371/journal.pone.0185484.
17. Nikoh N, Nakabachi A. Aphids acquired symbiotic genes via lateral gene transfer. BMC Biol. 2009;7:12 https://doi.org/10.1186/1741-7007-7-12.
18. Nikoh N, McCutcheon JP, Kudo T, Miyagishima S, Moran NA, Nakabachi A. Bacterial genes in the aphid genome: absence of functional gene transfer from Buchnera to its host. PLoS Genet. 2010;6:e1000827 https://doi.org/10.1371/journal.pgen.1000827.
19. Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, et al. A unified classification system for eukaryotic transposable elements. Nat Rev Genet. 2007;8:973–82 https://doi.org/10.1038/nrg2165.
20. Xie W, Chen C, Yang Z, Guo L, Yang X, Wang D, et al. Genome sequencing of the sweetpotato whitefly Bemisia tabaci MED/Q. GigaScience. 2017;6:1–7 https://doi.org/10.1093/gigascience/gix018.
21. Legeai F, Shigenobu S, Gauthier J-P, Colbourne J, Rispe C, Collin O, et al. AphidBase: a centralized bioinformatic resource for annotation of the pea aphid genome. Insect Mol Biol. 2010;19:5–12 https://doi.org/10.1111/j.1365-2583.2009.00930.x.
22. Giordano R, Donthu RK, Zimin AV, Julca Chavez IC, Gabaldon T, van Munster M, et al. Soybean aphid biotype 1 genome: insights into the invasive biology and adaptive evolution of a major agricultural pest. Insect Biochem Mol Biol. 2020;120:103334 https://doi.org/10.1016/j.ibmb.2020.103334.
23. Li Y, Park H, Smith TE, Moran NA. Gene family evolution in the pea aphid based on chromosome-level genome assembly. Mol Biol Evol. 2019;36: 2143–56 https://doi.org/10.1093/molbev/msz138.
24. Mathers TC, Chen Y, Kaithakottil G, Legeai F, Mugford ST, Baa-Puyoulet P, et al. Rapid transcriptional plasticity of duplicated gene clusters enables a clonally reproducing aphid to colonise diverse plant species. Genome Biol. 2017;18:27 https://doi.org/10.1186/s13059-016-1145-3.
25. Buchner P. Endosymbiosis of animals with plant microorganisms, vol. 7. New York: John Wiley & Sons; 1965.
26. Ponsen MB. A histological description of the alimentary tract and related organs of Phylloxeridae (Homoptera, Aphidoidea). Wageningen: Wageningen Agricultural University; 1997.
27. Vorwerk S, Martinez-Torres D, Forneck A. Pantoea agglomerans-associated bacteria in grape phylloxera (Daktulosphaira vitifoliae, Fitch). Agric For Entomol. 2007;9:57–64 https://doi.org/10.1111/j.1461-9563.2006.000319.x.
28. Kellow AV, Sedgley M, Van Heeswijck R. Interaction between Vitis vinifera and grape phylloxera: changes in root tissue during nodosity formation. Ann Bot. 2004;93:581–90 https://doi.org/10.1093/aob/mch082.
29. Johnson S, Hiltpold I, Turlings T. Behaviour and physiology of root herbivores, volume 45. 1st ed; 2013.
30. Zhao C, Nabity PD. Plant manipulation through gall formation constrains amino acid transporter evolution in sap-feeding insects. BMC Evol Biol. 2017;17 https://doi.org/10.1186/s12862-017-1000-5.
31. Arp AP, Hunter WB, Pelz-Stelinski KS. Annotation of the Asian citrus psyllid genome reveals a reduced innate immune system. Front Physiol. 2016;7 https://doi.org/10.3389/fphys.2016.00570.
32. Arp AP, Martini X, Pelz-Stelinski KS. Innate immune system capabilities of the Asian citrus psyllid, Diaphorina citri. J Invertebr Pathol. 2017;148:94–101 https://doi.org/10.1016/j.jip.2017.06.002.
33. Salcedo-Porras N, Guarneri A, Oliveira PL, Lowenberger C. Rhodnius prolixus: identification of missing components of the IMD immune signaling pathway and functional characterization of its role in eliminating bacteria. PLoS One. 2019;14:e0214794 https://doi.org/10.1371/journal.pone.0214794.
34. Davis GK. Cyclical parthenogenesis and viviparity in aphids as evolutionary novelties. J Exp Zoolog B Mol Dev Evol. 2012;318:448–59 https://doi.org/10.1002/jez.b.22441.
35. Mesquita RD, Vionette-Amaral RJ, Lowenberger C, Rivera-Pomar R, Monteiro FA, Minx P, et al. Genome of Rhodnius prolixus, an insect vector of Chagas disease, reveals unique adaptations to hematophagy and parasite infection. Proc Natl Acad Sci U S A. 2015;112:14936–41 https://doi.org/10.1073/pnas.1506226112.
36. Smadja C, Shi P, Butlin RK, Robertson HM. Large gene family expansions and adaptive evolution for odorant and gustatory receptors in the pea aphid, Acyrthosiphon pisum. Mol Biol Evol. 2009;26:2073–86 https://doi.org/10.1093/molbev/msp116.
37. 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: 1221–36 https://doi.org/10.1093/gbe/evy081.
38. Zhao J, Zhang Y, Fan D, Feng J. Identification and expression profiling of odorant-binding proteins and chemosensory proteins of Daktulosphaira vitifoliae (Hemiptera: Phylloxeridae). J Econ Entomol. 2017;110:1813–20 https://doi.org/10.1093/jee/tox121.

39. Ramsey JS, Rider DS, Walsh TK, Vos MD, Gordon KHJ, Ponnala L, et al. Comparative analysis of detoxification enzymes in Acyrthosiphon pisum and Myzus persicae. Insect Mol Biol. 2010;19:155–64 https://doi.org/10.1111/j.1365-2583.2009.00973.x.
40. Schama R, Pedrini N, Juárez MP, Nelson DR, Torres AQ, Valle D, et al. Rhodnius prolixus supergene families of enzymes potentially associated with insecticide resistance. Insect Biochem Mol Biol. 2016;69:91–104 https://doi.org/10.1016/j.ibmb.2015.06.005.
41. Grbić M, Van Leeuwen T, Clark RM, Rombauts S, Rouzé P, Grbić V, et al. The genome of *Tetranychus urticae* reveals herbivorous pest adaptations. Nature. 2011;479:487–92 https://doi.org/10.1038/nature10640.
42. Feyereisen R. Arthropod CYPomes illustrate the tempo and mode in P450 evolution. Biochim Biophys Acta BBA - Proteins Proteomics. 1814;2011:19–28 https://doi.org/10.1016/j.bbapap.2010.06.012.
43. Chen W, Hasegawa DK, Kaur N, Kliot A, Pinheiro PV, Luan J, et al. The draft genome of whitefly Bemisia tabaci MEAM1, a global crop pest, provides novel insights into virus transmission, host adaptation, and insecticide resistance. BMC Biol. 2016;14 https://doi.org/10.1186/s12915-016-0321-y.
44. Kozomara A, Griffiths-Jones S. miRBase: annotating high confidence microRNAs using deep sequencing data. Nucleic Acids Res. 2014;42:D68–73 https://doi.org/10.1093/nar/gkt1181.
45. Campbell EM, Ball A, Hoppler S, Bowman AS. Invertebrate aquaporins: a review. J Comp Physiol B. 2008;178:935–55 https://doi.org/10.1007/s00360-008-0288-2.
46. Jing X, White TA, Luan J, Jiao C, Fei Z, Douglas AE. Evolutionary conservation of candidate osmoregulation genes in plant phloem sap-feeding insects. Insect Mol Biol. 2016;25:251–8 https://doi.org/10.1111/imb.12215.
47. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30:772–80 https://doi.org/10.1093/molbev/mst010.
48. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses Bioinformatics 2009;25:1972–1973. https://doi.org/10.1093/bioinformatics/btp348.
49. Ronquist F, Huelsenbeck JP. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinforma Oxf Engl. 2003;19:1572–4.
50. Abascal F, Zardoya R, Posada D. ProtTest: selection of best-fit models of protein evolution. Bioinformatics. 2005;21:2104–5 https://doi.org/10.1093/bioinformatics/bti263.
51. Doherty CJ, Kay SA. Circadian control of global gene expression patterns. Annu Rev Genet. 2010;44:419–44 https://doi.org/10.1146/annurev-genet-102209-163432.
52. Barberà M, Collantes-Alegre JM, Martínez-Torres D. Characterisation, analysis of expression and localisation of circadian clock genes from the perspective of photoperiodism in the aphid Acyrthosiphon pisum. Insect Biochem Mol Biol. 2017;83:54–67 https://doi.org/10.1016/j.ibmb.2017.02.006.
53. Cortés T, Ortiz-Rivas B, Martínez-Torres D. Identification and characterization of circadian clock genes in the pea aphid Acyrthosiphon pisum. Insect Mol Biol. 2010;19:123–39 https://doi.org/10.1111/j.1365-2583.2009.00931.x.
54. Yuan Q, Metterville D, Briscoe AD, Reppert SM. Insect cryptochromes: gene duplication and loss define diverse ways to construct insect circadian clocks. Mol Biol Evol. 2007;24:948–55 https://doi.org/10.1093/molbev/msm011.
55. Barberà M, Martínez-Torres D. Identification of the prothoracicotropic hormone (Ptth) coding gene and localization of its site of expression in the pea aphid Acyrthosiphon pisum. Insect Mol Biol. 2017;26:654–64 https://doi.org/10.1111/imb.12326.
56. Keller O, Odronitz F, Stanke M, Kollmar M, Waack S. Scipio: using protein sequences to determine the precise exon/intron structures of genes and their orthologs in closely related species. BMC Bioinformatics. 2008;9:278 https://doi.org/10.1186/1471-2105-9-278.
57. Slater GSC, Birney E. Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics. 2005;6:31 https://doi.org/10.1186/1471-2105-6-31.
58. Birney E, Clamp M, Durbin R. GeneWise and Genomewise. Genome Res. 2004;14:988–95 https://doi.org/10.1101/gr.1865504.
59. Lee E, Helt GA, Reese JT, Munoz-Torres MC, Childers CP, Buels RM, et al. Web Apollo: a web-based genomic annotation editing platform. Genome Biol. 2013;14 https://doi.org/10.1186/gb-2013-14-8-r93.
60. Cao D, Liu Y, Walker WB, Li J, Wang G. Molecular characterization of the Aphis gossypii olfactory receptor gene families. PLoS One. 2014;9 https://doi.org/10.1371/journal.pone.0101187.
61. Katoh K, Kuma K, Toh H, Miyata T. MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res. 2005;33:511–8 https://doi.org/10.1093/nar/gki198.
62. Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol. 2010;59:307–21 https://doi.org/10.1093/sysbio/syq010.
63. Lefort V, Longueville J-E, Gascuel O. SMS: smart model selection in PhyML. Mol Biol Evol. 2017;34:2422–4 https://doi.org/10.1093/molbev/msx149.
64. Anisimova M, Gascuel O. Approximate likelihood-ratio test for branches: a fast, accurate, and powerful alternative. Syst Biol. 2006;55:539–52 https://doi.org/10.1080/10635150600755453.
65. Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, et al. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res. 2016;44:D279–85 https://doi.org/10.1093/nar/gkv1344.
66. Zhou J-J, Vieira FG, He X-L, Smadja C, Liu R, Rozas J, et al. Genome annotation and comparative analyses of the odorant-binding proteins and chemosensory proteins in the pea aphid Acyrthosiphon pisum. Insect Mol Biol. 2010;19:113–22 https://doi.org/10.1111/j.1365-2583.2009.00919.x.
67. Yamada KD, Tomii K, Katoh K. Application of the MAFFT sequence alignment program to large data—reexamination of the usefulness of chained guide trees. Bioinformatics. 2016;32:3246–51 https://doi.org/10.1093/bioinformatics/btw412.
68. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30:1312–3 https://doi.org/10.1093/bioinformatics/btu033.
69. Sawyer S. Statistical tests for detecting gene conversion. Mol Biol Evol. 1989;6:526–38 https://doi.org/10.1093/oxfordjournals.molbev.a040567.
70. Letunic I, Bork P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Res. 2016;44:W242–5 https://doi.org/10.1093/nar/gkw290.
71. Werck-Reichhart D, Feyereisen R. Cytochromes P450: a success story. Genome Biol 2000;1:reviews3003.1-reviews3003.9.
72. Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, Shah P, et al. Galaxy: a platform for interactive large-scale genome analysis. Genome Res. 2005;15:1451–5 https://doi.org/10.1101/gr.4086505.
73. Du Y-P, Zheng Q-L, Zhai H, Jiang E-S, Wang Z-Y. Selectivity of Phylloxera viticola Fitch (Homoptera: Phylloxeridae) to grape with different resistance and the identification of grape root volatiles. Acta Entomol Sin. 2009;52:537–43.
74. Zhang R, Wang B, Grossi G, Falabella P, Liu Y, Yan S, et al. Molecular basis of alarm pheromone detection in aphids. Curr Biol. 2017;27:55–61 https://doi.org/10.1016/j.cub.2016.10.013.
75. Pelosi P, Iovinella I, Felicioli A, Dani FR. Soluble proteins of chemical communication: an overview across arthropods. Front Physiol. 2014;5 https://doi.org/10.3389/fphys.2014.00320.
76. Xue W, Fan J, Zhang Y, Xu Q, Han Z, Sun J, et al. Identification and expression analysis of candidate odorant-binding protein and chemosensory protein genes by antennal transcriptome of Sitobion avenae. PLoS One. 2016;11 https://doi.org/10.1371/journal.pone.0161839.
77. Borden KLB. RING domains: master builders of molecular scaffolds?11Edited by P. E Wright J Mol Biol. 2000;295:1103–12 https://doi.org/10.1006/jmbi.1999.3429.
78. Mosavi LK, Cammett TJ, Desrosiers DC, Peng Z. The ankyrin repeat as molecular architecture for protein recognition. Protein Sci Publ Protein Soc. 2004;13:1435–48 https://doi.org/10.1110/ps.03554604.
79. Carafoli E, Krebs J. Why calcium? How calcium became the best communicator. J Biol Chem. 2016;291:20849–57 https://doi.org/10.1074/jbc.R116.735894.
80. Teixeira LK, Reed SI. Ubiquitin ligases and cell cycle control. Annu Rev Biochem. 2013;82:387–414 https://doi.org/10.1146/annurev-biochem-060410-105307.
81. Zhao C, Escalante LN, Chen H, Benatti TR, Qu J, Chellapilla S, et al. A massive expansion of effector genes underlies gall-formation in the wheat pest Mayetiola destructor. Curr Biol. 2015;25:613–20 https://doi.org/10.1016/j.cub.2014.12.057.
82. Pitino M, Hogenhout SA. Aphid protein effectors promote aphid colonization in a plant species-specific manner. Mol Plant-Microbe Interact. 2012;26:130–9 https://doi.org/10.1094/MPMI-07-12-0172-FI.
83. Jeffares DC, Penkett CJ, Bähler J. Rapidly regulated genes are intron poor. Trends Genet. 2008;24:375–8 https://doi.org/10.1016/j.tig.2008.05.006.

