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

Paula Escuer[1], Vadim Pisarenco[1], Angel Fernández-Ruiz[1], Joel Vizueta[1], Jose Sanchez-Herrero[2], Miquel Arnedo[1], Alejandro Sánchez-Gracia[1], and Julio Rozas[1]

[1]Universitat de Barcelona
[2]Institut d'Investigació en Ciències de la Salut Germans Trias i Pujol

May 10, 2021

Abstract

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; and scaffold N50 of 174.2 Mb), was performed using the chromosome conformation capture scaffolding technique, represents a continuity improvement of more than 4,500 times with respect to the previous version. The seven largest scaffolds or pseudochromosomes cover 87% of the total assembly size and match consistently with the seven chromosomes of the karyotype of this species, including the 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.

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, Jonathan A., Giribet, G., Harvey, M. S., Prendini, L., and Walter, 2004; Garb, Sharma, & Ayoub, 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), there are currently very few available genome sequences of species from this group (Garb et al., 2018; Thomas et al., 2020; Vizueta, Rozas, & Sánchez-Gracia, 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 Catalog, 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 (Wheeler et al., 2017; Kallal et al., 2020). The nocturnal ground-dwelling genus *Dysdera* Latreille 1804, which contains 286 species (World Spider Catalog, 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, Oromí, Múrria, Macías-Hernández, & Ribera, 2007; Arnedo, Oromí, & Ribera, 2001; Macías-Hernández, López, Roca-Cusachs, Oromí, & Arnedo, 2016; Vizueta, Macías-Hernández, Arnedo, Rozas, & Sánchez-Gracia, 2019). Shifts in dietary preferences have been identified as one of the main drivers of island diversification in this group (Řezáč, Pekár, Arnedo, Macías-Hernández, & Řezáčová, 2021). Indeed, *Dysdera* includes some of the few reported cases of stenophagy (i.e. prey specialization) across the mostly generalists spiders (Pekár, Líznarová, & Řezáč, 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), behavioral (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, & Lubin, 2008; Řezáč & Pekár, 2007; 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 eco-phenotypic 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, Gurdasani, Mentzer, Porter, & Sandhu, 2018; Saha, 2019).

We present the first high-quality chromosome-level assembly of the species *D. silvatica (D. silvatica* genome draft, 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.

Material and Methods

Sampling, DNA Sequencing and Genome assembly

The new genome assembly of *D. silvatica* was obtained using the previous genome assembly (denoted as version 1; Sánchez-Herrero et al., 2019), which was further scaffolded using Chicago and Hi-C libraries. Version 1 assembly 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 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 (Santa Cruz; California). 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) was generated using the HiRise scaffolding software (Putnam et al., 2016), and further polished with NextPolish (Hu, Fan, Sun, & Liu, 2020) using illumina short-read data from a single individual unequivocally identified as *D. silvatica* (see results and discussion section). We determined the completeness of the new assembly by applying the pipeline of Benchmarking Universal Single Copy Orthologs (BUSCO; v. 5.0.0; Seppey, Manni, & Zdobnov, 2019), searching the arachnida (odb10; 2934 genes), arthropoda (odb10; 1013 genes) and eukaryota (odb10; 255 genes) datasets.

Repetitive elements identification

We analyzed repetitive regions in the new assembly using a combination of a *de novo* transposable elements identification with RepeatModeler v1.0.11 (Smit, & Hubley; RepeatModeler), and a similarity-based search with RepeatMasker v.4.0.7 (Smit, Hubley, & Green; 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.

Structural and functional genome annotation

Structural annotation of the soft-masked genome was accomplished with BRAKER2 (Brůna, Hoff, Lomsadze, Stanke, & Borodovsky, 2021; Hoff, Lomsadze, Borodovsky, & Stanke, 2019; Stanke, Diekhans, Baertsch, & Haussler, 2008; Stanke, Schöffmann, Morgenstern, & Waack, 2006), 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, Lin, Mak, & Lowe, 2019).

