Genome Biology and Evolution Advance Access published April 28, 2011 doi:10.1093/gbe/evr033

Comparative Genomics of the Odorant-Binding and Chemosensory Protein Gene Families across the Arthropoda: Origin and evolutionary history of the chemosensory system

Authors and affiliations

Filipe G. Vieira and Julio Rozas*

Departament de Genètica, Facultat de Biologia, Universitat de Barcelona, Diagonal 645, 08028 Barcelona, Spain

Institut de Recerca de la Biodiversitat, Universitat de Barcelona, Spain

***Author for Correspondence:**
Julio Rozas, Departament de Genètica, Facultat de Biologia, Universitat de Barcelona, Barcelona, Spain,
Tel: (+34) 93 4021495,
Fax: (+34) 93 4034420

Emails:
Filipe G. Vieira (fgarret@ub.edu)
Julio Rozas (jrozas@ub.edu)

Running head: Arthropoda chemosensory multigene families

Abstract

Chemoreception is a biological process essential for the survival of animals, as it allows the recognition of important volatile cues for the detection of food, egg-laying substrates, mates or predators, among other purposes. Furthermore, its role in pheromone detection may contribute to evolutionary processes such as reproductive isolation and speciation. This key role in several vital biological processes makes chemoreception a particularly interesting system for studying the role of natural selection in molecular adaptation. Two major gene families are involved in the perireceptor events of the chemosensory system: the odorant-binding protein (OBP) and chemosensory protein (CSP) families. Here, we have conducted an exhaustive comparative genomic analysis of these gene families in twenty Arthropoda species. We show that the evolution of the OBP and CSP gene families is highly dynamic, with a high number of gains and losses of genes, pseudogenes and independent origins of subfamilies. Taken together, our data clearly support the birth-and-death model for the evolution of these gene families with an overall high gene-turnover rate. Moreover, we show that the genome organization of the two families is significantly more clustered than expected by chance and, more important, that this pattern appears to be actively maintained across the *Drosophila* phylogeny. Finally, we suggest the homologous nature of the OBP and CSP gene families, dating back their MRCA (most recent common ancestor) to 380–420 Mya, and we propose a scenario for the origin and diversification of these families.

Keywords: OBP, CSP, birth-and-death, gene family evolution, olfactory system

Introduction

Chemoreception is a widely used mechanism across animal species for perception of the surrounding environment, from communication between conspecifics to detection of predators and location of food or hosts, playing a critical role in an organism's fitness (Krieger and Ross 2002; Matsuo et al. 2007; Asahina, Pavlenkovich, and Vosshall 2008; Whiteman and Pierce 2008; Smadja and Butlin 2009). Moreover, its role in reproduction may contribute to a number of evolutionary processes, such as reproductive isolation and speciation. Thus, understanding the evolution of genes involved in sensorial perception may provide valuable insight into the role of natural selection in molecular adaptation.

The first steps in the recognition of odorant signals (peripheral events) are accomplished by binding and membrane receptor proteins that recognize external ligands and translate this interaction into an electrical signal to the central nervous system. In the Insecta, there are three different types of chemosensory receptors, the odorant (OR), the gustatory (GR) and the Ionotropic (IR) receptors, which are located in the dendritic membrane of chemosensory neurons (Kaupp 2010). The dendrites of these neurons are positioned inside the sensilla, which is a hair-like hollow structure that is filled with an aqueous fluid, the sensillar lymph. The chemical signals enter the sensilla lumen through the sensilla pores of the chitin wall, diffuse through the lymph and activate the receptors [for a review, see (Sanchez-Gracia, Vieira, and Rozas 2009)]. The sensillar lymph is secreted by non-neuronal support cells and contains a variety of proteins, including the odorant-binding (OBP) and chemosensory (CSP) proteins (Vogt and Riddiford 1981; Steinbrecht 1998). These proteins are small (10 to 30 kDa), globular and highly abundant water-soluble proteins, characterized by a specific domain of six α-helices, joined by either two or three disulfide bonds (Leal, Nikonova, and Peng 1999; Tegoni, Campanacci, and Cambillau 2004). Although the full range of functions

of these molecules has not been well established, there is increasing evidence of their importance in chemosensory perception (Pophof 2004; Xu et al. 2005; Grosse-Wilde, Svatos, and Krieger 2006; Matsuo et al. 2007). Most likely, OBP and CSP proteins are involved in the solubilization and transport of odorants, which are generally hydrophobic (Kaissling 2001; Leal et al. 2005). Recent studies, however, have revealed that OBP and CSP genes are not restricted to the olfactory tissues and may, in fact, participate in other physiological functions such as olfactory coding and stimulus inactivation (Kaissling 2001; Graham et al. 2003; Pophof 2004; Findlay et al. 2008) [for a review, see (Pelosi et al. 2006)]. Despite carrying out a similar physiological role, vertebrate OBPs are not homologous to their insect counterparts and actually differ in structure and size (Pelosi and Maida 1990). In fact, these genes belong to a large superfamily of carrier proteins, the lipocalins, that usually consist of a β-barrel structure and a carboxy-terminal α-helix (Flower 1996).

Comprehensive analysis of the complete genome sequences of *Drosophila* and a number of other insects (*Anopheles gambiae*, *Bombyx mori*, *Tribolium castaneum* and *Apis mellifera*) has revealed that the OBP and CSP gene repertoires differ markedly across species. In fact, the OBP family comprises from 21 (in *A. mellifera*) to 66 genes (in *A. gambiae*), whereas the CSP gene family ranges from 4 members (in *Drosophila*) to 22 (in *B. mori*) (Foret and Maleszka 2006; Foret, Wanner, and Maleszka 2007; Gong et al. 2007; Vieira, Sanchez-Gracia, and Rozas 2007; Gong et al. 2009; Kirkness et al. 2010). Interestingly, these genes are unevenly distributed throughout the genome, with many of them (69% of the OBP genes in *Drosophila*) being arranged in small clusters (from 2 to 6 OBP genes) (Vieira, Sanchez-Gracia, and Rozas 2007). The *Drosophila* OBP gene family has been classified into several phylogenetic subfamilies on the basis of distinctive structural features, functional information and phylogenetic relationships: the Classic, Minus-C, Plus-C, Dimer, PBP/GOBP, ABPI and

ABPII, CRLBP and D7 subfamilies (Hekmat-Scafe et al. 2002; Valenzuela et al. 2002; Vieira, Sanchez-Gracia, and Rozas 2007; Gong et al. 2009; Kirkness et al. 2010). Interestingly, these subfamilies are unequally distributed across arthropods, even among the dipterans, and they are totally absent in some species. In contrast, the CSP gene family is much more conserved across insects, without distinctive phylogenetic clades. It has been suggested that the OBP and CSP gene families may have shared a MRCA (most recent common ancestor) near the origin of the arthropods, though the evidence for this is controversial (Pelosi, Calvello, and Ban 2005; Zhou et al. 2006).

In the present study, we used the complete genome sequence data from twenty Arthropoda species to conduct a fine and exhaustive comparative genomic analysis of the OBP and CSP gene families. In particular, we aimed to gain insights into the origin and evolutionary fate of OBP and CSP duplicates and to determine their role in the adaptive process. Our exhaustive analysis allowed us to identify new genes and several gene contractions and expansions in different lineages. Interestingly, we also identified two OBP genes that are present in almost all of the analyzed species, indicating a putative critical role in chemosensation. Overall, our results are clearly consistent with the birth-and-death (BD) evolutionary model (Nei and Rooney 2005), with estimates for the birth (β) and death (δ) rates of $\beta = 0.0049$ and $\delta = 0.0010$ for OBP, and $\beta = 0.0028$ and $\delta = 0.0007$ for CSP. We also found that the organization of the members of these gene families into clusters is not a by-product of their tandem origin but, instead, is actively maintained by natural selection. Finally, we point to the homologous nature of the OBP and CSP gene families, estimating their MRCA to have occurred 380-420 Mya, and we propose a scenario for the origin and diversification of these two families.

Materials and Methods

Genomic Data

Genome sequence data and gene annotations were downloaded from public data repositories: *Drosophilidae* (release FB2008_08) from FlyBase (Drysdale 2008), *Anopheles gambiae* (release AgamP3.46) from Ensembl (Flicek et al. 2008), *Bombyx mori* (release April/2008) from SilkDB (Wang et al. 2005), *Tribolium castaneum* (release V3.0) from BeetleBase (Wang et al. 2007), *Apis mellifera* (release 4.0) from NCBI [ftp://ftp.ncbi.nih.gov/genomes], *Pediculus humanus* (release PhumU1.1) and *Ixodes scapularis* (release IscaW1.1) from VectorBase (Lawson et al. 2007), *Acyrthosiphon pisum* (release June/2008) from AphidBase [http://www.aphidbase.com] and *Daphnia pulex* (release jgi060905) from wFleaBase [http://iubio.bio.indiana.edu/daphnia].

Gene Identification

We identified putative OBP and CSP members through several rounds of exhaustive searches using information from already known OBP and CSP proteins as queries (Foret and Maleszka 2006; Foret, Wanner, and Maleszka 2007; Gong et al. 2007; Vieira, Sanchez-Gracia, and Rozas 2007; Flicek et al. 2008; Gong et al. 2009; Zhou et al. 2010). First, we searched the preliminary predicted gene set using BLASTp (Altschul et al. 1997) (BLOSUM45 matrix with an e-value threshold of 10^{-5}), HMMER [http://hmmer.wustl.edu/] (e-value domain threshold of 10^{-5}) and HHsearch (Soding 2005) (e-value threshold of 10^{-5}). The HMMER and HHsearch searches used PFAM (Finn et al. 2006) PBP/GOBP (for OBP; PF01395) and OS-D (for CSP; PF03392) HMM profiles. Furthermore, because OBP family members are highly divergent, we also built four extra custom HMM profiles (used in all HMMER and HHsearch

searches). We built these profiles after clustering all known OBP protein sequences (only *D. melanogaster* and *D. mojavensis* from the *Drosophila* genus) with BLASTClust [ftp://ftp.ncbi.nih.gov/genomes] (e-value threshold of 10^{-5}, length coverage "-L" of 0.5 and score density "-S" of 0.6). We selected the four clusters with the highest numbers of sequences, aligned the clusters separately with MAFFT (Katoh et al. 2005) (E-INS-i with BLOSUM30 matrix, 10000 maxiterate and offset "0") and, for each cluster, built an HMM profile using HMMER. Because HHsearch only makes comparisons between HMM profiles, it was necessary to transform the proteome of each species into a set of HMM profiles. For this, we clustered the proteomes for each species separately with BLASTClust and built an HMM profile from each cluster separately. We followed a similar HMM profile building approach as described above, with the exception of the BLASTClust parameters (e-value threshold of 10^{-6}, length coverage "-L" of 0.7 and score density "-S" of 1.0). All profiles used by HHsearch included secondary structure information predicted with PSIPRED (McGuffin, Bryson, and Jones 2000). Second, we searched the raw DNA sequence data using tBLASTn (BLOSUM45 with e-value threshold of 10^{-3}), EXONERATE (Slater and Birney 2005) (50% of the maximum store threshold) and HMMER (e-value domain threshold of 10^{-10}). For the latter analysis, we searched against all 6-frames using PFAM's and our four custom HMM profiles as queries. All searches were performed exhaustively until no new hit was found, adding always all newly identified members to the queries.

We manually checked all putative positive hits, specifically looking for the presence of a signal peptide [predicted by PrediSi (Hiller et al. 2004)], the characteristic "cysteine domain" (Pelosi et al. 2006; Vieira, Sanchez-Gracia, and Rozas 2007) and a secondary structure including six α-helices [predicted by PSIPRED (McGuffin, Bryson, and Jones 2000)]. We used the Artemis (Rutherford et al. 2000) genome annotator with the putative splice sites

predicted by Genesplicer (Pertea, Lin, and Salzberg 2001) to assist with the annotation process.