84. Luo H, Nijveen H. Understanding and identifying amino acid repeats. Brief Bioinform. 2014;15:582–91 https://doi.org/10.1093/bib/bbt003.
85. Craig A, Ewan R, Mesmar J, Gudipati V, Sadanandom A. E3 ubiquitin ligases and plant innate immunity. J Exp Bot. 2009;60:1123–32 https://doi.org/10.1093/jxb/erp059.
86. Banfield MJ. Perturbation of host ubiquitin systems by plant pathogen/pest effector proteins. Cell Microbiol. 2015;17:18–25 https://doi.org/10.1111/cmi.12385.
87. Stuart J. Insect effectors and gene-for-gene interactions with host plants. Curr Opin Insect Sci. 2015;9:56–61 https://doi.org/10.1016/j.cois.2015.02.010.
88. Nabity PD. Insect-induced plant phenotypes: revealing mechanisms through comparative genomics of galling insects and their hosts. Am J Bot. 2016;103:979–81 https://doi.org/10.3732/ajb.1600111.
89. Will T, Tjallingii WF, Thönnessen A, van Bel AJE. Molecular sabotage of plant defense by aphid saliva. Proc Natl Acad Sci U S A. 2007;104:10536–41 https://doi.org/10.1073/pnas.0703535104.
90. Hattori M, Nakamura M, Komatsu S, Tsuchihara K, Tamura Y, Hasegawa T. Molecular cloning of a novel calcium-binding protein in the secreted saliva of the green rice leafhopper Nephotettix cincticeps. Insect Biochem Mol Biol. 2012;42:1–9 https://doi.org/10.1016/j.ibmb.2011.10.001.
91. Zhao C, Rispe C, Nabity PD. Secretory RING finger proteins function as effectors in a grapevine galling insect. BMC Genomics. 2019;20:923. https://doi.org/10.1186/s12864-019-6313-x.
92. Lilley CJ, Maqbool A, Wu D, Yusup HB, Jones LM, Birch PRJ, et al. Effector gene birth in plant parasitic nematodes: neofunctionalization of a housekeeping glutathione synthetase gene. PLoS Genet. 2018;14:e1007310 https://doi.org/10.1371/journal.pgen.1007310.
93. Boulain H, Legeai F, Guy E, Morlière S, Douglas NE, Oh J, et al. Fast evolution and lineage-specific gene family expansions of aphid salivary effectors driven by interactions with host-plants. Genome Biol Evol. 2018;10:1554–72 https://doi.org/10.1093/gbe/evy097.
94. Downie DA. Locating the sources of an invasive pest, grape phylloxera, using a mitochondrial DNA gene genealogy. Mol Ecol. 2002;11:2013–26 https://doi.org/10.1046/j.1365-294X.2002.01584.x.
95. Tello J, Mammerler R, Čajić M, Forneck A. Major outbreaks in the nineteenth century shaped grape phylloxera contemporary genetic structure in Europe. Sci Rep. 2019;9:1–11 https://doi.org/10.1038/s41598-019-54122-0.
96. Planchon J, Lichtenstein J. Le Phylloxéra (de 1854 à 1873, résumé pratique et scientifique); 1873.
97. Lund KT, Riaz S, Walker MA. Population structure, diversity and reproductive mode of the grape phylloxera (Daktulosphaira vitifoliae) across its native range. PLoS One. 2017;12:e0170678 https://doi.org/10.1371/journal.pone.0170678.
98. Hebert PDN, Stoeckle MY, Zemlak TS, Francis CM. Identification of birds through DNA barcodes. PLoS Biol. 2004;2 https://doi.org/10.1371/journal.pbio.0020312.
99. Johnston JS, Bernardini A, Hjelmen CE. Genome size estimation and quantitative cytogenetics in insects. In: Brown SJ, Pfrender ME, editors. Insect genomics methods Protoc. New York: Springer New York; 2019. p. 15–26. https://doi.org/10.1007/978-1-4939-8775-7_2.
100. Bonnivard E, Catrice O, Ravaux J, Brown SC, Higuet D. Survey of genome size in 28 hydrothermal vent species covering 10 families. Genome. 2009;52:524–36 https://doi.org/10.1139/G09-027.
101. Sunnucks P, Hales DF. Numerous transposed sequences of mitochondrial cytochrome oxidase I-II in aphids of the genus Sitobion (Hemiptera: Aphididae). Mol Biol Evol. 1996;13:510–24 https://doi.org/10.1093/oxfordjournals.molbev.a025612.
102. Boetzer M, Pirovano W. SSPACE-LongRead: scaffolding bacterial draft genomes using long read sequence information. BMC Bioinformatics. 2014;15:211 https://doi.org/10.1186/1471-2105-15-211.
103. Boetzer M, Pirovano W. Toward almost closed genomes with GapFiller. Genome Biol. 2012;13:R56 https://doi.org/10.1186/gb-2012-13-6-r56.
104. Holt C, Yandell M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics. 2011;12:491 https://doi.org/10.1186/1471-2105-12-491.
105. Keller O, Kollmar M, Stanke M, Waack S. A novel hybrid gene prediction method employing protein multiple sequence alignments. Bioinformatics. 2011;27:757–63 https://doi.org/10.1093/bioinformatics/btr010.
106. Korf I. Gene finding in novel genomes. BMC Bioinformatics. 2004;5:59 https://doi.org/10.1186/1471-2105-5-59.
107. Rispe C, Legeai F, Papura D, Bretaudeau A, Hudaverdian S, Le Trionnaire G, et al. De novo transcriptome assembly of the grapevine phylloxera allows identification of genes differentially expressed between leaf- and root-feeding forms. BMC Genomics. 2016;17 https://doi.org/10.1186/s12864-016-2530-8.
108. Misof B, Liu S, Meusemann K, Peters RS, Donath A, Mayer C, et al. Phylogenomics resolves the timing and pattern of insect evolution. Science. 2014;346:763–7 https://doi.org/10.1126/science.1257570.
109. Buels R, Yao E, Diesh CM, Hayes RD, Munoz-Torres M, Helt G, et al. JBrowse: a dynamic web platform for genome visualization and analysis. Genome Biol. 2016;17:66 https://doi.org/10.1186/s13059-016-0924-1.
110. Mungall CJ, Emmert DB. A Chado case study: an ontology-based modular schema for representing genome-associated biological information. Bioinformatics. 2007;23:i337–46 https://doi.org/10.1093/bioinformatics/btm189.
111. Sanderson L-A, Ficklin SP, Cheng C-H, Jung S, Feltus FA, Bett KE, et al. Tripal v1.1: a standards-based toolkit for construction of online genetic and genomic databases. Database 2013;2013. https://doi.org/10.1093/database/bat075.
112. Bretaudeau A, Dunn N, Gladman S, Grüning B, Rasche H, Seemann T. Galaxy Genome Annotation project: integrating Galaxy and GMOD for genome annotation . F1000Research 2018;7. https://doi.org/10.7490/f1000research.1116180.1.
113. Laetsch DR, Blaxter ML. BlobTools: interrogation of genome assemblies. F1000Research 2017;6:1287. https://doi.org/10.12688/f1000research.12232.1.
114. 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;69:313–9 https://doi.org/10.1016/j.ympev.2012.08.023.
115. Laslett D, Canbäck B. ARWEN: a program to detect tRNA genes in metazoan mitochondrial nucleotide sequences. Bioinformatics. 2008;24:172–5 https://doi.org/10.1093/bioinformatics/btm573.
116. Quesneville H, Bergman CM, Andrieu O, Autard D, Nouaud D, Ashburner M, et al. Combined evidence annotation of transposable elements in genome sequences. PLoS Comput Biol. 2005;1 https://doi.org/10.1371/journal.pcbi.0010022.
117. Flutre T, Duprat E, Feuillet C, Quesneville H. Considering transposable element diversification in de novo annotation approaches. PLoS One. 2011;6:e16526 https://doi.org/10.1371/journal.pone.0016526.
118. Huerta-Cepas J, Capella-Gutiérrez S, Pryszcz LP, Marcet-Houben M, Gabaldón T. PhylomeDB v4: zooming into the plurality of evolutionary histories of a genome. Nucleic Acids Res. 2014;42:D897–902 https://doi.org/10.1093/nar/gkt1177.
119. Landan G, Graur D. Heads or tails: a simple reliability check for multiple sequence alignments. Mol Biol Evol. 2007;24:1380–3 https://doi.org/10.1093/molbev/msm060.
120. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32:1792–7 https://doi.org/10.1093/nar/gkh340.
121. Lassmann T, Sonnhammer EL. Kalign – an accurate and fast multiple sequence alignment algorithm. BMC Bioinformatics. 2005;6:298 https://doi.org/10.1186/1471-2105-6-298.
122. Wallace IM, O'Sullivan O, Higgins DG, Notredame C. M-Coffee: combining multiple sequence alignment methods with T-Coffee. Nucleic Acids Res. 2006;34:1692–9 https://doi.org/10.1093/nar/gkl091.
123. Gascuel O. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. Mol Biol Evol. 1997;14:685–95 https://doi.org/10.1093/oxfordjournals.molbev.a025808.
124. Akaike H. Information theory and an extension of the maximum likelihood principle. Proc. 2nd Int. Symp. Inf. Theory, Petrov, CN, Csaki, F; 2009.
125. Huerta-Cepas J, Serra F, Bork P. ETE 3: reconstruction, analysis, and visualization of phylogenomic data. Mol Biol Evol. 2016;33:1635–8 https://doi.org/10.1093/molbev/msw046.
126. Price MN, Dehal PS, Arkin AP. FastTree 2 – approximately maximum-likelihood trees for large alignments. PLoS One. 2010;5 https://doi.org/10.1371/journal.pone.0009490.
127. Le SQ, Gascuel O. An improved general amino acid replacement matrix. Mol Biol Evol. 2008;25:1307–20 https://doi.org/10.1093/molbev/msn067.
128. Wehe A, Bansal MS, Burleigh JG, Eulenstein O. DupTree: a program for large-scale phylogenetic analyses using gene tree parsimony. Bioinformatics. 2008;24:1540–1 https://doi.org/10.1093/bioinformatics/btn230.

129. Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, et al. InterProScan 5: genome-scale protein function classification. Bioinformatics. 2014;30: 1236–40 https://doi.org/10.1093/bioinformatics/btu031.
130. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. Nat Genet. 2000;25:25–9 https://doi.org/10.1038/75556.
131. Al-Shahrour F, Díaz-Uriarte R, Dopazo J. FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. Bioinformatics. 2004;20:578–80 https://doi.org/10.1093/bioinformatics/btg455.
132. Supek F, Bošnjak M, Škunca N, Šmuc T. REVIGO summarizes and visualizes long lists of gene ontology terms. PLoS One. 2011;6 https://doi.org/10.1371/journal.pone.0021800.
133. Aury J-M, Jaillon O, Duret L, Noel B, Jubin C, Porcel BM, et al. Global trends of whole-genome duplications revealed by the ciliate Paramecium tetraurelia. Nature. 2006;444:171 https://doi.org/10.1038/nature05230.
134. Castresana J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol. 2000;17:540–52 https://doi.org/10.1093/oxfordjournals.molbev.a026334.
135. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 2007;24:1586–91 https://doi.org/10.1093/molbev/msm088.
136. Vellozo AF, Véron AS, Baa-Puyoulet P, Huerta-Cepas J, Cottret L, Febvay G, et al. CycADS: an annotation database system to ease the development and update of BioCyc databases. Database. 2011;2011 https://doi.org/10.1093/database/bar008.
137. Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics. 2005;21:3674–6 https://doi.org/10.1093/bioinformatics/bti610.
138. Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M. KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic Acids Res. 2007;35:W182–5 https://doi.org/10.1093/nar/gkm321.
139. Claudel-Renard C, Chevalet C, Faraut T, Kahn D. Enzyme-specific profiles for genome annotation: PRIAM. Nucleic Acids Res. 2003;31:6633–9 https://doi.org/10.1093/nar/gkg847.
140. Caspi R, Billington R, Fulcher CA, Keseler IM, Kothari A, Krummenacker M, et al. The MetaCyc database of metabolic pathways and enzymes. Nucleic Acids Res. 2018;46:D633–9 https://doi.org/10.1093/nar/gkx935.
141. Karp PD, Latendresse M, Paley SM, Krummenacker M, Ong QD, Billington R, et al. Pathway tools version 19.0 update: software for pathway/genome informatics and systems biology. Brief Bioinform. 2016;17:877–90 https://doi.org/10.1093/bib/bbv079.
142. Baa-Puyoulet P, Parisot N, Febvay G, Huerta-Cepas J, Vellozo AF, Gabaldón T, et al. ArthropodaCyc: a CycADS powered collection of BioCyc databases to analyse and compare metabolism of arthropods. Database. 2016;2016 https://doi.org/10.1093/database/baw081.
143. Steffensen JL, Dufault-Thompson K, Zhang Y. PSAMM: a portable system for the analysis of metabolic models. PLoS Comput Biol. 2016;12:e1004732 https://doi.org/10.1371/journal.pcbi.1004732.
144. Thiele I, Vlassis N, Fleming RMT. fastGapFill: efficient gap filling in metabolic networks. Bioinformatics. 2014;30:2529–31 https://doi.org/10.1093/bioinformatics/btu321.
145. Waterhouse RM, Kriventseva EV, Meister S, Xi Z, Alvarez KS, Bartholomay LC, et al. Evolutionary dynamics of immune-related genes and pathways in disease-vector mosquitoes. Science. 2007;316:1738–43 https://doi.org/10.1126/science.1139862.
146. Rebers JE, Riddiford LM. Structure and expression of a Manduca sexta larval cuticle gene homologous to Drosophila cuticle genes. J Mol Biol. 1988;203: 411–23 https://doi.org/10.1016/0022-2836(88)90009-5.
147. Ioannidou ZS, Theodoropoulou MC, Papandreou NC, Willis JH, Hamodrakas SJ. CutProtFam-Pred: detection and classification of putative structural cuticular proteins from sequence alone, based on profile Hidden Markov Models. Insect Biochem Mol Biol. 2014;52:51–9 https://doi.org/10.1016/j.ibmb.2014.06.004.
148. Gallot A, Rispe C, Leterme N, Gauthier J-P, Jaubert-Possamai S, Tagu D. Cuticular proteins and seasonal photoperiodism in aphids. Insect Biochem Mol Biol. 2010;40:235–40 https://doi.org/10.1016/j.ibmb.2009.12.001.
149. Nicholson SJ, Nickerson ML, Dean M, Song Y, Hoyt PR, Rhee H, et al. The genome of Diuraphis noxia, a global aphid pest of small grains. BMC Genomics. 2015;16 https://doi.org/10.1186/s12864-015-1525-1.
150. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol. 2011;7:539 https://doi.org/10.1038/msb.2011.75.
151. Dereeper A, Guignon V, Blanc G, Audic S, Buffet S, Chevenet F, et al. Phylogeny.fr: robust phylogenetic analysis for the non-specialist. Nucleic Acids Res. 2008;36:W465–9 https://doi.org/10.1093/nar/gkn180.
152. Labunskyy VM, Hatfield DL, Gladyshev VN. Selenoproteins: molecular pathways and physiological roles. Physiol Rev. 2014;94:739–77 https://doi.org/10.1152/physrev.00039.2013.
153. Chapple CE, Guigó R. Relaxation of selective constraints causes independent selenoprotein extinction in insect genomes. PLoS One. 2008;3:e2968 https://doi.org/10.1371/journal.pone.0002968.
154. Santesmasses D, Mariotti M, Guigó R. Computational identification of the selenocysteine tRNA (tRNASec) in genomes. PLoS Comput Biol. 2017;13: e1005383 https://doi.org/10.1371/journal.pcbi.1005383.
155. Otero L, Romanelli-Cedrez L, Turanov AA, Gladyshev VN, Miranda-Vizuete A, Salinas G. Adjustments, extinction, and remains of selenocysteine incorporation machinery in the nematode lineage. RNA. 2014;20:1023–34 https://doi.org/10.1261/rna.043877.113.
156. Mariotti M, Guigó R. Selenoprofiles: profile-based scanning of eukaryotic genome sequences for selenoprotein genes. Bioinformatics. 2010;26:2656–63 https://doi.org/10.1093/bioinformatics/btq516.
157. Villarroel CA, Jonckheere W, Alba JM, Glas JJ, Dermauw W, Haring MA, et al. Salivary proteins of spider mites suppress defenses in Nicotiana benthamiana and promote mite reproduction. Plant J. 2016;86:119–31 https://doi.org/10.1111/tpj.13152.
158. Oates CN, Denby KJ, Myburg AA, Slippers B, Naidoo S. Insect gallers and their plant hosts: from omics data to systems biology. Int J Mol Sci. 2016:1–14.
159. Kofler R, Orozco-terWengel P, Maio ND, Pandey RV, Nolte V, Futschik A, et al. PoPoolation: a toolbox for population genetic analysis of next generation sequencing data from pooled individuals. PLoS One. 2011;6:e15925 https://doi.org/10.1371/journal.pone.0015925.
160. Raynal L, Marin J-M, Pudlo P, Ribatet M, Robert CP, Estoup A. ABC random forests for Bayesian parameter inference. Bioinformatics. 2019;35:1720–8 https://doi.org/10.1093/bioinformatics/bty867.
161. Rispe C, Legeai F, Nabity PD, Fernández R, Arora AK, Baa-Puyoulet P, et al. The genome sequence of the grape phylloxera provides insights into the evolution, adaptation and invasion routes of an iconic pest. NCBI accession number PRJNA588186. https://www.ncbi.nlm.nih.gov/bioproject/PRJNA588186. 2020.
162. Rispe C, Legeai F, Nabity PD, Fernández R, Arora AK, Baa-Puyoulet P, et al. The genome sequence of the grape phylloxera provides insights into the evolution, adaptation and invasion routes of an iconic pest. NCBI accession number PRJNA588387. https://www.ncbi.nlm.nih.gov/bioproject/PRJNA588387. 2020.
163. Rispe C, Legeai F, Nabity PD, Fernández, R., Arora AK, Baa-Puyoulet, et al. The genome sequence of the grape phylloxera provides insights into the evolution, adaptation and invasion routes of an iconic pest. Aphidbase repository. https://bipaa.genouest.org/sp/daktulosphaira_vitifoliae/download/ 2020.

