The Insect Chemoreceptor Superfamily in Drosophila pseudoobscura: Molecular Evolution of Ecologically-Relevant Genes Over 25 Million Years

Author: Robertson, Hugh M.

Source: Journal of Insect Science, 9(18) : 1-14

Published By: Entomological Society of America

URL: https://doi.org/10.1673/031.009.1801

The insect chemoreceptor superfamily in *Drosophila pseudoobscura*: Molecular evolution of ecologically-relevant genes over 25 million years

Hugh M. Robertson

Department of Entomology, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

Abstract

The insect chemoreceptor superfamily, consisting of the odorant receptor (Or) and gustatory receptor (Gr) families, exhibits patterns of evolution ranging from highly conserved proteins to lineage-specific gene subfamily expansions when compared across insect suborders and orders. Here their evolution across the timespan of 25 million years is examined which yield orthologous divergences ranging from 5–50%. They also reveal the beginnings of lineage-specific gene subfamilies as multiple duplications of particular gene lineages in either or both *Drosophila melanogaster* and *D. pseudoobscura* (Frolova and Astaurov) (Diptera: Drosophilidae). Gene losses and pseudogenes are similarly evident in both lineages, and even in closer comparisons of *D. melanogaster* with *D. yakuba*, leaving these species with roughly similar numbers of chemoreceptors despite considerable gene turnover. The large range of divergences and gene duplications provide abundant raw material for studies of structure and function in this novel superfamily, which contains proteins that evolved to bind specific ligands that mediate much of the ecology and mating behavior of insects.

Keywords: odorant receptor, gustatory receptor, smell, taste, *Drosophila melanogaster*
Abbreviations: Or: odorant receptor, a recently expanded family within the insect chemoreceptor superfamily whose members to date have all been shown to function as olfactory receptors, except for Or83b which is co-expressed with other Ors and appears to have a "chaperone" function as a heterodimeric partner with all the specific Ors, **Gr**: gustatory receptor, members of the Gr family represent most of the protein diversity in the insect chemoreceptor superfamily, and include both functional gustatory receptors and olfactory receptors, e.g. Gr21a and Gr63a form a heterodimeric receptor for carbon dioxide
Correspondence: hughrobe@uiuc.edu
Received: 16 August 2007 | **Accepted:** 10 February 2008 | **Published:** 8 May 2009

ISSN: 1536-2442 | Vol. 9, Number 18

Cite this paper as:
Robertson HM. 2009. The insect chemoreceptor superfamily in *Drosophila pseudoobscura*: Molecular evolution of ecologically-relevant genes over 25 million years. 14pp. *Journal of Insect Science* 9:18, available online: insectscience.org/9.18

Introduction

The molecular basis of insect olfaction and gustation became amenable to study through discovery of two large families of genes that encode candidate chemoreceptor proteins in *Drosophila melanogaster* (Clyne et al. 1999; Vosshall et al. 1999; Clyne et al. 2000; Scott et. al. 2001; Dunipace et al. 2001; Robertson et al. 2003). These proteins have at least seven transmembrane domains and although once thought to be similar to the known G-protein-coupled receptor superfamilies, they share no sequence similarity with them (e.g. Hill et al. 2002) and appear to have the reverse membrane topology (Benton et al. 2006; Wistrand et al. 2006; Lundin et al. 2007). Genes encoding related receptors have been recognized in the genomes of other insects, including the mosquito *Anopheles gambiae* (Hill et al. 2002), the moths *Heliothis virescens* (Krieger et al. 2004) and *Bombyx mori* (Sakurai et al. 2004; Krieger et al. 2005; Wanner et al. 2007a), the red flour beetle *Tribolium castaneum* (Engsontia et al. 2008; Tribolium Genome Sequencing Consortium 2008), and the honey bee *Apis mellifera* (Robertson and Wanner 2006).

In *D. melanogaster*, the insect chemoreceptor gene superfamily consists of 60 odorant receptor (Or) and 60 gustatory receptor (Gr) genes encoding 62 and 68 different proteins through alternative splicing of long exons encoding the N-termini in some genes to one or more short exons encoding shared C-termini (Robertson et al. 2003). Examination of the molecular evolution of these genes indicates that the superfamily is old, at least as old as arthropods and perhaps older given the presence of five distantly related *gur* proteins encoded in the nematode *Caenorhabditis elegans* genome (unpublished results). There are many highly divergent gene lineages within the superfamily, with the Or family being a particularly highly expanded lineage, and these genes are now distributed throughout the *Drosophila* genome. In addition, signals of recent gene family evolution are apparent, including recent duplication of genes, either in tandem (e.g. Or22a/b) or near each other (Or19a/b), polymorphism of pseudogenes (Or85e, Gr22b, and Gr22d are pseudogenes with single obvious defects in the sequenced Canton-S-based genome, while they are intact in the Oregon-R genome), and apparent movement of genes from tandem duplicated series to elsewhere in the genome (e.g. Gr5a and Gr61a from the Gr64a-f cluster) (Robertson et al. 2003).

Comparison of this *D. melanogaster* chemoreceptor repertoire with that encoded by the *Anopheles gambiae* mosquito genome sequence revealed that on this long timescale of approximately 250 Myr, there are few simple orthologous relationships, mostly involving a few highly conserved genes such as DmOr83b, Gr21a, and Gr63a (Hill et al. 2002). Both families reveal several complicated potentially orthologous relationships of one:many, many:one, and many:many genes, while the majority of the evolution involves differential gene subfamily lineage expansions and losses in these two highly divergent subordinal fly lineages. The availability of a draft genome sequence for another congeneric drosophilid fly, *D. pseudoobscura* (Frolova and Astaurov) (Diptera: Drosophilidae) (Richards et al. 2005), provides an opportunity to examine the patterns and processes of molecular evolution of these ecologically-relevant, and therefore presumably fairly rapidly evolving, genes on the timescale of approximately 25 Myr, or 10 times shorter than the *Drosophila: Anopheles* (Cyclorrhapha: Nematocera) subordinal comparison. An additional ten *Drosophila* species genomes have now also been sequenced (*Drosophila* 12 Genomes Consortium 2007). Additional comparisons with one of them, *D. yakuba*, which was the first to become available and represents a lineage roughly 10 Myr old from *D. melanogaster*, reinforces conclusions about evolutionary relationships of these genes.

Materials and Methods

Gene models were built manually in the text editor of PAUP*v4.0b10 (Swofford 2002) using the output of TBLASTN searches of the 27 February 2003 *D. pseudoobscura* assembly available from the Baylor Human Genome Sequencing Center as guides. All DmOr and DmGr proteins described in Robertson et al. (2003) were employed as queries. These gene models were checked against the final draft assembly of 23 August 2003 (Freeze 2.0), as well as the unassembled reads in the Trace Archive at NCBI when necessary. Automated gene models were obtained from version 2.0 released from FlyBase in October 2005. The initial draft of the *D. yakuba* genome assembly was accessed from GenBank in August 2004, and gene models were updated using the DroYak2.1 assembly from the Washington University Genome Sequencing Center in August 2007.

Encoded proteins were aligned separately for each family using CLUSTALX (Jeanmougin et al. 1998) at default settings. For phylogenetic analysis the highly divergent N- and C-termini, that is, beyond the TM domains, and an internal segment corresponding to the long insertions in Or83b and Gr66a, respectively, were excluded. As a result, the low divergence of the most highly conserved proteins is slightly exaggerated in the trees, because these regions contain the few differing amino acids for some of them, e.g. Or83b and Gr21a. Corrected distances were calculated in TREE-PUZZLE v5.0 (Schmidt et al. 2002) using the BLOSUM62 amino acid matrix, and distance trees were estimated in PAUP*v4.0b10 using tree-bisection-and-reconnection branch swapping. Support for branches was obtained from 1000 bootstrap replications of uncorrected distance analysis using neighbor-joining.

Results

Gene model annotation difficulties

Draft genomes have sequence gaps between contigs that can disrupt gene models, and several such situations existed for the *D. pseudoobscura* assembly. Specifically, the DpOr genes were all intact, however an internal gap in DpGr2a and the C-termini of DpGr10b and 85a were obtained from unassembled reads that covered these gaps, although the DpGr10b model depends on a single low quality read. In the 2.0 assembly release, DpGr85a is now also separately assembled in the 4.5kb contig Unknown_group_751. Five DyOr and Gr genes were terminated by ends of contigs in the first genome release, however all are intact in the latest genome assembly.

For the Or and Gr families of roughly similar size in *D. pseudoobscura*, 18 and 8 proteins were precisely correctly annotated in FlyBase, respectively (Appendices 1 and 2), as part of the automated annotation effort (Richards et al. 2005). Existing annotations for 23 Ors and 24 Grs required some modification, commonly minor N- or C-terminal extensions to reach appropriate start and stop codons, but sometimes involving missed exons or unspliced open in-frame introns. In a few cases a single gene was annotated in a region that encodes multiple genes or transcripts, e.g. GA12528 represents the five transcripts that are hypothesized to be produced from the alternatively-spliced Gr28b locus. Eleven genes in each family were represented by GA placeholders in the *D. pseudoobscura* genome browser but no annotation was available for them, while 20 Ors and 16 Grs had no GA identifier associated with them. Some of the latter are pseudogenes and some were truncated by ends of contigs so they could not be annotated, and the remainder are instances of genes without *D. melanogaster* orthologs and hence might more easily have been missed by the orthology-based automated annotation process. The automated annotation rate was nevertheless rather low at just over 50%, perhaps because of the high divergences of some of these orthologous pairs. All of these gene models were communicated to FlyBase in 2005, and the encoded proteins are available in a supplementary file (chemoreceptor_proteins.txt).

