Comparative Genomics Uncovers Unique Gene Turnover and Evolutionary Rates in a Gene Family Involved in the Detection of Insect Cuticular Pheromones

Montserrat Torres-Oliva[1,2], Francisca C. Almeida[1,3], Alejandro Sánchez-Gracia[1,*] and Julio Rozas[1,*]

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

[2]Present Address: Georg-August-University Göttingen, Johann-Friedrich-Blumenbach Institute for Zoology and Anthropology, Department of Developmental Biology, Ernst-Caspari-House (GZMB), Justus-von-Liebig-Weg 11, Göttingen, Germany

[3]Present Address: Consejo Nacional de Investigaciones Científicas y Tecnológicas (CONICET-IEGEBA), Universidad de Buenos Aires, Departamento de Ecología, Genética y Evolución, Av. Intendente Güiraldes y Costanera Norte s/n, Pabellón II, Ciudad Universitaria, Capital Federal, Argentina

*Corresponding author: E-mail: elsanchez@ub.edu or jrozas@ub.edu.

Accepted: April 27, 2016

Abstract

Chemoreception is an essential process for the survival and reproduction of animals. Many of the proteins responsible for recognizing and transmitting chemical stimuli in insects are encoded by genes that are members of moderately sized multigene families. The members of the CheB family are specialized in gustatory-mediated detection of long-chain hydrocarbon pheromones in *Drosophila melanogaster* and play a central role in triggering and modulating mating behavior in this species. Here, we present a comprehensive comparative genomic analysis of the CheB family across 12 species of the *Drosophila* genus. We have identified a total of 102 new *CheB* genes in the genomes of these species, including a functionally divergent member previously uncharacterized in *D. melanogaster*. We found that, despite its relatively small repertory size, the CheB family has undergone multiple gain and loss events and various episodes of diversifying selection during the divergence of the surveyed species. Present estimates of gene turnover and coding sequence substitution rates show that this family is evolving faster than any known *Drosophila* chemosensory family. To date, only other insect gustatory-related genes among these families had shown evolutionary dynamics close to those observed in CheBs. Our findings reveal the high adaptive potential of molecular components of the gustatory system in insects and anticipate a key role of genes involved in this sensory modality in species adaptation and diversification.

Key words: Chemosensory proteins, CheB gene family, functional divergence, birth and death evolution, positive selection.

Introduction

Chemoreception is a critical biological process, essential for the survival, reproduction, and social behavior of animals. In insects, the chemosensory system is extremely specific and sensitive, allowing the detection and discrimination of a great number of chemical cues through olfactory and gustatory sensory modalities. In general, olfaction allows the recognition of volatile and long-distance molecules that confers on animals the ability to detect food, predators, and mates, whereas taste, on the other hand, allows a short-distance detection of soluble substances, which can induce responses related to feeding behaviors, courtship, and reproduction.

In insects, the first contact between the chemical signals and their receptors takes place inside specialized hair-like porous structures (the sensilla), in an aqueous environment surrounding sensory neurons (sensillar lymph) (Steinbrecht 1996; Carey and Carlson 2011). The signalling molecules enter through pores, are solubilized and transported across the lymph aided by molecular transporters (binding proteins), and interact with specific chemoreceptors anchored on the dendritic membrane of the sensory neuron, which in turn activate the corresponding signalling cascade (Shanbhag et al. 2001). Both the soluble binding proteins and the transmembrane receptors involved in these events are encoded

by multigene families (Pelosi et al. 2006; Touhara and Vosshall 2009). Soluble binding proteins include the Odorant Binding Protein (OBP), Chemosensory Protein (CSP), Chemosensory Protein A and B (CheA and CheB) and might also include other recently discovered families (NPC2), whereas the chemoreceptor gene families are represented by the Olfactory (OR), Gustatory (GR) and Ionotropic (IR) Receptors (Sánchez-Gracia et al. 2009; Pelosi et al. 2014).

The *D. melanogaster* CheB gene family is a moderately sized family with 12 members described to date (Xu et al. 2002). Proteins encoded by this gene family are small proteins (192–226 amino acids long) and characterized by a specific protein domain (DM11 from InterPro; Zdobnov and Apweiler 2001). Their secondary structure is similar to that of the Myeloid Differentiation-like protein family, a superfamily of lipid-binding proteins present in all eukaryotes. More specifically the secondary structure of the CheBs resembles that of the human Ganglioside M2 activator protein (GM2-AP), a soluble protein that binds to GM2 glycolipid, whose absence causes Tay-Sachs neurodegenerative disease (Starostina et al. 2009). Consistent with a role as extracellular ligand-binding proteins, all known CheB proteins have a hydrophobic amino-terminus of 15–20 residues (likely encoding a signal peptide) and are specifically expressed in secretory cells that surround gustatory neurons (Park et al. 2006).

It has been proposed that the CheB proteins are involved in the detection of cuticular long-chain hydrocarbons, which are very important pheromones that modulate vital and complex behaviors, such as mating and aggressiveness (Touhara and Vosshall 2009). The *CheB42a* gene of *D. melanogaster*, the first member of this family identified in insects (Xu et al. 2002), is expressed specifically in a set of gustatory sensilla of male front legs, the organs involved in the courtship-activating pheromone perception (Begg and Hogben 1946; Robertson 1983). Mutant males lacking this protein attempt to copulate sooner and more frequently with females than control males. Furthermore, these mutants also copulate more frequently with other males that express female specific pheromones, but not with females lacking these compounds (Park et al. 2006).

Although all *D. melanogaster CheB* genes are expressed predominantly (or exclusively) in gustatory organs, they exhibit specific expression patterns (often nonoverlapping) and, in some cases, sexual dimorphism. According to the expression patterns, the *CheB* genes have been classified into those specifically expressed in male front legs and those exhibiting a preferential expression in wings of both sexes (Starostina et al. 2009). Since human GM2-AP acts as a coreceptor in the glycolipid degradation pathway, it has been suggested (Pikielny 2010) that CheB proteins might work as coreceptors of pheromone-degrading enzymes. The same authors proposed that alternatively, CheB proteins could be necessary for correct detection and processing of cuticular hydrocarbons (CHCs) by facilitating their diffusion across the inner lumen of the gustatory sensilla and the activation of the specific membrane associated receptors.

The CheB gene family was discovered upon the isolation of the CheB42a cDNA in a subtractive library of front leg RNA of *D. melanogaster* males. In the same experiment, a cDNA encoding CheA29a was concomitantly isolated, prompting the discovery of the CheA family. All eight *D. melanogaster* CheA proteins, similarly to CheBs, contain a putative signal peptide region and at least two of them are preferentially expressed in chemosensory sensilla of male appendages (Xu et al. 2002), which suggests they may also have a role in male specific pheromone response. Despite these resemblances, CheA and CheB proteins have no apparent primary sequence similarity to each other, therefore defining either two separate families or two very divergent novel subfamilies. The CheB family members share a high degree of sequence similarity (minimum similarity between two members is 30%) and are clustered in the genome, while CheA proteins have lower sequence conservation (minimum similarity between two members is 21%) and show more isolated chromosomal locations (Xu et al. 2002).

Here, we present a comprehensive comparative genomic analysis of the *CheB* family repertoire across several insect species, including a careful reannotation and curation of known *CheB* genes in 12 *Drosophila* genomes (Clark et al. 2007). We also conducted an exhaustive search for putative, novel family members, searching for putative unidentified genes or genes that have not been recognized as members of the CheB family in current genome annotations. The curated data was then used to study the origin of the CheB family, to estimate the number of gene gains and losses and the turnover rates, and to analyze coding sequence evolution and functional divergence. We found that the CheB family is noticeably more dynamic (it shows higher birth and death rates) and exhibits lower selective constraints than the other binding protein families involved in chemosensation in insects (OBP, CSP, and NPC2). Although still higher, the evolutionary rate of the CheB family is much closer to those estimated for GRs and divergent IRs, both involved in taste perception, than to those estimated for families involved in olfaction. These findings indicate that in insects, gustatory proteins are more commonly involved in physiological processes causing accelerated rates of evolution, postulating for this chemosensory modality an important role in promoting adaptation and, potentially, speciation.

Materials and Methods

Databases

The nucleotide and protein sequences of the 12 *CheB* genes previously identified in the genome of *D. melanogaster* were downloaded from FlyBase (dos Santos et al. 2014) (release 6.03 of the genome annotation). We also retrieved from this

database the FASTA and GFF3 files with genome sequences and annotations of the following *Drosophila* species: *D. ananassae* (release 1.3), *D. erecta* (release 1.3), *D. grimshawi* (release 1.3), *D melanogaster* (release 6.03), *D. mojavensis* (release 1.3), *D. persimilis* (release 1.3), *D. pseudoobscura* (release 3.2), *D. sechellia* (release 1.3), *D. simulans* (release 1.4), *D. virilis* (release 1.2), *D. willistoni* (release 1.3), and *D. yakuba* (release 1.3). The corresponding genomic sequences, proteins and annotations of *A. aegypti* (AaegL3 assembly and AaegL3.3 annotations), *A. gambiae* (AgamP4 chromosome arm sequences, AgamP4.2 annotations), and *C. quinquefasciatus* (CpipJ2 scaffolds, CpipJ2.2 annotations) were obtained from VectorBase (Giraldo-Calderón et al. 2014), *Bombyx mori* (Silkworm_glean_pep annotation) from SilkDB v2.0 (Duan et al. 2010), and *Tribolium castaneum* (Glean.prot.51906 annotation, version 3.0) from BeetleBase (Kim et al. 2010). The annotated proteins of *Apis mellifera* (Amel_release1_OGS_pep.fa) were downloaded from BeeBase (Weinstock et al. 2006), of *Nasonia vitripennis* (Nvit_OGSv1.2_pep.fa) from NasoniaBase (Werren et al. 2010), of *Acyrthosiphon pisum* (ACYPI.proteins.v2.0.fa) from AphidBase (Legeai et al. 2010), and of *Pediculus humanus* (phumanus.PEPTIDES-PhumU2.1.fa; Kirkness et al. 2010) from VectorBase.