Publisher's Note

B

Smelling in the dark: Phylogenomic insights into the chemosensory system of a subterranean beetle

Received: 10 November 2020 | Revised: 22 March 2021 | Accepted: 26 March 2021

DOI: 10.1111/mec.15921

ORIGINAL ARTICLE

MOLECULAR ECOLOGY WILEY

Smelling in the dark: Phylogenomic insights into the chemosensory system of a subterranean beetle

Pau Balart-García[1] | **Alexandra Cieslak**[1] | **Paula Escuer**[2] | **Julio Rozas**[2] | **Ignacio Ribera**[1] | **Rosa Fernández**[1]

[1]Institute of Evolutionary Biology (CSIC - Universitat Pompeu Fabra), Barcelona, Spain

[2]Department of Genetics, Microbiology and Statistics, Institut de Recerca de la Biodiversitat (IRBio), Universitat de Barcelona, Barcelona, Spain

Correspondence
Pau Balart-García and Rosa Fernández, Institute of Evolutionary Biology (CSIC - Universitat Pompeu Fabra), Barcelona, Spain.
Email: pau.balart@ibe.upf-csic.es (PB-G); rosa.fernandez@ibe.upf-csic.es (RF)

Funding information
Ministerio de Economía y Competitividad and the Ministerio de Ciencia of Spain, Grant/Award Number: CGL2016-76705-P, PID2019-108824GA-I00, CGL2016-75255, PID2019-103947GB, BES-2017-081050, BES-2017-081740 and RyC2017-22492; Marie Sklodowska-Curie, Grant/Award Number: 747607

Abstract

The chemosensory system has experienced relevant changes in subterranean animals, facilitating the perception of specific chemical signals critical to survival in their particular environment. However, the genomic basis of chemoreception in cave-dwelling fauna has been largely unexplored. We generated de novo transcriptomes for antennae and body samples of the troglobitic beetle *Speonomus longicornis* (whose characters suggest an extreme adaptation to a deep subterranean environment) in order to investigate the evolutionary origin and diversification of the chemosensory gene repertoire across coleopterans through a phylogenomic approach. Our results suggested a diminished diversity of odourant and gustatory gene repertoires compared to polyphagous beetles that inhabit surface habitats. Moreover, *S. longicornis* showed a large diversity of odourant-binding proteins, suggesting an important role of these proteins in capturing airborne chemical cues. We identified a gene duplication of the ionotropic coreceptor IR25a, a highly conserved single-copy gene in protostomes involved in thermal and humidity sensing. In addition, no homologous genes to sugar receptors or the ionotropic receptor IR41a were detected. Our findings suggest that the chemosensory gene repertoire of this cave beetle may result from adaptation to the highly specific ecological niche it occupies, and that gene duplication and loss may have played an important role in the evolution of gene families involved in chemoreception. Altogether, our results shed light on the genomic basis of chemoreception in a cave-dwelling invertebrate and pave the road towards understanding the genomic underpinnings of adaptation to the subterranean lifestyle at a deeper level.

KEYWORDS
chemosensory proteins, Coleoptera, transcriptomics, troglobitic fauna

1 | INTRODUCTION

Major lifestyle transitions in insects, such as the conquest of terrestrial habitats, flight or host-plant interactions, are often accompanied by dramatic shifts in their sensory systems (Almudi et al., 2020; Anholt, 2020; Missbach et al., 2015; Vieira & Rozas, 2011; Wang, Pentzold, et al., 2018). Subterranean specialization has also offered opportunities for evolutionary innovation in the way animals interact with this particular environment (Cartwright et al., 2017). While adapting to subterranean niches, different species,

Ignacio Ribera and Rosa Fernández are Senior authors.

ranging from fishes to insects, have evolved highly convergent alternatives to live in perpetual darkness in habitats exhibiting specific biotic and abiotic factors (i.e., limited nutrient sources, constant temperature and humidity). Evolutionary regressions (e.g., loss of eyes and pigmentation), elaborated elements (e.g., hypertrophy of extra-optic sensory structures) and other physiological changes (e.g., modified life cycles) have been reported as possible adaptations for many obligate subterranean fauna (Culver & Pipan, 2019). Likewise, it is conceivable that the subterranean selective pressures have driven adaptive shifts in other sensory systems, including the chemosensory systems of subterranean animals. For instance, some studies on cavefish pointed out an enhancement of chemosensory systems from a morphological point of view (i.e., visible differences in taste buds and olfactory neural bulbs) when compared to surface populations (Parzefall, 2001; Yamamoto et al., 2009; Yang et al., 2016). In subterranean arthropods, elongation of antennae and body appendages have also been attributed to enhanced sensory capabilities (Turk et al., 1996). Nevertheless, the evolution of the chemosensory system in subterranean fauna from a molecular perspective remains widely unexplored.

Environmental chemical signals are enormously diverse in nature. Animals have developed a wide diversity of mechanisms to perceive and interpret specific cues essential to their evolutionary success (Nei et al., 2008). In insects, these chemicals include palatable nutrient or repellent odours and tastes, pheromones, warning signals of predators and those indicating optimal substrates for oviposition, and various others (Joseph & Carlson, 2015). The chemosensory system in insects is distributed morphologically at the interface between the environment and the dendrites of the peripheral sensory neurons, where different chemosensory proteins act in parallel for the signal transduction to the brain centers in which the information is processed (Dippel et al., 2016; Joseph & Carlson, 2015). To capture this complex information, insects use three large and divergent families of transmembrane chemoreceptor proteins: gustatory receptors (GRs), odourant receptors (ORs) and ionotropic receptors (IRs; Benton et al., 2009; Clyne et al., 1999; Gao & Chess, 1999; Sánchez-Gracia et al., 2009; Vosshall et al., 1999). GRs, which detect nonvolatile compounds, probably represent the oldest chemosensory receptors (Eyun et al., 2017), being distributed in several taste organs along the entire body including mouth pieces, legs, wing margins and other specialized structures such as vaginal plate sensilla in abdomens of female flies (Stocker, 1994). Airborne chemical particles are perceived in the head appendages by the ORs, an insect-specific chemoreception gene family thought to have originated from the GR gene family (Robertson, 2019; Robertson et al., 2003; Thoma et al., 2019). ORs work with the functionally universal odourant receptor coreceptor (ORCO), which is highly conserved in winged insects (i.e., Paleoptera and Neoptera; Brand et al., 2018). Moreover, IRs derived from the ionotropic glutamate receptor gene (IGluRs) superfamily in protostomes (Benton, 2015; Vosshall & Stocker, 2007) and mediate responses to many organic acids and amines, including pheromones and nutrient odours (Benton et al., 2009). In insects there are other gene families that also participate in chemosensory functions, such as the sensory neuron membrane proteins (SNMPs; Grimaldi & Engel, 2005; Missbach et al., 2014; Nichols & Vogt, 2008). The odourant-binding proteins (OBPs) and chemosensory proteins (CSPs) also play a key role for chemoreception in terrestrial insects, besides other physiological roles. The stable and compact structure of OBPs and CSPs make them versatile soluble proteins relevant for signal transduction of small hydrophobic compounds such as pheromones and odourants (Pelosi et al., 2014, 2018; Roys, 1954; Stürckow, 1970).

As in other large gene families encoding ecologically relevant proteins, constant birth-and-death dynamics may play an important role in their evolution in arthropods (Nei & Rooney, 2005; Sánchez-Gracia et al., 2009; Vieira et al., 2007). A general positive correlation has been observed when comparing the chemosensory gene diversity across species and the complexity of chemical signal in the ecological niche they occupy; several studies reported contrasting patterns of gene family expansions and gene losses when exploring the chemosensory gene repertoires in extreme specialist and generalist species, with the latter usually exhibiting larger expansions of gene families involved in chemoreception (Andersson et al., 2019; Kirkness et al., 2010; Li et al., 2018; McBride, 2007; McBride & Arguello, 2007; Ngoc et al., 2016). However, the evolution of the chemosensory gene families in subterranean species are still largely unexplored, hampering our understanding on how these animals perceive their particular environment.

Cave beetles represent ideal models to shed light on the genomic basis of chemoreception in subterranean environments. The Leptodirini tribe is a speciose lineage of scavenger beetles that represents one of the most impressive radiations of subterranean organisms. Several lineages within Leptodirini (estimated to have colonized subterranean habitats ca. 33 Ma (Ribera et al., 2010)) acquired morphological and physiological traits typically associated with troglobitic adaptations. Their modifications include complete lack of eyes and optic lobes, depigmentation, membranous wings, elongation of antennae and legs (Deleurance, 1963; Jeannel, 1924; Luo et al., 2019) and loss of thermal acclimation capacity (Pallarés et al., 2020; Rizzo et al., 2015). They also exhibit modified life cycles as a key innovation for their subterranean specialization (Cieslak et al., 2014; Delay, 1978). One of the highly modified species of the Leptodirini tribe is *Speonomus longicornis* Saulcy, 1872 (Coleoptera, Polyphaga, Leiodidae; Figure 1a). This obligate cave-dwelling beetle is completely blind, depigmented, possesses enlarged antennae (Jeannel, 1924) with a high sensilla density (Figure 1b) and it has a contracted life cycle, comprising a single larval-instar during its development in which the larvae remain practically quiescent like the pupal stage (Glaçon, 1953). The troglobitic characters of this species suggest an extreme adaptation to the deep subterranean environment.

This study aims to characterize the chemosensory gene repertoire of *S. longicornis*. The aims of this project were (i) to pinpoint genes putatively involved in chemoreception in the cave beetle *S. longicornis* through a transcriptomic approach, and (ii) to explore

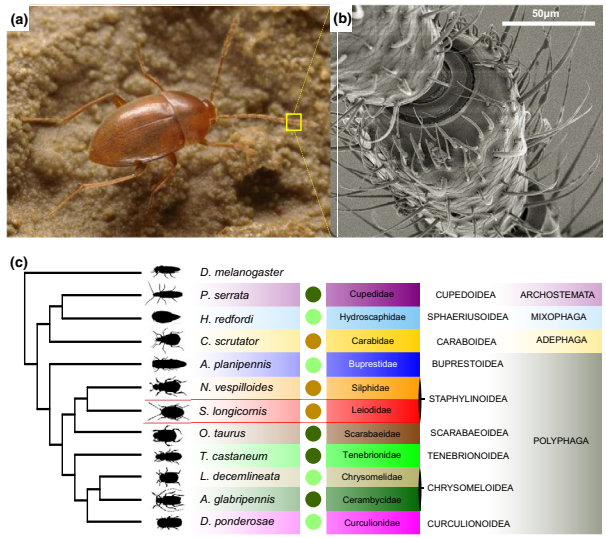


FIGURE 1 *Speonomus longicornis* and its phylogenetic position within Coleoptera. (a) *Speonomus longicornis*. (b) Scanning electron microscopy image of the antennal sensilla of *S. longicornis* (voucher IBE-AI531). (c) Simplified phylogeny showing the relationships of the studied species, adapted from McKenna et al. (2019). Coloured circles illustrate the dietary habits of the species: darkgreen corresponds to polyphagous herbivores, lightgreen to oligophagous herbivores and brown to nonphytophagous species [Colour figure can be viewed at wileyonlinelibrary.com]

how such genes evolved in the broader phylogenetic context of beetle and insect evolution. Our study therefore aims to provide the first characterization of the chemosensory gene repertoire of an obligate cave-dwelling species.

2 | MATERIALS AND METHODS

2.1 | Sample collection and preservation

Thirty specimens of *Speonomus longicornis* were collected in 2016 at the type locality: Grotte de Portel cave, in the Plantaurel massif at the French region of Ariège (43°01′51″N, 1°32′22″E). All specimens were manually captured and kept alive inside a thermo-box during the stay at the cave. Once sampling was completed, all individuals were placed in an 8 ml tube and flash-frozen in liquid nitrogen at the cave entrance in order to prevent stress-related alterations in gene expression levels and to minimize RNA degradation during transportation to the laboratory, where the samples were stored at −80°C until RNA extraction.

2.2 | RNA extraction

All steps were performed in cold and RNase free conditions. Several specimens were pooled in each sample prior to RNA extraction in order to obtain sufficient tissue for an efficient extraction. We did not examine the sex of the specimens to minimize the manipulation in order to avoid RNA degradation. Nevertheless, no significant sexual dimorphism has previously been found in the chemosensory system of other coleopterans (Dippel et al., 2016; Wu et al., 2016).

The specimens were split into three groups of 10 individuals each, representing biological replicates. Since chemosensory structures are mainly concentrated in the antennae (see Section 1), they were dissected from each specimen. Therefore, the experimental design included three biological replicates representing two conditions: antennae and the rest of the body.

The isolation of total RNA was performed by phenol/chloroform extraction with a lysis through guanidinium thiocyanate buffer following the protocol of Sambrook et al. (1989) with minor modifications (i.e., not using 2-mercaptoethanol). A first quality check was done by size separation in a 1% TBE agarose gel chromatography. Total RNA yield was quantified by an RNA assay in a Qubit fluorometer (Life Technologies).

2.3 | cDNA library construction and next-generation sequencing

For the antennae samples, a low-input RNA sequencing protocol was followed. mRNA sequencing libraries were prepared following the SMARTseq2 protocol (Picelli et al., 2013) with some modifications. Briefly, RNA was quantified using the Qubit RNA HS Assay Kit (Thermo Fisher Scientific). Reverse transcription with an input material of 2 ng was performed using SuperScript II (Invitrogen) in the presence of oligo-dT30VN (1 µM; 5′-AAGCAGTGGTATC AACGCAGAGTACT30VN-3′), template-switching oligonucleotides (1 µM) and betaine (1 M). The cDNA was amplified using

the KAPA Hifi Hotstart ReadyMix (Roche), 100 nM ISPCR primer (5′-AAGCAGTGGTATCAACGCAGAGT-3′) and 15 cycles of amplification. Following purification with Agencourt Ampure XP beads (1:1 ratio; Beckmann Coulter), product size distribution and quantity were assessed on a Bioanalyzer High Sensitivity DNA Kit (Agilent). The amplified cDNA (200 ng) was fragmented for 10 min at 55°C using Nextera XT (Illumina) and amplified for 12 cycles with indexed Nextera PCR primers. The library was purified twice with Agencourt Ampure XP beads (0.8:1 ratio) and quantified on a Bioanalyzer using a high sensitivity DNA kit.

For the samples containing the rest of the body, total RNA was assayed for quantity and quality using the Qubit RNA BR Assay kit (Thermo Fisher Scientific) and RNA 6000 Nano Assay on a Bioanalyser 2100 (Agilent). The RNASeq libraries were prepared from total RNA using the KAPA Stranded mRNA-Seq Kit for Illumina (Roche) with minor modifications. Briefly, after poly-A based mRNA enrichment from 500 ng of total RNA, the mRNA was fragmented. The second strand cDNA synthesis was performed in the presence of dUTP to achieve strand specificity. The blunt-ended double stranded cDNA was 3′ adenylated and Illumina single indexed adapters (Illumina) were ligated. The ligation product was enriched with 15 PCR cycles and the final library was validated on an Agilent 2100 Bioanalyser with the DNA 7500 assay.

The libraries were sequenced on an Illumina HiSeq 2500 platform in paired-end mode with a read length of 2 × 76 bp. Image analysis, base calling and quality scoring of the run were processed using the manufacturer's software REAL TIME ANALYSIS (RTA 1.18.66.3) and followed by generation of FASTQ sequence files by CASAVA. cDNA libraries and mRNA sequencing were performed at the National Center of Genomic Analyses (CNAG).

2.4 | Sequence processing, decontamination and de novo assembly

Raw reads for all samples were downloaded in FASTQ format. The quality of the raw reads was assessed and visualized using FASTQC version 0.11.8 (www.bioinformatics.babraham.ac.uk). For each data set, remaining Illumina adaptors were removed and low-quality bases were trimmed off according to a threshold average quality score of 30 based on a Phred scale with TRIMMOMATIC version 0.38 (Bolger et al., 2014). Filtered paired-end reads were validated through a FASTQC visualization.