Gene Clustering Analysis

We have tested, by computer simulations, whether OBP or CSP genes are actually physically closer in the chromosomes than expected by chance. This analysis was conducted separately for each species and gene family (either OBP or CSP), excluding species with poorly assembled genomes or families with less than ten members. Specifically, we computed for each genome a statistic based on the average physical distance (in base pairs) between neighboring genes (within a given chromosome). This observed value was contrasted against the null empirical distribution of this statistic generated by computer simulations (based on 10,000 replicates). In each replicate we randomly chose a fixed number of genes (the same number than that observed OBP or CSP members in a particular genome) and calculated the statistic (Table 2).

To try to gain insight into the biological meaning of such chromosome clusters we analyzed whether the observed OBP clusters are more conserved across the phylogeny than expected by chance. The analysis was conducted using the MCMuSeC algorithm (Ling, He, and Xin 2009) which examines, using the "gene teams" model (Luc et al. 2003), the distribution pattern of gene clusters across the phylogeny. The method uses as statistic the branch length score (BLS) to measure the evolutionary time (the total lengths of the phylogenetic tree) where the gene cluster is conserved. Therefore, the longer the BLS value the more likely it will be under functional constraint. For such analysis we used the cluster definition as in (Vieira, Sanchez-Gracia, and Rozas 2007). The statistical significance of the test was obtained by comparing the observed BLS value (for each OBP cluster) against the null empirical

distribution of the same cluster size generated by computer simulations (based on 1,000,000 replicates).

Phylogenetic Analysis

We performed a phylogenetic analysis including all complete OBP and CSP genes and partial coding sequences with more than 85 and 78 amino acids, respectively (the size of the smallest full CDS in each gene family). Because the signal peptide portion of OBPs has a high substitution rate, we removed these regions [identified using the PrediSi program (Hiller et al. 2004)] before conducting the analyses. The protein sequences were multiply aligned using MAFFT v6.624b (Katoh et al. 2005) (E-INS-i with BLOSUM30 matrix, 10000 maxiterate and offset "0"). We estimated the phylogenetic relationships by maximum likelihood using the software RAxML v7.2.3 (Stamatakis 2006), assuming the WAG evolutionary model (Whelan and Goldman 2001) and fixing the amino acid frequencies ("-f d -e 0.0001 -d -N 30 -m PROTGAMMAWAG"). The genetic distances (number of amino acid changes per site) were estimated using MEGA software (Tamura et al. 2007) with the pairwise deletion option and assuming the JTT evolutionary model (Jones, Taylor, and Thornton 1992).

We inferred the OBP and CSP orthology groups using the OrthoMCL software (inflation of 1.5 and e-value threshold of 10^{-5}), which is based on reciprocal best hits within and between proteomes. These orthology relationships were used to estimate the OBP and CSP birth (β) and death (δ) rates (events per gene and per million years) by maximum likelihood (Librado, Vieira and Rozas; unpublished results) using the divergence times from Tamura, Subramanian, and Kumar (2004) and Hedges, Dudley, and Kumar (2006). Briefly, for each orthology group, we inferred the number of genes in each internal node using those numbers in extant species, and the phylogenetic branch lengths. This information allow us to further

estimate the number of gene gain and loss events in each phylogenetic branch, and the global birth and death rates following equations (1) and (2) in Vieira, Sanchez-Gracia, and Rozas (2007). The half-life for a gene to be lost from the genome ($t_{½}$) was estimated assuming that the death rate follows an exponential decay curve. In particular, $t_{1/2} = \frac{-\ln(0.5)}{\delta}$

For all analyses, customized in-house scripts were written in Perl, with extra modules from BioPerl (Stajich et al. 2002); these scripts are available upon request.

Results

Identification and Characterization of OBP and CSP Genes

We performed exhaustive and manually curated searched that allowed us to identify the complete set of putative functional OBP and CSP genes across the twenty Arthropoda species analyzed (Figure 1), improving currently published data. In addition, we also found some scattered fragments that likely correspond to incomplete sequences and pseudogenes (Table 1). Almost all of the identified genes have the characteristic hallmarks of the OBP and CSP gene families: the signal peptide, the 6 α-helix pattern and the highly conserved cysteine profile. However, despite the highly conserved secondary structure of OBP proteins, the OBP family members are highly divergent (average per-site amino acid divergence of d = 2.99; sequence identity of 16.71%), exhibiting a wide range of gene lengths (from 85 to 329 amino acids) and cysteine profiles. The CSP gene family shows lower divergence values (d = 1.51, with overall identity of 34.04%), with the four-cysteine profile (forming the two disulfide bridges) being completely conserved and exhibiting fairly constant gene lengths (60% of the mature proteins have lengths between 97 and 119 amino acids).

In spite of the intensive analyses that have been previously conducted in *D. melanogaster*, our HMM-based searches allowed the identification of a new OBP gene (*Obp73a*). It is likely that the high divergence of this gene from the other OBP members prevented its previous identification by similarity-based methods. Interestingly, this gene has a 1:1 orthology, not only in the 12 *Drosophila* genomes, but also in almost all insect species analyzed (except in Hymenoptera). In fact, there are only two OBP members with clear orthology relationships across insects: Obp73a and Obp59a (Zhou et al. 2010). This high conservation across a large number of arthropod species suggests a critical function for these proteins.

Chromosomal Organization

We studied the evolutionary meaning of the organization in chromosome clusters of the OBP and CSP genes. We have found that within species OBP and CSP genes are physically closer in the genome (significantly clustered) than expected by chance ($p < 0.0064$ and $p < 0.0008$, respectively) (Table 2). In contrast, the OR and GR gene families of *D. melanogaster*, which have a similar number of genes to OBP, are more scattered across the genome (Robertson, Warr, and Carlson 2003), and do not exhibit such clear structuring ($p = 0.194$ and $p = 0.023$ for OR and GR, respectively).

These chromosome clusters, however, could be just a consequence of the origin of genes by tandem gene duplication, rather than having some functional significance. To gain insight into the functional meaning of this clustering we have analyzed whether these clusters have been maintained throughout evolution despite of the breaks produced by inevitable chromosomal rearrangements. This analysis was conducted using only OBP data from the 12 *Drosophila* species because there are few orthologous clusters among species sharing large divergence

times. Our results show that OBP genes are significantly clustered across the *Drosophila* evolution ($p < 0.033$), suggesting the existence of some functional constraints maintaining the clusters (Quijano et al. 2008).

Phylogenetic Analysis

Our phylogenetic analysis shows that the evolution of OBP and CSP gene families is highly dynamic, though to a lesser degree in the CSP gene family, exhibiting a number of taxa-specific subfamilies, several branch-specific expansions and almost no groups of orthologous genes shared across *Arthropoda* (Figures 2-4).

The *Drosophila* OBP gene family has been classified into several groups on the basis of distinctive structural features, functional information and phylogenetic relationships: the Classic, Minus-C, Plus-C, Dimer, PBP/GOBP, ABPI and ABPII (formerly known as ABPX), CRLBP and D7 subfamilies (Hekmat-Scafe et al. 2002; Valenzuela et al. 2002; Vieira, Sanchez-Gracia, and Rozas 2007; Gong et al. 2009). The Atypical subfamily, which has so far been identified only in mosquitoes (Xu, Zwiebel, and Smith 2003; Zhou et al. 2008), is in fact a Dimer OBP clade (Supplemental Figure 1). These proteins have a double domain profile that most likely originated from a fusion of two Classic OBP genes. Our results show that the basal OBP group seems to be the Classic, whereas all other groups are internal clades of the Classic subfamily which is, in fact, paraphyletic (Figure 3). The Plus-C subfamily, present in all Hexapoda species, has been lost in the *Hymenoptera*. Interestingly, some subgroups of the Classic subfamily, such as Dimer, Minus-C and CRLBP, appear to have had independent origins. The Dimer OBP originated independently in the *Culicidae* and *Drosophilidae* lineages, the Minus-C appeared in the *Drosophilidae*, *Bombyx*/*Tribolium* and *Apis* lineages, while the CRLBP members are highly scattered across the tree and appear to lack any

phylogenetic meaning. Furthermore, we also identified in *A. gambiae* a putative new OBP member (*AgamOBP78*) of the D7 subfamily, a widespread subfamily in blood-sucking *Diptera* (Valenzuela et al. 2002).

The CSP gene family consistently has fewer members than the OBP family, exhibiting only two lineage-specific expansions (in *B. mori* and *T. castaneum*; Figure 4). The genes in this family also exhibit lower genetic distances, although its members are present across all Arthropoda species, including Crustacea (*D. pulex*) and Chelicerata (*I. scapularis*). Overall, the CSP gene family has an evolutionary pattern that is less dynamic than the OBP family, with fewer and more conserved members that are not grouped into distinctive phylogenetic clades.

We observed that the number of groups of orthologous genes that are shared among different species quickly decreases with increasing divergence time (Figure 2). For example, the number of groups of orthologous genes ranges from 34 OBP and 3 CSP within the genus *Drosophila*, to 2 OBP and 2 CSP across *Hexapoda*, and no OBP nor CSP groups shared across all of the Arthropoda. Noticeably, only two OBP genes have orthologs across all insects except in Hymenoptera: *Obp59a* and *Obp73a*.

Despite the high divergence that is seen among paralogs, some genes have unexpected features that may indicate important functions or, alternatively, that may be the result of misannotation. For instance, the *Obp59a* gene has an unusually long sequence and a unique cysteine pattern. *BmorOBP41*, a Plus-C subfamily member, has a pattern of cysteine residues that is unusual for this family (Figure 3). Furthermore, we also identified three CSP genes

(*TcasCSP6*, *ApisCSP1* and *ApisCSP9*) with a markedly different secondary structure (Figure 4).

Common Origin of OBP and CSP

The common origin of the OBP and CSP gene families is a controversial issue (Pelosi, Calvello, and Ban 2005; Zhou et al. 2006). To attempt to detect a putative remote homology between the OBP and CSP gene families, we performed a series of similarity searches using different approaches. With a standard BLASTp-based search (e-value threshold <1) we did not detect any significant similarity. Using more powerful approaches, like HMM-based analyses (HMMER software), together with PFAM (Finn et al. 2006) and our four specific custom profiles (see Methods) allowed us to detect some slight indications of sequence similarity between the PFAM CSP profile (OS-D: PF03392) and the OBP *TcasOBP16* (e-value of 0.0049), but the analysis also detected some false positives (data not shown). Since the degree of functional constraint on the tertiary structure of proteins is probably higher than their primary structure we studied the similarity among OBP and CSP protein structures to gain insight into their putative remote homology. For that we generated rigid structural alignments using FATCAT (Ye and Godzik 2004) between all OBP and CSP proteins present in the RCSB Protein Data Bank (www.pdb.org) (Berman et al. 2000). We found that the majority of OBP-CSP structure alignments are statistically significant (p = 0.0089 for the lowest p-value) (Figure 5; Table 3). Moreover, using OBP and CSP protein sequences as a query in additional BLASTp searches against all PDB sequences, we detect no proteins (other than OBP and CSP) with significant structural similarity (on the top scoring 10 hits).

Birth-and-Death Evolution

Overall, our phylogenetic analyses showed that the OBP and CSP families fit well with a BD evolutionary model (Figure 3, 4 and 6) based on the following results: (i) phylogenetic trees based on orthologous genes fit well with the accepted species phylogeny; (ii) there is no evidence of gene conversion between paralogous genes (data from Drosophila); (iii) paralogous genes have higher divergence times compared with orthologs; (iv) several gene gain and loss events can be identified in numerous phylogeny lineages; (v) several nonfunctional members (pseudogenes) were found (mainly in the terminal branches); (vi) many orthology groups can be seen among closely related species, and this number gradually decreases with increasing divergence times; and (vii) there is an uneven phylogenetic subfamily distribution across species. Hence, OBP and CSP genes appear to have evolved independently from the time of their origin by gene duplication until their loss by deletion or transiently as pseudogenes.