Homology searches

We searched for homologs of *D. silvatica* in the genome data of representatives of a broad taxonomic range across the order Araneae, namely *Acanthoscurria geniculata* (C. L. 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*S. dumicola* Pocock, 1898 (Eresidae), and the Araneoidea*Parasteatoda tepidariorum* (C. L. 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 Sanchez-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, Lima, Guerrero, Piesman, & Black IV, 2005), the mite *Tetranychus urticae* C. L. Koch, 1836 (Grbić et al., 2011), and the spiders *Stegodyphus mimosarum* (Sanggaard et al., 2014) and *Parasteatoda tepidariorum*(Schwager et al., 2017).

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, Frıas-Lopez, et al., 2020; Vizueta et al., 2018). For this task, we used the pipeline BITACORA (Vizueta, Escuer, Sánchez-Gracia, & Rozas, 2020; Vizueta, Sánchez-Gracia, & Rozas, 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 re-annotated 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).

Cluster definition and analysis

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 analyzing the distribution of pairwise physical distances between the members of a particular gene family and scaffold. We classified the paralogous copies as "clustered" and "non clustered". 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 cut-offC_L value following Vieira, Sánchez-Gracia, & Rozas (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 = 0.0004 (Poisson distribution, $\lambda = 0.0301$); this p -value is even lower for the *Gr* family. Thus, the selected g guarantees conservativeC_L lengths for the two chemoreceptor families. Pairwise physical distances between gene family copies were processed with the R package ComplexHeatmap (Gu, Eils, & Schlesner, 2016), and plotted as heatmaps to facilitate the visualization of gene clustering across scaffolds.

Phylogenetic and evolutionary analyses

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, 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 (MSAp).

Physical versus 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, Stecher, Peterson, & Tamura, 2012), using the JTT substitution model (Jones, Taylor, & Thornton, 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, 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:

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

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

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

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:

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

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

where n is the number of copies in cluster k , anddij 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.

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).

Results and Discussion

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 4 scaffolds, respectively, also demonstrate the high continuity of the assembly. The seven largest scaffolds (or pseudochromosomes), including the larger scaffold that likely 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 37,198 putative tRNAs; this annotation includes the 90% and 95% of the BUSCO arthropoda and eukaryote data sets, respectively, evidencing the completeness of the annotation (Table 2). 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 homologs 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 (63.5%), 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 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 (2,200 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 dataset 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 more similar than those obtained in Sánchez-Herrero et al. (2019), reflecting that the vast improvement in assembly continuity does not affect orthology inference quality.

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 4,500 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, Oromí & Ribera, 1997, a phylogenetically close relative to *D. silvatica* also endemic from La Gomera (Arnedo et al., 2007; see also Adrián-Serrano, Lozano-Fernandez, Pons, Rozas, & Arnedo (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.

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, likely resulted from a constant and prolonged 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 homolog 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, Baits, Walden, Wada-Katsumata, & Schal, 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 likely 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.

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, 5C, S2A and S2C). These clusters, which harbor 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 highC_{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 < 0.05 for nearly all pseudochromosomes) but also at the whole genome levels (p -values < 0.001 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 paralogs that are retained long enough (i.e., those that are not lost by genetic drift or purifying selection), continuously diverged at the sequence, and likely at the functional level (at least in terms of ligand specificity or signaling characteristics). We expect, therefore, that over time, evolutionary distances of these retained copies increase with physical distance, just as we have found (Figures 5B, 5D, S2B and S2D). 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.

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, Frıas-Lopez, et al., 2020; Vizueta et al., 2018). However, after 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 co-receptors (*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 (likely 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 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 non-selective 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, Frıas-Lopez, 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.

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, e.g. 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. melanogasteriGluRs* 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*).

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.

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

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. Sanchez-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

DATA AVAILABILITY STATEMENT

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 QLNU02000000. This project repository includes raw data, sequencing libraries information, and current version assembly. Protein sequence data of all chemoreceptor proteins (including incomplete fragments) identified in this study are provided in the Supplementary Material online.

Tables

Table 1. Genome assembly statistics

Table 2. Genome annotation statistics

Table 3. Chromosome level statistics

Figure Legends

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 *D. silvatica* genome are denoted with an asterisk.