Gene Identification and Reannotation

We identified the complete set of *CheB* genes and pseudogenes in the 12 surveyed *Drosophila* species by performing several rounds of exhaustive searches. First, we used BLASTP (threshold *E-value* of 10^{-5}) to search against the annotated proteins of these species using the 12 *D. melanogaster* CheB proteins as queries. We then repeated the search using the newly identified CheBs as queries. Second, we performed a TBLASTN (threshold *E-value* of 10^{-5}) search against the genome sequences to identify putative nonannotated CheB genes and used Artemis r.13 (Rutherford et al. 2000) to annotate the newly identified genes. For that, we incorporate information from the GeneSplicer (Pertea et al. 2001) analysis and from the annotations available in FlyBase. In doubtful cases, we refined the annotation assisted with BLASTN, BLASTP, and MAFFT v.6.857 (Katoh et al. 2002) and consulted trace archives to detect putative sequencing errors. Third, we searched for remote homologs of this multigene family using HMMER v.3.0 (Durbin et al. 1998) (threshold *E-value* of 10^{-5}), and both the available PFAM-HMM profiles and CheB-specific HMM profiles built from our data (following Vieira and Rozas 2011) as queries. We used the same nomenclature criteria as in Vieira and Rozas (2011) to name all members of the *CheB* family identified in the surveyed insects.

We also created two HMM profiles, one using all CheB genes identified in *D. melanogaster* (13 genes) and *D. grimshawi* (5 genes) and the other including all CheA annotated in FlyBase for these two species (8 in *D. melanogaster* and 3 in *D. grimshawi*), which were used to identify all peptides of these two families annotated in the other nine insect species.

Protein Structure Predictions

We used SignalP 3.0 (Bendtsen et al. 2004) to predict signal peptide in the sequence of all 114 CheB proteins. We determined the secondary structure of CheB proteins with PROMALS3D (Pei et al. 2008), independently for each orthologous group. The obtained hallmarks were confirmed with PSI-PRED v.3.0 (Buchan et al. 2013), using at least one protein per orthologous group. In addition, we also investigated the presence of the DM11 domain in our proteins using InterProScan v.4.8 (Zdobnov and Apweiler 2001). Protein modelling and analysis of functional and mutational features was performed for one representative protein from each orthogroup (CheB15a, CheB38a, CheB42a, and CheB74a; see Results section), using the Phyre2 web portal (Kelley et al. 2015) with the intensive mode (i.e., the final model is a combination of template modelling and *ab initio* folding simulation). The predicted model and relevant amino acids from our functional divergence and positive selection analyses were viewed in Swiss-PdbViewer version 4.1 (Guex and Peitsch 1997).

Homologous Relationships and Phylogenetic Analysis

We used the program MAFFT to generate the multiple sequence alignments (MSA). All Maximum Likelihood (ML) phylogenetic trees were obtained with RAxML v.7.2.8 software (Stamatakis 2006) with the PROTGAMMAWAG substitution model and the amino acid-based MSAs. Node support values were estimated based on 500 ML bootstrap replicates in RAxML.

For the ML analysis including the 114 *Drosophila* CheBs *Drosophila* CheBs (total MSA), we excluded the putative signal peptide region. This tree was used to determine the precise orthologous/paralogous relationships among members of the CheB family by contrasting the gene tree with the species tree, as described in Almeida et al. (2014). This strategy was also followed to define the focal orthogroups for the analysis of gene turnover rates (see details in next section).

Estimation of Birth and Death Rates

We estimated the gene birth and death (BD) rates of the CheB family using the gene tree vs. species tree (GT/ST) reconciliation method (Goodman et al. 1979) as described in Almeida et al. (2014). Briefly, we used information from the phylogenetic analysis to determine the set of orthologous groups that descend from the same copy among the present in the ancestor of the *Drosophila* genus (focal orthogroups). For each of these orthogroups, we estimated the number of duplications and losses and the number of ancestral copies in internal nodes based on the GT/ST reconciliation approach. The total number of losses included the number of pseudogenes. With

this information we estimated the birth (β) and death (δ) rates by applying equations (1) and (2) in Almeida et al. (2014) and the formula in Vieira et al. (2007).

Analysis of Coding Sequence Evolution and Functional Divergence

We estimated the impact of natural selection on the CheB coding regions, using the *codeml* program (implemented in the package PAML; Yang 2007). For that, we estimated the average ω (i.e., the nonsynonymous (d_N) to synonymous substitution (d_S) rate ratio) and fitted site-specific (Yang 2000) and branch-specific models (Yang 1998) to each of the 13 orthologous groups identified in the phylogenetic analysis (see Results section). First, we compared some predefined nested models to study the distribution of selective constraints across amino acid sites: (1) a test of heterogeneity across sites, M0 vs. M3 (with $K = 3$ categories), (2) a direct test of the presence of positive selected amino acid sites, M1 vs. M2, and (3) a test of positive selection on some sites but fitting a beta distribution of ω values across sites, M7 vs. M8. We also used *codeml* to investigate lineage specific selective pressures, by comparing the fit of two different branch models to the data: the FR (free ratios) model, where each branch has a different ω, and the Mspec model, where specialist species are allowed to evolve under distinctive ω in comparison to nonspecialist species (see also Almeida et al. 2014). In order to detect CheB amino acid positions subject to episodic selection, we used the mixed effects model of evolution (MEME; Murrell et al. 2012), a method included in the HYPHY software (Pond et al. 2005) . We automatized many of these analyses by using custom-made in-house Perl scripts that, in some cases, used BioPerl modules (Stajich et al. 2002).

We analyzed protein family evolution and functional divergence using the ML framework implemented in the program DIVERGE v.3 (Gu et al. 2013). Using this program, we estimated the coefficients of type I (i.e., a measure of the levels of site-specific rate shift after gene duplication) and type II (i.e., an estimate of the amount of radical amino acid changes fixed between duplicates) functional divergence between specific CheB subfamilies in a phylogenetic framework. Then, we identified and mapped candidate functionally diverged sites onto the 3D protein models.

Results

Gene Identification and Reannotation

We first set out to identify all members of the CheB family present in the sequenced genomes of the 12 *Drosophila* species in Clark et al. (2007). A first BLASTP search using the 12 CheB proteins previously identified in *D. melanogaster* as queries identified 100 additional putative members of this family in the 12 surveyed *Drosophila* genomes, including one previously uncharacterized protein in *D. melanogaster* (CG13002). A second round of BLASTP searches using all 112 putative CheB proteins identified two additional proteins among the predicted peptides of *D. ananassae*. In addition, TBLASTN searches against the genomic sequences of the 12 species allowed identifying 5 putative *CheB* genes that had not been predicted in the used annotation releases (one in *D. ananassae*, two in *D. willistoni*, one in *D. mojavensis*, and one in *D. virilis*).

HMM-based searches using the CheB profile revealed some additional putative remote homologues (*E*-value = 10^{-5}). Nevertheless, none of these putative CheB shows the characteristic DM11 domain. Actually, most of them present the specific profile associated with CheA protein family (DUF1091), suggesting that CheA and CheB could be distantly related protein families. In any case, these additional proteins were not included in our analyses of the CheB family.

After performing an exhaustive process of manual reannotation, we corrected the CDS of 26 of the identified genes, mainly affecting the location of the splicing sites, some artificial gene fusions, and missing exons (see details in supplementary table S1A, Supplementary Material online). Our final data set comprises 114 putatively functional genes and 5 pseudogenes across the 12 surveyed *Drosophila* species, with a mean protein length of 199 amino acids (from 189 to 226 amino acids). Two of these 114 genes are likely partial sequences (gene fragments) because they are truncated at the beginning of a chromosome scaffold. We included these incomplete sequences in the phylogenetic analysis but not in functional divergence and selective constraints analyses.

Remarkably, we identified (BLASTP, *E*-value 8×10^{-8} to *DmelCheB42b*) a novel distant member of the CheB family in the genome of *D. melanogaster* (CG13002), which has orthologs in all other 11 surveyed species (1:1 orthologs). This gene is the only member of this family located in a sex chromosome (it is located on the X chromosome in all 12 species). Following the same rationale used in the nomenclature of the other *CheB* genes, we named this protein CheB15a because of its cytological location in *D. melanogaster*. We found that *Drosophila CheB* genes are organized in small chromosomal clusters in the 12 genomes, most of them with two to four members (supplementary fig. S1, Supplementary Material online).