To gain insight into the specific BD dynamics of these families it is important to quantify the magnitude of this process. Previous reports have addressed this issue using automatic annotations, surveying a set of too closely related species, or applying less accurated statistical models (Hahn et al. 2005; Demuth et al. 2006; Guo and Kim 2007; Hahn, Han, and Han 2007; Vieira, Sanchez-Gracia, and Rozas 2007). Here, we have estimated BD rates using a manually curated dataset covering several species across the Arthropoda phylum, and using more accurate gene turnover models, which allowed us to separately estimate birth (β) and death (δ) rates. Our BD estimates for the OBP gene family are $\beta = 0.0049$ and $\delta = 0.0010$, whereas for the CSP family they are $\beta = 0.0028$ and $\delta = 0.0007$ (Figure 6).

Discussion

OBP and CSP Gene Family Evolution

The OBP and CSP gene families exhibit a highly dynamic evolutionary history. For instance, the number of members of these families is quite variable across Arthropoda species [OBP ranges from 0 to 83 genes and CSP from 1 to 22 in (Table 1)], and its members are highly diverse, with divergent proteins exhibiting a wide range of gene lengths and encoding different cysteine profiles. As a result, and despite the exhaustive studies that have been performed in recent years, we have still been able to identify a new OBP member (*Obp73a*) in the 12 *Drosophila* species which, in addition, is conserved across Arthropoda (except in Hymenoptera). Interestingly, there are only two genes with a clear 1:1 orthology relationship across insects: *Obp73a* and *Obp59a*. This conservation pattern is highly suggestive, reminiscent of the *Or83b* gene, an essential and highly conserved OR member present in all sequenced Arthropoda species (Larsson et al. 2004).

The OBP and CSP genes in *Drosophila*, *A. gambiae, Aedes aegypti*, *B. mori* and *T. castaneum* are frequently organized in clusters (Zhou et al. 2006; Foret, Wanner, and Maleszka 2007; Gong et al. 2007; Zhou et al. 2008; Gong et al. 2009). However, no stringent statistical analysis has been conducted to determine their evolutionary significance. We have found that the members of these families are actually significantly clustered across the genome and, moreover, that the OBP cluster distribution has been maintained across the *Drosophila* evolution. This conservation across ~400 Myrs of evolution (the total branch lengths) suggests the action of natural selection in preventing cluster brake up. Indeed, this conservation could be explained by the existence of shared regulatory elements among members (Boutanaev et al. 2002; Gong et al. 2007; Matsuo et al. 2007; Quijano et al. 2008).

Since chromosomal rearrangement breakpoints are unevenly distributed across the genome, the current clustering of OBP genes might also reflect the existence of the so-called fragile regions, regions with a propensity to breakage (Pevzner and Tesler 2003; von Grotthuss, Ashburner, and Ranz 2010). This feature, nevertheless, would not provide the best explanation since our null empirical distribution already reflects the actual spatial distribution of genes in the genomes. The OBP clusters, therefore, likely have a functional meaning.

Our phylogenetic analysis uncovered a highly dynamic mode of OBP and CSP gene family evolution, although to a lesser extent for the CSP family. Both families exhibit lineage-specific expansions and a high number of orthology groups at short evolutionary times that gradually disappear with increasing divergence (Figure 2). Our results also indicate that the Dimer and Minus-C OBP subfamilies are polyphyletic and, therefore, have no phylogenetic significance. The striking fact that a similar cysteine pattern arose independently several times during the evolution of these genes is intriguing and suggests that these conformations may be advantageous. Because OBP genes form dimers *in vitro* (Andronopoulou et al. 2006), the Dimer OBP gene structure might be functionally equivalent to two single-domain OBP genes. In the case of Minus-C, the loss of one disulfide bridge might also have functional relevance, as it could generate a more flexible structure (like CSPs) (Angeli et al. 1999; Leal, Nikonova, and Peng 1999; Scaloni et al. 1999).

Overall, our results clearly support the birth-and-death model of evolution for these two gene families. Hence, the model of evolution described for the OBP family of *Drosophila* also holds for the evolution of OBP and CSP families and for both short and long period of times (across *Arthropoda*). The BD model, therefore, is neither incidental nor specific to the *Drosophila* genus but rather it is a more general model of evolution. Interestingly, the

estimated birth rates of both families are higher than that estimated for the whole *Drosophila* genome ($\lambda = 0.0012$) (Hahn, Han, and Han 2007), reflecting a highly dynamic evolution. Indeed, the half-life estimates of a given gene ($t_{1/2}$) are $t_{1/2} = 693$ Myr and $t_{1/2} = 990$ Myr for the OBP and CSP genes, respectively. Nevertheless, and in spite of using complete genome data, our current estimates should be viewed with caution. The species we surveyed belong to a phylogenetic tree with some large branches (e.g. branches leading to *T. castaneum* or *B. mori*) that can lead to inaccurate estimates. In the future, these estimates can be further improved by using genome information from species that are more homogeneously distributed across the tree.

Current rates of birth and death suggest a very high gene turnover rate, placing gene gain and loss events as one of the most important processes in the evolution of these gene families. These high rates can have a significant adaptive value, due to the function of these families in the contact with the exterior environment. During adaptation to a changing environment, newly arisen genes can play an important role as raw material for the action of natural selection. The actual OBP and CSP family sizes would result from a balance between the effect of the stochastic BD process [or random genomic drift (Nei 2007)], the maintenance of a core number of genes required for basal chemosensory performance, and the requirement of newly arisen genes which diverged into species-specific activities.

Origin and evolutionary history of the chemosensory system

The putatively remote homology between OBP and CSP proteins suggest that these gene families belong to a larger superfamily of general binding proteins. The OBP and CSP gene families, together with the two major chemosensory receptor families (OR and GR), show a suggestive parallel distribution across Arthropoda. OBP and OR genes are found only in

Hexapoda, whereas CSP and GR genes have been identified in all major Arthropoda groups: Hexapoda, Crustacea, Myriapoda (just CSP) and Chelicerata (Pelosi et al. 2006; Wanner et al. 2007; Wanner and Robertson 2008; Penalva-Arana, Lynch, and Robertson 2009; Sanchez-Gracia, Vieira, and Rozas 2009; Smadja et al. 2009). This suggests that the OBP and OR gene families originated after the Hexapoda–Crustacea split (~470 Mya), whereas the CSP and GR families were already present in the MRCA of these two groups and Chelicerata (~700 Mya) (Hedges, Dudley, and Kumar 2006). Because the earliest fossil evidence of terrestrial animal activity that that has been found comes from the Ordovician [~425 Mya (Labandeira 2005)], the common ancestor of these three groups is expected to be aquatic. This scenario agrees with other studies proposing the independent terrestrialization of Hexapoda, Chelicerata and Myriapoda lineages (380-420 Mya; Figure 1) (Ward et al. 2006).

According to our results, the aquatic ancestor of the extant major Arthropoda groups would have had chemoreceptors tuned to the perception of soluble components (proto-GR) and also a generic gene family of binding proteins (proto-CSP) with diverse physiological roles. The colonization of the hostile terrestrial environment by Hexapoda, Chelicerata and Myriapoda (but not Crustacea) led to diverse adaptations. For example, Arthropoda species overcame the challenges of water supply and desiccation by the development of an impermeable cuticle. Because the neurons must be connected with the exterior, they developed a porous sensillar cuticular wall and, to avoid desiccation, they also developed an aqueous lumen around their chemosensory neurons. The new aerial environment also changed the perceived chemical signals from essentially hydrophilic (in aqueous solution) to mainly hydrophobic (in gaseous phase) molecules (Freitag et al. 1998). Hence, two major problems emerged with terrestrialization: i) the new aqueous lumen prevented the access of hydrophobic molecules to chemoreceptors, and ii) likely the chemoreceptors were unable to perform a fine detection of

these new molecules. The origin of new specialized protein families to mediate the transport and detection of these new hydrophobic odorants, solved these problems. Generalist binding proteins might have evolved and further specialized to bind odorants and pheromones and, in parallel, the ancestral aquatic-specific receptors evolved into a new class of receptors specialized for sensing airborne compounds (olfactory receptors). Because the split of the four major Arthropoda groups occurred before their terrestrial colonization, the evolutionary novelty representing the origin of the odorant-binding molecules and olfactory receptors must have occurred independently in the Hexapoda, Myriapoda and Chelicerata lineages. These independent origins imply that these molecules might have evolved from different ancestral gene families: while in Hexapoda a proto-CSP gene family would have given rise to the OBP genes, in the other two groups might have derived from different (and still unknown) ancestral proteins. A similar scenario would have occurred with the olfactory receptors, which likely evolved from the GR family in the Hexapoda, and from other protein families in the two other taxa. This hypothesis would explain the presence of GR but absence of OBP and OR (even pseudogenes) in the *Daphnia* (Penalva-Arana, Lynch, and Robertson 2009) and *Ixodes* (unpublished results) genomes. Nevertheless, the reasons might be different: while in *Ixodes* olfactory genes probably evolved from different ancestral families, Crustacea remained largely aquatic with no need for airborne detection.

This scenario is further supported by a number of convergent evolution cases affecting the olfactory system. The IR, a new and structurally divergent chemoreceptor gene family, has recently been discovered in *Drosophila* (Benton et al. 2009; Brigaud et al. 2009; Croset et al. 2010). The robber crab (*Birgus latro*) is an attractive example of the changes that have occurred during the adaptive process to the terrestrial environment. This land-living crustacean has developed a complex olfactory sense with organs very similar to the insect

sensilla (Stensmyr et al. 2005; Krieger et al. 2010). Another example occurs in the vertebrate olfactory system. In spite of having equivalent physiological functions, vertebrates exhibit phylogenetically unrelated chemoreceptor and odorant-binding molecules. Vertebrate receptors belong to the GPCR family, whereas OBP genes belong to a large superfamily of carrier proteins, the lipocalins (Flower 1996; Pelosi et al. 2006; Nei, Niimura, and Nozawa 2008). Curiously, GPCR and lipocalins are also present in *Hexapoda*, though with different biochemical functions. In *Drosophila*, GPCRs function as neurotransmitters and hormone receptors or in axon guidance during embryonic nervous system development (Brody and Cravchik 2000; Sanchez et al. 2000); lipocalins function as salivary anticlotting proteins in *Rodnius prolixus* (Montfort, Weichsel, and Andersen 2000) while the anticlotting proteins of blood sucking Diptera belong to the D7 OBP subfamily (Valenzuela et al. 2002).

Taking all data together, we can hypothesize a scenario for the evolution of the chemosensory system (Figure 7). We can assume the existence of some general molecule-binding and receptor genes before the Vertebrata–Arthropoda split [~900 Myr (Hedges, Dudley, and Kumar 2006)], such as proto-lipocalins and proto-OBP/CSP or proto-GPCR and proto-GR genes, among others ("A" in Figure 7). After the split, the two taxa developed functionally equivalent gustatory receptor proteins tuned for soluble chemicals: the GR in Arthropoda ("B" in Figure 7) and gustatory-GPCR in Vertebrata ("C" in Figure 7). These two lineages later terrestrialized [380-420 Mya (Arthropoda) and ~340 Mya (Vertebrata)] (Ward et al. 2006) and the new selective pressures led to the independent functional diversification of existing gene families to mediate the transport and detection of volatile molecules. In Crustacea most lineages remained aquatic with no need for such evolutionary innovations ("D" in Figure 7). The new odorant binding and transport activities were taken over by olfactory lipocalins in vertebrates, OBP/CSP in Hexapoda and likely by some (but unknown)

binding protein family in Chelicerata (“E” in Figure 7). A parallel scenario could have occurred during chemoreceptor evolution (“F” in Figure 7): the GR would have evolved into the Hexapoda OR [as proposed by (Robertson, Warr, and Carlson 2003; Penalva-Arana, Lynch, and Robertson 2009)], gustatory-GPCR into vertebrate olfactory-GPCR receptors, and some unknown receptor gene family into the Chelicerata olfactory chemoreceptors. Interestingly, and further supporting this idea, mammals have experienced the reverse adaptive changes during the transition from a terrestrial to a fully aquatic habitat (Hayden et al. 2010) with large-scale pseudogenizations resulting in major reductions (in some cases total) of the OR repertoire (“G” in Figure 7) (McGowen, Clark, and Gatesy 2008). The diversification of olfactory-binding and receptor gene families in Arthropoda and Vertebrata seems to have occurred at roughly the same time, after the terrestrialization of each taxon. The nearly contemporary but independent origin of basic molecular elements of the olfactory system suggests a coevolution process between these gene families (OBP with OR; olfactory GPCR with lipocalins). In this sense, it is highly suggestive the similar distribution pattern of selective constraints (Sanchez-Gracia, Vieira, and Rozas 2009) and birth-and-loss rates (Sanchez-Gracia et al. 2011) between Hexapoda OBP and OR genes (but not between OBP and GR genes).