Figure 2. Long-range contact heat-map 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 color of each square gives the number of read pairs within that bin. The seven largest scaffolds, which likely 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).

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 *D. 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, *P. tepidariorum* (Ptep), *S. 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 orthologs), 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 4 species, respectively, that are not included in the two previous categories. The blue bar shows other more complex homologous relationships.

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.

Figure 5. Genome organization and relationships between physical and evolutionary distances of the members of the *Gr* and*Ir* families on the *D. silvatica* pseudochromosome 1. A and C) Heatmaps illustrating the distribution of physical distances (in units of 100 kb) 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. Colored and grey points show distances within and outside genomic clusters, respectively. Different clusters are depicted in different colors.

Figure 6. Phylogenetic relationships among the members of the*Gr* family of *D. silvatica* and *D. melanogaster* . The tree only includes the copies of *D. silvatica* classified as complete genes. The outer ring indicates the chromosome (Chr; in different colors) 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.

Figure 7. Phylogenetic relationships among the members of the*Ir/IGluR* family of *D. silvatica* and *D. 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 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* , respectively. The red triangles mark the *D. silvatica* genes with putative distant homologs 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.

Supplementary material

Supplementary Tables

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 multiple sequence alignments used in the analyses are provided in the Supplementary Material online.

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 distances (in units of 100 kb) 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. 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). 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.

Bibliography

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. doi: 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. doi: 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. doi: 10.1006/clad.2001.0168

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., . . . & Sherlock, G. (2000). Gene Ontology: tool for the unification of biology. *Nature Genetics* , *25* (1), 25–29.

Bleidorn, C. (2016). Third generation sequencing: technology and its potential impact on evolutionary biodiversity research.*Systematics and Biodiversity* , *14* (1), 1–8. doi: 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. doi: 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). doi: 10.1101/614032

Clifton, B. D., Jimenez, J., Kimura, A., Chahine, Z., Librado, P., Sanchez-Gracia, A., . . . 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. doi: 10.1093/molbev/msaa109

Coddington, Jonathan A., Giribet, G., Harvey, M. S., Prendini, L., and Walter, D. E. (2004). “Arachnida.” in Assembling the Tree of Life.*Oxford University Press* , 296–318.

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

Eddy, S. R. (2011). Accelerated profile HMM searches. *PLoS Computational Biology* , *7* (10). doi: 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. doi: 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. doi: 10.1016/j.cois.2017.11.005

Grbić, M., Van Leeuwen, T., Clark, R. M., Rombauts, S., Rouzé, P., Grbić, V., . . . Van De Peer, Y. (2011). The genome of Tetranychus urticae reveals herbivorous pest adaptations. *Nature* ,*479* (7374), 487–492. doi: 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. doi: 10.1093/bioinformatics/btw313

Hoff, K. J., Lomsadze, A., Borodovsky, M., & Stanke, M. (2019). Whole-genome annotation with BRAKER. *Methods in Molecular Biology* , *1962* (0), 65–95. doi: 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. doi: 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. doi: 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.

Jones, P., Binns, D., Chang, H. Y., Fraser, M., Li, W., McAnulla, C., . . . Hunter, S. (2014). InterProScan 5: Genome-scale protein function classification. *Bioinformatics* , *30* (9), 1236–1240. doi: 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* , 1–19. doi: 10.1111/cla.12439

Kanehisa, M., & Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research* , *28* (1), 27–30.