A reference de novo transcriptome assembly was constructed with TRINITY version 2.8.4, using paired read files and default parameters, including all replicates and conditions (Grabherr et al., 2011; Haas et al., 2013). BLOBTOOLS version 1.1.1 (Laetsch & Blaxter, 2017) was used to detect putative contamination from the assembled transcriptome. Transcripts were annotated using BLAST+ version 2.4.0 against the nonredundant (nr) database from NCBI with an expect value (*E*-value) cutoff of $1e^{-10}$ and reads were mapped to the reference transcriptome with BOWTIE2 version 2.3.5.1 (Langmead & Salzberg, 2012). Putative contaminants included transcripts with significant hits to viruses, fungi, bacteria or chordates, accounting for a total of 7.8% of the mapped sequences (see also Figure S1).

2.5 | Inference of candidate coding regions and transcriptome completeness assessment

To check completeness of the reference transcriptome, we searched for single copy universal genes in insects through benchmarking universal single-copy orthologues (BUSCO version 4.1: Simão et al., 2015) using the insecta database (insecta_odb10) and three different data sets as a query: (i) the assembled transcripts in TRINITY, (ii) the total predicted open reading frames (ORFs), and (iii) the longest isoform per ORF. The assembly was processed in TRANSDECODER version 5.4.0 to identify candidate ORFs within the transcripts using the universal genetic code (Haas et al., 2013). Only the longest ORFs (i.e., with a minimum length of 100 amino acids) of each transcript were retained as final candidate coding regions for further analyses.

2.6 | Chemosensory gene repertoire characterization

BITACORA version 1.0.0 (Vizueta et al., 2020) was used to curate annotations during the sequence similarity searches of the chemosensory gene families of interest. Curated protein databases containing chemoreceptor genes (ORs, GRs, IRs, SNMPs, OBPs and CSPs) of several arthropods were used in the BITACORA searches (Vizueta et al., 2016). An additional database was used for the ORs annotation, containing Coleoptera ORs based on the data sets from Mitchell et al. (2019). The "protein mode" pipeline of BITACORA was used to annotate all the predicted ORFs of the transcriptome (see previous section), combining BLAST and HMMER searches. All the predicted coding regions were used for BITACORA searches, retrieving a multifasta file for each of the chemosensory families. Dubious annotations were manually inspected (i.e., some ORFs received significant hits for both ORs and GRs, which were post-validated through Pfam searches). Results were filtered with customized Python scripts using BIOPYTHON version 1.76 SEQIO package (https://biopython.org [Cock et al., 2009]); in order to obtain final candidates, which were represented by the longest isoform per gene and thus achieving unique gene annotations.

2.7 | Expression levels quantification and differential gene expression analysis

SALMON version 0.10.2 (Patro et al., 2017) was used for indexing and quantification of transcript expression. Expression estimated counts were transformed into an expression matrix using a Perl script included in the TRINITY software (abundance_estimates_to_matrix.pl), which implements the trimmed mean of *M*-values normalization method (TMM). We also examined the data to ensure that the

biological replicates were well correlated using the "PtR" module included in TRINITY version 2.8.4 analysis toolkit, generating a variety of plots that allowed us to visually inspect the presence of strong outliers or batch effects that can affect the differential expression analysis. Differential expression analysis was conducted in the Bioconductor EDGER package (Robinson et al., 2010; Robinson & Oshlack, 2010). The Benjamini-Hochberg method was applied to control the false discovery rate (FDR; Benjamini & Hochberg, 1995). The significance value for multiple comparisons was adjusted to 0.001 FDR threshold cutoff and a four-fold change. Differentially expressed genes (up- and downregulated) in antennae and the rest of the body were plotted in heatmaps using the R scripts provided in the TRINITY software. The expression matrix was also interrogated in order to detect exclusively expressed genes in antennae (defined as genes that showed positive expression values in the three replicates of antennae and with expression values lower than 0.001 TMM in the rest of the body).

2.8 | Transcriptome characterization, gene ontology enrichment and visualization

The peptide predictions, including all isoforms, were used as input for EGGNOG-MAPPER version 4.5.1 (Huerta-Cepas et al., 2017), retrieving gene ontology (GO) terms for all the annotated transcripts. The GO annotations were subsequently filtered to discard those corresponding to nonanimal taxa (i.e., viruses, bacteria, fungi, plants, 10.86% of total GO annotations) and to eliminate the redundancy provided by the isoforms. All GO terms for each unique gene were retained and used in the subsequent analyses. GO enrichment analysis was performed using a two-tailed Fisher's test in FATIGO software (Al-Shahrour et al., 2007) to detect significant overrepresentation of the GO terms in the pairwise comparisons between the upregulated genes in antennae and the rest of the body, adjusting the *p*-value to .05.

GO enrichment analyses were visualized in the REVIGO web server (Supek et al., 2011), plotting the results in a "TREEMAP" graph using R, where the size of the rectangles is proportional to the enrichment *p*-value (abs_log10_pvalue) of the overrepresented GO terms.

2.9 | Phylogenetic inferences for the candidate chemosensory genes

Complete and partial annotated genes for *S. longicornis* (referred to as *Slon* in the figures) were included in the phylogenetic inferences in order to interrogate their phylogenetic relationships with chemosensory genes of other species, all of them based on genomic data. With this approach, the aim is to infer diversity patterns of the chemosensory repertoire of *S. longicornis* and to characterize each gene family more specifically, indicating with a higher confidence the putative function of these genes compared to analysis merely based on homology. Individual phylogenies for each chemosensory gene family as annotated by BITACORA (see above) were inferred using the following pipeline. Amino acid sequences were aligned using PASTA software version 1.7.8 (Mirarab et al., 2015). Poorly aligned regions were trimmed using TRIMAL version 1.2 (Capella-Gutiérrez et al., 2009) with the "-automated1" flag. Maximum likelihood phylogenetic inference was inferred with IQ-TREE version 2.0.4 (Nguyen et al., 2015). The mixture model LG + C20 + F + G was used with the site-specific posterior mean frequency model (PMSF; Wang et al., 2018) and the ultrafast bootstrap option (Hoang et al., 2018). A guide tree was inferred with FASTTREE2 under the LG model (Price et al., 2010). Results were visualized using the ITOL web interface (Letunic & Bork, 2019).

The ORs phylogeny included the coleopteran ORs obtained from Mitchell et al. (2019). These species (Figure 1c) included a range of ecological strategies. Considering the degree of dietary specialization, the phytophagous specialists (i.e., species feeding on a small number of plants or algae species usually belonging to the same botanical family, thus considered as oligophagous [Schoonhoven et al., 2005]) were represented by the aquatic beetle *Hydroscapha redfordi* (Myxophaga, Hydroscaphidae), the ash borer *Agrilus planipennis* (Polyphaga, Buprestidae), the Colorado potato beetle *Leptinotarsa decemlineata* (Polyphaga, Chrysomelidae) and the mountain pine beetle *Dendroctonus ponderosae* (Polyphaga, Curculionidae). By contrast, the phytophagous generalists (i.e., species feeding on several plants belonging to different families, thus considered as polyphagous) included the red flour beetle *Tribolium castaneum* (Polyphaga, Tenebrionidae), the wood borer *Anoplophora glabripennis* (Polyphaga, Cerambycidae), the reticulated beetle *Priacma serrata* (Archostemata, Cupedidae) and the dung beetle *Onthophagus taurus* (Polyphaga, Scarabaeidae). Moreover, the species set include two nonphytophagous beetles, the insectivorous *Calosoma scrutator* (Adephaga, Carabidae) and the burying beetle *Nicrophorus vespilloides* (Polyphaga, Silphidae) that feeds on vertebrate carrion (Vogel et al., 2017).

For the ORCO phylogeny, some additional ORCO sequences of coleopteran species and other taxa as outgroups (*Apis mellifera* and *Drosophila melanogaster*) were also included (see species and GenBank accessions at Table S1).

3 | RESULTS

3.1 | A high quality de novo transcriptome for *Speonomus longicornis* facilitates the annotation of its chemosensory gene repertoire

Since no reference genome is available for the focus species, a deeply sequenced de novo assembly transcriptome was constructed combining the paired-end reads from the six libraries (~411 million reads; ~356 million after trimming), obtaining a total of 245,131 transcripts (Figure S1). These transcripts include 177,711 unique predicted "genes" by TRINITY and 74,273 candidate ORFs (including all genes and isoforms). When filtering by the longest isoform per gene (which could be considered as a "proxy" for the total number of genes in the genome), we obtained a total of

20,956 ORFs. BUSCO analysis indicated a high completeness for the assembled transcriptome, with 97% of complete BUSCO genes compared to the insecta database (Figure S1), indicating that a mostly complete reference gene set was recovered and hence it was of enough quality to explore gene family evolution. These results should be interpreted with caution as they are based on transcriptomic data instead of high-quality genomes, and further reference-level genomic analyses may help clarify the evolutionary dynamics of chemosensory gene families with higher precision. However, this approach has been successfully applied in other studies with non-model organisms through the combination of genomic and high quality transcriptomic data (e.g., Fernández & Gabaldón, 2020; Vizueta, Escuer, et al., 2020). Further details about sequencing, assembly statistics, the completeness assessment and the putative contamination results are summarized in Figure S1.

3.2 | Differential gene expression analysis reveals chemosensory genes upregulated in the antennae

BITACORA searches identified a total of 205 chemosensory gene candidates for *S. longicornis* (Table 1, see also Table S2). The expression level distribution obtained in the transcript quantification steps was assessed in order to identify possible biases when comparing replicates and conditions (Figure S2). Results indicated that replicates are more similar to each other than between the different conditions. A total of 18,160 clusters of transcripts (reported as clusters of transcripts or "genes" by TRINITY, referred to as TRINITY genes hereafter) were detected as differentially expressed in antennae and body, with 8,949 TRINITY genes upregulated in antennae (Figure S3; Table S3). Out of the 205 candidate chemosensory genes as detected by BITACORA, 78 were detected as differentially expressed. From those, 49 genes were overexpressed in antennae (18 ORs, 17 OBPs, five GRs, five IRs/IGluRs, two SNMPs/CD36s and two CSPs; Table 1; Figure 2a). Furthermore, seven ORs were identified as exclusively expressed in antennae, including two additional genes not recovered as differentially expressed due to the disparity of the expression values between antennae replicates (Table S4).

3.3 | Gene ontology enrichment reveals upregulated chemosensory specificity in antennae

Out of the 74,273 predicted ORFs (including isoforms), 59.4% were annotated through EGGNOG-MAPPER, and from those only 23,555 of these annotations yielded associated GO terms, representing 31.7% of the total queried sequences from the assembled transcriptome. After filtering annotations from nonanimal taxa (including viruses, bacteria and fungi) 89.14% of the annotations were retained. Figure 2b depicts enriched GO terms for upregulated genes in antennae and in the rest of the body. In antennae, "sensation and perception of chemical stimulus" represent the most enriched category within the biological processes analysed and, in a minor proportion, some categories related to cilium activity. "Mechanosensory activity" terms are also overrepresented but in a minor proportion. More than half of the cellular component GO terms enriched in antennae correspond to "dendritic structures", and a high proportion of the overrepresented terms correspond to "extracellular and membrane structures". Regarding the molecular function category in antennae, "odourant-binding" and "odourant reception" terms occupy a large proportion of the enriched functions followed by other "binding and signal transduction" terms.

TABLE 1 Annotated chemosensory genes of *S. longicornis*

	ORs	GRs	IRs/IGluRs	SNMPs/CD36s	OBPs	CSPs
Total	50	36	53	20	39	7
Antennae	18	5	5	2	17	2

The number of overexpressed genes in antennae are indicated.

3.4 | Phylogenetic interrelationships of Coleoptera ORs and ORCOs

A total of 1,222 OR sequences were aligned and trimmed (see Section 2). The final length of the alignment was of 254 amino acid positions. To facilitate comparison, we retained the nomenclature used by Mitchell et al. (2019) to describe the phylogenetic groups and clades recovered in their phylogenetic analyses (i.e., groups 1, 2A, 2B, 3, 4, 5A, 5B, 6 and 7). Our results were overall congruent with those reported by Mitchell et al. (2019), with virtually all OR groups recovered with high support except group 6 and a different position for group 4, which was recovered as nested within group 3 (Figure 3a). While most of the genes fall into the same groups than in Mitchell et al. (2019), four genes (i.e., AplaOR1, PserOR120-121 and SlonOR34525c0g1) were not recovered for any of the previously proposed groups. The upregulated ORs in *S. longicornis* antennae were distributed among the different coleopteran OR groups, mostly clustered within group 1 and group 7 (with six and five genes, respectively). Exclusively expressed ORs in antennae are found in groups 1, 3, 4 and 7. The number of ORs is highly variable among these species (Figure 3b). *S. longicornis* and the other nonphytophagous species (i.e., *C. scutator*, *N. vespilloides*) exhibited relatively moderate OR repertoires and a similar distribution pattern (i.e., without representation in group 5A, moderate gene family expansions and the largest expansion in group 3; Figure 3b,d). All the ORCO sequences were recovered as a clade with high support and were used to root the tree, facilitating the identification of the ORCO candidate of *S. longicornis* (SlonORCO; Figure 3a,c).

The phylogeny of ORCOs (Figure 3c; with a final trimmed alignment of 478 amino acid positions) recovered clades for the different beetle families, but did not mirror the phylogeny of Coleoptera at the family level. For instance, all ORCOs of species of Cucujiformia were recovered in a clade and were subsequently clustered into their corresponding families. The same pattern was

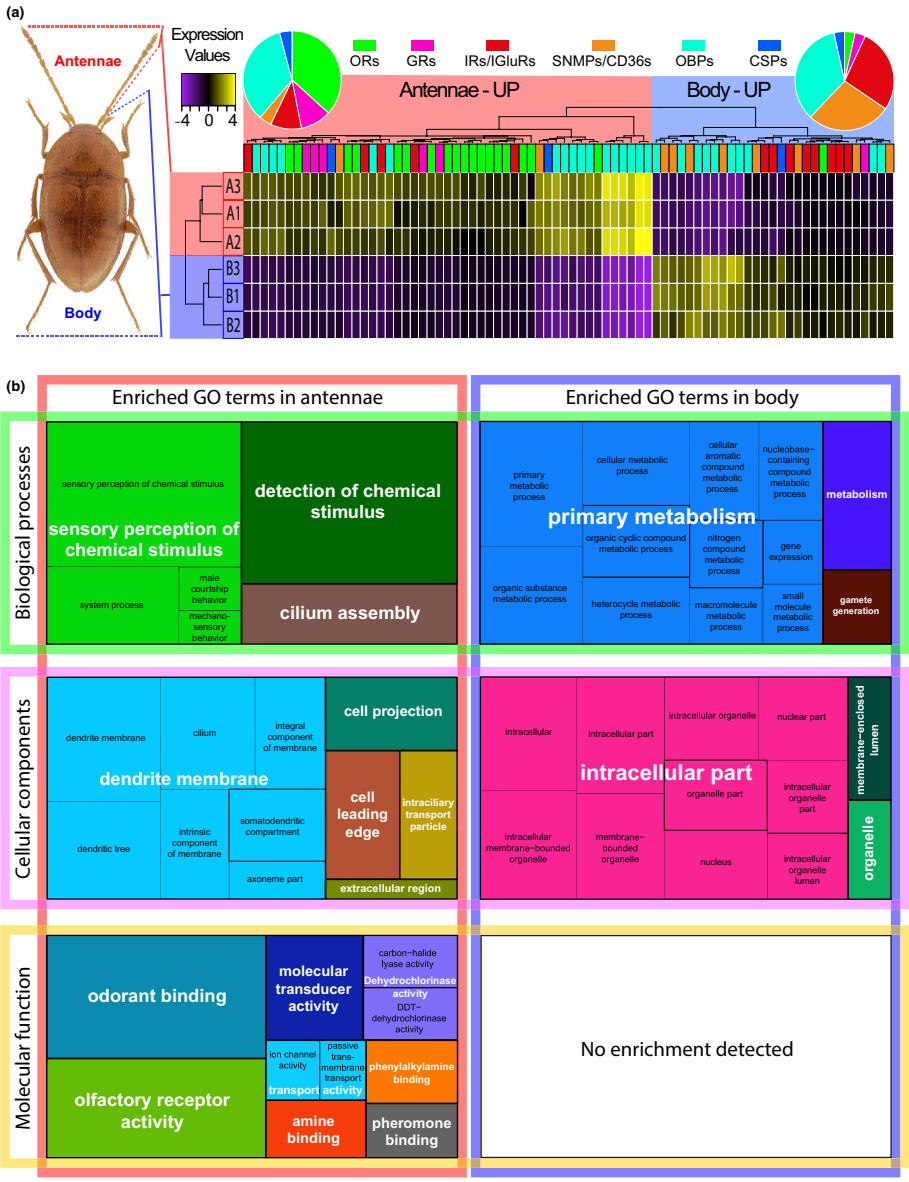


FIGURE 2 Differentially expressed genes in antennae and body in *S. longicornis*. (a) Heatmap of chemosensory genes of *S. longicornis* differentially expressed in antennae and the rest of the body. (b) Gene ontology (GO) treemaps for the differentially expressed genes in antennae versus the rest of the body. Biological process, molecular function and cellular component enriched GO terms are shown [Colour figure can be viewed at wileyonlinelibrary.com]