Funding

This work was supported by the Ministerio de Ciencia y Innovación (Spain) [grants BFU2007-62927 and BFU2010-15484] and the Comissió Interdepartamental de Recerca i Innovació Tecnològica (Spain) [grant number 2009SGR-1287]. FGV was supported by the predoctoral fellowship from the “Fundação para a Ciência e a Tecnologia” (Portugal) [SFRH/BD/22360/2005].

Acknowledgments

We would like to thank P. Librado for allowing us the use of an early version of the BadiRate program and for all his help, support and valuable discussions. We are also grateful to A. Sánchez-Gracia and J. M. Ranz for their constructive input and comments on the manuscript.

References

Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res **25**:3389-3402.

Andronopoulou, E., V. Labropoulou, V. Douris, D. F. Woods, H. Biessmann, and K. Iatrou. 2006. Specific interactions among odorant-binding proteins of the African malaria vector Anopheles gambiae. Insect Mol Biol **15**:797-811.

Angeli, S., F. Ceron, A. Scaloni, M. Monti, G. Monteforti, A. Minnocci, R. Petacchi, and P. Pelosi. 1999. Purification, structural characterization, cloning and immunocytochemical localization of chemoreception proteins from Schistocerca gregaria. Eur J Biochem **262**:745-754.

Asahina, K., V. Pavlenkovich, and L. B. Vosshall. 2008. The survival advantage of olfaction in a competitive environment. Curr Biol **18**:1153-1155.

Benton, R., K. S. Vannice, C. Gomez-Diaz, and L. B. Vosshall. 2009. Variant ionotropic glutamate receptors as chemosensory receptors in Drosophila. Cell **136**:149-162.

Berman, H. M., J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. 2000. The Protein Data Bank. Nucleic Acids Res **28**:235-242.

Boutanaev, A. M., A. I. Kalmykova, Y. Y. Shevelyov, and D. I. Nurminsky. 2002. Large clusters of co-expressed genes in the Drosophila genome. Nature **420**:666-669.

Brigaud, I., N. Montagne, C. Monsempes, M. C. Francois, and E. Jacquin-Joly. 2009. Identification of an atypical insect olfactory receptor subtype highly conserved within noctuids. FEBS J **276**:6537-6547.

Brody, T., and A. Cravchik. 2000. Drosophila melanogaster G protein-coupled receptors. J Cell Biol **150**:F83-88.

Croset, V., R. Rytz, S. F. Cummins, A. Budd, D. Brawand, H. Kaessmann, T. J. Gibson, and R. Benton. 2010. Ancient protostome origin of chemosensory ionotropic glutamate receptors and the evolution of insect taste and olfaction. PLoS Genet **6**.

Demuth, J. P., T. De Bie, J. E. Stajich, N. Cristianini, and M. W. Hahn. 2006. The evolution of mammalian gene families. PLoS One **1**:e85.

Drysdale, R. 2008. FlyBase : a database for the Drosophila research community. Methods Mol Biol **420**:45-59.

Findlay, G. D., X. Yi, M. J. Maccoss, and W. J. Swanson. 2008. Proteomics reveals novel Drosophila seminal fluid proteins transferred at mating. PLoS Biol **6**:e178.

Finn, R. D., J. Mistry, B. Schuster-Bockler, S. Griffiths-Jones, V. Hollich, T. Lassmann, S. Moxon, M. Marshall, A. Khanna, R. Durbin, S. R. Eddy, E. L. Sonnhammer, and A. Bateman. 2006. Pfam: clans, web tools and services. Nucleic Acids Res **34**:D247-251.

Flicek, P., B. L. Aken, K. Beal, B. Ballester, M. Caccamo, Y. Chen, L. Clarke, G. Coates, F. Cunningham, T. Cutts, T. Down, S. C. Dyer, T. Eyre, S. Fitzgerald, J. Fernandez-Banet, S. Graf, S. Haider, M. Hammond, R. Holland, K. L. Howe, K. Howe, N. Johnson, A. Jenkinson, A. Kahari, D. Keefe, F. Kokocinski, E. Kulesha, D. Lawson, I. Longden, K. Megy, P. Meidl, B. Overduin, A. Parker, B. Pritchard, A. Prlic, S. Rice, D. Rios, M. Schuster, I. Sealy, G. Slater, D. Smedley, G. Spudich, S. Trevanion, A. J. Vilella, J. Vogel, S. White, M. Wood, E. Birney, T. Cox, V. Curwen, R. Durbin, X. M. Fernandez-Suarez, J. Herrero, T. J. Hubbard, A. Kasprzyk, G. Proctor, J. Smith, A. Ureta-Vidal, and S. Searle. 2008. Ensembl 2008. Nucleic Acids Res **36**:D707-714.

Flower, D. R. 1996. The lipocalin protein family: structure and function. Biochem J **318 (Pt 1)**:1-14.

Foret, S., and R. Maleszka. 2006. Function and evolution of a gene family encoding odorant binding-like proteins in a social insect, the honey bee (Apis mellifera). Genome Res **16**:1404-1413.

Foret, S., K. W. Wanner, and R. Maleszka. 2007. Chemosensory proteins in the honey bee: Insights from the annotated genome, comparative analyses and expressional profiling. Insect Biochem Mol Biol **37**:19-28.

Freitag, J., G. Ludwig, I. Andreini, P. Rossler, and H. Breer. 1998. Olfactory receptors in aquatic and terrestrial vertebrates. J Comp Physiol [A] **183**:635-650.

Gong, D. P., H. J. Zhang, P. Zhao, Y. Lin, Q. Y. Xia, and Z. H. Xiang. 2007. Identification and expression pattern of the chemosensory protein gene family in the silkworm, Bombyx mori. Insect Biochem Mol Biol **37**:266-277.

Gong, D. P., H. J. Zhang, P. Zhao, Q. Y. Xia, and Z. H. Xiang. 2009. The odorant binding protein gene family from the genome of silkworm, Bombyx mori. BMC Genomics **10**:332.

Graham, L. A., D. Brewer, G. Lajoie, and P. L. Davies. 2003. Characterization of a subfamily of beetle odorant-binding proteins found in hemolymph. Mol Cell Proteomics **2**:541-549.

Grosse-Wilde, E., A. Svatos, and J. Krieger. 2006. A pheromone-binding protein mediates the bombykol-induced activation of a pheromone receptor in vitro. Chem Senses **31**:547-555.

Guo, S., and J. Kim. 2007. Molecular evolution of Drosophila odorant receptor genes. Mol Biol Evol **24**:1198-1207.

Hahn, M. W., T. De Bie, J. E. Stajich, C. Nguyen, and N. Cristianini. 2005. Estimating the tempo and mode of gene family evolution from comparative genomic data. Genome Res **15**:1153-1160.

Hahn, M. W., M. V. Han, and S. G. Han. 2007. Gene family evolution across 12 Drosophila genomes. PLoS Genet **3**:e197.

Hayden, S., M. l. Bekaert, T. A. Crider, S. Mariani, W. J. Murphy, and E. C. Teeling. 2010. Ecological adaptation determines functional mammalian olfactory subgenomes. Genome Research **20**:1-9.

Hedges, S. B., J. Dudley, and S. Kumar. 2006. TimeTree: a public knowledge-base of divergence times among organisms. Bioinformatics **22**:2971-2972.

Hekmat-Scafe, D. S., C. R. Scafe, A. J. McKinney, and M. A. Tanouye. 2002. Genome-wide analysis of the odorant-binding protein gene family in Drosophila melanogaster. Genome Res **12**:1357-1369.

Hiller, K., A. Grote, M. Scheer, R. Munch, and D. Jahn. 2004. PrediSi: prediction of signal peptides and their cleavage positions. Nucleic Acids Res **32**:W375-379.

Jones, D. T., W. R. Taylor, and J. M. Thornton. 1992. The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci **8**:275-282.

Kaissling, K. E. 2001. Olfactory perireceptor and receptor events in moths: a kinetic model. Chem Senses **26**:125-150.

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

Kaupp, U. B. 2010. Olfactory signalling in vertebrates and insects: differences and commonalities. Nat Rev Neurosci **11**:188-200.

Kirkness, E. F., B. J. Haas, W. Sun, H. R. Braig, M. A. Perotti, J. M. Clark, S. H. Lee, H. M. Robertson, R. C. Kennedy, E. Elhaik, D. Gerlach, E. V. Kriventseva, C. G. Elsik, D. Graur, C. A. Hill, J. A. Veenstra, B. Walenz, J. M. Tubio, J. M. Ribeiro, J. Rozas, J. S. Johnston, J. T. Reese, A. Popadic, M. Tojo, D. Raoult, D. L. Reed, Y. Tomoyasu, E. Krause, O. Mittapalli, V. M. Margam, H. M. Li, J. M. Meyer, R. M. Johnson, J. Romero-Severson, J. P. Vanzee, D. Alvarez-Ponce, F. G. Vieira, M. Aguade, S. Guirao-Rico, J. M. Anzola, K. S. Yoon, J. P. Strycharz, M. F. Unger, S. Christley, N. F. Lobo, M. J. Seufferheld, N. Wang, G. A. Dasch, C. J. Struchiner, G. Madey, L. I. Hannick, S. Bidwell, V. Joardar, E. Caler, R. Shao, S. C. Barker, S. Cameron, R. V. Bruggner, A. Regier, J. Johnson, L. Viswanathan, T. R. Utterback, G. G. Sutton, D. Lawson, R. M. Waterhouse, J. C. Venter, R. L. Strausberg, M. R. Berenbaum, F. H. Collins, E. M. Zdobnov, and B. R. Pittendrigh. 2010. Genome sequences of the human body louse and its primary endosymbiont provide insights into the permanent parasitic lifestyle. Proc Natl Acad Sci U S A **107**:12168-12173.

Krieger, J., R. E. Sandeman, D. C. Sandeman, B. S. Hansson, and S. Harzsch. 2010. Brain architecture of the largest living land arthropod, the Giant Robber Crab Birgus latro (Crustacea, Anomura, Coenobitidae): evidence for a prominent central olfactory pathway? Front Zool **7**:25.

Krieger, M. J., and K. G. Ross. 2002. Identification of a major gene regulating complex social behavior. Science **295**:328-332.

Labandeira, C. C. 2005. Invasion of the continents: cyanobacterial crusts to tree-inhabiting arthropods. Trends Ecol Evol **20**:253-262.

Larsson, M. C., A. I. Domingos, W. D. Jones, M. E. Chiappe, H. Amrein, and L. B. Vosshall. 2004. Or83b encodes a broadly expressed odorant receptor essential for Drosophila olfaction. Neuron **43**:703-714.

Lawson, D., P. Arensburger, P. Atkinson, N. J. Besansky, R. V. Bruggner, R. Butler, K. S. Campbell, G. K. Christophides, S. Christley, E. Dialynas, D. Emmert, M. Hammond, C. A. Hill, R. C. Kennedy, N. F. Lobo, M. R. MacCallum, G. Madey, K. Megy, S. Redmond, S. Russo, D. W. Severson, E. O. Stinson, P. Topalis, E. M. Zdobnov, E. Birney, W. M. Gelbart, F. C. Kafatos, C. Louis, and F. H. Collins. 2007. VectorBase: a home for invertebrate vectors of human pathogens. Nucleic Acids Res **35**:D503-505.

Leal, W. S., A. M. Chen, Y. Ishida, V. P. Chiang, M. L. Erickson, T. I. Morgan, and J. M. Tsuruda. 2005. Kinetics and molecular properties of pheromone binding and release. Proc Natl Acad Sci U S A **102**:5386-5391.