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. doi: 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. doi: 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. doi: 10.1093/bioinformatics/bts507

Lee, E., Helt, G. A., Reese, J. T., Munoz-Torres, M. C., Childers, C. P., Buels, R. M., . . . Lewis, S. E. (2013). Web Apollo: A web-based genomic annotation editing platform. *Genome Biology* ,*14* (8). doi: 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. doi: 10.1093/bioinformatics/btl529

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), 11–23. doi: 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. doi: 10.1093/molbev/msaa015

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G. A., Sonnhammer, E. L. L., ... Bateman, A. (2021). Pfam: The protein families database in 2021. *Nucleic Acids Research* , *49* (D1), D412–D419. doi: 10.1093/nar/gkaa913

Ni, L. (2021). The Structure and Function of Ionotropic Receptors in*Drosophila* . *Frontiers in Molecular Neuroscience* ,*13* (February), 1–11. doi: 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. doi: 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. doi: 10.1093/hmg/ddy177

Putnam, N. H., Connell, B. O., Stites, J. C., Rice, B. J., Hartley, P. D., Sugnet, C. W., ... Rokhsar, D. S. (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. doi: 10.1101/gr.193474.115.Freely

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

Řezáč, Milan, & 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

Řezáč, Milan, 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

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. doi: 10.1002/jez.b.22797

Saha, S. (2019). Long range sequencing and validation of insect genome assemblies. In Humana Press (Ed.), *Methods in Molecular Biology*(Vol. 1858). Springer New York. doi: 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. doi: 10.1093/gigascience/giz099

Sanggaard, K. W., Bechsgaard, J. S., Fang, X., Duan, J., Dyrlund, T. F., Gupta, V., ... Wang, J. (2014). Spider genomes provide insight into composition and evolution of venom and silk. *Nature Communications* , *5* (May). doi: 10.1038/ncomms4765

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

Seppey, M., Manni, M., & Zdobnov, E. M. (2019). *BUSCO: Assessing Genome Assembly and Annotation Completeness BT - Gene Prediction: Methods and Protocols* . Retrieved from 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. doi: 10.1093/molbev/msu235

Sheffer, M. M., Hoppe, A., Krehenwinkel, H., Uhl, G., Kuss, A. W., Jensen, L., ... 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. doi: 10.1093/gigascience/giaa148

Smit AF, H. R. (n.d.-a). *RepeatMasker-4.0* . Retrieved from http://www.repeatmasker.org

Smit AF, H. R. (n.d.-b). *RepeatModeler-1.0* . Retrieved from 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. doi: 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. doi: 10.1186/1471-2105-7-62

Thomas, G. W. C., Dohmen, E., Hughes, D. S. T., Murali, S. C., Poelchau, M., Glastad, K., ... Richards, S. (2020). Gene Content Evolution in the Arthropods. *Genome Biology* , *21* (15), 1–14. doi: 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. doi: 10.1111/phen.12192

Ullmann, A. J., Lima, C. M. R., Guerrero, F. D., Piesman, J., & Black IV, W. C. (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. doi: 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. doi: 10.1186/gb-2007-8-11-r235

Vizueta, J., Escuer, P., Fıas-Lopez, C., Guirao-Rico, S., Hering, L., Mayer, G., ... Sanchez-Gracia, A. (2020). Evolutionary history of major chemosensory gene families across panarthropoda. *Molecular Biology and Evolution* , *37* (12), 3601–3615. doi: 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 *Methods in Enzymology* (1st ed., Vol. 642). Elsevier Inc. doi: 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. doi: 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. doi: 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. doi: 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. doi:

10.1111/1755-0998.13202

Wheeler, W. C., Coddington, J. A., Crowley, L. M., Dimitrov, D., Goloboff, P. A., Griswold, C. E., . . . 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. doi: 10.1111/cla.12182

