On the origin of the odorant receptor gene family in insects: evidence from the dragonfly *Ladona fulva* and the mayfly *Ephemera danica*

Hugh M. Robertson

Department of Entomology, University of Illinois at Urbana-Champaign, Urbana, IL61801, USA

Running title: Origin of the insect odorant receptor gene family

Keywords: chemoreceptor; olfactory receptor; smell

Abstract

The odorant receptor (OR) gene family encodes the major olfactory receptors of insects. It evolved from a lineage of the older gustatory receptor (GR) family, and in most insects consists of a single gene encoding a conserved odorant receptor co-receptor (Orco) and several to hundreds of specific ORs that mediate the specificity and sensitivity of most of insect olfaction. Previous work has suggested that the family originated within the wingless insects or Apterygota, between the Archaeognatha (a bristletail with no expressed ORs) and the Zygentoma (a firebrat with at least three apparent Orco relatives - TdomOrco1-3). Examination of the OR family in the dragonfly *Ladona fulva* and the mayfly *Ephemera danica*, along with the published damselfly *Calopteryx splendens*, reveals that both of these paleopteran lineages have a single Orco gene. The odonates only have a few specific ORs, while the mayfly has about 50 ORs. Phylogenetic analysis reveals that the specific ORs in these two paleopteran lineages form two major clades, and that TdomOrco3 belongs in one of these clades. TdomOrco1 might also be a specific OR, leaving TdomOrco2 as the sole Orco ortholog in the firebrat. This finding implies that the entire Orco/OR system evolved before zygentomans.

Introduction

The odorant receptor (OR) family in insects was first recognized in the fledgling *Drosophila melanogaster* genome sequence (Clyne et al. 1999; Vosshall et al. 1999), and soon extended to the full complement of 60 genes encoding 62 receptors with completion of the genome sequence (Vosshall et al. 2000; Robertson et al. 2003). Enormous progress has been made since then in understanding this ecologically important gene family in insects, including their likely structure, gene family expansions and contractions in diverse insects, their diverse ligands, and their roles in fly and other insect biology (Leal 2013; Benton 2015; Hopf et al. 2015; Joseph and Carlson 2015; Haverkamp et al. 2018). Scott et al. (2001) suggested that the OR family was related to the gustatory receptor (GR) family (Clyne et al. 2000), an observation confirmed by Robertson et al. (2003). On the basis of the pattern of molecular evolution in *D. melanogaster*, Robertson et al. (2003) suggested that the OR family might have evolved from within the much older GR family concomitant with the evolution of terrestriality in insects. The GR family extends back to basal animals (Saina et al. 2015; Robertson 2015; Eyun et al. 2017), but the ORs are clearly much younger because they have not been found in non-insect arthropods. Thus the genome sequences of the crustaceans *Daphnia pulex* (Penalva-Arana et al. 2009) and *Eurytemora affinis* (Eyun et al. 2017), the centipede *Strigamia maritima* (Chipman et al. 2014; Almeida et al. 2015), and the chelicerates *Metaseiulus occidentalis* (Hoy et al. 2016), *Ixodes scapularis* (Gulia-Nuss et al. 2016), and *Tetranychus urticae* (Ngoc et al. 2016) reveal only members of the GR family, and

the unrelated and similarly ancient Ionotropic Receptor (IR) family (Rytz et al. 2013; Eyun et al. 2017; Rimal and Lee 2018).

Efforts to understand more precisely the origin of the OR family within hexapods were greatly advanced by the findings of Missbach et al. (2014) who sequenced chemosensory organ transcriptomes for two basal apterygotes, the bristletail *Lepismachilis y-signata* (Archaeognatha) and the firebrat *Thermobia domestica* (Zygentoma). From the firebrat they obtained transcripts encoding three proteins they named TdomOrco1-3 for their apparent relationship to the odorant receptor co-receptor protein of neopteran insects, known as Orco (Vosshall and Hansson 2011). Orco is present in all other studied insects to date as a single gene and the protein is a partner with each of the other "specific" ORs (Benton et al. 2006). However, they could not find ORs or Orco relatives in the bristletail, instead finding only members of the IR family. Given evidence that IRs serve olfactory roles in terrestrial crustaceans (Groh-Lunow 2015), they argued that basal terrestrial hexapods and insects used IRs for all of their olfaction, as all studied insects still do for a subset of olfactory sensitivities (Rytz et al. 2013; Rimal and Lee 2017), with these three Orco relatives evolving from a GR lineage between the Archaeognatha and Zygentoma. Missbach et al. (2014) left off with the observation that "the existence of three Orco types remains mysterious".