Leal, W. S., L. Nikonova, and G. Peng. 1999. Disulfide structure of the pheromone binding protein from the silkworm moth, Bombyx mori. FEBS Lett **464**:85-90.

Letunic, I., and P. Bork. 2007. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics **23**:127-128.
Ling, X., X. He, and D. Xin. 2009. Detecting gene clusters under evolutionary constraint in a large number of genomes. Bioinformatics **25**:571-577.
Luc, N., J. L. Risler, A. Bergeron, and M. Raffinot. 2003. Gene teams: a new formalization of gene clusters for comparative genomics. Comput Biol Chem **27**:59-67.
Matsuo, T., S. Sugaya, J. Yasukawa, T. Aigaki, and Y. Fuyama. 2007. Odorant-binding proteins OBP57d and OBP57e affect taste perception and host-plant preference in Drosophila sechellia. PLoS Biol **5**:e118.
McGowen, M. R., C. Clark, and J. Gatesy. 2008. The Vestigial Olfactory Receptor Subgenome of Odontocete Whales: Phylogenetic Congruence between Gene-Tree Reconciliation and Supermatrix Methods. Syst Biol **57**:574-590.
McGuffin, L. J., K. Bryson, and D. T. Jones. 2000. The PSIPRED protein structure prediction server. Bioinformatics **16**:404-405.
Montfort, W. R., A. Weichsel, and J. F. Andersen. 2000. Nitrophorins and related antihemostatic lipocalins from Rhodnius prolixus and other blood-sucking arthropods. Biochim Biophys Acta **1482**:110-118.
Nei, M. 2007. The new mutation theory of phenotypic evolution. Proc Natl Acad Sci U S A **104**:12235-12242.
Nei, M., Y. Niimura, and M. Nozawa. 2008. The evolution of animal chemosensory receptor gene repertoires: roles of chance and necessity. Nat Rev Genet **9**:951-963.
Nei, M., and A. P. Rooney. 2005. Concerted and birth-and-death evolution of multigene families. Annu Rev Genet **39**:121-152.
Pelosi, P., M. Calvello, and L. Ban. 2005. Diversity of odorant-binding proteins and chemosensory proteins in insects. Chem Senses **30 Suppl 1**:i291-292.
Pelosi, P., and R. Maida. 1990. Odorant-binding proteins in vertebrates and insects: similarities and possible common function. Chem. Senses **15**:205-215.
Pelosi, P., J. J. Zhou, L. P. Ban, and M. Calvello. 2006. Soluble proteins in insect chemical communication. Cell Mol Life Sci **63**:1658-1676.
Penalva-Arana, D. C., M. Lynch, and H. M. Robertson. 2009. The chemoreceptor genes of the waterflea Daphnia pulex: many Grs but no Ors. BMC Evol Biol **9**:79.
Pertea, M., X. Lin, and S. L. Salzberg. 2001. GeneSplicer: a new computational method for splice site prediction. Nucleic Acids Res **29**:1185-1190.
Pevzner, P., and G. Tesler. 2003. Human and mouse genomic sequences reveal extensive breakpoint reuse in mammalian evolution. Proc Natl Acad Sci U S A **100**:7672-7677.
Pophof, B. 2004. Pheromone-binding proteins contribute to the activation of olfactory receptor neurons in the silkmoths antheraea polyphemus and Bombyx mori. Chem Senses **29**:117-125.
Quijano, C., P. Tomancak, J. Lopez-Marti, M. Suyama, P. Bork, M. Milan, D. Torrents, and M. Manzanares. 2008. Selective maintenance of Drosophila tandemly arranged duplicated genes during evolution. Genome Biol **9**:R176.
Robertson, H. M., C. G. Warr, and J. R. Carlson. 2003. Molecular evolution of the insect chemoreceptor gene superfamily in Drosophila melanogaster. Proc Natl Acad Sci U S A **100 Suppl 2**:14537-14542.
Rutherford, K., J. Parkhill, J. Crook, T. Horsnell, P. Rice, M. A. Rajandream, and B. Barrell. 2000. Artemis: sequence visualization and annotation. Bioinformatics **16**:944-945.
Sanchez-Gracia, A., F. G. Vieira, F. C. Almeida, and J. Rozas. 2011. Comparative Genomics of the Major Chemosensory Gene Families in Arthropods *in* E. L. Sciences, ed. John Wiley & Sons, Ltd, Chichester.

Sanchez-Gracia, A., F. G. Vieira, and J. Rozas. 2009. Molecular evolution of the major chemosensory gene families in insects. Heredity **103**:208-216.
Sanchez, D., M. D. Ganfornina, S. Torres-Schumann, S. D. Speese, J. M. Lora, and M. J. Bastiani. 2000. Characterization of two novel lipocalins expressed in the Drosophila embryonic nervous system. Int J Dev Biol **44**:349-359.
Scaloni, A., M. Monti, S. Angeli, and P. Pelosi. 1999. Structural analysis and disulfide-bridge pairing of two odorant-binding proteins from Bombyx mori. Biochem Biophys Res Commun **266**:386-391.
Slater, G. S., and E. Birney. 2005. Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics **6**:31.
Smadja, C., and R. K. Butlin. 2009. On the scent of speciation: the chemosensory system and its role in premating isolation. Heredity **102**:77-97.
Smadja, C., P. Shi, R. K. Butlin, and H. M. Robertson. 2009. Large gene family expansions and adaptive evolution for odorant and gustatory receptors in the pea aphid, Acyrthosiphon pisum. Mol Biol Evol:msp116.
Soding, J. 2005. Protein homology detection by HMM-HMM comparison. Bioinformatics **21**:951-960.
Stajich, J. E., D. Block, K. Boulez, S. E. Brenner, S. A. Chervitz, C. Dagdigian, G. Fuellen, J. G. Gilbert, I. Korf, H. Lapp, H. Lehvaslaiho, C. Matsalla, C. J. Mungall, B. I. Osborne, M. R. Pocock, P. Schattner, M. Senger, L. D. Stein, E. Stupka, M. D. Wilkinson, and E. Birney. 2002. The Bioperl toolkit: Perl modules for the life sciences. Genome Res **12**:1611-1618.
Stamatakis, A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics **22**:2688-2690.
Steinbrecht, R. A. 1998. Odorant-binding proteins: expression and function. Ann N Y Acad Sci **855**:323-332.
Stensmyr, M. C., S. Erland, E. Hallberg, R. Wallen, P. Greenaway, and B. S. Hansson. 2005. Insect-like olfactory adaptations in the terrestrial giant robber crab. Curr Biol **15**:116-121.
Tamura, K., J. Dudley, M. Nei, and S. Kumar. 2007. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol Biol Evol **24**:1596-1599.
Tamura, K., S. Subramanian, and S. Kumar. 2004. Temporal patterns of fruit fly (Drosophila) evolution revealed by mutation clocks. Mol Biol Evol **21**:36-44.
Tegoni, M., V. Campanacci, and C. Cambillau. 2004. Structural aspects of sexual attraction and chemical communication in insects. Trends Biochem Sci **29**:257-264.
Valenzuela, J. G., R. Charlab, E. C. Gonzalez, I. K. de Miranda-Santos, O. Marinotti, I. M. Francischetti, and J. M. Ribeiro. 2002. The D7 family of salivary proteins in blood sucking diptera. Insect Mol Biol **11**:149-155.
Vieira, F. G., A. Sanchez-Gracia, and J. Rozas. 2007. Comparative genomic analysis of the odorant-binding protein family in 12 Drosophila genomes: purifying selection and birth-and-death evolution. Genome Biol **8**:R235.
Vogt, R. G., and L. M. Riddiford. 1981. Pheromone binding and inactivation by moth antennae. Nature **293**:161-163.
von Grotthuss, M., M. Ashburner, and J. M. Ranz. 2010. Fragile regions and not functional constraints predominate in shaping gene organization in the genus Drosophila. Genome Res **20**:1084-1096.
Wang, J., Q. Xia, X. He, M. Dai, J. Ruan, J. Chen, G. Yu, H. Yuan, Y. Hu, R. Li, T. Feng, C. Ye, C. Lu, S. Li, G. K. Wong, H. Yang, Z. Xiang, Z. Zhou, and J. Yu. 2005. SilkDB: a knowledgebase for silkworm biology and genomics. Nucleic Acids Res **33**:D399-402.

Wang, L., S. Wang, Y. Li, M. S. R. Paradesi, and S. J. Brown. 2007. BeetleBase: the model organism database for Tribolium castaneum. Nucl. Acids Res. **35**:D476-479.

Wanner, K. W., A. R. Anderson, S. C. Trowell, D. A. Theilmann, H. M. Robertson, and R. D. Newcomb. 2007. Female-biased expression of odourant receptor genes in the adult antennae of the silkworm, Bombyx mori. Insect Mol Biol **16**:107-119.

Wanner, K. W., and H. M. Robertson. 2008. The gustatory receptor family in the silkworm moth Bombyx mori is characterized by a large expansion of a single lineage of putative bitter receptors. Insect Mol Biol **17**:621-629.

Ward, P., C. Labandeira, M. Laurin, and R. A. Berner. 2006. Confirmation of Romer's Gap as a low oxygen interval constraining the timing of initial arthropod and vertebrate terrestrialization. Proc Natl Acad Sci U S A **103**:16818-16822.

Whelan, S., and N. Goldman. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol Biol Evol **18**:691-699.

Whiteman, N. K., and N. E. Pierce. 2008. Delicious poison: genetics of Drosophila host plant preference. Trends Ecol Evol **23**:473-478.

Xu, P., R. Atkinson, D. N. Jones, and D. P. Smith. 2005. Drosophila OBP LUSH is required for activity of pheromone-sensitive neurons. Neuron **45**:193-200.

Xu, P. X., L. J. Zwiebel, and D. P. Smith. 2003. Identification of a distinct family of genes encoding atypical odorant-binding proteins in the malaria vector mosquito, Anopheles gambiae. Insect Mol Biol **12**:549-560.

Ye, Y., and A. Godzik. 2004. FATCAT: a web server for flexible structure comparison and structure similarity searching. Nucleic Acids Res **32**:W582-585.

Zhou, J. J., X. L. He, J. A. Pickett, and L. M. Field. 2008. Identification of odorant-binding proteins of the yellow fever mosquito Aedes aegypti: genome annotation and comparative analyses. Insect Mol Biol **17**:147-163.

Zhou, J. J., Y. Kan, J. Antoniw, J. A. Pickett, and L. M. Field. 2006. Genome and EST analyses and expression of a gene family with putative functions in insect chemoreception. Chem Senses **31**:453-465.

Zhou, J. J., F. G. Vieira, X. L. He, C. Smadja, R. Liu, J. Rozas, and L. M. Field. 2010. Genome annotation and comparative analyses of the odorant-binding proteins and chemosensory proteins in the pea aphid Acyrthosiphon pisum. Insect Mol Biol **19 Suppl 2**:113-122.

Table 1 – Number of OBP and CSP genes in the Arthropoda species analyzed

	OBP			CSP		
	Putative Functional		Pseudogenes	Putative Functional		Pseudogenes
	Complete Sequence	Low Coverage Sequence[(a)]		Complete Sequence	Low Coverage Sequence[(a)]	
Dmel	52	0	0	4	0	0
Dsim	52	0	0	4	0	0
Dsec	51	0	1	4	0	0
Dyak	55	0	0	4	0	0
Dere	50	0	2	4	0	0
Dana	50	0	2	3	0	1
Dpse	45	0	2	4	0	0
Dper	45	0	2	4	0	0
Dwil	62	0	2	4	0	0
Dmoj	43	0	0	4	0	0
Dvir	41	0	1	4	0	0
Dgri	46	0	3	4	0	0
Agam	81 (66)	2 (0)	0	8 (7)	0	0
Bmor	43 (44)	3 (0)	1 (0)	19 (16)	3 (2)	2 (0)
Tcas	49 (46)	0	1 (0)	19 (20)	0	1 (0)
Amel	21	0	0	6	0	0
Phum	4	1	1	6	1	1
Apis	14 (11)	4	1 (0)	10	2 (3)	1 (0)
Dpul	0	0	0	3	0	0
Isca	0	0	0	1	0	0

[(a)]genes with truncated CDS due to incomplete genome assembly

The four-letter code used for the species is: *Drosophila melanogaster* (Dmel), *D. simulans* (Dsim), *D. sechellia* (Dsec), *D. erecta* (Dere), *D. yakuba* (Dyak), *D. ananassae* (Dana), *D. pseudoobscura* (Dpse), *D. persimilis* (Dper), *D. willistoni* (Dwil), *D. mojavensis* (Dmoj), *D. virilis* (Dvir) and *D. grimshawi* (Dgri), *Anopheles gambiae* (Agam), *Bombyx mori* (Bmor), *Tribolium castaneum* (Tcas), *Apis mellifera* (Amel), *Pediculus humanus* (Phum), *Acyrthosiphon pisum* (Apis), *Daphnia pulex* (Dpul) and *Ixodes scapularis* (Isca). The numbers of the OBP and CSP genes reported in previous works are given in parenthesis (only in cases with discrepancies).