World Spider Catalog. (2021). World Spider Catalog. doi: 11 June 2015

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

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

Hosted file

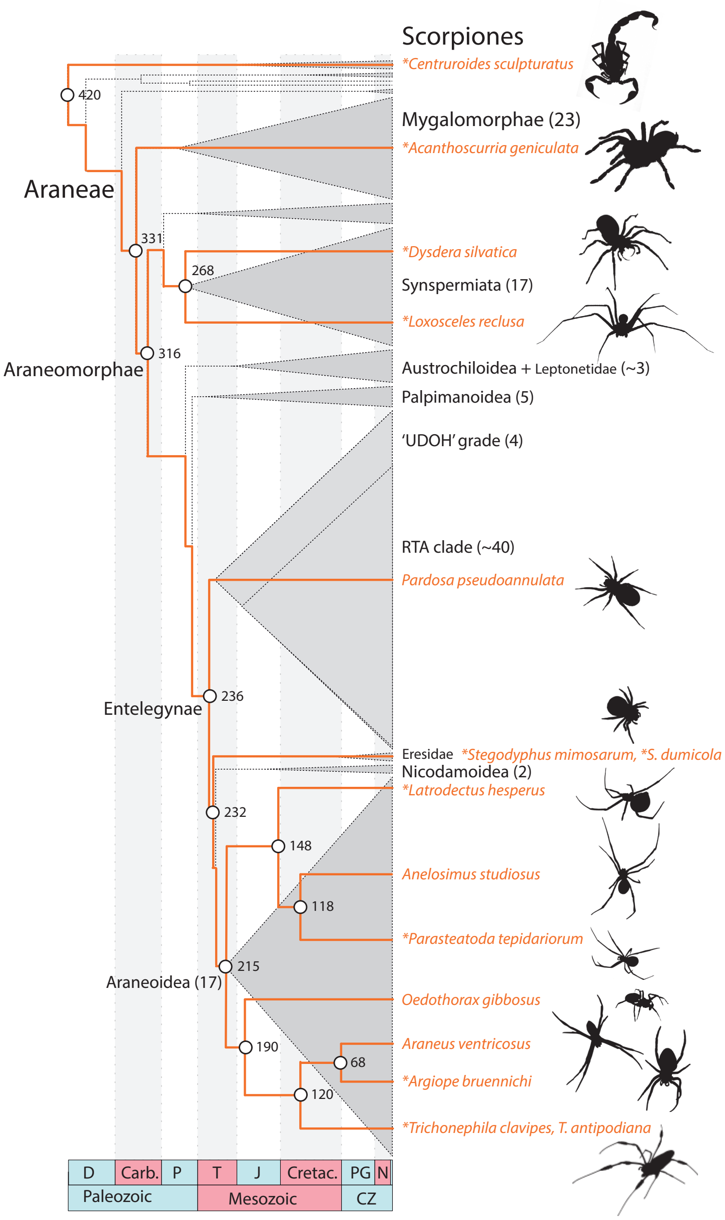
20210409_Table1.xlsx available at https://authorea.com/users/413060/articles/521514-chromosome-scale-assembly-of-the-canary-island-endemic-spider-dysdera-silvatica-arachnida-araneae-sheds-light-on-the-origin-and-genome-structure-of-chemoreceptor-gene-families-in-spiders