Comparisons with the *D. pseudoobscura* and *D. yakuba* genes also allowed improvement of several *D. melanogaster* gene models, most of which were incorporated in FlyBase with publication of Robertson et al. (2003). Additional subsequent changes included recognition that the N-termini of DmGr21a and 63a are probably shorter than earlier predicted (see also Robertson and Kent 2009), while DmOr65b/c each likely have a short N-terminal extension to the existing annotation. These changes have been made in FlyBase. In addition, the version of DmOr98b in the sequenced genome is a pseudogene, see below.

The Or family

Comparisons of the *D. melanogaster* and *D. pseudoobscura* Or genes are shown in Appendix 1 and Figure 1. Identification of orthologous relationships was based on a combination of reciprocal best BLASTP matches, simple sister relationships in phylogenetic analysis (Figure 1), and examination of microsynteny with neighboring loci. In the Or family, there are 49 simple orthologs, with amino acid identities ranging from 94% for the highly conserved Or83b protein, which is known to be highly conserved throughout the endopterygote insects (Krieger et al. 2003), to around 50% for several relationships (e.g. Or19a/b, 23a, 65b/c, 67a, and 69aA/B). Most of the Ors are approximately twice as divergent as the average orthologous comparison for the entire proteome (Richards et al. 2005), confirming that these genes/proteins are among the more rapidly evolving portion of the genome.

Processes of gene family evolution that led to the major subfamily expansions and losses seen in the comparison with *A. gambiae* are evident on a smaller scale in the comparison with *D. pseudoobscura*. In addition to the recent duplication of the DmOr19a/b lineage, the comparison with *D. pseudoobscura* reveals that the DmOr22a/b duplication is younger than this species split and therefore specific to the *D. melanogaster* lineage. Remarkably, the orthologous locus in *D. pseudoobscura* has undergone two independent duplications leading to two apparently functional genes and a pseudogene. This comparison also reveals that DmOr65b/c were duplicated since the species split, while the orthologous locus in *D. pseudoobscura* has also seen repeated duplications yielding five apparently functional genes, and a N-terminal fragment not included in Figure 1 or Appendix 1. In addition there are two duplications and one triplication of other loci (Or42a, 49a, 98a) in the *D. pseudoobscura* lineage.

Balancing these gene duplications are several gene losses, including the orthologs of DmOr65a and 85a from *D. pseudoobscura*, while the ortholog of DmOr7a has been reduced to a fragmentary pseudogene. Reciprocally, *D. pseudoobscura* has three genes that appear to have been lost from *D. melanogaster* (DpOr33N, Or56N, and OrN). The first two of these are neighbors of genes with clear orthologs in *D. melanogaster*, and hence are given names reflecting these relationships, while the last one is a highly divergent lineage with no clear relationship to any DmOr. *D. melanogaster* also does not have an alternatively spliced exon in the Or69a locus that is a pseudogene in *D. pseudoobscura* (DpOr69aP). Examination of *D. yakuba* reveals that all three of these DpOr genes were lost before the *D. yakuba-D. melanogaster* split, but the DpOr69aP exon is intact in *D. yakuba*, so it was relatively recently lost from *D. melanogaster* and also became defective in *D. pseudoobscura*.

Comparison with *D. pseudoobscura*, as well as the *D. yakuba* and *D. simulans* draft genome assemblies, also led to

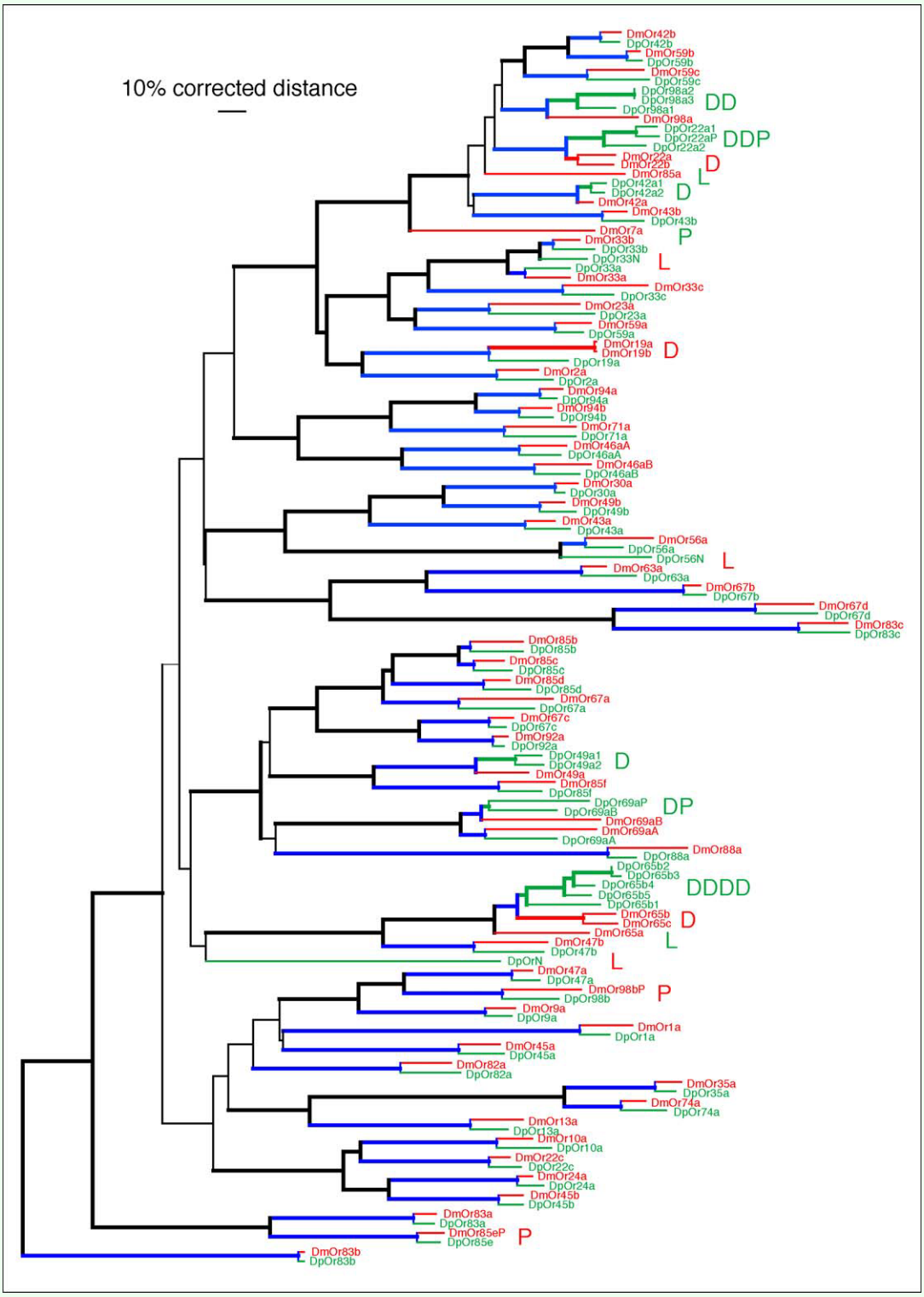

Figure 1. Phylogenetic relationships of the Or family in *Drosophila melanogaster* and *D. pseudoobscura*. This corrected distance tree was rooted by declaring the Or83b lineage as the outgroup, based on its position at the base of the Or family in analyses of the entire chemoreceptor superfamily (Robertson et al. 2003). DmOr proteins are shown in red and DpOr proteins in green. Proposed orthologous relationships, both simple and complicated, are indicated by blue branches connecting the Dm and Dp proteins. Letters on the right of the protein names highlight significant gene family evolutionary events: D – gene duplication; L – gene loss; P – pseudogene. DpOr7a is a highly degenerate pseudogene ortholog of DmOr7a and was not included in the tree analysis. Heavier lines indicate branches supported by at least 70% of 1000 bootstrap replications of uncorrected distance analysis. The terminal branches for Or33a were redrawn based on analyses of reduced datasets that included the *D. yakuba* proteins. Removal of extreme N- and C-termini plus an internal region of great length differences from the alignment for the phylogenetic analysis causes the most similar proteins to appear even more similar in this tree than they really are.

recognition that DmOr98b is a polymorphic pseudogenic allele. A single base deletion in the sequenced genome causes a frameshift near the end of the second exon. Amplification of this region from genomic DNA of pooled animals of the New Jersey and Ives strains (for unknown reasons it would not amplify from the Oregon-R strain used previously to examine polymorphic Or and Gr pseudogenes), revealed that the New Jersey strain is also fixed for this single base deletion, but that the Ives strain is polymorphic. Amplifications from single Ives strain animals revealed six homozygotes for the deletion, three homozygotes for an intact allele, and three heterozygotes. The intact allele encodes the 20aa sequence MLISYQRTGELQPKFPFPSV at the end of exon 2, with the deletion removing an adenine in the third codon position of the first glutamine. Examination of the original traces from the Celera Whole Genome Shotgun *D. melanogaster* genome project reveals that this gene was polymorphic even within this strain, with 2 of 17 reads crossing the region having the intact allele.