Table 2 – P-values of the chromosomal clusters analysis.

	OBP			CSP		
	Observed	**Dist. Average**	**p-value**	**Observed**	**Dist. Average**	**p-value**
Dmel	1322846.7	2026501.8	< 0.0001	-	-	-[a]
Dsim	1268476.5	2006452.9	< 0.0001	-	-	-[a]
Dsec	719946.2	1478286.5	< 0.0001	-	-	-[a]
Dyak	1229276.6	1993428.8	< 0.0001	-	-	-[a]
Dere	1271164.5	2057290.3	0.0004	-	-	-[a]
Dana	1245187.1	1982158.1	0.0004	-	-	-[a]
Dpse	1102439.9	2146580.2	< 0.0001	-	-	-[a]
Dper	331607.1	1467330.8	< 0.0001	-	-	-[a]
Dwil	306109.5	1680312.9	< 0.0001	-	-	-[a]
Dmoj	1528514.0	2917407.1	< 0.0001	-	-	-[a]
Dvir	774464.4	2483148.4	< 0.0001	-	-	-[a]
Dgri	1112257.6	2172345.2	< 0.0001	-	-	-[a]
Agam	2478428.4	3020631.7	0.0064	-	-	-[a]
Bmor	124328.2	22207012.0	0.001	33582.7	1835740.8	0.0008
Tcas	2267791.0	3736858.6	0.0052	832475.9	5529128.4	< 0.0001
Amel	-	-	-[b]	-	-	-[a]
Phum	-	-	-[a]	-	-	-[a]
Apis	-	-	-[b]	-	-	-[b]
Dpul	-	-	-[a]	-	-	-[a]
Isca	-	-	-[a]	-	-	-[a]

[a] Species not analyzed for having less than ten gene members

[b] Species not analyzed for having a fragmented genome (probably due to poor coverage or assembling)

P-values calculated by computer simulations. The species four-letter code is as in Table 1.

Table 3 – Structural alignments between OBP and CSP proteins.

	OBP								**CSP**		
	BmorGOBP2	ApolPBP1	AgamD7r4	BmorPBP	AgamOBP1	AmelASP2	BmorPBP	Lush	BmorCSP1	SgreCSP4	MbraCSP2
	2WC5	**2JPO**	**2QEV**	**2P70**	**2ERB**	**1TUJ**	**1GM0**	**1T14**	**2JNT**	**2GVS**	**1K19**
2WC5		$1.66x10^{-07}$	$1.39x10^{-05}$	$2.18x10^{-14}$	$5.05x10^{-08}$	$3.39x10^{-06}$	$8.51x10^{-06}$	$1.21x10^{-07}$	NS	NS	NS
2JPO			$5.41x10^{-04}$	$1.29x10^{-08}$	$2.12x10^{-05}$	$6.04x10^{-05}$	$1.90x10^{-14}$	$1.76x10^{-05}$	NS	NS	**$4.80x10^{-02}$**
2QEV				$1.89x10^{-05}$	$9.01x10^{-05}$	$5.15x10^{-05}$	$3.85x10^{-04}$	$4.67x10^{-06}$	**$4.04x10^{-02}$**	**$3.30x10^{-02}$**	**$3.81x10^{-02}$**
2P70					$3.90x10^{-08}$	$1.21x10^{-07}$	$3.26x10^{-08}$	$2.52x10^{-08}$	**$2.72x10^{-02}$**	NS	**$4.40x10^{-02}$**
2ERB						$3.43x10^{-06}$	$1.12x10^{-04}$	$2.59x10^{-10}$	**$4.39x10^{-02}$**	**$4.35x10^{-02}$**	**$3.31x10^{-02}$**
1TUJ							$5.66x10^{-05}$	$1.62x10^{-06}$	**$1.61x10^{-02}$**	NS	**$4.23x10^{-02}$**
1GM0								$2.14x10^{-04}$	NS	NS	NS
1T14									**$3.15x10^{-02}$**	**$1.16x10^{-02}$**	**$3.21x10^{-02}$**
2JNT										$6.01x10^{-06}$	$6.80x10^{-05}$
2GVS											$6.84x10^{-06}$
1K19											

P-values obtained using FATCAT; statistical significant values between OBP and CSP structures are depicted in bold; NS: non significant.

FIGURE LEGENDS

Fig. 1 - Accepted tree topology for the Arthropoda species surveyed.

Blue shadowed boxes depict an aquatic environment. Divergence times are given in millions of years (Tamura, Subramanian, and Kumar 2004; Hedges, Dudley, and Kumar 2006). Right: number of members of the OBP and CSP gene families classified into subfamilies, and the presence of the OR and GR gene families.

Fig. 2 – OBP orthologous groups shared across species.

Venn diagrams indicate the inferred number of groups of orthologous genes (OG) shared among different insect species. A) Drosophila, B) Diptera, C) Diptera and Lepidoptera, D) Diptera, Lepidoptera and Coleoptera, E) Endopterygota and F) Hexapoda.

Fig. 3 - Phylogenetic relationships of the OBP proteins.

Unrooted phylogenetic tree of OBP protein sequences from *Drosophila melanogaster* and *D. mojavensis* (red branches), *Anopheles gambiae* (blue branches), *Bombyx mori* (brown branches), *Tribolium castaneum* (green branches), *Apis mellifera* (orange branches), *Pediculus humanus* (pink branches) and *Acyrthosyphon pisum* (cyan branches). Inner and outer rings indicate phylogenetic subfamilies (Classic in black, Minus-C in green, Plus-C in blue, Dimer in red, D7 in yellow, ABPI in cyan, ABPII in grey and PBP/GOBP in pink) and the secondary structure information (box: α-helix; arrow: β-sheet), respectively. The scale bar represents 1 amino acid substitution per site. The image was created using the iTOL web server (Letunic and Bork 2007).

Fig. 4 - Phylogenetic relationships of the CSP proteins.

Unrooted phylogenetic tree of CSP protein sequences from *Drosophila melanogaster* and *D. mojavensis* (red branches), *Anopheles gambiae* (blue branches), *Bombyx mori* (brown branches), *Tribolium castaneum* (green branches), *Apis mellifera* (orange branches), *Pediculus humanus* (pink branches), *Acyrthosyphon pisum* (cyan branches) and *Daphnia pulex* (black lines). Outer ring indicates the secondary structure information (box: α-helix; arrow: β-sheet). The scale bar represents 1 amino acid substitution per site. The tree was displayed using the iTOL web server (Letunic and Bork 2007).

Fig. 5 - Tertiary structure alignments.

Representation of the significant alignments between OBP and CSP structures. PDB protein structures are represented as nodes in yellow (OBP) and green (CSP). Significant alignments are depicted as edges between nodes; edge thickness and color range (ranging from grey, blue to red) indicate increasing significance levels.

Figure 6 - OBP and CSP gene gains and losses.

The inferred numbers of genes at each phylogenetic node are depicted in red. Values above and below the branches indicate the number of gene gains and losses, respectively. Subfamily gains (▲) and losses (✕) are color-coded (Classic in black, Minus-C in green, Plus-C in blue, Dimer in red, D7 in yellow, ABPI in cyan, ABPII in grey and PBP/GOBP in pink). OBP: odorant-binding protein; CSP: chemosensory protein

Fig. 7 - Putative Scenario for the evolution of the Chemosensory System

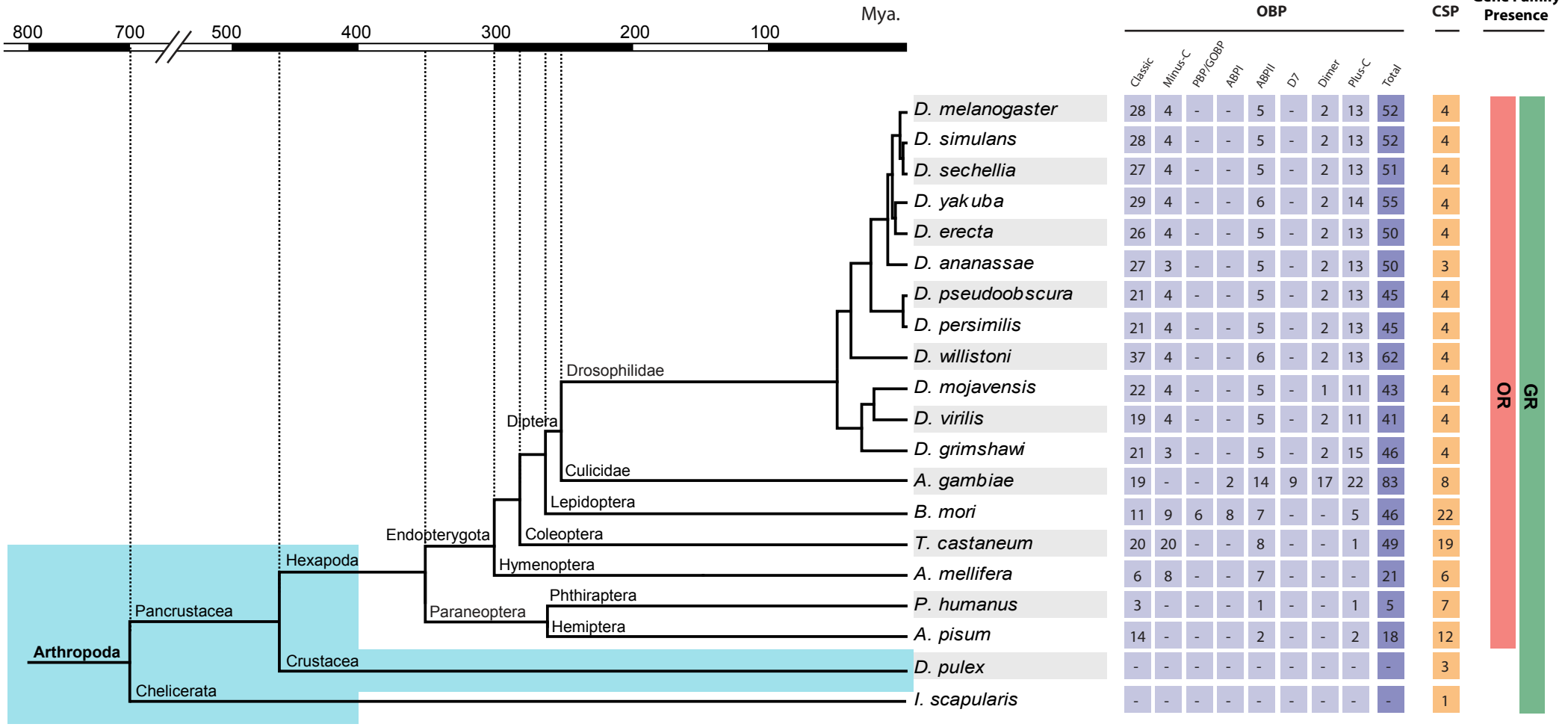
Shaded in blue boxes represent the aquatic lifestyle. Right: Presence or absence of the chemosensory gene families in extant species. Branch lengths are not to scale. Letters from A-F stand for the different evolutionary events (see text).