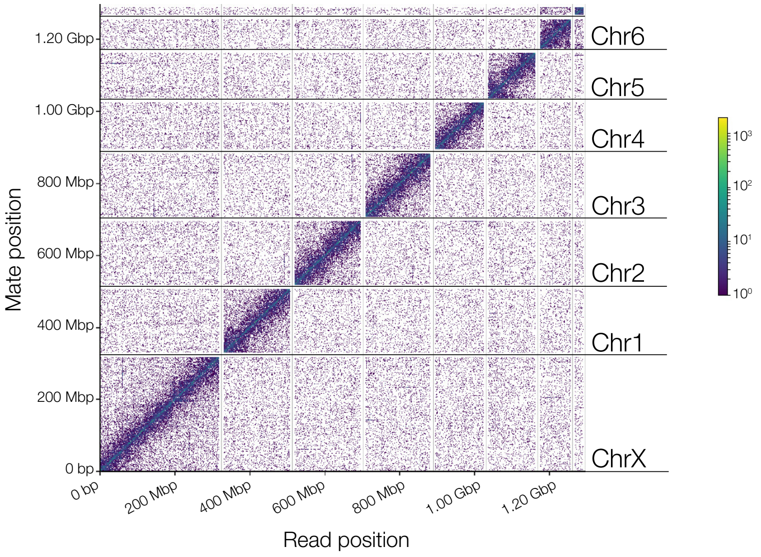
Hosted file

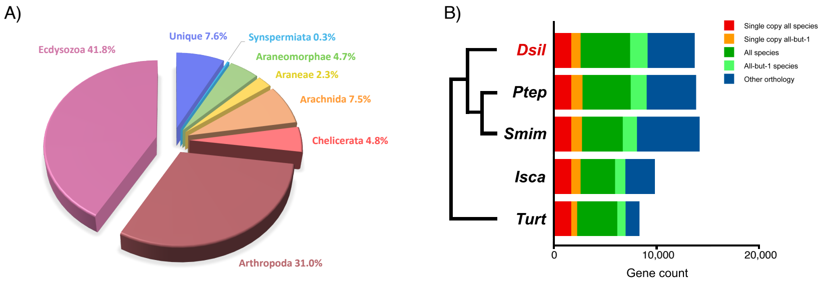
20210409_Table2.xlsx available at https://authorea.com/users/413060/articles/521514-chromosome-scale-assembly-of-the-canary-island-endemic-spider-dysdera-silvatica-arachnida-araneae-sheds-light-on-the-origin-and-genome-structure-of-chemoreceptor-gene-families-in-spiders

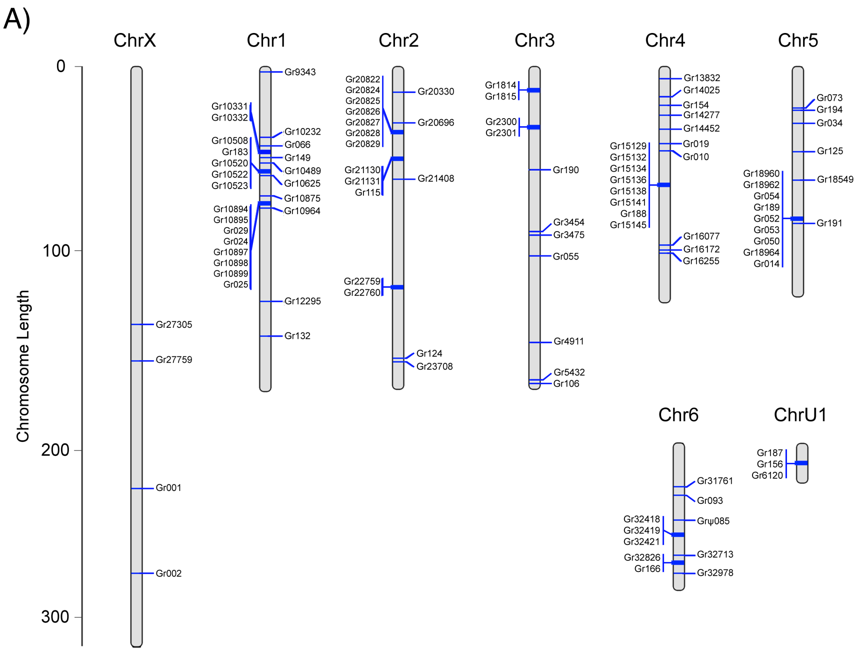
Hosted file

20210409_Table3.xlsx available at https://authorea.com/users/413060/articles/521514-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


Scorpiones
*Centruroides sculpturatus
420
Araneae
Mygalomorphae (23)
*Acanthoscurria geniculata
331
268
*Dysdera silvatica
Synspermiata (17)
*Loxosceles reclusa
316
Araneomorphae
Austrochiloidea + Leptonetidae (~3)
Palpimanoidea (5)
'UDOH' grade (4)
RTA clade (~40)
Pardosa pseudoannulata
236
Entelegynae
Eresidae *Stegodyphus mimosarum, *S. dumicola
Nicodamoidea (2)
*Latrodectus hesperus
232
148
Anelosimus studiosus
118
*Parasteatoda tepidariorum
215
Araneoidea (17)
Oedothorax gibbosus
190
Araneus ventricosus
68
*Argiope bruennichi
120
*Trichonephila clavipes, T. antipodiana
D
Carb.
P
T
J
Cretac.
PG
N
Paleozoic
Mesozoic
CZ