CheB Protein Structure

We assessed whether the distinctive residues and characteristic secondary structure of CheB proteins are also present in the newly identified members. We found that the amino acid sequence of all 114 CheB proteins contains the four characteristic cysteines conserved across the family in the same relative position, with the single exception of *DsecCheB38a*, which is missing the third cysteine (top part of fig. 1*A*). Moreover, we also checked the presence of the two identified motifs highly conserved between the CheB and the human protein GM2-

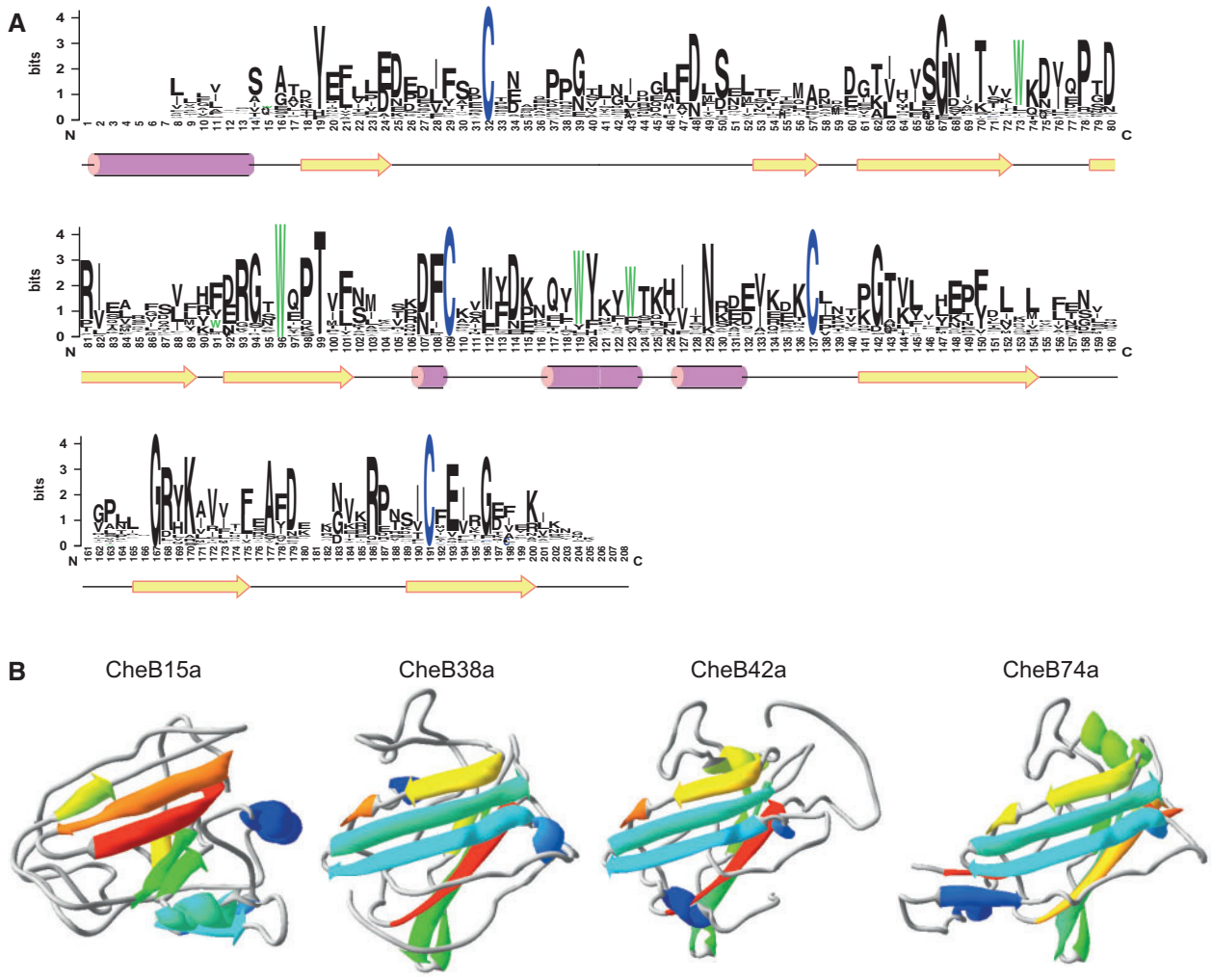

Fig. 1.—Structural analysis of the *Drosophila* CheB proteins. (*A*) Amino acid conservation through the sequence in the 114 CheB proteins (top), with cysteine residues (*C*) depicted in blue and tryptophan residues (W) in green. Below is shown the consensus of the predicted secondary structures for the 114 CheB proteins (bottom). Pink cylinders and yellow arrows indicate α helices and β sheets, respectively. (*B*) Three-dimensional (3D) model structure of the DmelCheB15a, DmelCheB38a, DmelCheB42a and DmelCheB74a proteins. In all cases, colours represent secondary structure succession (from dark blue to red).

AP in Starostina et al. (2009). We found that motif I ((KR)-X-X-X-G-X-W, 12 or 30 residues before the second cysteine) is present in all 114 CheB (positions 90 to 96), while motif II (G-X-(YWH)-(KR); 12 or 20 residues before the 4th cysteine) is also relatively well conserved (positions 167–170). However, while the three first conserved cysteine residues in CheB are also present in GM2-AP, the last one is specific to the *Drosophila* CheB proteins.

We found that secondary structure is also well conserved among *Drosophila* CheB proteins (bottom part of fig. 1*A*). SignalP predicted a signal peptide in 97 of the 114 CheB proteins, while PROMALS 3D predicted a signal peptide in all the proteins. After the signal peptide region all CheB proteins have four to five β-sheets before the second cysteine (the 1st cysteine is between the first and second β-sheets, in a coiled region), followed by a number of short α-helix flanked by the second and third cysteines (PROMALDS3D tends to predict two and PSI-PRED, three). Finally, after the third cysteine, there are three to four β-sheets, the last of them including the fourth cysteine. Noticeably, this conserved secondary structure across the genus remains very similar to that of GM2-AP, which similarly contains signal peptide in the N-terminal end, followed by five β-sheets (without unstructured region), one α-helix and three final β-sheets like CheB.

We applied a combination of multiple template homology-based structure prediction and folding simulation to obtain a 3D structure model of four representative CheB proteins (fig. 1*B*). In all proteins but CheB15a, the structure of

Ganglioside M2 activator protein (GM2-AP; PDB 2AG4) was selected as the unique template for homology modelling. Using this template, 75%, 70%, and 73% of the CheB38a, CheB42a, and CheB74a residues, respectively, were modelled with >90% of confidence. In the case of CheB15a, only 58% of residues were modelled using GM2-AP (residues 85–207), and a second template (an uncharacterized protein of an *Exiguobacterium*; PDB 2Q9K) was also used for modelling, resulting however in very low confidence values (<60%) for a substantial part of the model. On the other hand, all residues modelled by *ab initio* folding simulations (most corresponding to disordered regions, ~50 residues on average), have very low confidence values in the four final models. Overall, the four predicted models show that the β-sheets are grouped together forming two sheets that face each other in the interior of the protein structures surrounding the unstructured region and creating a characteristic globular structure. The putative α-helix present in the mature proteins is located facing the solvent, covering the upper area of the structure, while the position and length of the loop containing the signal peptide (or transmembrane) helix varies considerably among models. This configuration creates wide grooves of variable length inside the structures in all models, as well as several other small cavities (supplementary fig. S2, Supplementary Material online).

Model quality and functional prediction analyses, including pocket detection and mutational sensitivity results, evidence certain differences across 3D models (considering only the scores calculated for the regions with high confidence values), with the CheB15a model being the most different model by large. The fpocket2 program (Le Guilloux et al. 2009) in Phyre2 Inspector predicts a quite large binding pocket inside the cavities of CheB38a and CheB42a, composed by atoms of at least 48 and 32 amino acids, respectively, and smaller (or partial) pockets in CheB74a and CheB15a, with only 24–25 residues identified as part of the putative pockets (supplementary fig. S2, Supplementary Material online). In all cases, the amino acid positions predicted to be part of these binding pockets show significantly larger mutational effects than the rest of residues in the protein (supplementary fig. S3, Supplementary Material online).

Phylogenetic Analysis

To determine the evolutionary history of the CheB family, we first obtained the phylogenetic relationships among their members in the *Drosophila* genus (fig. 2). In this analysis we used the predicted mature proteins (i.e., we discarded the highly variable signal peptide region). Most of the orthologous that can be defined based on the *CheB* gene tree include members from species of the *melanogaster* group (*D. melanogaster*, *D. simulans*, *D. sechellia*, *D. yakuba*, and *D. erecta*) and only in some cases also include a *D. ananassae* (*CheB42c*) or *D. persimilis* and *D. pseudoobscura* (*CheB74a*, *CheB53a*, *CheB98a*, and *CheB42b*) copies (supplementary table S1B, Supplementary Material online). The only *CheB* with members in all 12 species is *CheB15a* (the newly characterized *CheB* gene). The tree also shows that the species-specific duplications of *D. willistoni* and *D. mojavensis* cluster in two clearly defined monophyletic clades. The results of the GENCONV analysis (Sawyer 1989) (supplementary table S2, Supplementary Material online) suggest that gene conversion might account for the inferred homologous relationships between the *D. mojavensis* paralogs *DmojCheB42aL3-5* and *DmojCheB42aL3-6*.