In addition to these polymorphic and fragmentary pseudogenes in each genome, there are fragmentary pseudogene copies of the Or98a locus in each genome, however they do not appear to represent orthologous duplicates as they are not microsyntenic with each other (these are not shown in Appendix 1 or Figure 1). Thus there has been additional gene degeneration in each fly lineage leaving only fragments that will presumably eventually be lost completely from these genomes. Remarkably, a potential ortholog of one of these fragments remains intact in *D. yakuba*.

The Gr family

The Gr family contains most of the protein diversity in the insect chemoreceptor superfamily (Robertson et al. 2003), from which the Ors are in reality a single highly expanded lineage. By most protein family criteria, the Gr family would be split into several families, including the highly divergent DmGr21a/63a lineage which form a heterodimeric olfactory receptor for carbon dioxide (Jones et al. 2007; Kwon et al. 2007) and a lineage of candidate sugar receptors related to the trehalose receptor DmGr5a (Chyb et al. 2003; Jiao et al. 2007; Slone et al. 2007; Dahanukar et al. 2007) (see top of Figure 2). The deeper divergences within the Gr family are also reflected in the lack of bootstrap support for most basal relationships (thin branches in Figure 2, versus Figure 1).

Analysis of the Gr family revealed patterns of evolution similar to the Ors (Appendix 2 and Figure 2). There are 54 pairs of apparently simple 1:1 orthologs. Their amino acid identities cover a similar range to the Ors, from 96% for the perfectly colinear Gr21a proteins to around 50% and multiple length differences for several proteins like Gr9a, Gr85a, and the Gr58 and Gr59 sets. The identification of Gr21a and 63a as a heterodimeric receptor for carbon dioxide (Jones et al. 2007; Kwon et al. 2007), which is presumably a very difficult molecule for proteins to interact with, might explain why so few amino acid substitutions are allowable that maintain function. Functional studies will be necessary to determine whether rapidly evolving receptors with around 50% amino acid identity like the Gr58 and Gr59 sets still detect the same ligands in both species - one can envisage that amino acid divergences in the range of 50% along with multiple length differences might indicate that the ligand specificity of these apparently orthologous but rapidly evolving receptors has changed.

There are several examples of gene birth in the Gr family in each species lineage, specifically DmGr22b/c, DmGr36a/b/c, and DmGr98c/d, while in *D. pseudoobscura* there are two duplications (Gr47b1/2 and Gr39a2/3) and one triplication (Gr59a1/2/3). Not all of these are simple examples however, primarily because some of these duplications are old and hence it is not clear whether in fact an ortholog has been lost from the other species instead. In the DmGr22a-f region, the DmGr22f gene I treated as having been lost from *D. pseudoobscura*, but some phylogenetic analyses suggest it is duplicated in the *D. melanogaster* lineage. The DpGr59a1-3 triplication is also complicated, with microsynteny analysis suggesting that this locus has undergone multiple duplications and gene losses in each lineage so that the simple "orthologous" comparisons to DmGr59a used in Appendix 2 might not be appropriate.

These Gr gene duplications are again balanced by gene losses, including the orthologs of DmGr5a, 22f, 92a, 93d, and 98c/d from *D. pseudoobscura*, while the orthologs of DpGr39a1 and a5, and of DpGr93N have been lost from *D. melanogaster*. In the case of DmGr5a, which has been shown to be a receptor for trehalose (e.g. Chyb et al. 2003), this means that *D. pseudoobscura* has either lost the ability to detect this sugar, or that this ability has been replaced with function of another receptor (see also Dahanukar et al. 2007). Comparisons with *D. yakuba* helped resolve several phylogenetic relationships, but also demonstrated that even more gene loss has occurred in this family, because it has four Grs which have been lost at least from *D. melanogaster* and sometimes also *D. pseudoobscura* (three of these are in the already complicated Gr59, 93, and 98 lineages, and the fourth is related to Gr85a).

While there are two Gr pseudogenes in the sequenced *D. melanogaster* strain (Gr22b and d), both of which are polymorphic in the species (Robertson et al. 2003), there are four pseudogenes in the sequenced *D. pseudoobscura* strain (Gr22d, 47a, 64e, and 98a). Remarkably Gr22d has independently become a pseudogene in each species. DpGr47a is a fragmentary pseudogene, but the lesions in the other three genes are single defects, so like Gr22b/d in *D. melanogaster*, these three pseudogenes in *D. pseudoobscura* might be polymorphic in the species. Indeed DpGr98a might not even be a pseudogene if the real

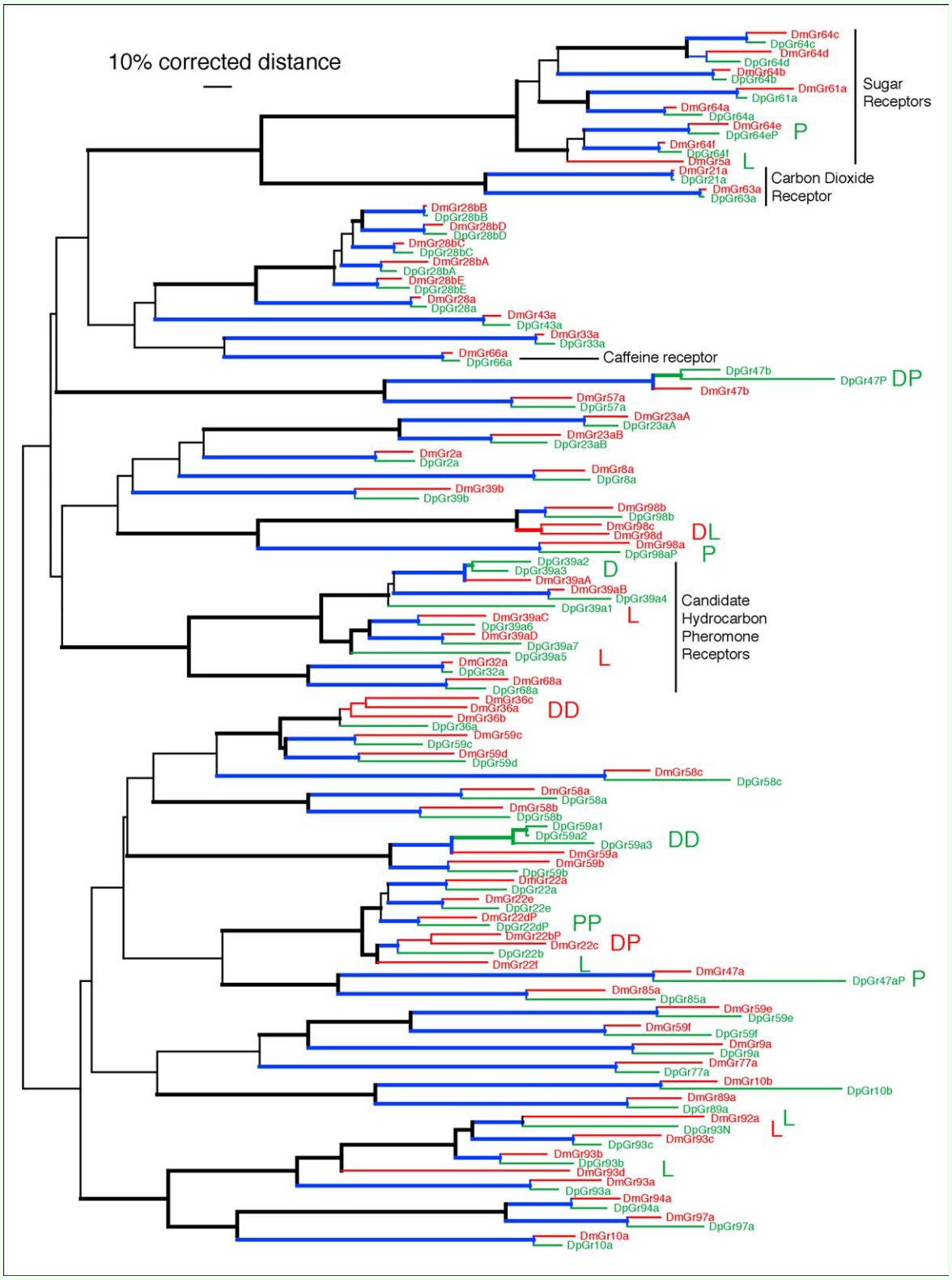

Figure 2. Phylogenetic relationships of the Gr family in *Drosophila melanogaster* and *D. pseudoobscura*. This corrected distance tree was rooted at the midpoint in the absence of a simple outgroup. Figure conventions are as in Figure 1 legend. Terminal branches were redrawn for Gr22b/c/f, 36a/b/c, 47b, and 64c/d. DpGr47P, 47aP, and 10b have extra long branches because in the first two instances parts of these proteins are missing, and in the latter, parts were reconstructed from a single read of poor quality.

start codon is considerably internal to that alignable with DmGr98a. There is also a recently duplicated pseudogene copy of DpGr47b that is truncated for the C-terminus.