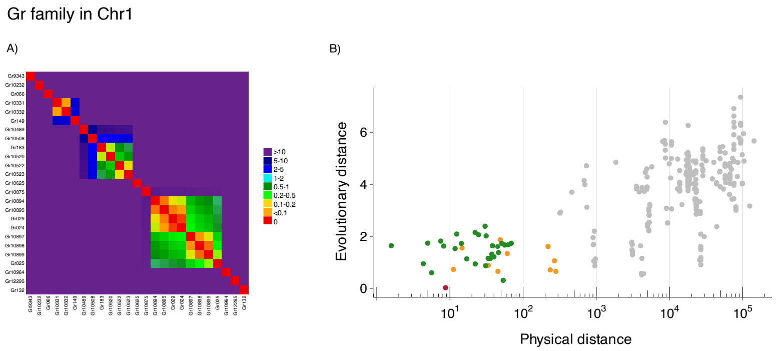

Mate position
1.20 Gbp
1.00 Gbp
800 Mbp
600 Mbp
400 Mbp
200 Mbp
0 bp
0 bp
200 Mbp
400 Mbp
600 Mbp
800 Mbp
1.00 Gbp
1.20 Gbp
Read position
Chr6
Chr5
Chr4
Chr3
Chr2
Chr1
ChrX
10^3
10^2
10^1
10^0


A)
Unique 7.6%
Synspermiata 0.3%
Araneomorphae 4.7%
Araneae 2.3%
Arachnida 7.5%
Chelicerata 4.8%
Arthropoda 31.0%
Ecdysozoa 41.8%
B)
Dsil
Ptep
Smim
Isca
Turt
Single copy all species
Single copy all-but-1
All species
All-but-1 species
Other orthology
0
10,000
20,000
Gene count