Finally, the phylogenetic analysis reveals four clear focal orthogroups (i.e., clusters of several ortholog groups descendant from a common ancestral copy), which would correspond to the four copies present in the ancestor of the 12 *Drosophila* species (fig. 2, outside circle).To study the origin of the *CheB* gene family, we expanded the phylogenetic analysis by including some non-*Drosophila*, insect sequences. In searches based on sequence similarity or HMM profiles, we identified eight putative *CheB* genes in *Aedes aegypti*, 17 in *Culex quinquefasciatus*, three in *Anopheles gambiae*, two in *Bombyx mori*, and four in *Tribolium castaneum* (supplementary table S3, Supplementary Material online); nevertheless, we fail to detect copies of this family in *Apis mellifera*, *Nasonia vitripennis*, *Acyrthosiphon pisum*, and *Pediculus humanus*. To assess the relationships among insect CheBs, we ran a phylogenetic analysis including all members of this family. Since our analyses suggest that *CheB* and *CheA* gene families could be phylogenetically related (see above), we also included the *D. melanogaster* members of the latter family in the analysis. Although bootstrap support is moderate for many nodes, the tree topology of the best ML tree indicates that some nondipteran copies previously identified in our searches by using CheB proteins as queries, may actually be members of the CheA family (brown clade in fig. 3). These linages are closely related to the *Drosophila* CheA copies and seven of them (out of nine) show the specific domain signature of CheA family (IPR010512). On the other hand, the tree also shows a monophyletic clade of 22 mosquito sequences (supported in 63% of bootstrap replicates) where three of them have the typical signature of CheB family (IPR006601), suggesting that this clade could represent other dipteran members of this family. Although the three remaining sequences (two from *T. castaneum* and one from *B. mori*; in blue in fig. 3) have the Ganglioside M2 (gm2) activator signature (Superfamily 2.70.220.10 in the CATH Protein Structure Classification Database; Sillitoe et al. 2015), we failed to detect a specific IPR domain signature, precluding their classification as members of either of these two families.

BD Rates

Given the small family size of the *Drosophila* CheB family, we estimated the gene turnover rates (gene gain and gene loss rates) under the GS/ST reconciliation framework (fig. 4).

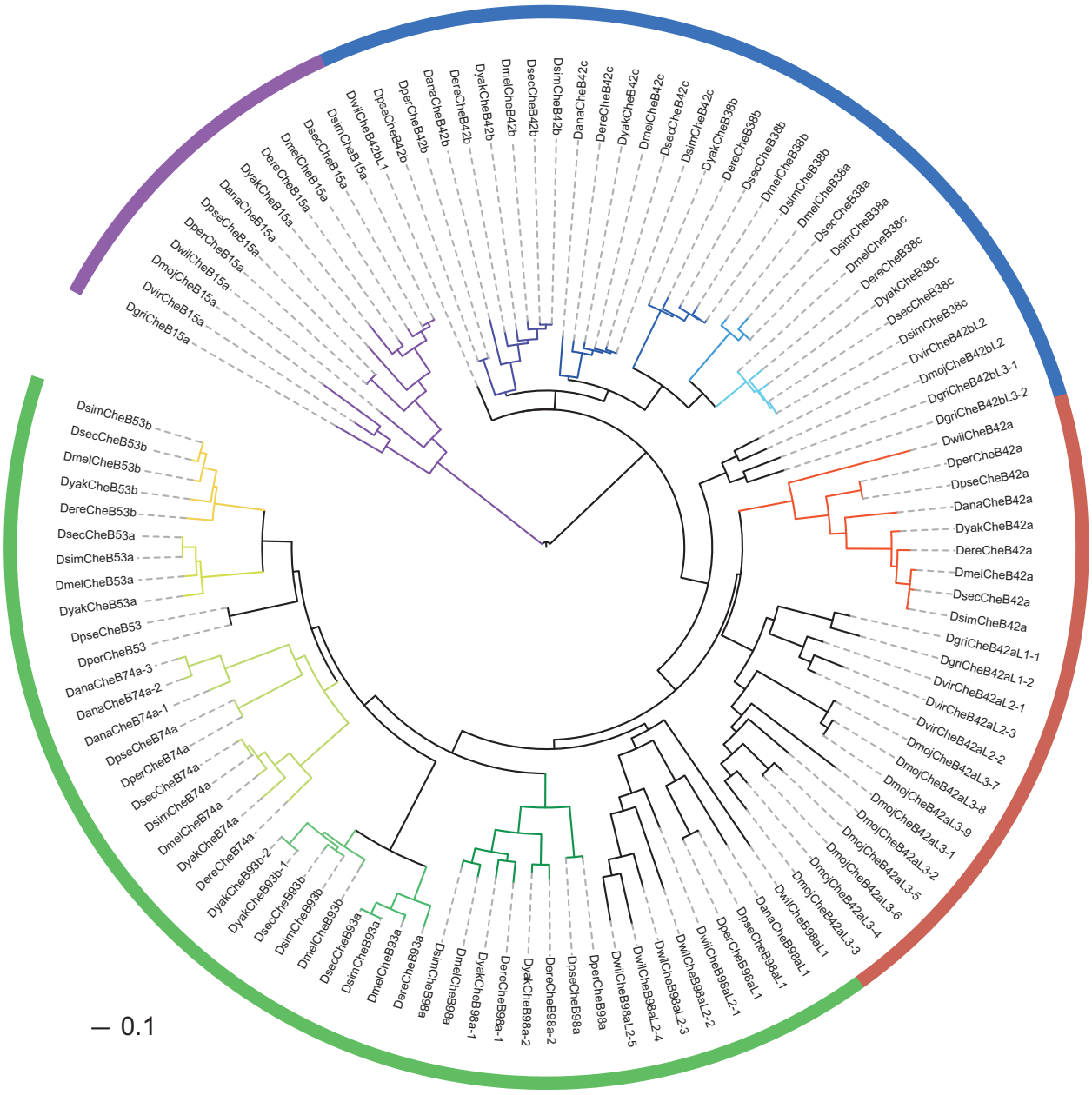

Fig. 2.—Maximum Likelihood phylogenetic tree of the *Drosophila* CheB proteins. Coloured branches indicate the clearly identifiable orthologous relationships with high support (basal nodes are supported by 88%-100% bootstrap replicates). Colours in the outside circle indicate the four focal orthogroups (i.e., groups of sequences that likely descend from the same ancestral copy) inferred from phylogenetic analyses and used to calculate gene turnover rates. Gene nomenclature follows the scheme proposed in Vieira et al. (2007).

We reconciled the gene trees for each of the four defined orthogroups taking into account the observed pseudogenes and also incorporating synteny data information. Using the equations 1 and 2 in Almeida et al. (2014), we estimated the gene birth β and δ death rates as $\beta = 0.024$ and $\delta = 0.008$ events per million year per ancestral gene content, respectively. These values are very close to those estimated using the (Vieira et al. 2007) formulas ($\beta = 0.021$ and $\delta = 0.009$), suggesting a relatively homogeneous distribution of turnover rates across lineages (Almeida et al. 2014). Given that Almeida et al. (2014) found that overall turnover rates (mostly death rates) of the other *Drosophila* chemosensory families are largely affected by the distinctive gene family evolution in the *D. sechellia* lineage, we re-estimated the CheB BD rates excluding *D. sechellia* data. The new estimated rates ($\beta = 0.026$ and $\delta = 0.005$) clearly