Gene movement

Richards et al. (2005) noted that while there were a large number of chromosomal rearrangements breaking up syntenic blocks of genes between these two species, the vast majority were intra-chromosome arm events, with few examples of translocations or transpositions between chromosome arms. This is also clearly evident from the locations of the *D. pseudoobscura* Or and Gr genes in Appendices 1 and 2. For example, in the entire Gr family every gene is still on the equivalent chromosome arm or Muller element (see Richards et al. 2005 for details of Muller elements), albeit usually changed in location along the arm. This means that the names of the genes in *D. melanogaster*, which indicate their chromosomal location, have little meaning in *D. pseudoobscura*, but I have chosen to use the same names for the orthologous *D. pseudoobscura* genes because most interest in these *D. pseudoobscura* genes derives from their comparison with the *D. melanogaster* genes and proteins.

The Or family has several examples of inter-chromosomal movement, however. Thus DpOr13a is on chromosome 4 instead of its expected location on XL; this appears to have been a retrotransposition mediated by reverse transcription of a mRNA, because DpOr13a is intronless and positioned between two genes whose orthologs in *D. melanogaster* are in 35A (left end of synteny block 626 in the Dp genome browser at FlyBase). Other examples of Or gene movement between chromosome arms are Or67a and Or92a, and the Or65b and Or98a expansions in *D. pseudoobscura*. None of these appear to involve retrotransposition because both *D. melanogaster* and *D. pseudoobscura* genes share introns. The direction of transposition cannot be determined by microsynteny analysis alone for Or67a and Or92a. For Or67a this single gene in synteny block 190 moved; for Or92a the *D. pseudoobscura* gene is 30kb in from the end of 12.5Mbp scaffold XL_1e with no neighboring orthologs so nothing can be said about it. In the case of the DpOr65b and DpOr98a expansions in *D. pseudoobscura*, it appears that these transpositions occurred in the *D. pseudoobscura* lineage, perhaps concomitantly with the duplications of these genes, because in each case there is one copy of the DpOr that is microsyntenic with the single DmOr ortholog.

Discussion

The overall pattern of evolution of these chemoreceptor genes in the recent *Drosophila* lineage appears to be a balance of birth/duplication and death/loss of genes. Thus 25 have become pseudogenes or were lost in the lineages leading to *D. melanogaster* and *D. pseudoobscura* (Table 1), although some pseudogenes are polymorphic, while 22 have been born through gene duplication, leaving each species with roughly the same number of encoded proteins. This stability of total functional Or number despite considerable gene turnover was recently reported by Nozawa and Nei (2007) and Guo and Kim (2007) for comparisons across most of the 12 available *Drosophila* species, although the Hawaiian *D. grimshawi* appears to have a relative expansion of this family. Remarkably, however, both the Ors and Grs show great acceleration of pseudogenization in the specialist species *D. sechellia*, which is a sibling of the generalist *D. simulans*, revealing how an ecological revolution can rapidly change the evolutionary dynamics of these chemoreceptors (Dekker et al. 2006; McBride 2007).

Table 1. Gene duplications, pseudogenizations, and losses in Drosophila *melanogaster* and *D. pseudoobscura*.

	Duplications	Pseudogenes	Losses
Or genes			
melanogaster	3	2	4
pseudoobscura	11	2	2
Total	14	4	6
Gr genes			
melanogaster	4	2	3
pseudoobscura	4	5	5
Total	8	7	8
Superfamily			
melanogaster	7	4	7
pseudoobscura	15	7	7
Total	**22**	**11**	**14**

Alternatively-spliced isoforms are treated as separate genes. Gene fragments are treated as losses.

Comparisons of genes and their encoded proteins between *D. melanogaster* and *D. pseudoobscura* proved to be too distant to provide much useful information on the patterns of selection pressures (Richards et al. 2005). Simple analyses of Ka/Ks ratios of non-synonymous to synonymous changes reveal that all of these chemoreceptor genes are under strong purifying selection pressure across this 10–25 Myr timescale when considered across their entire lengths (results not shown). This is in agreement with the results of Tunstall et al. (2007) for 10 Ors and one Gr considered from 8–12 species between *D. melanogaster* and *D. pseudoobscura*. Their generation of additional sequences for each chemoreceptor lineage allowed them to apply more sophisticated tree-based maximum likelihood methods that yielded suggestions of positive selection on three genes in particular lineages, or on 12 codons within four genes. Guo and Kim (2007) extended these kinds of analyses to the entire Or family across all

12 species with genome sequences, however they found signals of positive selection in a different set of genes than Tunstall et al. (2007). More detailed analyses of the entire superfamily, but limited to the *D. melanogaster* species group, provide additional insights into patterns of selection on these genes (McBride et al. 2007).

This comparison across the past 25 Myr provides a window into the processes by which these two large, ecologically-relevant gene families have evolved over much longer timescales, leading to the largely species-specific gene lineages seen in comparisons with mosquitoes (Hill et al. 2002), moths (Wanner et al. 2007a), beetles (Engsontia et al. 2008; Tribolium Genome Sequencing Consortium 2008), and bees (Robertson and Wanner 2006). The lineages that are still orthologous in these distant comparisons, e.g. the DmOr83b "chaperone" (Krieger et al. 2003; Larsson et al. 2004; Neuhaus et al. 2005) and the DmGr21a/63a carbon dioxide heterodimer (Jones et al. 2007; Kwon et al. 2007; Lu et al. 2007; Robertson and Kent 2009) are the most highly conserved lineages within these *Drosophila* species. Another set of relatively well-conserved proteins are the candidate sugar receptors related to the trehalose receptor DmGr5a (Chyb et al. 2003; Jiao et al. 2007; Slone et al. 2007; Dahanukar et al. 2007), although remarkably that particular gene was lost from the *D. pseudoobscura* lineage. The only other lineages showing simple orthology out to the honey bee are the DmGr43a lineage of unknown ligand specificity and the alternatively spliced DmGr28a/b loci (Robertson and Wanner 2006), both of which are highly conserved between *D. melanogaster* and *D. pseudoobscura*. The DmGr28a/b proteins have recently been shown to be expressed in both chemosensory neurons and other neurons not obviously involved in gustation (Thorne and Amrein 2008). In stark contrast, several lineages in each gene family and sometimes each species, e.g. Or22, 65, 98a and Gr22, 36, 39, 59a, 98a, show lineage-specific expansions. Combined with the relatively high rate of gene loss, it is not hard to see how most of these insect chemoreceptors come to be quite different in more distant comparisons across orders of insects.

Examination of several other receptors of particular interest reveals various levels of conservation. DmOr67d, a gene model contributed by Robertson et al. (2003), has been shown to be a receptor for cis-vaccenyl acetate, a volatile chemical that serves as an aggregation, male sex, and mating deterrent pheromone (Ha and Smith 2006; van der Goes van Naters and Carlson 2007; Kurtovic et al. 2007; Ejima et al. 2007). Or67d is not particularly highly conserved, at 64% identity, so it will be of interest to determine whether the DpOr67d receptor also recognizes this ligand. Two other chemoreceptors of interest in the context of mating behavior are the sister DmGr68a and 32a proteins implicated in recognition of female-specific cuticular hydrocarbons by males during tapping and licking of the female, although their specific ligands have not been identified (Bray and Amrein 2003; Ebbs and Amrein 2007). Gr68a is not particularly highly conserved at 67% amino acid identity, however Gr32a is amongst the most well conserved at 85% identity. Perhaps even more interesting is the implication that the related Gr39a set of protein isoforms (Figures 2) might also be involved in pheromone perception (Ebbs and Amrein 2007). In addition to high divergence between the four orthologous proteins in this set (62–75% identity - Appendix 2 and Figure 2), this protein set has undergone considerable lineage-specific evolution, with two losses in the common ancestor of *D. melanogaster* and *D. yakuba*, and a duplication in *D. pseudoobscura* (Figure 3).