B

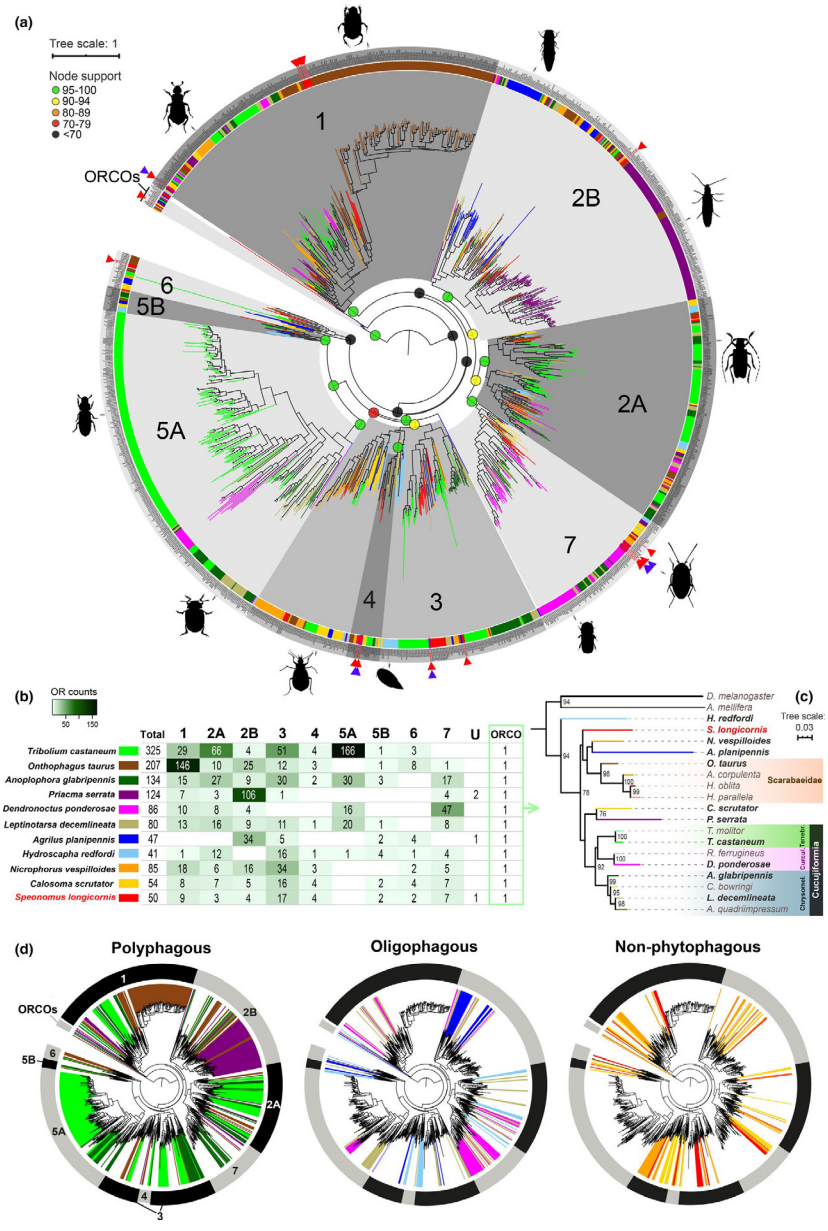

(a)
Tree scale: 1
Node support
95-100
90-94
80-89
70-79
<70
ORCOs
1
2B
2A
7
3
4
5A
5B
6
(b)
OR counts
50
150
Total 1 2A 2B 3 4 5A 5B 6 7 U ORCO
Tribolium castaneum 325 29 66 4 51 4 166 1 3 1
Onthophagus taurus 207 146 10 25 12 3 1 8 1 1
Anoplophora glabripennis 134 15 27 9 30 2 30 3 17 1
Priacma serrata 124 7 3 106 1 4 2 1
Dendronoctus ponderosae 86 10 8 4 16 47 1
Leptinotarsa decemlineata 80 13 16 9 11 1 20 1 8 1
Agrilus planipennis 47 34 5 2 4 1 1
Hydroscapha redfordi 41 1 12 16 1 1 4 1 4 1
Nicrophorus vespilloides 85 18 6 16 34 3 2 5 1
Calosoma scrutator 54 8 7 5 16 4 2 4 7 1
Speonomus longicornis 50 9 3 4 17 4 2 2 7 1 1
(c)
Tree scale: 0.03
D. melanogaster
A. mellifera
H. redfordi
S. longicornis
N. vespilloides
A. planipennis
O. taurus
A. corpulenta
H. obita
H. parallela
Scarabaeidae
C. scrutator
P. serrata
T. molitor
T. castaneum
Tenebr.
R. ferrugineus
D. ponderosae
A. glabripennis
C. bowringi
L. decemlineata
A. quadrimpressum
Curcul.
Chrysomel.
Cucujiformia
(d)
Polyphagous
Oligophagous
Non-phytophagous
ORCOs
1
2B
2A
7
3
4
5A
5B
6

FIGURE 3 Phylogeny of odourant receptors. (a) Maximum likelihood phylogenetic tree of odourant receptors (ORs) including OR sets of *S. longicornis* and other coleopterans from Mitchell et al. (2019), representing the proposed OR groups in grey ranges. Red triangles represent the upregulated genes in the antennae of *S. longicornis*. Purple triangles represent exclusively expressed genes in antennae. Species are colour coded as indicated in (b). (b) Number of OR genes of each OR group inferred for each species included in the phylogeny. "U" indicates unclassified ORs. (c) Maximum likelihood phylogeny of ORCO across coleopterans (see Methods and Table S1 for species codes). (d) Simplified representation of OR diversity recovered for each species, highlighting the OR repertoire of species with different feeding strategies [Colour figure can be viewed at wileyonlinelibrary.com]

observed for the ORCOs of the different species of Scarabaeidae. By contrast, the ORCO of *S. longicornis* did not cluster together with that of *N. vespilloides*, despite belonging to the same superfamily (i.e., Staphylinoidea).

3.5 | Phylogenetic inference of the annotated GRs

A total of 374 sequences were interrogated, resulting in a multiple sequence alignment of 271 amino acid positions. Notably, none of the candidate GRs from *S. longicornis* clustered together with GRs involved in the perception of fructose and other sugars in the other species (Figure 4). On the other hand, in several coleopteran GRs, including *S. longicornis*, we recovered three candidates that cluster with those of *D. melanogaster* involved in perception of CO_2, which are well characterized functionally (termed as GR1, GR2, GR3 in beetles, and GR21a, GR63a in *D. melanogaster*; Dippel et al., 2016; Jones et al., 2007; Kwon et al., 2007). One of these three candidate CO_2 receptors of *S. longicornis* was upregulated in antennae and was recovered as orthologous to the GR2 gene in beetles. We also identified a candidate bitter taste GR (SlonGR19567c0g1) that clustered together with strong support with previously identified as conserved bitter taste GRs for *A. planipennis* and *D. ponderosae* (Andersson et al., 2019). The rest of the genes were generally recovered in well-supported clades with species-specific differences in the extent of GR expansions. *S. longicornis* showed divergent GRs distributed along the tree exhibiting relatively small expansions (i.e., from two to five genes) and in general a relatively diminished gustatory repertoire, whereas the other species exhibit remarkable expansions and considerably larger repertoires. The functions of the other four upregulated GRs in antennae cannot be further predicted due to the lack of functional annotation of the genes they cluster together with.

3.6 | IR and IGluR phylogenies

Since IRs derive from IGluRs (see Introduction), our phylogenetic approach facilitated the initial annotations of both types of genes in *S. longicornis*. IGluRs are not directly associated with chemoreception but show a high sequence identity with the most conserved IRs (i.e., IR8a, IR25a). Three genes clustered together with N-methyl-D-aspartate receptors (NMDARs) and nine genes clustered with different IGluR clades, one upregulated in antennae (Figure 5a).

A total of 164 IR sequences were aligned and trimmed resulting in a multiple sequence alignment of 356 amino acid positions (Figure 5c). Several IRs of *S. longicornis* clustered together with conserved IRs in insects (i.e., IR8a, IR25a, IR93a, IR76b, IR21a, IR68a, IR40a, IR100, IR60a). No genes clustering together with IR41a were detected for *S. longicornis*. Several putative gene duplications were detected for *S. longicornis* (containing each from 2 to 5 IRs), more moderate in size than the large gene family expansions observed for *A. planipennis* and *D. ponderosae*, which included up to eight and 17 IRs, respectively. Our analyses revealed a gene duplication in IR25a, a highly conserved single copy gene virtually in all protostomes (with the exception of the parasitoid wasps *Nasonia vitripennis* and *Microplitis mediator* and the limpet *Lottia gigantea*, see Section 4). The copy of IR25a exhibiting the shortest branch in *S. longicornis* was upregulated in antennae (Figure 5a–c). In order to assess the robustness of our results, we followed four steps. First, all isoforms from both genes were visually inspected in an alignment (Figure S4), observing notable differences between the sequences belonging to the different genes and a high similarity for the isoforms of the same gene. Second, we used the expression values of each isoform in both genes to test whether only a low number of sequence reads from some of the isoforms were mapping back to one of the genes (which would indicate errors in the assembly), and tested if the expression levels were statistically different by means of a one-way ANOVA (Table S5). Expression levels were uniformly distributed across replicates with the exception of one isoform that was highly expressed in one replicate, indicating that the inference of the two genes was not a methodological artefact. Third, the final alignment for the phylogeny including all the taxa (using the longest isoforms as described in Section 2) was examined to test that they were not nonoverlapping fragmented genes. Fourth, both IR25a copies were annotated with HMMER to inspect the similarity of the domain profiles. These IR25a candidates (i.e., SlonIR11039c0g2 and SlonIR14393c1g1) shared 44% of identical residues whereas the conserved copy of *S. longicornis* (SlonIR14393c1g1) had between 69% to 72% of amino acid sequence identity with the IR25 candidates of the other coleopteran species. In addition, the protein annotation of the IR25a candidates of *S. longicornis* by HMMER resulted in highly similar domain profiles, suggesting their similarity at the structural level.

3.7 | OBP phylogeny

A total of 137 OBP sequences were included to explore the OBPs diversity of *S. longicornis*, resulting in a multiple sequence alignment of 110 amino acid positions after trimming. Our results suggest that OBPs in *S. longicornis* are relatively abundant compared to the other species, being the most diverse repertoire of this comparison after

B

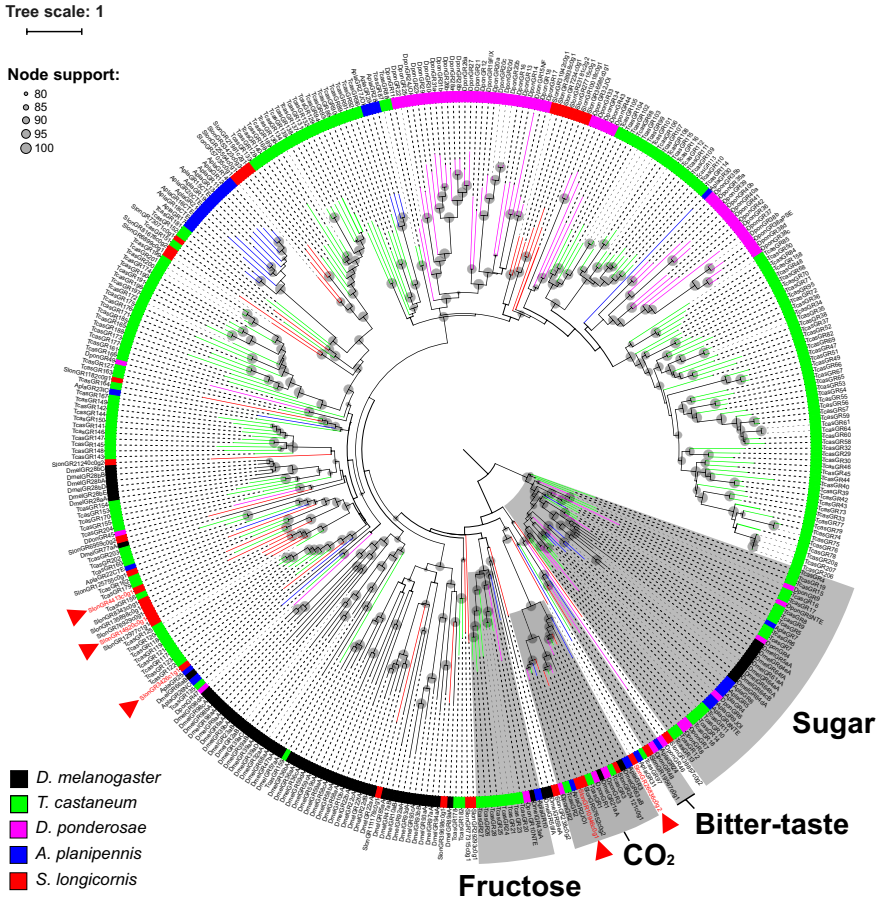
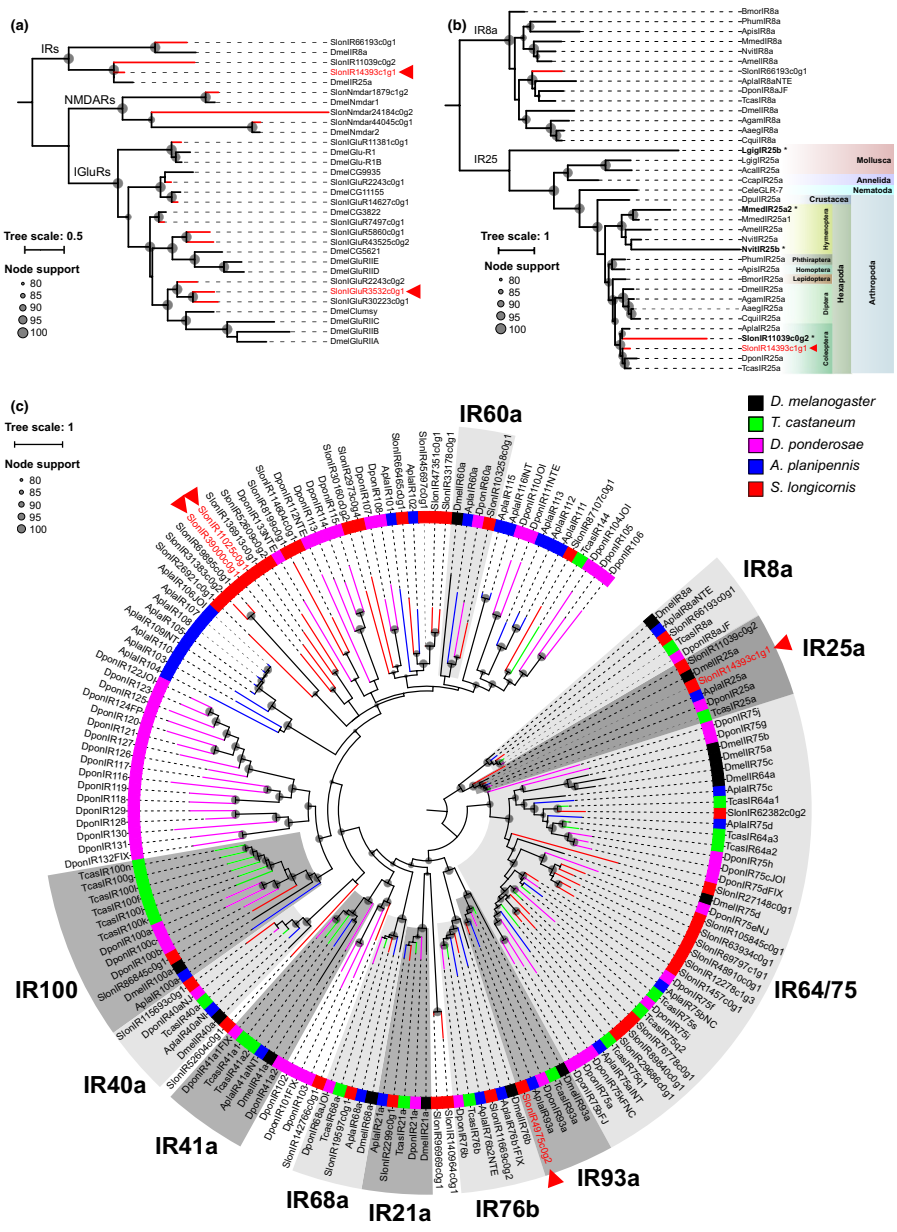