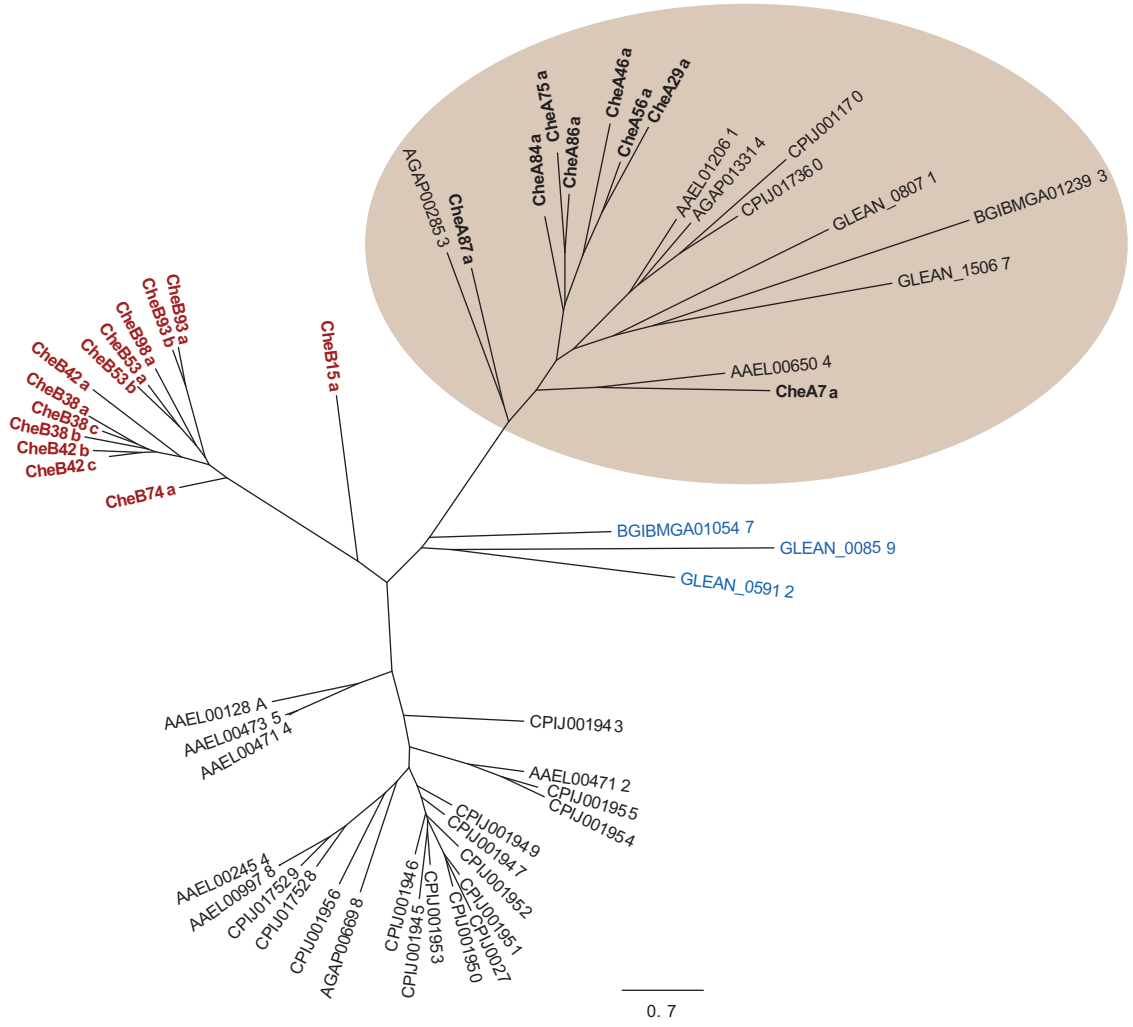

Fig. 3.—Maximum Likelihood phylogenetic tree of the insect CheA/B proteins. Red and black bold fonts highlight the CheB and CheA members of *D. melanogaster*, respectively. The putative insect CheA family members are shaded in brown. The three non-dipteran sequences missing the specific domain signature of any of these two families are in blue. Sequence names starting with AAEL, AGAP, BGIBM, CPIJ and GLEAN identifiers correspond to *Aedes aegypti*, *Anopheles gambiae*, *Bombyx mori*, *Culex quinquefasciatus* and *Tribolium castaneum* proteins, respectively.

confirm this lineage specific effect, especially on the death rate.

Impact of Natural Selection

We evaluated the impact of natural selection on the *CheB* multigene family applying codon substitution models (Bielawski and Yang 2003) to the 11 orthologous groups defined in the phylogenetic analysis (fig. 2). The average ratio between synonymous and non-synonymous substitution rates across the family is $\omega = 0.323$ (under model M0). Such value indicates that, although some codons may be evolving under neutral or even positive selection regimes, purifying selection is the main force driving protein sequence evolution in this family. In fact, this is a rather high value considering the current estimates for chemoreceptor families (ω values comprised between 0.05 and 0.22) (Shanbhag et al. 2001), or the average estimated for OBP genes ($\omega = 0.15$) (Vieira et al. 2007). *CheB74a* is the least constrained copy ($\omega = 0.558$) while *CheB42c* is the CheB member evolving under strongest selective constraints ($\omega = 0.175$). Nine out of the 13 likelihood ratio test (LRT) comparing the goodness of fit of M3 and M0 models gave significant results after controlling the false discovery rate (FDR) in these multiple comparisons, demonstrating that selective pressure is unevenly distributed across amino acid sites. Nevertheless, models including positively selected amino acid sites (M2 and M8) do not fit the data significantly better than models assuming neutral evolution (M1 and M7).

We also analyzed how selective constraints are distributed across lineages by comparing the fit of two models, a model

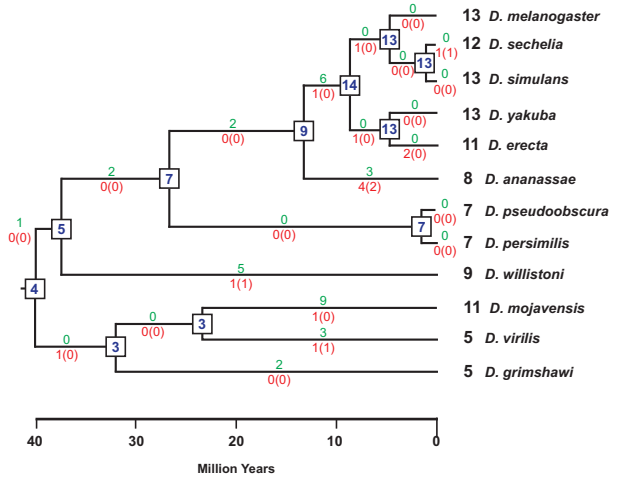

Fig. 4.—Birth and death evolution of CheB family in *Drosophila*. (*A*) Reconstruction of the gene turnover history of the CheB family during the diversification of the 12 *Drosophila* species. The estimated number of genes in each ancestral node, the number of gene gains and the number of gene losses are indicated in blue, green and red, respectively. In parenthesis is shown the number of pseudogenization events estimated in each lineage.

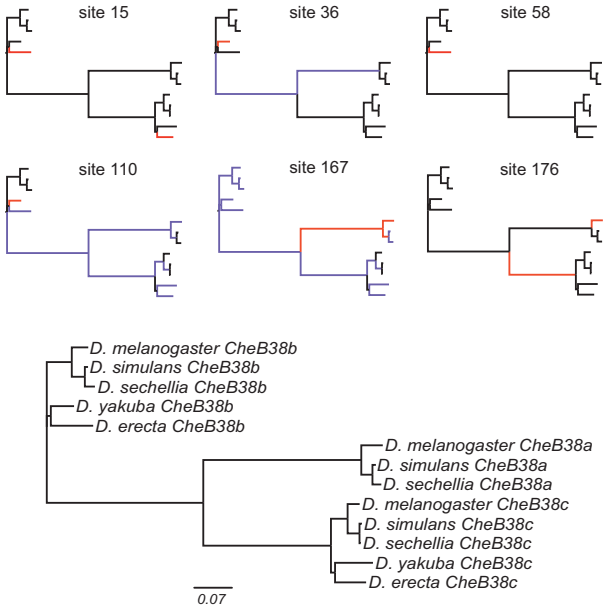

Fig. 5.—Example of the episodic evolution inferred to have occurred during the diversification of *CheB38* paralogs. (*A*) In red, black and blue, the lineages experiencing positive, neutral and negative selection shifts in each of the predicted sites. (*B*) Unrooted phylogenetic tree fitted to the codon alignment of *CheB38* paralogs.

where ω is allowed to freely vary across all branches of the tree (FR) and a model where all lineages are assumed to have the same ω (M0). This analysis shows that the FR model fits the data significantly better only for *CheB38c* and *CheB42a* genes after FDR correction (LRT, *P*-value = 0.005 and 0.004, respectively). Moreover, the results are not solely explained by significant changes in the functional constraints acting on these two copies in specialist lineages (the LRT comparing the fit of M0 and of a model where ω is allowed to vary only in specialist species, Mspe, is not significant; the *P*-value of this test was >0.05 in all orthologous groups); therefore, we should contemplate a more complex scenario to understand the evolution of CheB family coding regions during the diversification of these *Drosophila* species. To do this, we applied the powerful mixed effects model of evolution (MEME; Murrell et al. 2012), which allows the detection of the amino acid sites involved in episodic (lineage specific) diversifying selection.

We found 24 events of episodic positive selection in at least 20 different amino acid sites (site-by-site LRT, *P*-value < 0.05) and affecting a large number of lineages. Even considering only lineages with a high Bayes Factor (BF > 10), we found some sites that were involved in independent events in different lineages (supplementary table S4, Supplementary Material online). Remarkably, 16 of these selective events are recorded in lineages of the *melanogaster* subgroup, where they have been especially relevant during the diversification of *CheB74a* and *CheB38a*, *CheB38b*, and *CheB38c* copies (supplementary table S4, Supplementary Material online; fig. 5). All positive selection events predicted in *CheB38* paralogs (*CheB38* genes are only present in species of the *melanogaster* subgroup) and four out of the nine events found in *CheB74a* orthologs are recorded in lineages of this subgroup. Furthermore, some of these positively selected sites are also under strong purifying selection in other lineages, which suggests that they bear important functional roles.

Among the 20 sites with significant evidence of episodic selection, 14 could be mapped in some of the 3D models (supplementary fig. S4, Supplementary Material online). The selected positions are broadly distributed in different parts of the 3D structure, with only four of them (GLU32, VAL53, and SER75 in CheB38a and THR76 in CheB74a) predicted to be part of the binding pockets or corresponding to an amino acid immediately adjacent to one of them, suggesting that direct ligand-binding properties may not be a major target for adaptive changes. Most of the positively selected sites, however, have low or very low solvent accessibility (all except the two sites in CheB15a and one site in CheB74a) and five show moderate to large mutational effects as predicted in Phyre2 inspector analysis (TRP93 and ARG127 in CheB38a, VAL144 and GLU151 in CheB42a, and VAL73 in CheB74a), indicating that they are likely functionally important amino acids despite not being part of these predicted binding pockets.