Many of the remaining Grs are likely to be bitter taste receptors and DmGr66a has been shown to detect caffeine (Thorne et al. 2004; Wang et al. 2004; Marella et al. 2006; Ebbs and Amrein 2007). Again this is a relatively highly conserved protein with 84% identity so is likely to have the same ligand specificity in other *Drosophila* species. Most DmOr ligand specificities have now been established (e.g. Hallem and Carlson 2006), however there are few obvious relationships between the level of conservation of the receptors and the characteristics of their ligands. A possible pattern beyond their ligands is that receptors involved in formation of heterodimers, that is Or83b, Gr21a, Gr63a, and Gr66a, show higher conservation than most others, presumably reflecting in part the

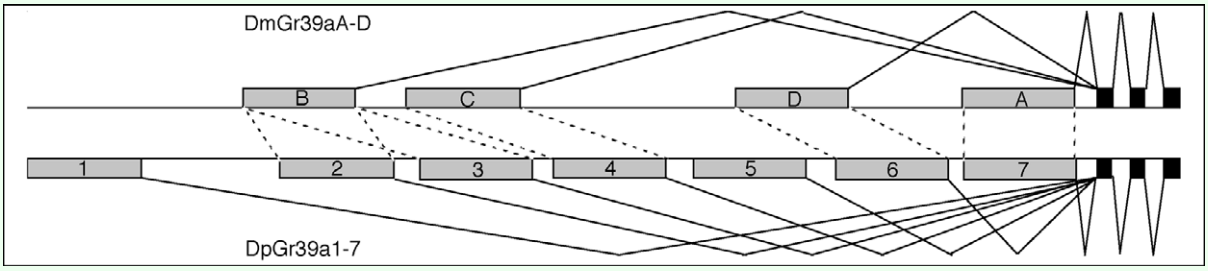

Figure 3. Schematic diagram of the Gr39a loci that are hypothesized to be alternatively-spliced. Grey boxes indicate the long alternatively-spliced N-terminal exons, while the three black boxes indicate the shared C-terminal exons. Two upstream exons present in *pseudoobscura* (1 and 5) were lost from *Drosophila melanogaster* (and *D. yakuba*), while one was duplicated in *D. pseudoobscura* (2/3).

need to maintain many residues to sustain dimerization with other receptors. In addition to misexpression of olfactory receptors in particular empty olfactory sensory neurons followed by single-sensillum recordings (e.g. Dobritsa et al. 2003; Hallem and Carlson 2006; Jones et al. 2007; Kwon et al. 2007), several different methods for studying ligand specificity in heterologous expression systems are now available (e.g. Chyb et al. 2003; Neuhaus et al. 2005; Kiely et al. 2007; Wanner et al. 2007b). The varying levels of amino acid divergence in related *Drosophila* species might provide a useful resource for studies of structure and function in this novel superfamily of proteins using both endogenous and heterologous expression systems.

Acknowledgments

I thank Kim Walden and Jaclyn Wegner for sequencing the New Jersey and Ives strain DmOr98a loci, Roman Arguello for discussions about the *pseudoobscura* and *D. yakuba* gene models, and the Washington University Genome Sequencing Center for making versions of the *D. yakuba* genome assembly public prior to publication. This work was supported by NIH grant RO1AI056081.

References

Bray S, Amrein H. 2003. A putative Drosophila pheromone receptor expressed in male-specific taste neurons is required for efficient courtship. *Neuron* 39(6): 1019-1029.

Benton R, Sachse S, Michnick SW, Vosshall LB. 2006. Atypical membrane topology and heteromeric function of *Drosophila* odorant receptors in vivo. *PLoS Biology* 4(2): e20

Chyb S, Dahanukar A, Wickens A, Carlson JR. 2003. Drosophila Gr5a encodes a taste receptor tuned to trehalose. *Proceedings of the National Academy of Sciences USA* 100(Suppl 2): 14526-14530.

Clyne PJ, Warr CG, Freeman MR, Lessing D, Kim J, Carlson JR. 1999. A novel family of divergent seven-transmembrane proteins: candidate odorant receptors in *Drosophila*. *Neuron* 22(2): 327-338.

Clyne PJ, Warr CG, Carlson JR. 2000. Candidate taste receptors in *Drosophila*. *Science* 287(5459): 1830-1834.

Dahanukar A, Lei YT, Kwon JY, Carlson JR. 2007. Two Gr genes underlie sugar reception in *Drosophila*. *Neuron* 56(3): 503-516.

Dekker T, Ibba I, Siju KP, Stensmyr MC, Hansson BS. 2006. Olfactory shifts parallel superspecialism for toxic fruit in *Drosophila melanogaster* sibling, *D. sechellia*. *Current Biology* 16(1): 101-109.

Dobritsa AA, van der Goes van Naters W, Warr CG, Steinbrecht RA, Carlson JR. 2003. Integrating the molecular and cellular basis of odor coding in the *Drosophila* antenna. *Neuron* 37(5): 827-841.

Drosophila 12 Genomes Consortium 2007. Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature* 450(7167): 203-218.

Dunipace L, Meister S, McNealy C, Amrein H. 2001. Spatially restricted expression of candidate taste receptors in the *Drosophila* gustatory system. *Current Biology* 11(11): 822-835.

Ebbs ML, Amrein H. 2007. Taste and pheromone perception in the fruit fly *Drosophila melanogaster*. *Pflugers Archives* 454(5): 735-747.

Ejima A, Smith BP, Lucas C, van der Goes van Naters W, Miller CJ, Carlson JR, Levine JD, Griffith LC. 2007. Generalization of courtship learning in *Drosophila* is mediated by cis-vaccenyl acetate. *Current Biology* 17(7): 599-605.

Engsontia P, Sanderson A, Cobb M, Walden KKO, Robertson HM, Brown S. 2008. The red flour beetle's large nose: an expanded odorant receptor gene family in *Tribolium castaneum*. *Insect Biochemistry and Molecular Biology* 38: 387-397.

Guo S, Kim J. 2007. Molecular evolution of *Drosophila* odorant receptor genes. *Molecular Biology and Evolution* 24(5): 1198-1207.

Ha TS, Smith DP. 2006. A pheromone receptor mediates 11-cis-vaccenyl acetate-induced responses in *Drosophila*. *Journal of Neuroscience* 26(34): 8727-8733.

Hallem EA, Carlson JR. 2006. Coding of odors by a receptor repertoire. *Cell* 125(1): 143-160.

Hill CA, Fox AN, Pitts RJ, Kent LB, Tan PL, Chrystal MA, Cravchik A, Collins FH, Robertson HM, Zwiebel LJ. 2002. G protein-coupled receptors in *Anopheles gambiae*. *Science* 298(5591): 176-178.

Jeanmougin F, Thompson JD, Gouy M, Higgins DG, Gibson TJ. 1998. Multiple sequence alignment with Clustal X. *Trends in Biochemical Sciences* 23(10): 403-405.

Jones WD, Cayirlioglu P, Kadow IG, Vosshall LB. 2007. Two chemosensory receptors together mediate carbon dioxide detection in *Drosophila*. *Nature* 445(7123): 86-90.

Kiely A, Authier A, Kralicek AV, Warr CG, Newcomb RD. 2007. Functional analysis of a *Drosophila melanogaster* olfactory receptor expressed in Sf9 cells. *Journal of Neuroscience Methods* 159(2): 189-194.

Krieger J, Grosse-Wilde E, Gohl T, Dewer YM, Raming K, Breer H. 2004. Genes encoding candidate pheromone receptors in a moth (*Heliothis virescens*). *Proceedings of the National Academy of Sciences USA* 101(32): 11845-11850.

Krieger J, Grosse-Wilde E, Gohl T, Breer H. 2005. Candidate pheromone receptors of the silkmoth *Bombyx mori*. *European Journal of Neuroscience* 21(8): 2167-2176.

Krieger J, Klink O, Mohl C, Raming K, Breer H. 2003. A candidate olfactory receptor subtype highly conserved across different insect orders. *Journal of Comparative Physiology A Neuroethology and Sensory Neural Behavioral Physiology* 189(7): 519-526.

Kwon JY, Dahanukar A, Weiss LA, Carlson JR. 2007. The molecular basis of CO_2 reception in *Drosophila*. *Proceedings of the National Academy of Sciences USA* 104(9): 3574-3578.

Larsson MC, Domingos AI, Jones WD, Chiappe ME, Amrein H, Vosshall LB. 2004. Or83b encodes a broadly expressed odorant receptor essential for *Drosophila* olfaction. *Neuron* 43(5): 703-714.

Lundin C, Käll L, Kreher SA, Kapp K, Sonnhammer EL, Carlson JR, Heijne G, Nilsson I. 2007. Membrane topology of the Drosophila OR83b odorant receptor. *FEBS Letters* 581(29): 5601-5604.

Marella S, Fischler W, Kong P, Asgarian S, Rueckert E, Scott K. 2006. Imaging taste responses in the fly brain reveals a functional map of taste category and behavior. *Neuron* 49(2): 285-295.

McBride CS. 2007. Rapid evolution of smell and taste receptor genes during host specialization in *Drosophila sechellia*. *Proceedings of the National Academy of Sciences USA* 104(12): 4996-5001.

McBride CS, Arguello JR, O'Meara BC. 2007. Five *Drosophila* genomes reveal nonneutral evolution and the signature of host specialization in the chemoreceptor superfamily. *Genetics* 177(3): 1395-1416.

Neuhaus EM, Gisselmann G, Zhang W, Dooley R, Stortkuhl K, Hatt H. 2005. Odorant receptor heterodimerization in the olfactory system of *Drosophila melanogaster*. *Nature Neuroscience* 8(1): 15-17.

Nozawa M, Nei M. 2007. Evolutionary dynamics of olfactory receptor genes in *Drosophila* species. *Proceedings of the National Academy of Sciences USA* 104(17): 7122-7127.

Richards S, Liu Y, Bettencourt BR, Hradecky P, Letovsky S, Nielsen R, Thornton K, Hubisz MJ, Chen R, Meisel RP, Couronne O, Hua S, Smith MA, Zhang P, Liu J, Bussemaker HJ, van Batenburg MF, Howells SL, Scherer SE, Sodergren E, Matthews BB, Crosby MA, Schroeder AJ, Ortiz-Barrientos D, Rives CM, Metzker ML, Muzny DM, Scott G, Steffen D, Wheeler DA, Worley KC, Havlak P, Durbin KJ, Egan A, Gill R, Hume J, Morgan MB, Miner G, Hamilton C, Huang Y, Waldron L, Verduzco D, Clerc-Blankenburg KP, Dubchak I, Noor MA, Anderson W, White KP, Clark AG, Schaeffer SW, Gelbart W, Weinstock GM, Gibbs RA. 2005. Comparative genome sequencing of *Drosophila pseudoobscura*: chromosomal, gene, and cis-element evolution. *Genome Research* 15(1): 1-18.