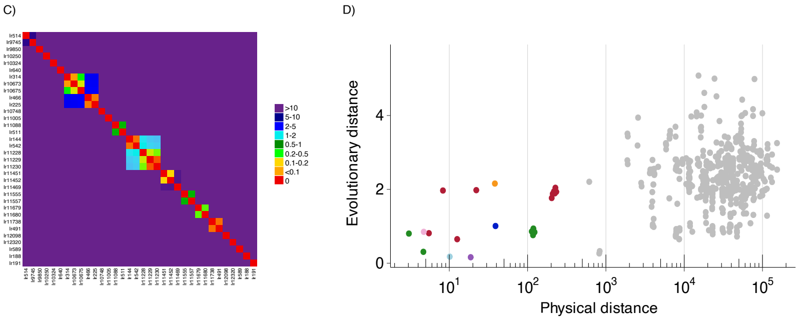


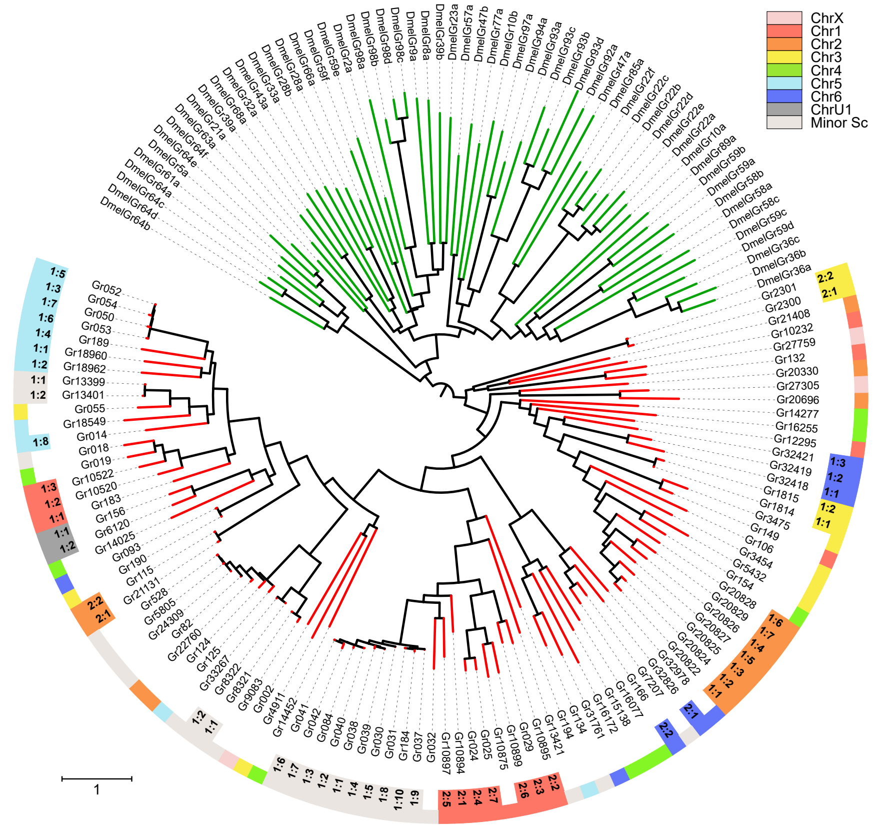
Hosted file

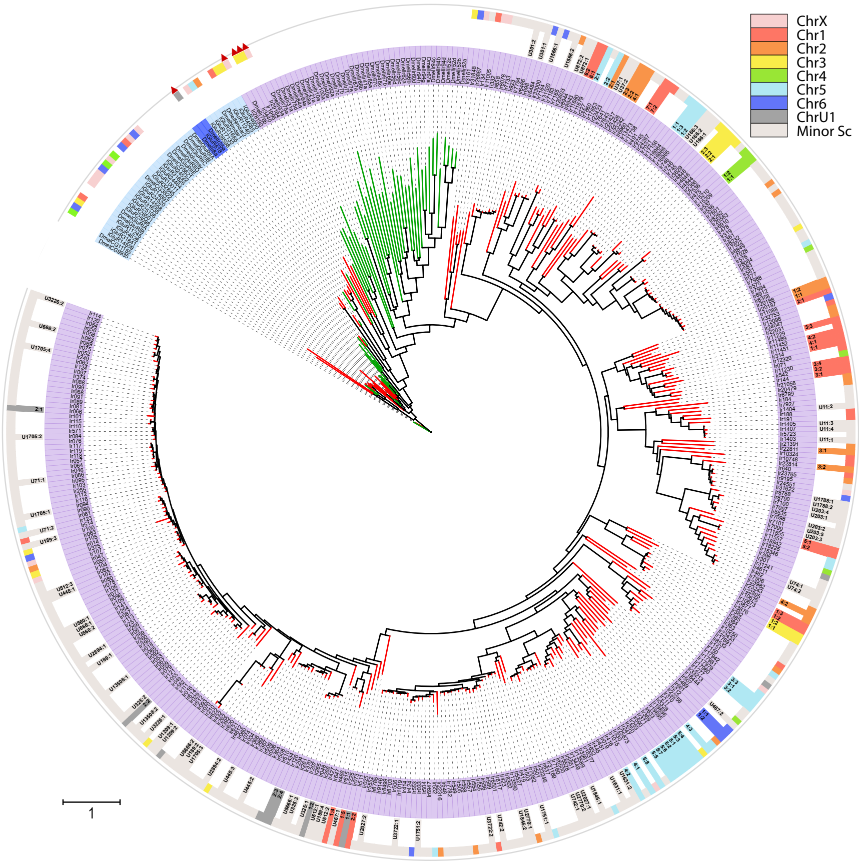
`Fig4B_IRs_chromosomes.pdf` available at https://authorea.com/users/413060/articles/521514-chromosome-scale-assembly-of-the-canary-island-endemic-spider-dysdera-silvatica-arachnida-araneae-sheds-light-on-the-origin-and-genome-structure-of-chemoreceptor-gene-families-in-spiders



Ir family in Chr1






ChrX
Chr1
Chr2
Chr3
Chr4
Chr5
Chr6
ChrU1
Minor Sc
U3226:2
U666:2
U1705:4
2:1
U1705:2
U71:1
U1705:1
U71:2
U189:3
U11:2
U11:3
U11:4
U11:1
3:1
3:2
1