Functional Divergence Analysis

To gain insight into the role of functional divergence in the evolution of CheBs during the diversification of the *Drosophila* genus, we applied the methods implemented in the software DIVERGE (Gu et al. 2013). The analysis was conducted with a

special focus on CheB15a, the novel divergent member identified in this study. This analysis requires information from at least two functionally homogeneous groups of proteins originated by a duplication event; for that we decided to use the inferred orthology as a proxy of common function. Nevertheless, given the highly dynamic nature of the CheB family (see above), very few genes can be recognized as clear orthologs in a reasonable number of species, which is crucial to perform a suitable DIVERGE analysis. For this reason, the analysis could be performed only with the CheB42a and CheB74a proteins which show orthologs in at least eight species of the *melanogaster* + *obscura* groups (fig. 2). The ML estimates of the coefficient of type I functional divergence (θ_I) between CheB15a and CheB74a ($\theta_I = 0.456 \pm 0.108$), and between CheB15a and CheB42a ($\theta_I = 0.760 \pm 0.120$) indicate the presence of amino acid positions that shifted their functional constraints after the duplication event from their ancestor (LRT, P-value $< 10^{-5}$ in both cases), and identify 16 of these sites. Interestingly, the comparison between CheB42a and CheB74a also detected statistically significant type I amino acid patterns ($\theta_I = 0.412 \pm 0.122$; LTR, P-value $= 8.6 \times 10^{-4}$).

Of the top 10 predicted sites under the posterior cut-off of 0.75 (FDR < 15%) of each comparison, four are shared in the two comparisons involving CheB15a (ILE97, ASN102, SER194, and ASP196 positions of the DmelCheB15a protein), and only one is common to the CheB15a/CheB74a and CheB42a/CheB74a comparisons (position ARG127 of the DmelCheB15a protein). Moreover, these results are consistent with the functional distance analysis (CheB15a is the duplicate with highest functional distance, $d_F = 1.095$) and with the asymmetric test for type I functional divergence, where the highest asymmetry delta variation is found between CheB15a and CheB42a proteins ($\Delta = 0.052$).

In contrast to the positively selected sites, the five relevant functionally divergent positions identified in DIVERGE map close together onto the four 3D models, on the area just above the structure formed by the faced β-sheets (fig. 6). Moreover, many of these sites are part of (or immediately by) the predicted binding pockets (ILE97, SER194, ASP197 in DmelCheB15a, ALA175 in DmelCheB42a and TRP77, ALA178 and ASP180 in DmelCheB74; the majority of which with low or very low predicted accessibility) or have moderate to large mutational effects predicted by Phyre2 inspector (the rest of positions except ASN102 in DmelCheB15a and ASP80 in DmelCheB42a), indicating a more important role of the amino acids located in this protein region in the early diversification of CheB proteins.

Discussion

Evolutionary Origin of the CheB Gene Family

To infer the evolutionary origin of a rapid evolving gene family, it is essential to include the members of other closely related families in the analysis. Here, we have performed comprehensive searches to identify putative remote homologs of the *CheA* and *CheB* genes in insect genomes, and present the phylogenetic analysis including all novel and previously identified members of these families. Nevertheless, three of the nondipteran sequences identified lack the specific domain signatures of these families (highlighted in blue in fig. 3), precluding the unambiguous rooting of the tree and, consequently, the determination of the precise timing of the CheA/CheB duplication. The incorporation of the human GM2-AP to the phylogenetic reconstruction does not provide information about the root location (results not shown), probably because the huge evolutionary distance between GM2-AP and the CheA/B members. Even so, given that some mosquito copies have the characteristic DM11 domain of CheBs, we can assert that the origin of the CheB gene family should be traced, at least, back to the emergence of dipterans (>250 Mya). Noticeably, it was also very difficult to infer with certainty the orthologous relationships among the putative CheB members in Culicidae species. All these observations are consistent with a high evolutionary rate of the CheB gene family, not only in the *Drosophila* genus but also in some of the other insects in which CHCs also function as sex attractants and/or in species recognition. The most plausible explanations of the apparent absence of CheB genes in other insect species that also use CHC as sex attractant and species recognition (as in Hymenopterans) are the high gene turnover rates of this family combined with its relative small repertory size, especially in ancestral nodes (fig. 4). These characteristics make this family prone to lineage extinction. Species lacking CheBs could have co-opted different binding proteins (as e.g., CSPs) to perform similar functions (as it has been found in some ant species; Ozaki et al. 2005).

The CheB Family is the Most Dynamic *Drosophila* Chemosensory Family

Almeida et al. (2014) demonstrated that the most accurate framework to estimate gene BD rates in moderate sized gene families is the gene tree/species tree reconciliation method. This approach allows taking advantage of pseudogenes and synteny information, while avoiding ML convergence problems associated to the limited amounts of data in full probabilistic methods. Moreover, by using this method our BD estimates can be directly compared with those obtained by Almeida et al. (2014) for the rest of chemosensory families. Despite the small size of CheB family, BD estimates are, on average, more than six (birth rates) and two (death rates) times higher than those obtained for the other chemosensory families. In fact, the CheB family turnover rates are higher than those estimated for the GRs and the divergent subgroup of IRs, the two chemosensory families with the highest gene turnover rates among the ones analyzed by Almeida et al. (2014). Consistent with the rapid gene birth-and-death

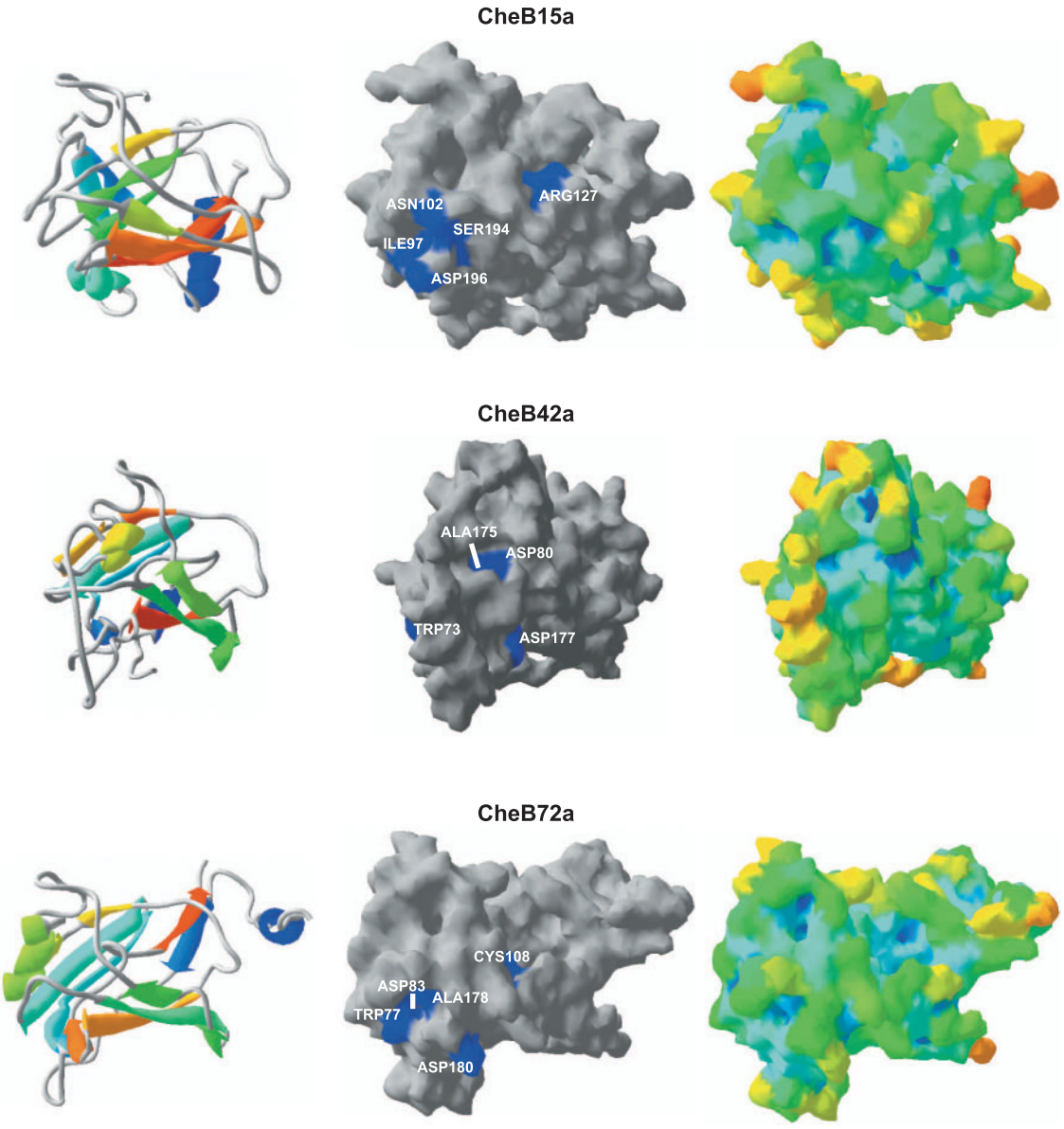

Fig. 6.—3D structure mapping of functionally diverged sites. (Left) Model orientation coloured as the secondary structure succession, from blue (C-terminal) to red (N-terminal). (Middle) Predicted protein surface from Swiss-pdb viewer with the location of diverged sites (in blue). (Right) Predicted protein surface coloured by solvent accessibility from Swiss-pdb viewer, from yellow (highly accessible) to blue (lower accessibility).

evolution, we detected a very small number of proper *CheB* orthologous groups and a large number of inparalogs, which generate large species-specific clades with long branches (e.g. *D. mojavensis* and *D. willistoni* duplications; fig. 2).