Recently the genome of an odonate, the damselfly *Calopteryx splendens*, revealed that this insect, belonging to an order previously thought to be anosmic but now known to be capable of olfaction (Piersanti et al. 2014), encodes a single Orco protein and five specific ORs (Ioannidis et al. 2017). Phylogenetic analysis of the OR family suggested that one of the three named Orco proteins from *T. domestica*, TdomOrco3, might be a specific OR. If this is correct, then the entire Orco/OR system evolved before the Zygentoma, which would explain the "mystery" of three apparent Orco types. In an effort to illuminate this issue further I examined the OR families in the genome sequences of another odonate, the dragonfly *Ladona fulva*, and a mayfly *Ephemera danica*

Materials and Methods

TBLASTN searches of the dragonfly *L. fulva* and mayfly *E. danica* genome sequences available from the i5k pilot project at the Human Genome Sequencing Center at Baylor College of Medicine and available at the i5k Workspace@NAL website (Poelchau et al. 2014) were performed using the *C. splendens* ORs and the *T. domestica* Orcos as queries, along with diverse ORs from other insects. These odonate and mayfly ORs are particularly divergent, so for exhaustive searches, E values were raised to 1000 and word size reduced to 2, and the amino acid sequences of the last two most-conserved exons were augmented with LQ before them, and when appropriate VS after, to mimic consensus splice site sequences of flanking phase-0 introns. Gene models were built in the Apollo browser available at the i5k Workspace@NAL, with problematic models and pseudogenes worked up in a text editor. Pseudogenes were translated as best possible, employing Z for stop codons and X for frameshifts, and only included if they encoded at least half the amino acids of a typical OR, and the same criterion was used to exclude short gene fragments. All protein sequences are available in the supplementary file, and transcripts of intact gene models are available from the i5k Workspace@NAL. Proteins were aligned in ClustalX v2.1 (Larkin et al. 2007) and the alignment was trimmed with TrimAl v1.4 (Capella-Gutierrez et al. 2009) using the "gappyout" option. Phylogenetic analysis was

conducted using maximum likelihood in PHyML v3.0 (Guindon et al. 2010), and the resultant tree organized and colored in FigTree v1.4.2 (http://tree.bio.ed.ac.uk/software/figtree/).

Results

The dragonfly *L. fulva* genome contains a single Orco gene, and three genes encoding specific ORs (LfulOr1-3), the smallest known OR family repertoire for an insect with a genome sequence and comparable to that of *C. splendens* (Ioannidis et al. 2017). All four genes have full-length models, although in the absence of transcript support their confidence is solely based on comparative evidence of gene structures and sequence similarity. The Orco protein shares some features with the *C. splendens* protein compared with neopteran Orco proteins, specifically a slightly longer N-terminus and a longer first intracellular loop. LfulOr1/2 have gene structures similar to those of CsplOr1-5, specifically they have introns in phases 2-2-0-0-0, with the second exon being the longest. The last four introns are in shared locations and phases as those of CsplOr1-5, as well as the entire OR family as originally inferred from *D. melanogaster* (Robertson et al. 2003). In contrast, LfulOr3 is rather divergent and has a gene structure with introns in phases 0-0-1-2-0-0-0, the last four again being apparently ancestral. *C. splendens* does not have a comparable OR gene, so appears to have lost this OR lineage.

In contrast, the *E. danica* genome encodes an OR family ten times larger than these two odonates, with a single Orco gene and at least 46 specific ORs. The gene models for these ORs were sometimes difficult to build, because they are sometimes highly divergent, some exons are missing in gaps in the assembly, and some genes, much like LfulOr3, have eight or even nine short exons. Ten of these 46 models are apparent pseudogenes with obvious pseudogenizing mutations like stop codons and/or frameshifting insertions or deletions. Another four models required repair of the genome assembly, while eight remain as partial models with exons missing. Approximately six gene fragments remain in the assembly that might represent intact genes in the genome. Given the divergence and small size of most exons, it is also possible that some highly divergent genes have evaded detection, although the phylogenetic analysis below suggests that the major clades have been discovered. Most genes share the ancestral 2-0-0-0 intron locations. EdanOr1-34 have only those four introns, while EdanOr35-46 have three additional introns splitting up the usually long first exon, in phases 1-2-1.