Robertson HM, Kent LB. 2009. Evolution of the gene lineage encoding the carbon dioxide receptor in insects. *Journal of Insect Science*. 9: 19, available online at: http://insectscience.org/9.19

Robertson HM, Wanner KW. 2006. The chemoreceptor superfamily in the honey bee, *Apis mellifera*: expansion of the odorant, but not gustatory, receptor family. *Genome Research* 16(11): 1395-1403.

Robertson HM, Warr CG, Carlson JR. 2003. Molecular evolution of the insect chemoreceptor gene superfamily in *Drosophila melanogaster*. *Proceedings of the National Academy of Sciences USA* 100(Suppl 2): 14537-14542.

Sakurai T, Nakagawa T, Mitsuno H, Mori H, Endo Y, Tanoue S, Yasukochi Y, Touhara K, Nishioka T. 2004. Identification and functional characterization of a sex pheromone receptor in the silkmoth *Bombyx mori*. *Proceedings of the National Academy of Sciences USA* 101(47): 16653-16658.

Schmidt HA, Strimmer K, Vingron M, von Haeseler A. 2002. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics* 18(3): 502-504.

Scott K, Brady R, Cravchik A, Morozov P, Rzhetsky A, Zuker C, Axel R. 2001. A chemosensory gene family encoding candidate gustatory and olfactory receptors in *Drosophila*. *Cell* 104(5): 661-673.

Slone J, Daniels J, Amrein H. 2007. Sugar receptors in *Drosophila*. *Current Biology* 17(20): 1809-1816.

Swofford DL. 2001. *PAUP*: Phylogenetic Analysis Using Parsimony and Other Methods, Version 4*. Sinauer Press.

Thorne N, Amrein H. 2008. Atypical expression of *Drosophila* gustatory receptor genes in sensory and central neurons. *The Journal of Comparative Neurology* 506(4): 548-568.

Thorne N, Chromey C, Bray S, Amrein H. 2004. Taste perception and coding in *Drosophila*. *Current Biology* 14(12): 1065-1079.

Tribolium Genome Sequencing Consortium 2008. The genome of the model beetle and pest *Tribolium castaneum*. *Nature* 452: 949-955.

Tunstall NE, Sirey T, Newcomb RD, Warr CG. 2007. Selective pressures on *Drosophila* chemosensory receptor genes. *Journal of Molecular Evolution* 64(6): 628-636.

van der Goes van Naters W, Carlson JR. 2007. Receptors and neurons for fly odors in *Drosophila*. *Current Biology* 17(7): 606-612.

Vosshall LB, Amrein H, Morozov PS, Rzhetsky A, Axel R. 1999. A spatial map of olfactory receptor expression in the *Drosophila* antenna. *Cell* 96(5): 725-736.

Wanner KW, Anderson AR, Trowell SC, Theilmann DA, Robertson HM, Newcomb RD. 2007a. Female-biased expression of odourant receptor genes in the adult antennae of the silkworm, *Bombyx mori*. *Insect Molecular Biology* 16(1): 107-119.

Wanner KW, Nichols AS, Walden KKO, Brockmann A, Luetje CW, Robertson HM. 2007b. A honey bee odorant receptor for the queen substance 9-oxo-2-decenoic acid. *Proceedings of the National Academy of Sciences USA* 104(36): 14383-14388.

Wang Z, Singhvi A, Kong P, Scott K. 2004. Taste representations in the *Drosophila* brain. *Cell* 117(7): 981-991.

Wistrand M, Kall L, Sonnhammer EL. 2006. A general model of G protein-coupled receptor sequences and its application to detect remote homologs. *Protein Science* 15(3): 509-521.

Jiao Y, Moon SJ, Montell C. 2007. A *Drosophila* gustatory receptor required for the responses to sucrose, glucose, and maltose identified by mRNA tagging. *Proceedings of the National Academy of Sciences USA* 104(35): 14110-14115.

Appendix 1. Comparison of Or family genes and proteins in *Drosophila melanogaster* and *D. pseudoobscura.*

Dm protein[a]	Dm CG[b]	Dm aa	Dp protein[c]	Dp GA[d]	Dp aa[e]	% ID[f]	Dp location[g]	Comments on genes[h]
Or1a	17867	392	Or1a	14708	388	70	XL_1e, 4.174	2 shared introns
Or2a	3206	397	Or2a	16647m	389	67	XL_1a, 0.154	1 shared intron
Or7a	10759	413	Or7aP	NA	135	-	XL_1e, 3.991	Degenerate pseudogene in Dp
Or9a	15302	392	Or9a	13635m	392	78	XL_1e, 5.696	2 shared introns; Dm gained intron
Or10a	17867	406	Or10a	14703m	404	70	XL_1e, 6.051	4 shared introns
Or13a	12697	418	Or13a	NA	417	68	4_4, 5.229	0 shared introns; Dp lost 4 introns
Or19a/b	18859/32825	387	Or19a	17168n	404	51	XL_1e, 7.187	2 shared introns; duplicated in Dm
Or22a/b	12193/4231	397	Or22a1	11469n	397	62/58	4_3, 10.044	2 shared introns; duplicated in both
Or22a/b	12193/4231	397	Or22a2	18049n	397	65/63	4_3, 10.094	2 shared introns; duplicated in both
Or22a/b	12193/4231	397	Or22aP	NA	360	54/50	4_3, 10.092	2 shared introns; duplicated in both
Or22c	15377	402	Or22c	13684	397	79	4_3, 8.850	5 shared introns
Or23a	9880	379	Or23a	22094	386	51	4_4, 1.978	1 shared intron
Or24a	11767	398	Or24a	11185	399	85	4_3, 1.745	4 shared introns
Or30a	13106	377	Or30a	12048m	377	88	4_5, 1.414	6 shared introns
-	-	-	Or33N	NA	380	-	4_3, 9.257	1 intron; Dm lost gene
Or33a	16960	378	Or33a	14239n	379	64	4_3, 9.258	1 shared intron
Or33b	16961	379	Or33b	14240	380	75	4_3, 9.261	1 shared intron
Or33c	5006	384	Or33c	18589m	389	60	4_3, 9.262	1 shared intron
Or35a	17868	409	Or35a	14704n	409	82	4_4, 2.171	3 shared introns
Or42a	17250	406	Or42a1	NA	409	80	3, 11.832	2 shared introns; duplicated in Dp
Or42a	17250	406	Or42a2	14414	407	85	3, 11.830	2 shared introns; duplicated in Dp
Or42b	12754	399	Or42b	11791	399	85	3, 11.827	2 shared introns
Or43a	1854	376	Or43a	14981m	378	74	3, 1.392	5 shared introns
Or43b	17853	403	Or43b	14700m	407	77	3, 13.390	2 shared introns
Or45a	1978	378	Or45a	15169m	381	71	3, 9.299	1 shared intron
Or45b	12931	396	Or45b	11917	398	81	3, 15.888	5 shared introns
Or46aA	17848-PA	381	Or46aA	14697-PAn	385	70	3, 6.875	3 shared introns; alternative splice
Or46aB	17848-PB	384	Or46aB	14697-PBn	390	68	3, 6.877	1 shared intron; alternative splice
Or47a	13225	385	Or47a	12137m	388	82	3, 11.025	3 shared introns
Or47b	13206	412	Or47b	12120m	413	60	3, 14.068	5 shared introns
Or49a	13158	396	Or49a1	12084n	394	62	3, 7.498	3 shared introns; duplicated in Dp
Or49a	13158	396	Or49a2	NA	399	66	3, 0.609	3 shared introns; duplicated in Dp
Or49b	17584	375	Or49b	14566	375	78	3, 14.763	3 shared introns
Or56a	12501	419	Or56a	11666m	417	66	3, 12.785	5 shared introns
-	-	-	Or56N	NA	416	-	3, 12.783	5 introns; Dm lost gene
Or59a	9820	379	Or59a	22057	379	77	3, 9.067	1 shared intron
Or59b	3569	398	Or59b	17527	398	89	3, 9.036	2 shared introns
Or59c	17226	411	Or59c	14401m	404	65	3, 9.034	1 shared intron
Or63a	9969	420	Or63a	22157m	422	75	XR_8, 3.780	8 shared introns; Dp lost an intron
Or65a	32401	417	-	-	-	-	-	4 introns; Dp lost gene
Or65b/c	32402/32403	406/410	Or65b1	16875m	422	51/49	XR_3a, 0.362	4 shared introns; duplicated in both
Or65b/c	32402/32403	406/410	Or65b2	NA	420	50/53	3, 3.083	4 shared introns; duplicated in both
Or65b/c	32402/32403	406/410	Or65b3	NA	420	49/52	3, 3.080	4 shared introns; duplicated in both
Or65b/c	32402/32403	406/410	Or65b4	NA	420	47/51	3, 3.078	4 shared introns; duplicated in both
Or65b/c	32402/32403	406/410	Or65b5	NA	416	51/54	3, 3.074	4 shared introns; duplicated in both
Or67a	12526	407	Or67a	NA	422	53	2, 11.413	2 shared introns

Appendix 1 (cont.)