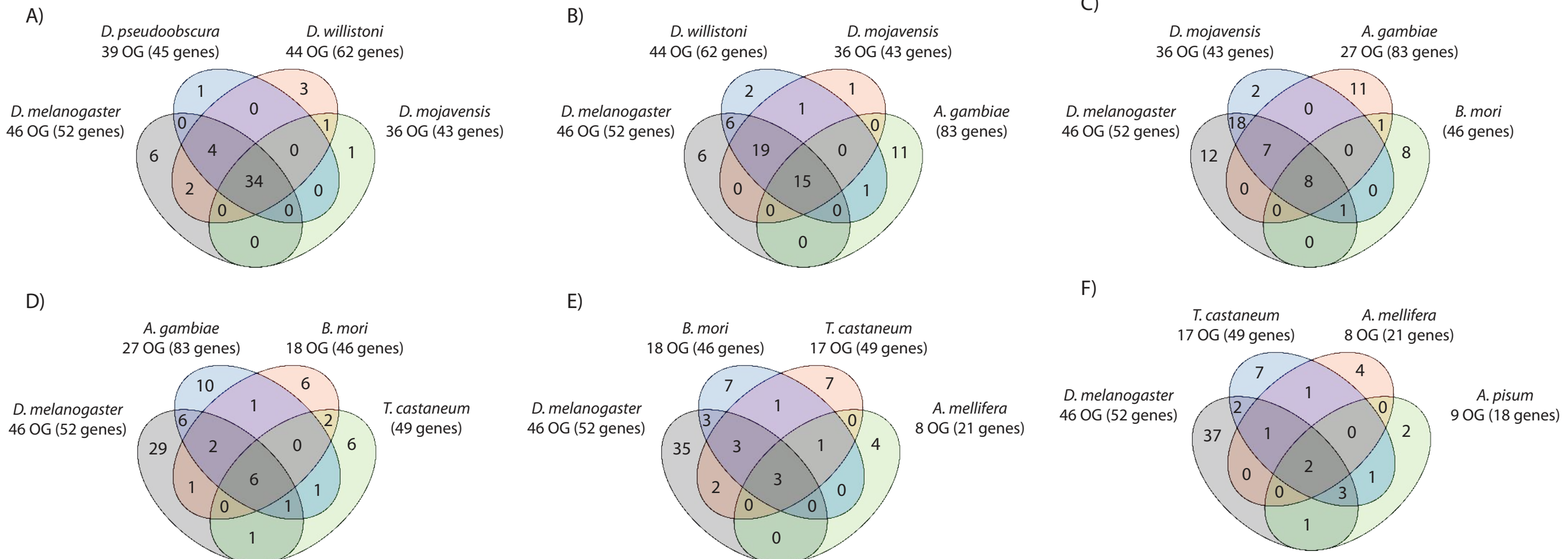
Supplementary Material

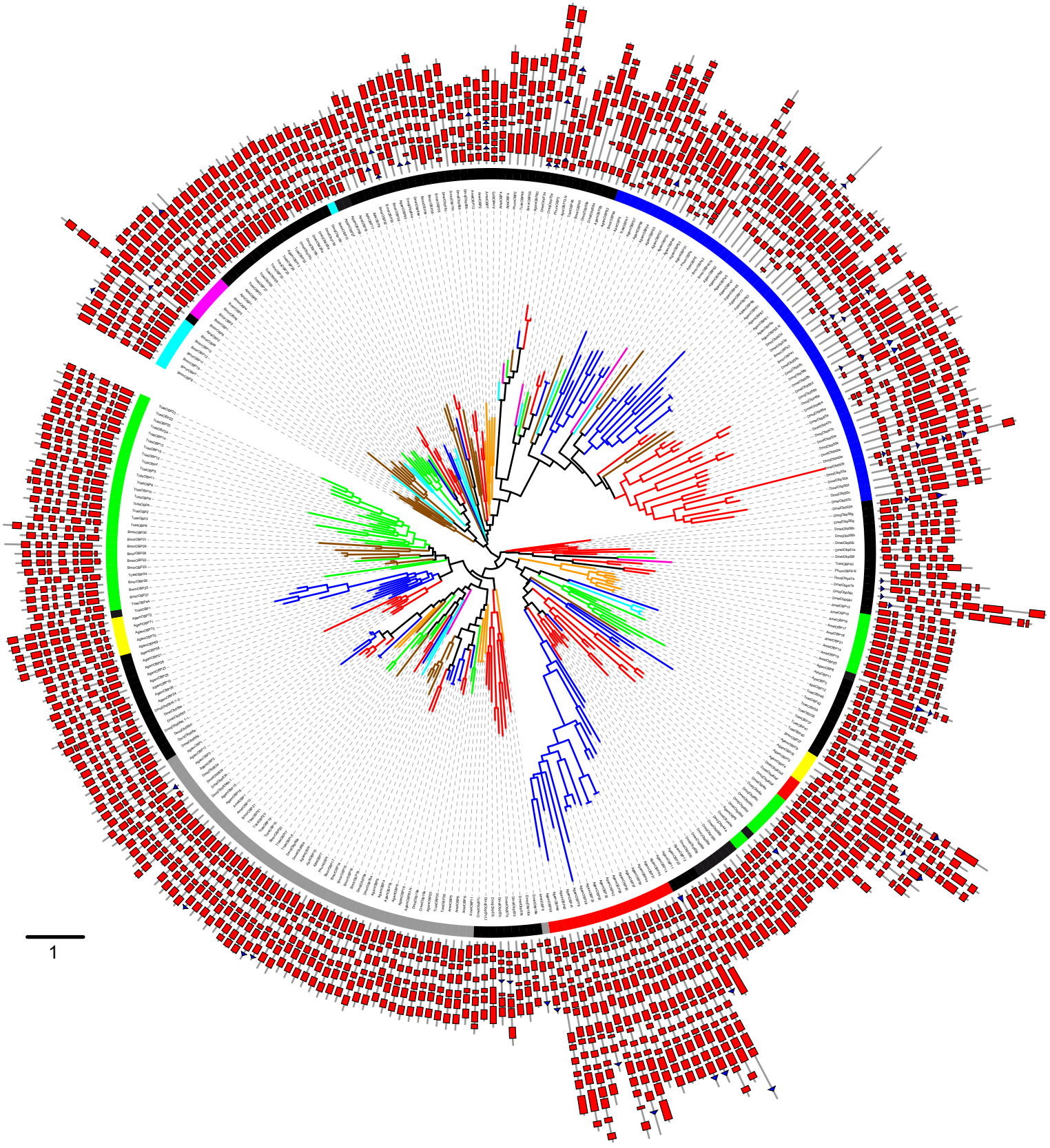
Supplementary File S1. OBP protein sequences in FASTA format.

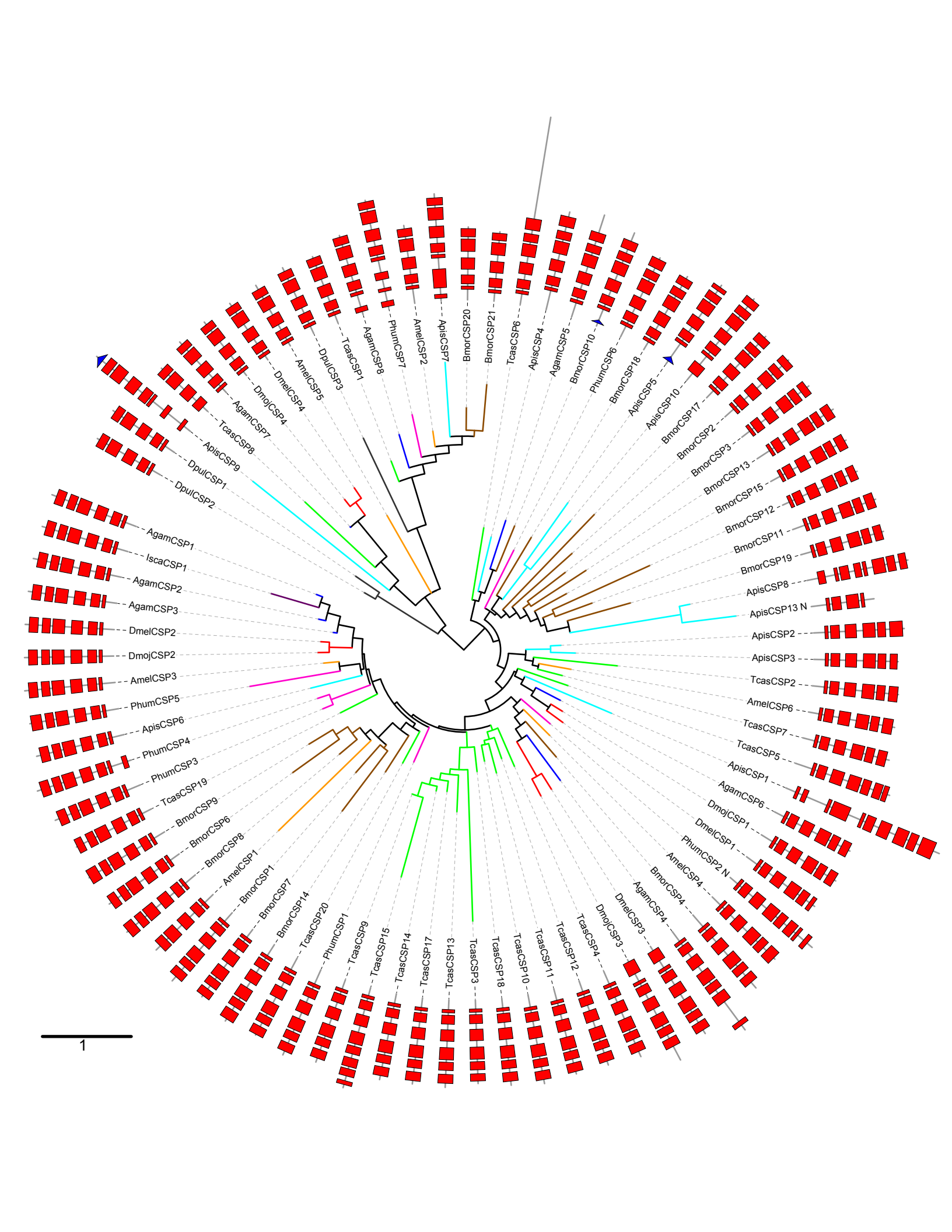
Supplementary File S2. CSP protein sequences in FASTA format.

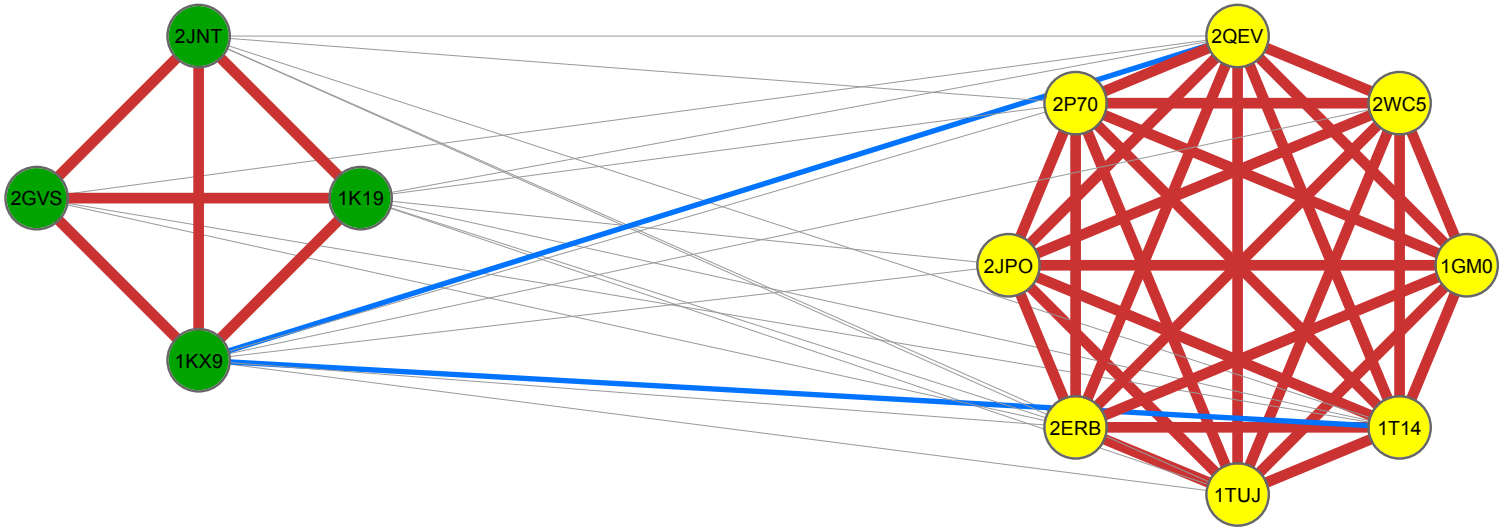
Supplementary Figure S1. Dot-plot between an Atypical (AgamOBP43) and a Classic (AgamOBP14) *A. gambiae* OBP. It is clearly seen the double domain structure of the Atypical OBP.