FIGURE 4 Phylogeny of gustatory receptors. Maximum likelihood phylogenetic tree of gustatory receptors (GRs) including GR sets of *S. longicornis*, other coleopterans from Andersson et al. (2019) and conserved GR sequences of *D. melanogaster*. Grey ranges represent well supported GR clades, indicating the proposed functions in the other species. Red triangles represent upregulated genes in the antennae of *S. longicornis* [Colour figure can be viewed at wileyonlinelibrary.com]

FIGURE 5 Phylogeny of ionotropic receptors. (a) Maximum likelihood phylogenetic tree of ionotropic glutamate receptors (IGluRs) including sequences of *S. longicornis* and *D. melanogaster*. (b) Maximum likelihood phylogenetic tree of the ionotropic receptors clades IR25a and IR8a (the later used to root the tree), including the candidate genes for *S. longicornis* (*Slon*), beetle sequences retrieved from Andersson et al. (2019) and Dippel et al. (2016) and sequences from an expanded invertebrate taxon sampling obtained from Croset et al. (2010) and Wang et al. (2015). Species codes are as follows: *D. melanogaster* (*Dmel*), *Aedes aegypti* (*Aaeg*), *Culex quinquefasciatus* (*Cqui*), *Anopheles gambiae* (*Agam*), *Bombyx mori* (*Bmor*), *Apis mellifera* (*Amel*), *Nasonia vitripennis* (*Nvit*), *Microplitis mediator* (*Mmed*), *Acyrthosiphon pisum* (*Apis*), *Pediculus humanus* (*Phum*), *Daphnia pulex* (*Dpul*), *Caenorhabditis elegans* (*Cele*), *Capitella capitata* (*Ccap*), *Aplysia californica* (*Acal*) *and Lottia gigantea* (*Lgig*). Asterisks indicate the divergent copies of IR25 candidates. (c) Maximum likelihood phylogenetic tree of ionotropic receptors (IRs) including sequences of *S. longicornis*, other coleopterans from Andersson et al. (2019) and sequences of *D. melanogaster*. Grey ranges represent the conserved IR clades, based on the annotations of the other species. In all trees, red triangles represent upregulated genes in the antennae of *S. longicornis* [Colour figure can be viewed at wileyonlinelibrary.com]


(a)
IRs
NMDARs
IGluRs
SlonIR66193c0g1
DmelIR8a
SlonIR11039c0g2
SlonIR14393c1g1
DmelIR25a
SlonNmdar1879c1g2
DmelNmdar1
SlonNmdar24184c0g2
SlonNmdar44045c0g1
DmelNmdar2
SlonIGluR11381c0g1
DmelGlu-R1
DmelGlu-R1B
DmelCG9935
SlonIGluR2243c0g1
DmelCG11155
SlonIGluR14627c0g1
DmelCG3822
SlonIGluR7497c0g1
SlonIGluR5860c0g1
SlonIGluR43525c0g2
DmelCG5621
DmelGluRIIE
DmelGluRIID
SlonIGluR2243c0g2
SlonIGluR3532c0g1
SlonIGluR30223c0g1
DmelClumsy
DmelGluRIIC
DmelGluRIIB
DmelGluRIIA
Tree scale: 0.5
Node support
80
85
90
95
100
(b)
IR8a
IR25
BmorIR8a
PhumIR8a
ApisIR8a
MmedIR8a
NvitIR8a
AmelIR8a
SlonIR66193c0g1
DponIR8aJF
TcasIR8a
DmelIR8a
AgamIR8a
AaegIR8a
CquiIR8a
LgigIR25b *
LgigIR25a
AcalIR25a
CcapIR25a
CeleGLR-7
DpulIR25a
MmedIR25a2 *
MmedIR25a1
AmelIR25a
NvitIR25a
NvitIR25b *
PhumIR25a
ApisIR25a
BmorIR25a
DmelIR25a
AgamIR25a
AaegIR25a
CquiIR25a
AplaIR25a
SlonIR11039c0g2 *
SlonIR14393c1g1
DponIR25a
TcasIR25a
Mollusca
Annelida
Nematoda
Crustacea
Hymenoptera
Phthiraptera
Homoptera
Lepidoptera
Diptera
Hexapoda
Coleoptera
Arthropoda
Tree scale: 1
Node support
80
85
90
95
100
(c)
Tree scale: 1
Node support
80
85
90
95
100
D. melanogaster
T. castaneum
D. ponderosae
A. planipennis
S. longicornis
IR60a
IR8a
IR25a
IR64/75
IR93a
IR76b
IR21a
IR68a
IR41a
IR40a
IR100

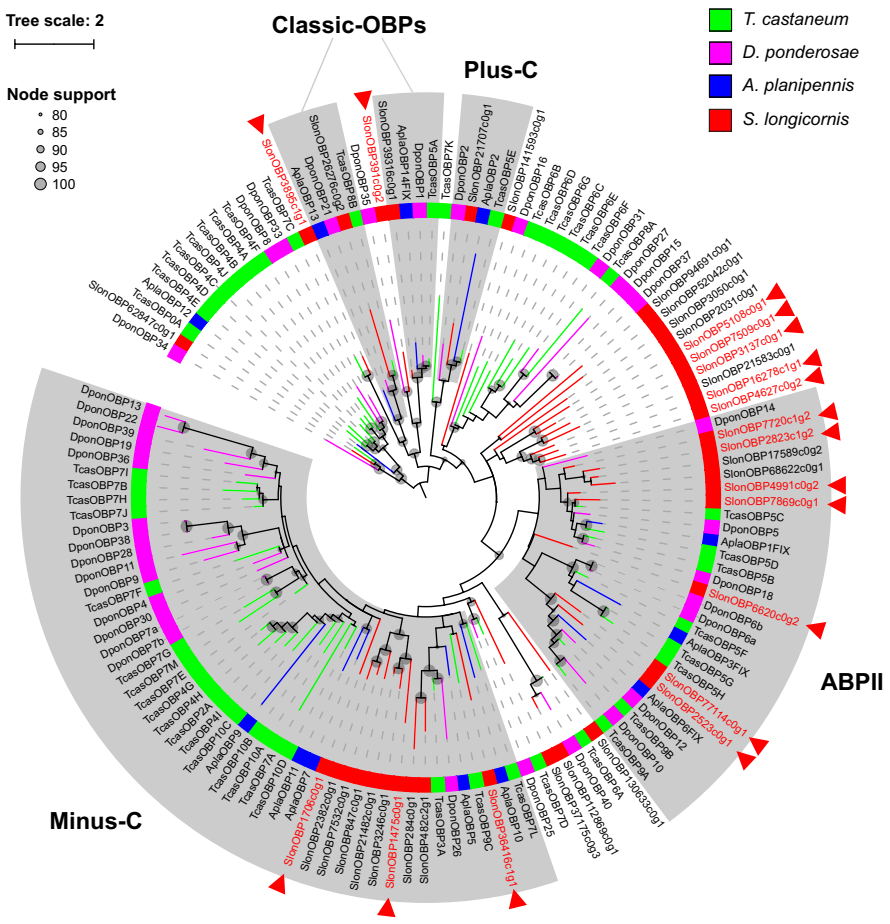


FIGURE 6 Phylogeny of OBPs. Maximum likelihood phylogenetic tree of OBPs including sequences of *S. longicornis* and other coleopterans from Andersson et al. (2019). Grey ranges represent the main OBP clades described in previous studies. Red triangles represent upregulated genes in the antennae of *S. longicornis* [Colour figure can be viewed at wileyonlinelibrary.com]

T. castaneum (Figure 6). Several OBP candidates of *S. longicornis* clustered together with the OBP subgroups described for the other species in Andersson et al. (2019; i.e., classic-OBPs, minus-C, plus-C and antennal binding proteins II [ABPII]). Only seven out of the 17 OBPs upregulated in antennae correspond to the ABPII clade. Furthermore, two relatively large OBP expansions include the majority of the upregulated OBPs, formed by (i) the minus-C clade (with three upregulated genes in antennae), and (ii) a specific *S. longicornis* OBP lineage of ten genes (with five upregulated genes in the antennae).

The phylogenomic characterization of other gene families (SNMP/CD36 and CSPs) was also explored. The results and discussion are included as Supporting Information (Results S1).

4 | DISCUSSION

A highly complete transcriptome for the cave-dwelling beetle *Speonomus longicornis* was generated in the present study (Figure

S1). Combining the differential gene expression and GO enrichment analyses with a curated annotation pipeline for the chemosensory related genes, we were able to explore the chemosensory gene repertoire of *S. longicornis*. Furthermore, the phylogenetic inferences for each of the chemosensory gene families offered the opportunity to compare the repertoire of genes involved in chemosensation in *S. longicornis* to other beetle species that occupy a wide variety of ecological niches, all of them in surface habitats.

The differential gene expression (Figure 2a) and GO enrichment analysis (Figure 2b) allowed us to identify upregulated genes in antennae (where chemosensory structures—sensilla—are highly concentrated) and compare the overall enriched functions in the antennae versus the rest of the body. As expected, olfaction was recovered as the most prominent function in antennae, representing more than half of the enriched terms in the molecular function category (odourant-binding and olfactory reception activities), indicating that ORs and OBPs are playing a major role in how cave beetles receive and process airborne cues. ORCO was upregulated in antennae, as expected since it is an essential component of the functional heterodimers that facilitate odourant reception combined with other ORs (Stengl & Funk, 2013). In addition, only seven ORs were observed as exclusively expressed in antennae (Table S4), suggesting that this gene family may include genes with high specificity in these appendages. All in all, these results highlight the importance of antennae in odourant perception in this cave beetle; further gene expression studies including additional structures such as mouth appendages would give more detailed insights for the rest of the identified ORs.

Concerning the OR phylogeny (Figure 3), our results are mostly consistent with what was found in Mitchell et al. (2019). Large expansions in several OR groups are highlighted in the polyphagous/generalist species *T. castaneum*, *O. taurus*, *A. glabripennis* and *P. serrata*, whereas the oligophagous/specialist *D. ponderosae*, *L. decemlineata*, *A. planipennis* and *H. redfordi* exhibit relatively reduced OR repertoires. Therefore, an apparent correlation between the host breadth and the OR diversity of herbivore Coleoptera is observed, clearly exemplified by the extent and distribution of OR diversity in the wood boring species (i.e., *A. glabripennis*, *D. ponderosae*, *A. planipennis*; Andersson et al., 2019). The insectivorous *C. scrutator* and the scavengers *N. vespilloides* and *S. longicornis* could be perceived as polyphagous/generalists but they showed an apparent low number of ORs and relatively smaller expansions compared to the rest of the polyphagous herbivores. Our results suggest a relatively reduced OR repertoire of *S. longicornis* compared to the other species, which may result from adaptation to the deep subterranean environment conditions: very limited in primary production, oligotrophic and inhabited by a relatively smaller number of species compared to surface habitats. Moreover the air remains still, saturated with water vapour and the potential evaporation rate is negligible for long time periods (Howarth & Moldovan, 2018). These features suggest an homogeneous habitat, probably less diverse in airborne odourant clues than surface habitats. An extreme contraction of chemosensory gene repertoire, particularly for ORs and OBPs, was observed in the fig wasp *Ceratosolen solmsi* when compared to other hymenopterans, possibly reflecting its high host-specificity (Xiao et al., 2013). Further research comparing high quality transcriptomes and genomes of surface and subterranean species or including a high quality reference genome for the target species would help to validate this conclusion.

Regarding gustatory perception (Figure 4), five GRs were significantly enriched in the antennae in *S. longicornis*, indicating a substantial gustatory role in these appendages. This result is consistent with what was found for GR expression levels of *T. castaneum*, where similar values in the maxillary palps and the antennae were reported (Dippel et al., 2016). Remarkably, no GRs associated with the perception of fructose and other sugars were detected in *S. longicornis* (at least clustering together with functionally annotated genes in *D. melanogaster*) indicating that either *S. longicornis* does not have receptors for these types of carbohydrates, or their evolutionary origin is different from that in other beetles. Further research including transcriptomic data of other structures involved in taste perception (e.g., mouth appendages) would help to better sustain the absence of these highly conserved sugar receptors with the current approach. In insects, the entire loss of sugar receptors has only been documented in some obligate blood feeders (i.e., *Glossina morsitans*, *Cimex lectularius* and *Pediculus humanus*; Benoit et al., 2016; Kirkness et al., 2010; Obiero et al., 2014). Further comparative studies including nonphytophagous beetles inhabiting surface habitats would help to test the hypothesis that a lack of sugar receptors may be directly associated with a strict subterranean lifestyle.

CO_2 perception may be crucial for *S. longicornis* to orientate within its habitat and to detect decomposing organic matter in the darkness, the main food source for this species. We detected three candidate GRs clustering together with highly conserved CO_2 receptors of insects (Robertson & Kent, 2009), among which only one candidate was significantly expressed in antennae (Figure 4). Our results suggest that CO_2 perception may not be restricted to a single chemosensory structure, congruent with what was found in *T. castaneum* after comparing different body structures (Dippel et al., 2016). These results in beetles are in contrast to what was found in well studied dipterans. For instance, *D. melanogaster* has only two CO_2 receptors that form functional heteromers significantly enriched in antennae (i.e., DmelGR21a and DmelGR63a; Dippel et al., 2016; Jones et al., 2007; Kwon et al., 2007), whereas *A. gambiae* has three CO_2 receptors that are upregulated in the mouthparts (AgamGR22-24; Pitts et al., 2011). Further studies exploring differential gene expression in different body parts are needed to deepen our understanding of CO_2 perception in *S. longicornis*.

The GR repertoire of *S. longicornis* was small and similar in size to that observed in the oligophagous *A. planipennis* (ash tree specialist). Our findings on the gustatory perception of this cavernicolous beetle may reflect the poor diversity of gustatory substances in the hypogean habitat compared to surface environments. A notably diminished odourant and gustatory capabilities were reported in *Drosophila sechellia*, an endemic species to the Seychelles that is specialized in feeding on a single plant species (McBride, 2007). Similar tendencies were found in independently specialized drosophilids

(McBride & Arguello, 2007) which showed a remarkable acceleration of gene loss when compared to other species. In contrast, the most remarkable GR family expansion in insects was found in *Periplaneta americana* (i.e. 522 GRs, with the most expanded clade (329 GRs) corresponding to potential bitter receptors), that been related to its omnivorous and opportunistic feeding habits. Moreover this species has also the most expanded OR repertoire of Blattodea species (Li et al., 2018). A large expansion of GRs has also been reported in the polyphagous moth *Helicoverpa armigera* compared to other lepidopterans, again finding bitter receptor GRs as the most expanded clade of the family (Xu et al., 2016).

Contrary to the specific role of IGluRs in synaptic communication, IRs have more diverse roles which in insects are often related to chemoreception (Koh et al., 2014; Rytz et al., 2013). While the most conserved IRs (e.g., IR8a and IR25a) act as coreceptors conferring multiple odour-evoked electrophysiological responses, more recently some insect IRs have been found to mediate specific stimulus forming heterodimers with more selectively expressed IR subunits (Abuin et al., 2011, 2019). For instance, in *D. melanogaster*, the highly conserved coreceptors IR93a and IR25a are coexpressed with IR21a, mediating physiological and behavioral responses to low temperatures (Knecht et al., 2017; Ni et al., 2016). In *S. longicornis* we found overexpression in the antennae of the candidate genes clustering together to *D. melanogaster* IR93a and IR25a, while the candidate IR21a (SlonIR2299c0 g1) was significantly underexpressed in these appendages (Figure 5c). By contrast, higher expression levels for IR21a of *D. melanogaster* were found in the antennae (Sánchez-Alcañiz et al., 2018) and the same was found for other Coleoptera (Bin et al., 2017; Dippel et al., 2016). However, we did not explore differential gene expression in different structures from the body, which could explain the differences in the observed results.