Phylogenetic analysis of these OR families along with the three named Orco proteins from *T. domestica* and Orco proteins from representative neopterans reveals that the specific ORs from these two odonates and the mayfly form two major clades (Figure 1). LfulOr1 and 2 are related to CsplOr1-5 along with EdanOr1-12. LfulOr3 is the sole odonate specific OR in the other major clade with EdanOr13-46. As suggested in Ioannidis et al. (2017), TdomOrco3 belongs confidently with these specific ORs, having strong support for clustering with them, and specifically with the first clade. Missbach et al. (2014) noted that TdomOrco3 has several amino acid changes from those conserved in Orcos. Given that this “division-of-labor” between a single Orco and a set of specific ORs had already originated by the time of the firebrat, TdomOrco1 might also be a specific OR. Missbach et al. (2014) performed in situ hybridizations to antennal sections with TdomOrco1, and found that it was expressed in only one or two olfactory sensory neurons per antennal segment, a pattern consistent with it being a “specific” OR. This would leave TdomOrco2 as the sole true Orco ortholog.

Discussion

The OR family plays a major role in the biology of insects so its evolutionary origins are of particular interest. While the observations of Missbach et al. (2014) appear to set the evolution of the family within the basal wingless insects, this conclusion is not ironclad. First, the findings here, confirming the proposal of Ioannidis et al. (2017) that at least one of the three named Orco proteins in the firebrat is a specific OR, leaving TdomOrco2 as the single true Orco ortholog, mean that this Orco/OR system had already evolved by the time of the firebrat, representing the Zygentoma. Second, the conclusion of Missbach et al. (2014) that the more ancient archaegnathan bristletail lineage does not have Orco or ORs, but instead relies entirely on IRs for olfaction, was based on a transcriptome. Until at least one and preferably several archaegnathan genomes are obtained it will remain an open question as to whether they indeed predate the origin of the OR family. Even then it remains possible that bristletails lost their Orco/OR genes, perhaps due to a shift in their chemical ecology that made them redundant.

There are several more basal hexapod lineages that might harbor the origins of the OR family if it indeed predates wingless insects, specifically the Collembola, Diplura, and Protura. Like a proposed bristletail genome project, projects on the genomes of representatives of each of these groups unfortunately had to be abandoned by the pilot i5k project (https://www.hgsc.bcm.edu/arthropods/i5k) because of their large size and complexity (S. Richards, personal communication). Nevertheless, genome sequences for three other collembolans, *Orchesella cincta*, *Folsomia candida*, and *Holacanthella duospinosa,* reveal GRs and IRs but do not contain Orco/OR genes (Faddeeva-Vakhrusheva et al. 2016, 2017; Wu et al. 2017), making it likely that the IR family, and perhaps some GRs, provided the first olfactory receptors in terrestrial hexapods, followed later by the evolution of the OR family. While this study confirms that the Orco/OR functional split predates the Zygentoma, it remains unclear precisely when this gene family of important olfactory receptors arose in insect evolution. The availability of third generation sequencing methods capable of generating long reads that allow improved assembly of difficult genomes such as other basal hexapod and insect lineages will hopefully soon further illuminate this question.

Acknowledgements

I thank Stephen Richards for permission to explore the *Ladona fulva* and *Ephemera danica* genome sequences from the i5k pilot project at the Human Genome Sequencing Center at the Baylor College of Medicine.

References

Almeida FC, Sanchez-Garcia A, Walden KK, Robertson HM, Rozas J. 2015. Positive selection in extra cellular domains in the diversification of *Strigamia maritima* chemoreceptors. *Front Ecol Evol*. 3:e79

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

Benton R. 2015. Multigene family evolution: perspectives from insect chemoreceptors. *Trends Ecol Evol.* 30:590-600.

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*. 25:1972-1973.

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

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* 22:327-338.