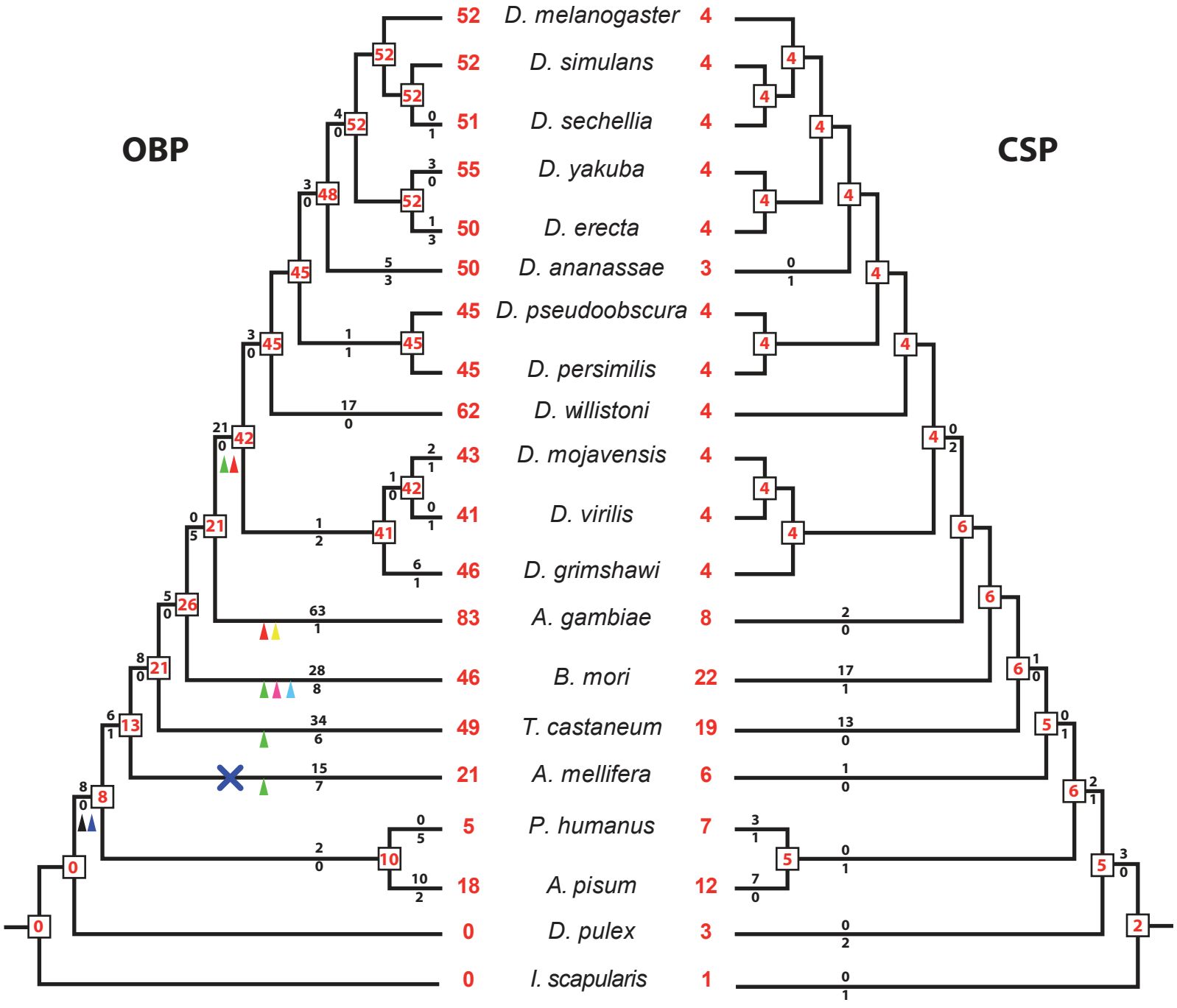

Mya.
800
700
500
400
300
200
100
Arthropoda
Pancrustacea
Chelicerata
Hexapoda
Crustacea
Endopterygota
Paraneoptera
Diptera
Hymenoptera
Phthiraptera
Hemiptera
Coleoptera
Lepidoptera
Culicidae
Drosophilidae
OBP
Classic
Minus-C
PBP/GOBP
ABPI
ABPII
D7
Dimer
Plus-C
Total
CSP
Gene Family Presence
OR
GR
D. melanogaster 28 4 - - 5 - 2 13 52 4
D. simulans 28 4 - - 5 - 2 13 52 4
D. sechellia 27 4 - - 5 - 2 13 51 4
D. yakuba 29 4 - - 6 - 2 14 55 4
D. erecta 26 4 - - 5 - 2 13 50 4
D. ananassae 27 3 - - 5 - 2 13 50 3
D. pseudoobscura 21 4 - - 5 - 2 13 45 4
D. persimilis 21 4 - - 5 - 2 13 45 4
D. willistoni 37 4 - - 6 - 2 13 62 4
D. mojavensis 22 4 - - 5 - 1 11 43 4
D. virilis 19 4 - - 5 - 2 11 41 4
D. grimshawi 21 3 - - 5 - 2 15 46 4
A. gambiae 19 - - 2 14 9 17 22 83 8
B. mori 11 9 6 8 7 - - 5 46 22
T. castaneum 20 20 - - 8 - - 1 49 19
A. mellifera 6 8 - - 7 - - - 21 6
P. humanus 3 - - - 1 - - 1 5 7
A. pisum 14 - - - 2 - - 2 18 12
D. pulex - - - - - - - - - 3
I. scapularis - - - - - - - - - 1

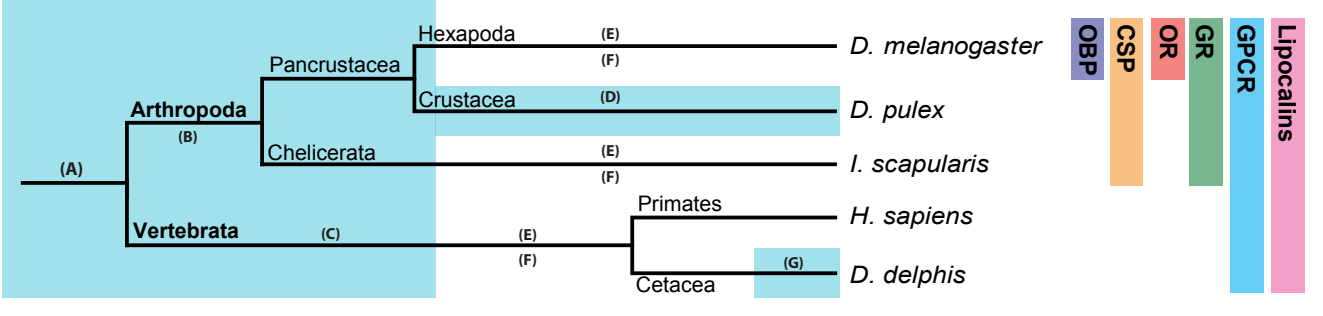

A)
D. pseudoobscura
39 OG (45 genes)
D. willistoni
44 OG (62 genes)
D. melanogaster
46 OG (52 genes)
D. mojavensis
36 OG (43 genes)
1
3
0
0
1
6
4
0
1
34
2
0
0
0
0
B)
D. willistoni
44 OG (62 genes)
D. mojavensis
36 OG (43 genes)
D. melanogaster
46 OG (52 genes)
A. gambiae
(83 genes)
2
1
1
6
19
0
11
15
0
1
0
0
0
C)
D. mojavensis
36 OG (43 genes)
A. gambiae
27 OG (83 genes)
D. melanogaster
46 OG (52 genes)
B. mori
(46 genes)
2
11
0
18
1
12
7
0
8
8
0
0
1
0
D)
A. gambiae
27 OG (83 genes)
B. mori
18 OG (46 genes)
D. melanogaster
46 OG (52 genes)
T. castaneum
(49 genes)
10
6
1
6
2
29
2
0
6
1
1
6
0
1
1
E)
B. mori
18 OG (46 genes)
T. castaneum
17 OG (49 genes)
D. melanogaster
46 OG (52 genes)
A. mellifera
8 OG (21 genes)
7
7
1
3
0
35
3
1
4
2
3
0
0
0
0
F)
T. castaneum
17 OG (49 genes)
A. mellifera
8 OG (21 genes)
D. melanogaster
46 OG (52 genes)
A. pisum
9 OG (18 genes)
7
4
1
2
0
37
1
0
2
0
2
1
0
3
1


1


AgamCSP1
IscaCSP1
AgamCSP2
AgamCSP3
DmelCSP2
DmojCSP2
AmelCSP3
PhumCSP5
ApisCSP6
PhumCSP4
PhumCSP3
TcasCSP19
BmorCSP9
BmorCSP6
BmorCSP8
AmelCSP1
BmorCSP1
BmorCSP7
BmorCSP14
TcasCSP20
PhumCSP1
TcasCSP9
TcasCSP15
TcasCSP14
TcasCSP17
TcasCSP13
TcasCSP3
TcasCSP18
TcasCSP10
TcasCSP11
TcasCSP12
TcasCSP4
DmojCSP3
DmelCSP3
AgamCSP4
BmorCSP4
AmelCSP4
PhumCSP2 N
DmelCSP1
DmojCSP1
AgamCSP6
ApisCSP1
TcasCSP5
TcasCSP7
AmelCSP6
TcasCSP2
ApisCSP3
ApisCSP2
ApisCSP13 N
ApisCSP8
BmorCSP19
BmorCSP11
BmorCSP12
BmorCSP15
BmorCSP13
BmorCSP3
BmorCSP2
BmorCSP17
ApisCSP10
ApisCSP5
BmorCSP18
PhumCSP6
BmorCSP10
AgamCSP5
ApisCSP4
TcasCSP6
BmorCSP21
BmorCSP20
ApisCSP7
AmelCSP2
PhumCSP7
AgamCSP8
TcasCSP1
DpulCSP3
AmelCSP5
DmelCSP4
DmojCSP4
AgamCSP7
TcasCSP8
ApisCSP9
DpulCSP1
DpulCSP2
1


2JNT
2GVS
1K19
1KX9
2QEV
2P70
2WC5
2JPO
1GM0
2ERB
1T14
1TUJ


OBP
CSP
52 D. melanogaster 4
52 D. simulans 4
51 D. sechellia 4
55 D. yakuba 4
50 D. erecta 4
50 D. ananassae 3
45 D. pseudoobscura 4
45 D. persimilis 4
62 D. willistoni 4
43 D. mojavensis 4
41 D. virilis 4
46 D. grimshawi 4
83 A. gambiae 8
46 B. mori 22
49 T. castaneum 19
21 A. mellifera 6
5 P. humanus 7
18 A. pisum 12
0 D. pulex 3
0 I. scapularis 1


Gene Family Presence
OBP
CSP
OR
GR
GPCR
Lipocalins
(A)
Arthropoda
(B)
Vertebrata
(C)
Pancrustacea
Chelicerata
Hexapoda
Crustacea
Primates
Cetacea
(D)
(E)
(F)
(G)
D. melanogaster
D. pulex
I. scapularis
H. sapiens
D. delphis