The results of our phylogenetic analysis suggest an evolutionary scenario where the ancestor of the 12 surveyed *Drosophila* species had (at least) four *CheB* copies. The highly dynamic evolution of the family, however, prevents the correct determination of true orthology/paralogy and the accurate estimation of the real ancestral family size. Remarkably, extant species have a relatively similar number of copies but only a few of them are in fact 1:1 orthologs. Actually, the high gene turnover rates have generated some lineage-specific expansions during the diversification of the genus, creating dissimilar repertories across species (i.e., same numbers but very different proteins). Given the anticipated importance of *CheB* genes in modulating *Drosophila* courtship behavior, natural selection likely played an important role in the evolution of family size. Nevertheless, the precise functional significance of the repertory differences between species remains unexplored. Additionally, our findings also corroborate the exceptional gene turnover rates

estimated for *D. sechellia* in other chemosensory families (Gardiner et al. 2008; Almeida et al. 2014), and point to ecological specialization as another important driver of gene turnover rates in the CheB family.

CheBs are the Fastest Evolving Proteins among *Drosophila* Chemosensory Families

Our functional constraints analyses also revealed high ω ratio values in the CheB family, even when comparing to the ω values estimated for other chemosensory-related families in *Drosophila*. These results, together with the estimated high turnover rates, suggest that the CheB family is, by large, the chemosensory gene family that evolved under the least selective constraints, followed by the other gustatory related families, the GRs and divergent IRs (Sánchez-Gracia et al. 2011). Although these results could be solely explained by differences in the strength of purifying selection acting on members of these chemosensory families, we found significant signs of pervasive episodic diversifying selection in the evolution of the *Drosophila* CheB family. The mapping of positive selected sites on 3D structures indicates that specific ligand-binding sites would not be major targets for adaptive changes in the members of this family (supplementary fig. S4, Supplementary Material online). These favored amino acid substitutions, however, could be responsible for other protein shape changes, as for example, the large differences observed across the predicted pocket sizes. These differences have been already observed among other lipid-binding proteins (Wright et al. 2005) and might represent differences in ligand-binding properties among CheB proteins. The small well-defined pockets of CheB15a and CheB74a would generate very specific binding sites, whereas the larger binding pockets predicted in CheB38a and CheB42a would be associated with lower affinities to specific compounds and thus to more promiscuous binding sites. On the other hand, we cannot rule out the possibility that some other features, such as protein stability or protein–protein interactions (mainly in CheB15a, the only case where the two only detected positively selected sites are highly exposed to the solvent), were the target of positive selection in some of these cases.

This scenario would be especially relevant in the diversification of *CheB74a* orthologs and in the divergence of *CheB38* paralogs after gene duplication. Interestingly, the genes of these two groups (except for *CheB38c*) are expressed only in male front legs, suggesting an active role of positive selection on female-specific pheromone detection by males. The importance of female sex pheromones in reproductive isolation between *Drosophila* species is well established in the *melanogaster* group. The highest variation in CHC profiles, mainly concerning sex dimorphism, occurs in the *melanogaster* subgroup (Bontonou and Wicker-Thomas 2014). Here, we found that 19 out of 24 events of episodic positive selection predicted in MEME analysis are recorded in lineages of the *melanogaster* group and that 16 of them involve specific *melanogaster* subgroup branches. As male CHC profiles are extremely similar across these species (the variation is principally found across females), positive selection on CheB proteins, especially those expressed only in male front legs, could be directly involved in preserving the sexual attraction of mates from the same group. A more exhaustive evaluation of the CHC profiles of the other *Drosophila* species are needed to confirm this hypothesis. This form of male vs. females coevolution can lead to reproductive isolation between populations by means of this divergent sexual selection and, ultimately, to selection-driven speciation.

Remarkably, the CheB family shows the same evolutionary pattern observed in other gustatory families, characterized by rapid evolutionary rates. Following the CheBs, the fastest evolving chemosensory families are those containing members related with gustatory perception, the GRs and divergent IRs (Sánchez-Gracia et al. 2011). The GR family, for example, has more members with strong indications of positive selection than ORs (one of the olfactory receptor families) in *Drosophila* (Gardiner et al. 2008). Yet, the CheB under the highest selective constraints (*CheB42c*) shows an estimated ω ($\omega = 0.175$) similar to the estimated average ω for the GRs and IRs ($\omega \sim 0.18$ and 0.21, respectively). The high evolutionary rates observed in the CheB family might be related to its role in CHC recognition. In fact, the CSP family shows patterns of rapid sequence evolution very similar to those of CheBs in some ant species, in which they bind the CHCs involved in nestmate recognition (Ozaki et al. 2005); interestingly, the CSP family is highly conserved in insects (Sánchez-Gracia et al. 2009; Vieira and Rozas, 2011), where they have been associated with chemoreception but often also involved in other nonsensory functions (Pelosi et al. 2014).

CheB15a, a Novel, Functionally Divergent Member of the *Drosophila* CheB Family

We have characterized a novel member of the CheB family in *Drosophila* (the orthogroup represented by the *D. melanogaster* CG13002 gene). Remarkably, *CheB15a* is the only member with 1:1 orthologs in the 12 surveyed *Drosophila* species, suggesting distinctive turnover dynamics and/or functional importance. Our functional divergence analysis demonstrates that the CheB15a protein has the largest functional distance from the other members of the family. In fact, all amino acid positions predicted to contribute significantly to the functional divergence between CheB15a and all other CheBs show the same rate shift pattern, i.e., they are highly variable between CheB15a copies but highly conserved across the other copies. These results suggest that the early changes in evolutionary rate posterior to the gene duplication event at the base of the *Drosophila* CheB clade were restricted to only one of the descendant copies: either a relaxation of functional constraints in CheB15a or an increase in the selective pressure

on the other CheBs. In this sense, most of the amino acids encoded by the significantly shifted positions shared between the two comparisons involving CheB15a had low or very low predicted solvent accessibility in the 3D structures (fig. 6) and map to the predicted pocket, suggesting that these functional constraint changes were probably associated with ligand binding properties. Finally, the 3D model of CheB15a shows fewer residues modelled when using the GM2-AP as a template and overall lower confidence values than the other predicted structures, envisaging some other additional differences in protein structure between this newly characterized member and the other CheBs.

The available *D. melanogaster* gene expression data (Robinson et al. 2013) also shows a distinctive expression pattern of the *CheB15a* gene. While all other CheB members are highly expressed in adult carcass (mainly in gustatory structures), *CheB15a* appears to be expressed exclusively in adult fat body (with low expression levels). It has been shown that fat body expresses some specific proteins that mediate courtship behavior and that some insect pheromones are synthetized from fatty acid precursors stored in this tissue (Arrese and Soulages 2010). We, therefore, hypothesize that CheB15a could be involved in pheromone precursor synthesis and/or storage, or participate in pheromone precursor transport between tissues. Further experiments would be necessary to test this interesting hypothesis that, if proven, would extend the functional roles known for the CheB family.

Conclusions

In recent years, numerous examples have been reported of the highly dynamic evolutionary nature of insect gustatory gene families. Gustatory families, systematically show higher BD rates and higher ω values (both among orthologs and paralogs) than the other chemosensory families, which are among the fastest evolving themselves if compared to averages across genomes (especially in recent duplicated copies; e.g., Sánchez-Gracia et al. 2011; Kulmuni et al. 2013; Almeida et al. 2014; Engsontia et al. 2014). The fact that most *Drosophila* CheB proteins may play a specialized role in gustatory detection of contact pheromones that modulate mating behavior makes the members of this family especially prone to sexual selection. This process may lead to functional divergence between orthologs and/or selective gene gains (through maintenance of duplicates) and/or selective gene losses of family members, affecting male vs. female coevolution and, potentially, speciation. We detected several events of episodic positive selection in the evolution of this family in *Drosophila*, especially in the *melanogaster* group, whose sexual dimorphism in CHCs is particularly pronounced. Moreover, we have identified a novel family member in *D. melanogaster*. As opposed to the pattern observed for positively selected sites, where the predicted pockets do not appear to be the direct targets of natural selection, important changes in ligand-binding properties may have driven the early functional diversification of the members of this family. Our findings on the CheB family confirm the evolutionary potential of the gustatory genes, placing this family as the most dynamic among all known chemosensory families in insects.