Dm protein[a]	Dm CG[b]	Dm aa	Dp protein[c]	Dp GA[d]	Dp aa[e]	% ID[f]	Dp location[g]	Comments on genes[h]
Or67b	14176	421	Or67b	12805	420	85	XR_8, 9.199	8 shared introns; Dp lost an intron
Or67c	14156	404	Or67c	12792m	406	84	XR_6, 8.633	3 shared introns
Or67d	14157	391	Or67d	12793m	390	64	XR_6, 8.657	2 shared introns; Dp lost an intron
Or69aA	33264-PA	393	Or69aA	NA	394	51	XR_6, 4.331	2 shared introns; alternative splice
-	-	-	Or69aP	NA	367	-	XR_6, 4.329	Pseudogene in Dp; Dm lost gene
Or69aB	33264-PB	393	Or69aB	NA	398	51	XR_6, 4.327	2 shared introns; alternative splice
Or71a	17871	378	Or71a	14707n	384	59	XR_6, 0.667	2 shared introns
Or74a	13726	404	Or74a	12488	404	75	XR_6, 9.324	3 shared introns
Or82a	31519	384	Or82a	16295	386	65	2, 24.840	5 shared introns
Or83a	10612	453	Or83a	10437m	457	80	2, 21.373	3 shared introns
Or83b	10609	486	Or83b	10435	488	94	2, 21.361	5 shared introns
Or83c	15581	397	Or83c	13827m	400	68	2, 15.898	3 shared introns; Dp lost an intron
Or85a	7454	397	-	-	-	-	-	2 introns; Dp lost gene
Or85b	11735	390	Or85b	11167	390	68	2, 23.307	2 shared introns
Or85c	17911	389	Or85c	14720m	389	77	2, 23.310	2 shared introns
Or85d	11742	412	Or85d	11171m	413	73	2, 23.352	2 shared introns
Or85eP	9700	467	Or85e	21973n	467	80	2, 17.229	2 shared introns; Dm pseudogene
Or85f	16755	392	Or85f	14128m	393	66	2, 3.196	3 shared introns
Or88a	14360	401	Or88a	12932m	400	68	2, 13.451	2 shared introns; Dp lost an intron
Or92a	17916	408	Or92a	NA	406	87	XL_1e, 0.030	2 shared introns
Or94a	17241	387	Or94a	14408m	387	84	2, 6.068	1 shared intron; Dm lost intron
Or94b	6679	383	Or94b	19774	383	76	2, 6.066	1 shared intron
Or98P	NA	69	-	-	-	-	-	Degenerate pseudogene in Dm only
Or98a	5540	397	Or98a1	18957n	393	54	2, 4.032	2 shared introns; duplicated in Dp
Or98a	5540	397	Or98a2	NA	398	56	3, 3.458	2 shared introns; duplicated in Dp
Or98a	5540	397	Or98a3	NA	384	55	3, 3.463	2 shared introns; duplicated in Dp
Or98a	5540	397	Or98aP	NA	106	-	3, 3.466	Degenerate pseudogene in Dp only
Or98bP	1867	383	Or98b	15044	386	61	2, 10.358	3 shared introns; Dm pseudogene
-	-	-	OrN	NA	423	-	XR_8, 1.239	4 introns; Dm lost gene

[a]The table is organized by position of the *Drosophila melanogaster* proteins along the chromosomes, as reflected by their names using the chromosomal locations of the genes, e.g. Or33a is the first gene in the 33rd division, which is on the left arm of chromosome 2. Where locations of genes unique to *D. pseudoobscura* are obvious they are interdigitated in the table, eg. DpOr33N is immediately upstream of and related to the Or33a-c orthologs. A single pseudoobscura gene with no obvious paralogs, DpOrN, is placed at the end of the table. DmOr85e and DmOr98b are polymorphic pseudogenes in melanogaster – see Robertson et al. (2003) for Or85e and text for Or98b. Intact versions are represented here and in all analyses.

[b]CG identifiers are provided for all the DmOrs, except degenerate pseudogenes, but the protein isoforms (PA, PB, etc) are only shown for the two alternatively spliced genes, Or46a and Or69a.

[c]The DpOrs are given names best reflecting their orthologous or paralogous relationships with the DmOrs. Names ending with a "P" are pseudogenes; with a "N" are genes present only in *D. pseudoobscura.*

[d]GA identifiers are given for DpOrs when available. Most of these are annotated genes/proteins. When my gene models differ from those in release 1.04 of the Dpse genome annotation in FlyBase as of July 2005 and published in Richards et al. (2005), then the identifier is followed by a "m" for "modified". Most of these differences are instances where a few N- and/or C-terminal amino acids are missing from the r1.04 versions, but some involve inclusion of open in-frame introns, missing C-terminal exons, or missed GC intron donor splice sites. Some GA identifiers are "place holders" showing where an anticipated orthologous gene model belongs in the Dp genome browser at FlyBase, in which case they are followed by a "n" for "new". GA identifiers are not available for some genes and all pseudogenes, indicated as "NA".

[e]Some DpOr gene models have N-terminal ORF extensions beyond a potential methionine start codon that aligns with the orthologous DmOr. These have not been included in these length figures, or in the FASTA files or alignments, but the lengths are noted here (Or65b1 – 91aa; Or85d – 52aa; Or85e – 48aa; Or85f – 72aa; Or98a2 – 17aa).

[f]Percent identities are based on the "BLAST 2 SEQUENCES" algorithm at NCBI which includes only alignable sequence, so highly divergent proteins that do not completely align full-length are in fact more divergent than they appear herein.

[g]Release 2.0 of the 140Mbp *pseudoobscura* genome is in 16 major superscaffolds, ranging from entire chromosome arms for 2 and 3, to 4 or 5 superscaffolds for arms XL, XR, and 4, plus 2650 unmapped small scaffolds grouped in a 10.7Mbp "chromosome U". The locations of the *D. pseudoobscura* genes are shown as the superscaffold, e.g. "XL_1e" means "XL_group1e", followed by the location in megabases according to the FlyBase pseudoobscura genome browser.

[h]Comments on genes include the number of shared introns, whether the gene was lost from, or duplicated in, one or other species, pseudogene status, and alternative splicing.

Appendix 2. Comparison of Gr family genes and proteins in *Drosophila melanogaster* and *D. pseudoobscura.*