Remarkably, our results suggest a putative duplication in the two genes annotated as IR25a, despite being a highly conserved gene with a single copy in virtually all protostomes (Figure 5b). Duplications have only been reported for the limpet *L. gigantea* and for two parasitoid wasp species: *N. vitripennis* and *M. mediator* (Croset et al., 2010; Wang et al., 2015). The inferred phylogeny suggested different origins for the observed duplication in the IR25a candidates. A lineage-specific duplication has been observed in *N. vitripennis* and *M. mediator*, where both copies of IR25a were retrieved as sister to each other, suggesting either recent duplication or gene conversion. The candidate duplication in *S. longicornis* could represent an ancestral duplication in Coleoptera that was retained in this cave beetle or, alternatively, a recent gene duplication followed by extensive sequence divergence. Our results represent the first report of a gene duplication observed in this highly conserved gene in Coleoptera, which may indicate that the evolutionary history of IR25a and its role in chemoreception may be more complex than originally considered across arthropods.

Cave beetles inhabit a medium where air tends to be still and the ambient temperature and humidity fluctuate only by tiny amounts over long periods, and therefore a sensitive thermal detection may have a selective advantage. On the contrary, it may also result in a loss of selection of thermal detection ability and/or a los ity to adapt to thermal extremes (i.e., climate variability hy Stevens, 1989). Physiological experiments on a close relativ (i.e., the cave-dwelling *Speophyes lucidulus*, Leoididae, Ch revealed an extreme sensitivity to small changes in ten incurred by antennal receptors (Corbière-Tichané & Loft that may be mediated by some of the inferred candi Consequently, other relevant IRs for *S. longicornis* may potentially related to humidity sensing. The functionall terized IR40a and IR68a in *D. melanogaster* have been se coexpressed with IR93a and IR25a in specialized sensory of the antennae performing hygrosensory responses (En Knecht et al., 2017). Through the phylogenetic analysis we the hygroreceptor candidates (IR40a and IR68a) for *S. l* (Figure 5c), although we did not find significant differenc expression values between antennae and the rest of the b rest of the candidate IRs annotated in *S. longicornis* (i.e., IR8 IR75a, IR64a, 100a and IR60a; see Figure 5c; Table S2) h shown to be potentially involved in taste and odour transc *D. melanogaster*, suggesting candidate odourant and gustat in *S. longicornis*.

Regarding OBPs, the GO enrichment analysis hi odourant-binding functions vastly enriched in antennae (F In addition, the differential gene expression analyses ide high number of upregulated genes (17 in antennae; Figure large number of OBPs annotated in *S. longicornis* sugges tively diverse repertoire with species-specific gene du and gene family expansions; this may indicate an importa these proteins in odourant perception in this subterranea Notably, less than half of the upregulated OBPs in anten tered together with the previously described "antenna proteins II" (ABPII) in Vieira and Rozas (2011; Figure 6). sults indicate that although ABPII were described as OBPs enriched in antennae, some genes of this clade may also entially expressed in some other body structures, as also *T. castaneum* (Dippel et al., 2014). The rest of the upregula in both conditions clustered together with the different OE described in previous studies (Andersson et al., 2019; Dip 2014; Vieira & Rozas, 2011). Further research including cl tives to this cave species with different ecological prefere allow us to test the hypothesis that subterranean specializ modified the chemosensory capabilities in Coleoptera.

5 | CONCLUSION

In this study, we characterized for the first time the chem gene repertoire of an obligate subterranean species, t dwelling coleopteran *S. longicornis*. We found relatively d odourant and gustatory repertoires compared to polypha leopterans inhabiting surface habitats and more similar t considered specialists based on their feeding habits. Co the selective pressures of the niche that *S. longicornis* occu

limited resources, poor diversity and heterogeneous distribution of food, among others), an optimized chemosensory repertoire in terms of diversity may result from its adaptation to the deep subterranean environment. In this obligate cave-dwelling beetle, we identified some putative gene losses (e.g., sugar GRs and IR41a). Furthermore, several gene duplications and gene family expansions were observed. A putative duplication of the gene IR25a was identified, which might potentially have facilitated adaptation to subterranean conditions in this cave beetle. Our study thus paves the way towards a better understanding of how subterranean animals perceive their particular environment.

ACKNOWLEDGEMENTS

This work was supported by the Ministerio de Economía y Competitividad and the Ministerio de Ciencia of Spain (CGL2016-76705-P to Ignacio Ribera, PID2019-108824GA-I00 to Rosa Fernández, and CGL2016-75255 and PID2019-103947GB to Julio Rozas). Pau Balart-García and Paula Escuer were supported by an FPI grant (Ministerio de Economía y Competitividad BES-2017-081050 and BES-2017-081740, respectively). Rosa Fernández was supported by a Marie Sklodowska-Curie grant (747607) and a Ramón y Cajal fellowship (Ministerio de Economía y Competitividad, RyC2017-22492). C. Vanderbergh and E. Ruzzier kindly provided the images for Figure 1a,b, respectively; the latter was obtained at the Electron Microscopy Service of the Institute of Marine Sciences (ICM) of Barcelona, Spain. This article is dedicated to the memory of our wonderful colleague, Ignacio Ribera, who worked intensively on this project until his very last days and recently sadly passed away.

AUTHOR CONTRIBUTIONS

Pau Balart-García conceived the study, generated, interpreted and analysed the data, prepared the figures and tables and wrote the first version of the manuscript. Alexandra Cieslak conceived the study, processed the samples and generated the data. Paula Escuer assisted in data analysis and interpretation. Julio Rozas provided computational resources and assisted in data interpretation. Ignacio Ribera conceived the study, provided resources and interpreted the data. Rosa Fernández conceived and supervised the study, analysed and interpreted the data, and wrote the first version of the manuscript. All authors contributed to the final version of the manuscript.

DATA AVAILABILITY STATEMENT

Raw reads have been deposited in the National Center for Biotechnology Information (NCBI; BioProject accession number PRJNA667243, BioSample accession number SAMN16362632). Assembled sequences, alignments and custom scripts have been deposited in github (https://github.com/MetazoaPhylogenomicsLab/Balart_Garcia_et_al_2021_MolecularEcology).

ORCID

Pau Balart-García https://orcid.org/0000-0002-8292-4674
Alexandra Cieslak https://orcid.org/0000-0003-3002-7995
Paula Escuer https://orcid.org/0000-0002-5941-0106
Julio Rozas https://orcid.org/0000-0002-6839-9148
Ignacio Ribera https://orcid.org/0000-0002-2791-7615
Rosa Fernández https://orcid.org/0000-0002-4719-664

REFERENCES

Abuin, L., Bargeton, B., Ulbrich, M. H., Isacoff, E. Y., Kellenbe
Benton, R. (2011). Functional architecture of olfactory
glutamate receptors. *Neuron*, *69*(1), 44–60.
Abuin, L., Prieto-Godino, L. L., Pan, H., Gutierrez, C., Huang,
& Benton, R. (2019). In vivo assembly and trafficking of
Ionotropic Receptors. *BMC Biology*, *17*(1), 34.
Almudi, I., Vizueta, J., Wyatt, C. D. R., de Mendoza, A., Ma
Firbas, P. N., Feuda, R., Masiero, G., Medina, P., Alcaina
Cruz, F., Gómez-Garrido, J., Gut, M., Alioto, T. S., Varga
C., Davie, K., Misof, B., González, J., Aerts, S., … Casares,
Genomic adaptations to aquatic and aerial life in mayfli
origin of insect wings. *Nature Communications*, *11*(1), 263
Al-Shahrour, F., Minguez, P., Tárraga, J., Medina, I., Alloza, E.,
D., & Dopazo, J. (2007). FatiGO +: A functional profilin
genomic data. Integration of functional annotation, regul
tifs and interaction data with microarray experiments. *Nu
Research*, *35*(Web Server issue), W91–W96.
Andersson, M. N., Keeling, C. I., & Mitchell, R. F. (2019). Gen
tent of chemosensory genes correlates with host range
boring beetles (*Dendroctonus ponderosae*, *Agrilus planip
Anoplophora glabripennis*). *BMC Genomics*, *20*(1), 690.
Anholt, R. R. H. (2020). Chemosensation and evolution of
host plant selection. *iScience*, *23*(1), 100799.
Benjamini, Y., & Hochberg, Y. (1995). Controlling the false disc
A practical and powerful approach to multiple testing. *Jou
Royal Statistical Society: Series B*, *57*(1), 289–300.
Benoit, J. B., Adelman, Z. N., Reinhardt, K., Dolan, A., Poe
Jennings, E. C., Szuter, E. M., Hagan, R. W., Gujar, H., Shukla
F., Mohan, M., Nelson, D. R., Rosendale, A. J., Derst, C.,
Wernig, S., Menegazzi, P., Wegener, C., … Richards, S. (201
features of a global human ectoparasite identified throug
ing of the bed bug genome. *Nature Communications*, *7*(1), 1
Benton, R. (2015). Multigene family evolution: Perspectives f
chemoreceptors. *Trends in Ecology & Evolution*, *30*(10), 59
Benton, R., Vannice, K. S., Gomez-Diaz, C., & Vosshall, L.
Variant ionotropic glutamate receptors as chemosensory
in *Drosophila*. *Cell*, *136*(1), 149–162.
Bin, S.-Y., Qu, M.-Q., Li, K.-M., Peng, Z.-Q., Wu, Z.-Z., & Lin, J.
Antennal and abdominal transcriptomes reveal chen
gene families in the coconut hispine beetle, *Brontispa
Scientific Reports*, *7*(1), 2809.
Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A fle
mer for Illumina sequence data. *Bioinformatics*, *30*(15), 21
Brand, P., Robertson, H. M., Lin, W., Pothula, R., Klingeman, W
Fuentes, J. L., & Johnson, B. R. (2018). The origin of th
receptor gene family in insects. *eLife*, *7*, e38340.
Capella-Gutiérrez, S., Silla-Martínez, J. M., & Gabaldón, T. (20
A tool for automated alignment trimming in large-scale
netic analyses. *Bioinformatics*, *25*(15), 1972–1973.
Cartwright, R. A., Schwartz, R. S., Merry, A. L., & Howell, M.
The importance of selection in the evolution of blindnes
fish. *BMC Evolutionary Biology*, *17*(1), 45.
Cieslak, A., Fresneda, J., & Ribera, I. (2014). Life-history specializ
not an evolutionary dead-end in Pyrenean cave beetles. *P
of the Royal Society B: Biological Sciences*, *281*(1781), 2013
Clyne, P. J., Warr, C. G., Freeman, M. R., Lessing, D., Kim, J.,
J. R. (1999). A novel family of divergent seven-transm

B

proteins: Candidate odorant receptors in *Drosophila*. *Neuron*, *22*(2), 327–338.

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., & de Hoon, M. J. L. (2009). Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, *25*(11), 1422–1423.

Corbière-Tichané, G., & Loftus, R. (1983). Antennal thermal receptors of the cave beetle *Speophyes lucidulus* Delar. *Journal of Comparative Physiology*, *153*(3), 343–351.

Croset, V., Rytz, R., Cummins, S. F., Budd, A., Brawand, D., Kaessmann, H., Gibson, T. J., & Benton, R. (2010). Ancient protostome origin of chemosensory ionotropic glutamate receptors and the evolution of insect taste and olfaction. *PLoS Genetics*, *6*(8), e1001064.

Culver, D. C., & Pipan, T. (2019). *The biology of caves and other subterranean habitats*. Oxford University Press.

Delay, B. (1978). Milieu souterrain et ecophysiologie de la reproduction et du développement des coléoptères Bathysciinae hypoges. *Mémoires De Biospéologie*, *5*, 1–349.

Deleurance, S. (1963). Recherches sur les coléoptères troglobies de la sous-famille des Bathysciinae. *Annales Des Sciences Naturelles (Paris) Zoologie*, *1*, 1–172.

Dippel, S., Kollmann, M., Oberhofer, G., Montino, A., Knoll, C., Krala, M., Rexer, K.-H., Frank, S., Kumpf, R., Schachtner, J., & Wimmer, E. A. (2016). Morphological and transcriptomic analysis of a beetle chemosensory system reveals a Gnathal Olfactory Center. *BMC Biology*, *14*(1), 90.

Dippel, S., Oberhofer, G., Kahnt, J., Gerischer, L., Opitz, L., Schachtner, J., Stanke, M., Schütz, S., Wimmer, E. A., & Angeli, S. (2014). Tissue-specific transcriptomics, chromosomal localization, and phylogeny of chemosensory and odorant binding proteins from the red flour beetle *Tribolium castaneum* reveal subgroup specificities for olfaction or more general functions. *BMC Genomics*, *15*(1), 1141.

Enjin, A. (2017). Humidity sensing in insects-from ecology to neural processing. *Current Opinion in Insect Science*, *24*, 1–6.

Eyun, S.-I., Soh, H. Y., Posavi, M., Munro, J. B., Hughes, D. S. T., Murali, S. C., Qu, J., Dugan, S., Lee, S. L., Chao, H., Dinh, H., Han, Y. I., Doddapaneni, H. V., Worley, K. C., Muzny, D. M., Park, E.-O., Silva, J. C., Gibbs, R. A., Richards, S., & Lee, C. E. (2017). Evolutionary history of chemosensory-related gene families across the Arthropoda. *Molecular Biology and Evolution*, *34*(8), 1838–1862.

Fernández, R., & Gabaldón, T. (2020). Gene gain and loss across the metazoan tree of life. *Nature Ecology & Evolution*, *4*(4), 524–533.

Gao, Q., & Chess, A. (1999). Identification of candidate *Drosophila* olfactory receptors from genomic DNA sequence. *Genomics*, *60*(1), 31–39.

Glaçon, S. (1953). The evolutive cycle of a cave coleopter *Speonomus longicornis* Saulcy. *Comptes Rendus Hebdomadaires Des Seances De L'academie Des Sciences*, *236*(25), 2443–2445.

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B. W., Nusbaum, C., Lindblad-Toh, K., … Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, *29*(7), 644–652.

Grimaldi, D., & Engel, M. S. (2005). *Evolution of the insects*. Cambridge University Press.

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., Couger, M. B., Eccles, D., Li, B. O., Lieber, M., MacManes, M. D., Ott, M., Orvis, J., Pochet, N., Strozzi, F., Weeks, N., Westerman, R., William, T., Dewey, C. N., … Regev, A. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols*, *8*(8), 1494–1512.

Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q., & Vinh, L. S. (2018). UFBoot2: Improving the ultrafast bootstrap approximation. *Molecular Biology and Evolution*, *35*(2), 518–522.

Howarth, F. G., & Moldovan, O. T. (2018). Where cave animals live. In O. T. Moldovan, Ľ. Kováč, & S. Halse (Eds.), *Cave ecology* (pp. 23–37). Springer International Publishing.

Huerta-Cepas, J., Forslund, K., Coelho, L. P., Szklarczyk, D., Jensen, L. J., von Mering, C., & Bork, P. (2017). Fast genome-wide functional annotation through orthology assignment by eggNOG-Mapper. *Molecular Biology and Evolution*, *34*(8), 2115–2122.

Jeannel, R. (1924). Monographie des Bathysciinae. *Archives De Zoologie Expérimentale Et Générale (Paris)*, *63*, 1–436.

Jones, W. D., Cayirlioglu, P., Kadow, I. G., & Vosshall, L. B. (2007). Two chemosensory receptors together mediate carbon dioxide detection in *Drosophila*. *Nature*, *445*(7123), 86–90.

Joseph, R. M., & Carlson, J. R. (2015). *Drosophila* chemoreceptors: A molecular interface between the chemical world and the brain. *Trends in Genetics*, *31*(12), 683–695.

Kirkness, E. F., Haas, B. J., Sun, W., Braig, H. R., Perotti, M. A., Clark, J. M., Lee, S. H., Robertson, H. M., Kennedy, R. C., Elhaik, E., Gerlach, D., Kriventseva, E. V., Elsik, C. G., Graur, D., Hill, C. A., Veenstra, J. A., Walenz, B., Tubio, J. M. C., Ribeiro, J. M. C., … Pittendrigh, B. R. (2010). Genome sequences of the human body louse and its primary endosymbiont provide insights into the permanent parasitic lifestyle. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(27), 12168–12173.