Supplementary Material

Supplementary figures S1-S4 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

This work was supported by the Ministerio de Economia y Competitividad of Spain (BFU2010-15484 and CGL2013-45211), and from the Comissió Interdepartamental de Recerca I Innovació Tecnològica of Catalonia, Spain (2009SGR-1287 and 2014SGR-1055). JR was partially supported by ICREA Academia (Generalitat de Catalunya), AS-G by a grant under the program Beatriu de Pinós (Generalitat de Catalunya, 2010-BP-B 00175) and FCA by the Juan de la Cierva fellowship (Ministerio de Economia y Competitividad of Spain).

Literature Cited

Almeida FC, Sánchez-Gracia A, Campos JL, Rozas J. 2014. Family size evolution in Drosophila chemosensory gene families: a comparative analysis with a critical appraisal of methods. Genome Biol Evol. 6:1669–1682.

Arrese EL, Soulages JL. 2010. Insect fat body: energy, metabolism, and regulation. Annu Rev Entomol. 55:207–225.

Begg M, Hogben L. 1946. Chemoreceptivity of *Drosophila melanogaster*. Proc R Soc London Ser B: Biol Sci. 133:1–19.

Bendtsen JD, Nielsen H, von Heijne G, Brunak S. 2004. Improved prediction of signal peptides: SignalP 3.0. J Mol Biol. 340:783–795.

Bielawski JP, Yang Z. 2003. Maximum likelihood methods for detecting adaptive evolution after gene duplication. J Struct Funct Genomics 3:201–212.

Bontonou G, Wicker-Thomas C. 2014. Sexual communication in the *Drosophila* genus. Insects 5:439–458.

Buchan DWA, Minneci F, Nugent TCO, Bryson K, Jones DT. 2013. Scalable web services for the PSIPRED Protein Analysis Workbench. Nucleic Acids Res. 41:W349–W357.

Carey AF, Carlson JR. 2011. Insect olfaction from model systems to disease control. Proc Natl Acad Sci U S A. 108:12987–12995.

Clark AG, et al. 2007. Evolution of genes and genomes on the Drosophila phylogeny. Nature 450:203–218.

Duan J, et al. 2010. SilkDB v2.0: a platform for silkworm (Bombyx mori) genome biology. Nucleic Acids Res. 38:D453–D456.

Durbin R, Eddy SR, Krogh A, Mitchison G. 1998. Biological sequence analysis: probabilistic models of proteins and nucleic acids. Cambridge (United Kingdom): Cambridge University Press.

Engsontia P, Sangket U, Chotigeat W, Satasook C. 2014. Molecular evolution of the odorant and gustatory receptor genes in lepidopteran insects: implications for their adaptation and speciation. J Mol Evol. 79:21–39.

Gardiner A, Barker D, Butlin RK, Jordan WC, Ritchie MG. 2008. Drosophila chemoreceptor gene evolution: selection, specialization and genome size. Mol Ecol. 17:1648–1657.

Giraldo-Calderón GI, et al. 2014. VectorBase: an updated bioinformatics resource for invertebrate vectors and other organisms related with human diseases. Nucleic Acids Res. 43((Database issue):D707–D713.
Goodman M, Czelusniak J, Moore GW, Romero-Herrera AE, Matsuda G. 1979. Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed from globin sequences. Syst Biol. 28:132–163.
Gu X, et al. 2013. An update of DIVERGE software for functional divergence analysis of protein family. Mol Biol Evol. 30:1713–1719.
Guex N, Peitsch MC. 1997. SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. Electrophoresis 18:2714–2723.
Le Guilloux V, Schmidtke P, Tuffery P. 2009. Fpocket: an open source platform for ligand pocket detection. BMC Bioinformatics 10:168.
Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 30:3059–3066.
Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJE. 2015. The Phyre2 web portal for protein modeling, prediction and analysis. Nat Protoc. 10:845–858.
Kim HS, et al. 2010. BeetleBase in 2010: revisions to provide comprehensive genomic information for *Tribolium castaneum*. Nucleic Acids Res. 38:D437–D442.
Kirkness EF, et al. 2010. Genome sequences of the human body louse and its primary endosymbiont provide insights into the permanent parasitic lifestyle. Proc Natl Acad Sci U S A. 107:12168–12173.
Kulmuni J, Wurm Y, Pamilo P. 2013. Comparative genomics of chemosensory protein genes reveals rapid evolution and positive selection in ant-specific duplicates. Heredity (Edinb) 110:538–547.
Legeai F, et al. 2010. AphidBase: a centralized bioinformatic resource for annotation of the pea aphid genome. Insect Mol Biol. 19:5–12.
Murrell B, et al. 2012. Detecting individual sites subject to episodic diversifying selection. PLoS Genet. 8:e1002764.
Ozaki M, et al. 2005. Ant nestmate and non-nestmate discrimination by a chemosensory sensillum. Science 309:311–314.
Park SK, et al. 2006. A Drosophila protein specific to pheromone-sensing gustatory hairs delays males' copulation attempts. Curr Biol. 16:1154–1159.
Pei J, Kim B-H, Grishin NV. 2008. PROMALS3D: a tool for multiple protein sequence and structure alignments. Nucleic Acids Res. 36:2295–2300.
Pelosi P, Iovinella I, Felicioli A, Dani FR. 2014. Soluble proteins of chemical communication: an overview across arthropods. Front Physiol. 5:320.
Pelosi P, Zhou J-J, Ban LP, Calvello M. 2006. Soluble proteins in insect chemical communication. Cell Mol Life Sci. 63:1658–1676.
Pertea M, Lin X, Salzberg SL. 2001. GeneSplicer: a new computational method for splice site prediction. Nucleic Acids Res. 29:1185–1190.
Pikielny CW. 2010. Drosophila CheB proteins involved in gustatory detection of pheromones are related to a human neurodegeneration factor. Vitam Horm. 83:273–287.
Pond SLK, Frost SDW, Muse SV. 2005. HyPhy: hypothesis testing using phylogenies. Bioinformatics 21:676–679.
Robertson HM. 1983. Chemical stimuli eliciting courtship by males in *Drosophila melanogaster*. Experientia 39:333–335.
Robinson SW, Herzyk P, Dow JAT, Leader DP. 2013. FlyAtlas: database of gene expression in the tissues of *Drosophila melanogaster*. Nucleic Acids Res. 41:D744–D750.
Rutherford K, et al. 2000. Artemis: sequence visualization and annotation. Bioinformatics 16:944–945.
Sánchez-Gracia A, Vieira FG, Almeida FC, Rozas J. 2011. Comparative genomics of the major chemosensory gene families in arthropods. Encycl. Life Sci. doi:10.1002/9780470015902.a0022848.
Sánchez-Gracia A, Vieira FG, Rozas J. 2009. Molecular evolution of the major chemosensory gene families in insects. Heredity (Edinb) 103:208–216.
Dos Santos G, et al. 2014. FlyBase: introduction of the *Drosophila melanogaster* Release 6 reference genome assembly and large-scale migration of genome annotations. Nucleic Acids Res. doi:10.1093/nar/gku1099.
Sawyer S. 1989. Statistical tests for detecting gene conversion. Mol Biol Evol. 6:526–538.
Shanbhag SR, Park SK, Pikielny CW, Steinbrecht RA. 2001. Gustatory organs of *Drosophila melanogaster*: fine structure and expression of the putative odorant-binding protein PBPRP2. Cell Tissue Res. 304:423–437.
Sillitoe I, et al. 2015. CATH: comprehensive structural and functional annotations for genome sequences. Nucleic Acids Res. 43:D376–D381.
Stajich JE, et al. 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.
Starostina E, Xu A, Lin H, Pikielny CW. 2009. A Drosophila protein family implicated in pheromone perception is related to Tay-Sachs GM2-activator protein. J Biol Chem. 284:585–594.
Steinbrecht RA. 1996. Structure and function of insect olfactory sensilla. Ciba Found Symp. 200:158–174. discussion 174–177.
Touhara K, Vosshall LB. 2009. Sensing odorants and pheromones with chemosensory receptors. Annu Rev Physiol. 71:307–332.
Vieira FG, Rozas J. 2011. Comparative genomics of the odorant-binding and chemosensory protein gene families across the Arthropoda: origin and evolutionary history of the chemosensory system. Genome Biol Evol. 3:476–490.
Vieira FG, 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 Biol. 8:R235.
Weinstock GM, et al. 2006. Insights into social insects from the genome of the honeybee *Apis mellifera*. Nature 443:931–949.
Werren JH, et al. 2010. Functional and evolutionary insights from the genomes of three parasitoid Nasonia species. Science 327:343–348.
Wright CS, Mi L-Z, Lee S, Rastinejad F. 2005. Crystal structure analysis of phosphatidylcholine-GM2-activator product complexes: evidence for hydrolase activity. Biochemistry 44:13510–13521.
Xu A, et al. 2002. Novel genes expressed in subsets of chemosensory sensilla on the front legs of male *Drosophila melanogaster*. Cell Tissue Res. 307:381–392.
Yang B. 2000. Statistical methods for detecting molecular adaptation. Trends Ecol Evol. 15:496–503.
Yang Z. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol Biol Evol. 15:568–573.
Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 24:1586–1591.
Zdobnov EM, Apweiler R. 2001. InterProScan – an integration platform for the signature-recognition methods in InterPro. Bioinformatics 17:847–848.

Associate editor: Richard Cordaux