Dm protein[a]	Dm CG[b]	Dm aa	Dp protein[c]	Dp GA[d]	Dp aa[e]	% ID[f]	Dp location[g]	Comments on genes[h]
Gr2a	18531	414	Gr2a	14976m	400	70	XL_1a, 4.680	3 shared introns
Gr5a	15779	444	-	-	-	-	-	6 introns; Dp lost gene
Gr8a	15371	385	Gr8a	13680m	390	67	XL_1e, 1.180	3 shared introns
Gr9a	32693	341	Gr9a	17078m	341	55	XL_3a, 1.186	2 shared introns
Gr10a	32664	408	Gr10a	NA	412	75	XL_1e, 6.054	1 shared intron
Gr10b	12622	373	Gr10b	11723m	379	44	XL_1e, 6.056	1 shared intron; Dp lost an intron
Gr21a	13948	447	Gr21a	12646n	447	96	4_3, 8.952	2 novel Dm introns, Dp lost an intron
Gr22a	31662	394	Gr22a	16376m	389	61	4_3, 8.515	1 shared intron
Gr22bP/c	NA/31929	386/383	Gr22b	16572n	386	47/43	4_3, 8.516	1 shared intron; duplicated in Dm
Gr22dP	NA	387	Gr22dP	NA	372	58	4_3, 8.518	1 shared intron; both pseudogenes
Gr22e	31936	389	Gr22e	16578	390	70	4_3, 8.524	1 shared intron
Gr22f	31932	378	-	-	-	-	-	1 intron; Dp lost gene
Gr23aA	15396-PA	383	Gr23aA	13700-PAn	388	75	4_4, 1.996	2 shared introns; alt. splice
Gr23aB	15396-PB	374	Gr23aB	13700-PBn	367	61	4_4, 1.997	2 shared introns; alt. splice
Gr28a	13787	450	Gr28a	12527n	448	88	4_1, 4.211	2 shared introns
Gr28bA	13788-PA	452	Gr28bA	12528m	452	77	4_1, 4.220	2 shared introns; alt. splice
Gr28bB	13788-PB	443	Gr28bB	12528m	444	96	4_1, 4.219	2 shared introns; alt. splice
Gr28bC	13788-PC	470	Gr28bC	12528m	464	87	4_1, 4.215	2 shared introns; alt. splice
Gr28bD	13788-PD	440	Gr28bD	12528m	440	84	4_1, 4.214	2 shared introns; alt. splice
Gr28bE	13788-PE	447	Gr28bE	12528m	445	79	4_1, 4.213	2 shared introns; alt. splice
Gr32a	14916	461	Gr32a	13351n	458	85	4_3, 6.466	3 shared introns
Gr33a	17213	475	Gr33a	14395m	484	86	4_3, 6.051	4 shared introns
Gr36a/b/c	31747/4/8	390–391	Gr36a	16444	392	45–48	4_4, 0.070	1 shared intron; Dp lost an intron
-	-	-	Gr39a1	NA	364	-	4_4, 3.427	3 introns; alt. splice; Dm lost
Gr39aB	31622-PB	372	Gr39a2	NA	371/369	64/64	4_4, 3.429	3 shared in.; alt. splice; Dp dup.
Gr39aB	31622-PB	372	Gr39a3	16340-PBm	371/369	64/64	4_4, 3.431	3 shared in.; alt. splice; Dp dup.
Gr39aC	31622-PC	381	Gr39a4	NA	382	75	4_4, 3.432	3 shared introns; alt. splice
-	-	-	Gr39a5	NA	385	-	4_4, 3.433	3 introns; alt. splice; Dm lost
Gr39aD	31622-PD	381	Gr39a6	NA	377	62	4_4, 3.434	3 shared introns; alt. splice
Gr39aA	31622-PA	371	Gr39a7	NA	392	64	4_4, 3.435	3 shared introns; alt. splice
Gr39b	31620	369	Gr39b	16339m	371	56	4_4, 3.499	2 shared introns
Gr43a	1712	427	Gr43a	14333m	427	81	3, 1.552	8 shared introns
Gr47a	12906	361	Gr47aP	11896m	220	-	3, 7.195	1 shared intron; Dp fragment pseudo
Gr47b	30030	414	Gr47b	15589n	423	65	3, 13.537	2 shared introns
Gr47b	30030	414	Gr47P	NA	257	63	3, 13.519	Dp pseudogene fragment
Gr57a	13441	416	Gr57a	12290	426	63	3, 8.572	1 shared intron
Gr58a	30396	395	Gr58a	15816n	409	54	3, 15.427	1 shared intron
Gr58b	13495	408	Gr58b	12328m	391	53	3, 15.427	1 shared intron
Gr58c	13491	412	Gr58c	12324m	411	55	3, 15.422	1 shared intron
Gr59a	30189	367	Gr59a1	15713n	371	47	3, 3.340	1 shared intron; duplicated in Dp
Gr59a	30189	367	Gr59a2	NA	371	47	3, 3.338	1 shared intron; duplicated in Dp
Gr59a	30189	367	Gr59a3	NA	374	44	3, 3.336	1 shared intron; duplicated in Dp
Gr59b	30191	366	Gr59b	15716m	362	53	3, 3.339	1 shared intron
Gr59c	30186	397	Gr59c	15710m	393	51	3, 3.315	1 shared intron
Gr59d	30330	390	Gr59d	15766m	393	48	3, 3.313	1 shared intron
Gr59e	33151	399	Gr59e	17326m	404	59	3, 9.573	1 shared intron; Dp lost an intron

Appendix 2 (cont.)

Dm protein[a]	Dm CG[b]	Dm aa	Dp protein[c]	Dp GA[d]	Dp aa[e]	% ID[f]	Dp location[g]	Comments on genes[h]
Gr59f	33150	406	Gr59f	17325m	408	59	3, 9.572	2 shared introns; first intron differs
Gr61a	13888	436	Gr61a	12601m	439	77	XR_6, 6.445	7 shared introns
Gr63a	14979	512	Gr63a	13400m	504	89	XR_6, 4.965	2 shared introns
Gr64a	32261	456	Gr64a	16796	459	78	XR_6, 6.444	6 shared introns
Gr64b	32257	406	Gr64b	16793	406	89	XR_6, 6.443	6 shared introns; Dp lost an intron
Gr64c	32256	419	Gr64c	16792m	416	80	XR_6, 6.441	4 shared introns
Gr64d	14987	429	Gr64d	17330n	421	64	XR_6, 6.439	4 shared introns
Gr64e	32258	460	Gr64eP	NA	461	74	XR_6, 6.438	8 shared introns; pseudogene in Dp
Gr64f	32255	469	Gr64f	16791m	469	84	XR_6, 6.436	6 shared introns
Gr66a	7189	530	Gr66a	20169m	528	84	XR_3a, 0.017	1 shared intron; Dp has extra intron
Gr68a	7303	389	Gr68a	20348m	393	67	XR_8, 4.366	intronless genes
Gr77a	32433	449	Gr77a	16898m	483	61	XR_8, 8.161	1 shared intron
Gr85a	31405	397	Gr85a	16233m	376	47	2, 5.143	1 shared intron; Dp incomplete
Gr89a	14901	362	Gr89a	13339n	362	68	2, 5.408	1 shared intron
Gr92a	31208	386	-	-	-	-	-	1 intron; Dp lost gene
Gr93a	13417	419	Gr93a	12269m	433	68	2, 6.513	1 shared intron
Gr93b	31336	401	Gr93b	16189m	402	63	2, 6.524	1 shared intron
Gr93c	31173	397	Gr93c	16064	397	65	2, 6.526	1 shared intron
Gr93d	31335	381	-	-	-	-	-	1 intron; Dp lost gene
-	-	-	Gr93N	16188n	405	-	2, 6.528	1 intron; Dm lost gene
Gr94a	31280	404	Gr94a	16146m	404	72	2, 5.997	intronless genes
Gr97a	31280	425	Gr97a	17226	427	64	2, 8.396	1 shared intron
Gr98a	13976	391	Gr98aP	12669	393	52	2, 8.974	2 shared introns; Dp pseudogene
Gr98b	31059	403	Gr98b	15973m	407	61	2, 9.018	3 shared introns
Gr98c/d	31060/1	408/412	-	-	-	-	-	3 introns; Dp lost gene

[a]The table is organized by position of the *melanogaster* proteins along the chromosomes, as reflected by their names using the chromosomal locations of the genes, e.g. Gr10a is the first Gr gene in the 10th division, which is on the X chromosome. The location of one gene unique to pseudoobscura DpGr93N, is obvious so it is interdigitated in the table. DmGr22b and 22d are polymorphic pseudogenes in melanogaster (see Robertson et al. 2003); the intact versions are used here and in analyses.

[b]CG identifiers are provided for all the DmGrs, except the two polymorphic pseudogenes, but the protein isoforms (PA, PB, etc) are only shown for the three alternatively spliced genes - Gr23a, 28b, and 39a. The letters designating the alternatively spliced forms of Gr39a are not in sequential order of the placement of the first exons on the chromosome because historically the last of the four exons was recognized first (see also Figure 3). To avoid confusion FlyBase nomenclature was used which therefore differs from the naming in Clyne et al. (2000) and in Robertson et al. (2003). pseudogenes; with a "N" are genes present only in pseudoobscura.

[c]The DpGrs are given names best reflecting their orthologous or paralogous relationships with the DmOrs. Names ending with a "P" are pseudogenes; with a "N" are genes present only in pseudoobscura.

[d]GA identifiers are given for DpGrs when available. Most of these are annotated genes/proteins. When my gene models differ from those in release 1.04 of the Dpse genome annotation in FlyBase as of July 2005 and published in Richards et al. (2005), then the identifier is followed by a "m" for "modified". Most of these differences are instances where a few N- and/or C-terminal amino acids are missing from the r1.04 versions, but some involve inclusion of open in-frame introns, missing C-terminal exons, or missed GC intron donor splice sites. Some GA identifiers are "place holders" showing where an anticipated orthologous gene model belongs in the Dp genome browser at FlyBase, in which case they are followed by a "n" for "new". GA identifiers are not available for some genes and all pseudogenes, indicated as "NA".

[e]Some DpGr gene models have N-terminal ORF extensions beyond a potential methionine start codon that aligns with the orthologous DmGr. These have not been included in these length figures, or in the FASTA files or alignments, but the lengths are noted here (Gr10b - 27aa, Gr21a - 55aa, Gr43a - 23aa, Gr61a - 15aa, Gr66a - 20aa, Gr89a – 18aa, and Gr98b - 53aa).

[f]Percent identities are based on the "BLAST 2 SEQUENCES" algorithm at NCBI which includes only alignable sequence, so highly divergent proteins that do not completely align full-length are in fact more divergent than they appear herein.

[g]Release 2.0 of the 140Mbp *pseudoobscura* genome is in 16 major superscaffolds, ranging from entire chromosome arms for 2 and 3, to 4 or 5 superscaffolds for arms XL, XR, and 4, plus 2650 unmapped small scaffolds grouped in a 10.7Mbp "chromosome U". The locations of the pseudoobscura genes are shown as the major fragment, e.g. "XL_1e" means "XL_group1e", followed by the location in megabases according to the FlyBase pseudoobscura genome browser.

[h]Comments on genes include the number of shared introns, whether the gene was lost from, or duplicated in, one or other species, pseudogene status, and alternative splicing. DpGr10b lost intron "XXXXXX"; DpGr36a lost intron "k"; DpGr59e lost intron "1"; the first intron in DpGr59f is phase 2 instead of phase 1 in DmGr59f and other related genes – alternatively DpGr59f might start later; DpGr64b lost intron "r"; DpGr66a has an extra intron – need to figure out if novel or shared with others; Gr68a and Gr94a are both intronless in both species, and both are in introns of other genes (the only other such example is Gr39a), so might have resulted from insertion of reverse transcribed copies of cDNAs in a common ancestor.