Knecht, Z. A., Silbering, A. F., Cruz, J., Yang, L., Croset, V., Benton, R., & Garrity, P. A. (2017). Ionotropic receptor-dependent moist and dry cells control hygrosensation in Drosophila. *eLife*, *6*(e26654), 1–11. https://doi.org/10.7554/elife.26654

Koh, T.-W., He, Z., Gorur-Shandilya, S., Menuz, K., Larter, N. K., Stewart, S., & Carlson, J. R. (2014). The *Drosophila* IR20a clade of ionotropic receptors are candidate taste and pheromone receptors. *Neuron*, *83*(4), 850–865.

Kwon, J. Y., Dahanukar, A., Weiss, L. A., & Carlson, J. R. (2007). The molecular basis of CO2 reception in *Drosophila*. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(9), 3574–3578.

Laetsch, D. R., & Blaxter, M. L. (2017). BlobTools: Interrogation of genome assemblies. *F1000Research*, *6*, 1287.

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–359.

Letunic, I., & Bork, P. (2019). Interactive Tree Of Life (iTOL) v4: Recent updates and new developments. *Nucleic Acids Research*, *47*(W1), W256–W259.

Li, S., Zhu, S., Jia, Q., Yuan, D., Ren, C., Li, K., Liu, S., Cui, Y., Zhao, H., Cao, Y., Fang, G., Li, D., Zhao, X., Zhang, J., Yue, Q., Fan, Y., Yu, X., Feng, Q., & Zhan, S. (2018). The genomic and functional landscapes of developmental plasticity in the American cockroach. *Nature Communications*, *9*(1), 1008.

Luo, X.-Z., Antunes-Carvalho, C., Wipfler, B., Ribera, I., & Beutel, R. G. (2019). The cephalic morphology of the troglobiontic cholevine species *Troglocharinus ferreri* (Coleoptera, Leiodidae). *Journal of Morphology*, *280*(8), 1207–1221.

McBride, C. S. (2007). Rapid evolution of smell and taste receptor genes during host specialization in *Drosophila sechellia*. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(12), 4996–5001.

McBride, C. S., & Arguello, J. R. (2007). Five *Drosophila* genomes reveal nonneutral evolution and the signature of host specialization in the chemoreceptor superfamily. *Genetics*, *177*, 1395–1416.

McKenna, D. D., Shin, S., Ahrens, D., Balke, M., Beza-Beza, C., Clarke, D. J., Donath, A., Escalona, H. E., Friedrich, F., Letsch, H., Liu, S., Maddison, D., Mayer, C., Misof, B., Murin, P. J., Niehuis, O., Peters, R. S., Podsiadlowski, L., Pohl, H., … Beutel, R. G. (2019). The evolution and

genomic basis of beetle diversity. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(49), 24729–24737.

Mirarab, S., Nguyen, N., Guo, S., Wang, L.-S., Kim, J., & Warnow, T. (2015). PASTA: Ultra-large multiple sequence alignment for nucleotide and amino-acid sequences. *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology*, *22*(5), 377–386.

Missbach, C., Dweck, H. K. M., Vogel, H., Vilcinskas, A., Stensmyr, M. C., Hansson, B. S., & Grosse-Wilde, E. (2014). Evolution of insect olfactory receptors. *eLife*, *3*(e02115), 1–22. https://doi.org/10.7554/elife.02115

Missbach, C., Vogel, H., Hansson, B. S., & 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. *Chemical Senses*, *40*(9), 615–626.

Mitchell, R. F., Schneider, T. M., Schwartz, A. M., Andersson, M. N., & McKenna, D. D. (2019). The diversity and evolution of odorant receptors in beetles (Coleoptera). *Insect Molecular Biology*, *29*(1), 77–91.

Nei, M., Niimura, Y., & Nozawa, M. (2008). The evolution of animal chemosensory receptor gene repertoires: Roles of chance and necessity. *Nature Reviews Genetics*, *9*(12), 951–963.

Nei, M., & Rooney, A. P. (2005). Concerted and birth-and-death evolution of multigene families. *Annual Review of Genetics*, *39*(1), 121–152.

Ngoc, P. C. T., Greenhalgh, R., Dermauw, W., Rombauts, S., Bajda, S., Zhurov, V., Grbić, M., Van de Peer, Y., Van Leeuwen, T., Rouzé, P., & Clark, R. M. (2016). Complex evolutionary dynamics of massively expanded chemosensory receptor families in an extreme generalist chelicerate herbivore. *Genome Biology and Evolution*, *8*(11), 3323–3339.

Nguyen, L.-T., Schmidt, H. A., von Haeseler, A., & Minh, B. Q. (2015). IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution*, *32*(1), 268–274.

Ni, L., Klein, M., Svec, K. V., Budelli, G., Chang, E. C., Ferrer, A. J., Benton, R., Samuel, A. D. T., & Garrity, P. A. (2016). The ionotropic receptors IR21a and IR25a mediate cool sensing in Drosophila. *eLife*, *5*(e13254), 1–12. https://doi.org/10.7554/eLife.13254

Nichols, Z., & Vogt, R. G. (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 Biochemistry and Molecular Biology*, *38*(4), 398–415.

Obiero, G. F. O., Mireji, P. O., Nyanjom, S. R. G., Christoffels, A., Robertson, H. M., & Masiga, D. K. (2014). Odorant and gustatory receptors in the tsetse fly *Glossina morsitans morsitans*. *PLoS Neglected Tropical Diseases*, *8*(4), e2663.

Pallarés, S., Colado, R., Botella-Cruz, M., Montes, A., Balart-García, P., Bilton, D. T., Sánchez-Fernández, D. (2020). Loss of heat acclimation capacity could leave subterranean specialists highly sensitive to climate change. *Animal Conservation*. https://doi.org/10.1111/acv.12654. [Epub ahead of print].

Parzefall, J. (2001). A review of morphological and behavioural changes in the cave molly, *Poecilia mexicana*, from Tabasco, Mexico. *Environmental Biology of Fishes*, *62*, 263–275.

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*, *14*(4), 417–419.

Pelosi, P., Iovinella, I., Felicioli, A., & Dani, F. R. (2014). Soluble proteins of chemical communication: An overview across arthropods. *Frontiers in Physiology*, *5*, 320.

Pelosi, P., Iovinella, I., Zhu, J., Wang, G., & Dani, F. R. (2018). Beyond chemoreception: Diverse tasks of soluble olfactory proteins in insects. *Biological Reviews of the Cambridge Philosophical Society*, *93*(1), 184–200.

Picelli, S., Björklund, Å. K., Faridani, O. R., Sagasser, S., Winberg, G., & Sandberg, R. (2013). Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nature Methods*, *10*(11), 1096–1098.

Pitts, R. J., Rinker, D. C., Jones, P. L., Rokas, A., & Zwiebel, L. J. (2011). Transcriptome profiling of chemosensory appendages in the malaria vector *Anopheles gambiae* reveals tissue- and sex-specific signatures of odor coding. *BMC Genomics*, *12*, 271.

Price, M. N., Dehal, P. S., & Arkin, A. P. (2010). FastTree 2–approximately maximum-likelihood trees for large alignments. *PLoS One*, *5*(3), e9490.

Ribera, I., Fresneda, J., Bucur, R., Izquierdo, A., Vogler, A. P., Salgado, J. M., & Cieslak, A. (2010). Ancient origin of a Western Mediterranean radiation of subterranean beetles. *BMC Evolutionary Biology*, *10*, 29.

Rizzo, V., Sánchez-Fernández, D., Fresneda, J., Cieslak, A., & Ribera, I. (2015). Lack of evolutionary adjustment to ambient temperature in highly specialized cave beetles. *BMC Evolutionary Biology*, *15*, 10.

Robertson, H. M. (2019). Molecular evolution of the major arthropod chemoreceptor gene families. *Annual Review of Entomology*, *64*(1), 227–242.

Robertson, H. M., & Kent, L. B. (2009). Evolution of the gene lineage encoding the carbon dioxide receptor in insects. *Journal of Insect Science*, *9*(19), 1–14.

Robertson, H. M., Warr, C. G., & Carlson, J. R. (2003). Molecular evolution of the insect chemoreceptor gene superfamily in *Drosophila melanogaster*. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(Suppl. 2), 14537–14542.

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, *26*(1), 139–140.

Robinson, M. D., & Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, *11*(3), R25.

Roys, C. (1954). Olfactory nerve potentials a direct measure of chemoreception in insects. *Annals of the New York Academy of Sciences*, *58*(2), 250–255.

Rytz, R., Croset, V., & Benton, R. (2013). Ionotropic receptors (IRs): Chemosensory ionotropic glutamate receptors in *Drosophila* and beyond. *Insect Biochemistry and Molecular Biology*, *43*(9), 888–897.

Sambrook, J., Fritsch, E. F., & Maniatis, T. (1989). *Molecular cloning: A laboratory manual*. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.

Sánchez-Alcañiz, J. A., Silbering, A. F., Croset, V., Zappia, G., Sivasubramaniam, A. K., Abuin, L., Sahai, S. Y., Münch, D., Steck, K., Auer, T. O., Cruchet, S., Neagu-Maier, G. L., Sprecher, S. G., Ribeiro, C., Yapici, N., & Benton, R. (2018). An expression atlas of variant ionotropic glutamate receptors identifies a molecular basis of carbonation sensing. *Nature Communications*, *9*(1), 4252.

Sánchez-Gracia, A., Vieira, F. G., & Rozas, J. (2009). Molecular evolution of the major chemosensory gene families in insects. *Heredity*, *103*(3), 208–216.

Schoonhoven, L. M., van Loon, J. J. A., & Dicke, M. (2005). *Insect-plant biology* (2nd ed., pp. 6–24). Oxford University Press.

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, *31*(19), 3210–3212.

Stengl, M., & Funk, N. W. (2013). The role of the coreceptor Orco in insect olfactory transduction. *Journal of Comparative Physiology A*, *199*(11), 897–909.

Stevens, G. C. (1989). The latitudinal gradient in geographical range: How so many species coexist in the tropics. *The American Naturalist*, *133*(2), 240–256.

Stocker, R. F. (1994). The organization of the chemosensory system in *Drosophila melanogaster*: A rewiew. *Cell and Tissue Research*, *275*(1), 3–26.

Stürckow, B. (1970). Responses of olfactory and gustatory receptor cells in insects. *Advances in Chemoreception*, *1*, 107–159.

Supek, F., Bošnjak, M., Škunca, N., & Šmuc, T. (2011). REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS One*, *6*(7), e21800.

Thoma, M., Missbach, C., Jordan, M. D., Grosse-Wilde, E., Newcomb, R. D., & Hansson, B. S. (2019). Transcriptome surveys in silverfish suggest a multistep origin of the insect odorant receptor gene family. *Frontiers in Ecology and Evolution*, *7*, 281.

Turk, S., Sket, B., & Sarbu, Ş. (1996). Comparison between some epigean and hypogean populations of *Asellus aquaticus* (Crustacea: Isopoda: Asellidae). *Hydrobiologia*, *337*(1–3), 161–170.

Vieira, F. G., & 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.

Vieira, F. G., Sánchez-Gracia, A., & Rozas, J. (2007). Comparative genomic analysis of the odorant-binding protein family in 12 *Drosophila* genomes: Purifying selection and birth-and-death evolution. *Genome Biology*, *8*(11), R235.

Vizueta, J., Escuer, P., Frías-López, C., Guirao-Rico, S., Hering, L., Mayer, G., Rozas, J., & Sánchez-Gracia, A. (2020). Evolutionary history of major chemosensory gene families across Panarthropoda. *Molecular Biology and Evolution*, *37*(12), 3601–3615.

Vizueta, J., Frías-López, C., Macías-Hernández, N., Arnedo, M. A., Sánchez-Gracia, A., & Rozas, J. (2016). Evolution of chemosensory gene families in arthropods: Insight from the first inclusive comparative transcriptome analysis across spider appendages. *Genome Biology and Evolution*, *9*(1), 178–196.

Vizueta, J., Sánchez-Gracia, A., & Rozas, J. (2020). BITACORA: A comprehensive tool for the identification and annotation of gene families in genome assemblies. *Molecular Ecology Resources*, *20*, 1445–1452.

Vogel, H., Shukla, S. P., Engl, T., Weiss, B., Fischer, R., Steiger, S., Heckel, D. G., Kaltenpoth, M., & Vilcinskas, A. (2017). The digestive and defensive basis of carcass utilization by the burying beetle and its microbiota. *Nature Communications*, *8*, 15186.

Vosshall, L. B., Amrein, H., Morozov, P. S., Rzhetsky, A., & Axel, R. (1999). A spatial map of olfactory receptor expression in the *Drosophila* antenna. *Cell*, *96*(5), 725–736.

Vosshall, L. B., & Stocker, R. F. (2007). Molecular architecture of smell and taste in *Drosophila*. *Annual Review of Neuroscience*, *30*, 505–533.

Wang, D., Pentzold, S., Kunert, M., Groth, M., Brandt, W., Pasteels, J. M., Boland, W., & Burse, A. (2018). A subset of chemosensory genes differs between two populations of a specialized leaf beetle after host plant shift. *Ecology and Evolution*, *8*(16), 8055–8075.

Wang, H.-C., Minh, B. Q., Susko, E., & Roger, A. J. (2018). Modeling site heterogeneity with posterior mean site frequency profiles accelerates accurate phylogenomic estimation. *Systematic Biology*, *67*, 216–235.

Wang, S.-N., Peng, Y., Lu, Z.-Y., Dhiloo, K. H., Gu, S.-H., Li, R.-J., & Guo, Y.-Y. (2015). Identification and expression analysis of putative chemosensory receptor genes in *Microplitis mediator* by antennal transcriptome screening. *International Journal of Biological Sciences*, *11*(7), 737–751.

Wu, Z., Bin, S., He, H., Wang, Z., Li, M., & Lin, J. (2016). Differential expression analysis of chemoreception genes in the striped Flea Beetle *Phyllotreta striolata* using a transcriptomic approach. *PLoS One*, *11*(4), e0153067.

Xu, W., Papanicolaou, A., Zhang, H.-J., & Anderson, A. (2016). Expansion of a bitter taste receptor family in a polyphagous insect herbivore. *Scientific Reports*, *6*, 23666.

Yamamoto, Y., Byerly, M. S., Jackman, W. R., & Jeffery, W. R. (2009). Pleiotropic functions of embryonic sonic hedgehog expression link jaw and taste bud amplification with eye loss during cavefish evolution. *Developmental Biology*, *330*(1), 200–211.

Yang, J., Chen, X., Bai, J., Fang, D., Qiu, Y., Jiang, W., Yuan, H., Bian, C., Lu, J., He, S., Pan, X., Zhang, Y., Wang, X., You, X., Wang, Y., Sun, Y., Mao, D., Liu, Y., Fan, G., … Shi, Q. (2016). The *Sinocyclocheilus* cavefish genome provides insights into cave adaptation. *BMC Biology*, *14*, 1.

Xiao, J. H., Yue, Z., Jia, L. Y., Yang, X. H., Niu, L. H., Wang, Z., Zhang, P., Sun, B. F., He, S. M., Li, Z., Xiong, T. L., Xin, W., Gu, H. F., Wang, B, Werren, J. H., Murphy, R. W, Wheeler, D., Niu, L. M., Ma, G. C., … Huang, D. W. (2013). Obligate mutualism within a host drives the extreme specialization of a fig wasp genome. *Genome Biology*, *14*(12), R141. http://dx.doi.org/10.1186/gb-2013-14-12-r141

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Balart-García P, Cieslak A, Escuer P, Rozas J, Ribera I, Fernández R. Smelling in the dark: Phylogenomic insights into the chemosensory system of a subterranean beetle. *Mol Ecol*. 2021;30:2573–2590. https://doi.org/10.1111/mec.15921

C
Funding

This work was supported by the Ministerio de Ciencia e Innovación of Spain (CGL2016-75255 and PID2019-103947GB) and the Comissió Interdepartamental de Recerca i Innovació Tecnològica of Catalonia, Spain (2017SGR83). Paula Escuer Pifarré was supported by a FPI grant (Ministerio de Ciencia e Innovación of Spain, BES-2017-081740).

Paula Escuer Pifarré
PhD Thesis 2022