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

Eyun SI, Soh HY, Posavi M, Munro JB, Hughes DS, Murali SC, Qu J, Dugan S, Lee SL, Chao H et al. 2017. Evolutionary history of chemosensory-related gene families across the Arthropoda. *Mol Biol Evol* 34:1838-1862.

Faddeeva-Vakhrusheva A, Derks MF, Anvar SY, Agamennone V, Suring W, Smit S, van Straalen NM, Roelofs D. 2016. Gene family evolution reflects adaptation to soil environmental stressors in the genome of the collembolan *Orchesella cincta*. *Genome Biol Evol* 8:2106-2117.

Faddeeva-Vakhrusheva A, Kraaijeveld K, Derks MF, Anvar SY, Agamennone V, Suring W, Kampfraath AA, Ellers J, Le Ngoc G, Gestel CA et al. 2017. Coping with living in the soil: the genome of the parthenogenetic springtail *Folsomia candida*. *BMC Genomics* 18:e493.

Groh-Lunow KC, Getahun MN, Grosse-Wilde E, Hansson BS. 2015. Expression of ionotropic receptors in terrestrial hermit crab's olfactory sensory neurons. *Front Cell Neuro* 8:e448.

Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* 59:307-321.

Gulia-Nuss M, Nuss AB, Meyer JM, Sonenshine DE, Roe RM, Waterhouse RM, Sattelle DB, De La Fuente J, Ribeiro JM, Megy K, et al. 2016. Genomic insights into the *Ixodes scapularis* tick vector of Lyme disease. *Nature Comm* 7:e10507.

Haverkamp A, Hansson BS, Knaden M. 2018. Combinatorial codes and labeled lines: How insects use olfactory cues to find and judge food, mates and oviposition sites in complex environments. Front Physiol 9:e49

Hopf TA, Morinaga S, Ihara S, Touhara K, Marks DS, Benton R. 2015. Amino acid coevolution reveals three-dimensional structure and functional domains of insect odorant receptors. *Nature Comm* 6:e6077.

Hoy MA, Waterhouse RM, Wu K, Estep AS, Ioannidis P, Palmer WJ, Pomerantz AF, Simao FA, Thomas J, Jiggins FM, et al. 2016. Genome sequencing of the phytoseiid predatory mite *Metaseiulus occidentalis* reveals completely atomized Hox genes and superdynamic intron evolution. *Genome Biol Evol* 8:1762-1775.

Ioannidis P, Simao FA, Waterhouse RM, Manni M, Seppey M, Robertson HM, Misof B, Niehuis O, Zdobnov EM. 2017. Genomic features of the damselfly *Calopteryx splendens* representing a sister clade to most insect orders. *Genome Biol Evol* 9:415-430.

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, et al. 2007. Clustal W and Clustal X version 2.0. *Bioinformatics* 23:2947-2948.

Leal WS. 2013. Odorant reception in insects: roles of receptors, binding proteins, and degrading enzymes. *Ann Rev Entomol* 58:373-391.

Missbach C, Dweck HK, Vogel H, Vilcinskas A, Stensmyr MC, Hansson BS, Grosse-Wilde E. 2014. Evolution of insect olfactory receptors. *Elife* 3:e02115.

Ngoc PC, Greenhalgh R, Dermauw W, Rombauts S, Bajda S, Zhurov V, Grbić M, Van de Peer Y, Van Leeuwen T, Rouze P et al. 2016. Complex evolutionary dynamics of massively expanded chemosensory receptor families in an extreme generalist chelicerate herbivore. *Genome Biol Evol* 8:3323-3339.

Peñalva-Arana DC, Lynch M, Robertson HM. 2009. The chemoreceptor genes of the waterflea *Daphnia pulex*: many Grs but no Ors. *BMC Evol Biol* 9:e79.

Piersanti S, Frati F, Conti E, Rebora M, Salerno G. 2014. The sense of smell in Odonata: An electrophysiological screening. *J Insect Physiol* 70:49-58.

Poelchau M, Childers C, Moore G, Tsavatapalli V, Evans J, Lee CY, Lin H, Lin JW, Hackett K. 2014. The i5k Workspace@ NAL—enabling genomic data access, visualization and curation of arthropod genomes. *Nucl Acids Res* 43:D714-D719.

Rimal S, Lee Y. 2018. The multidimensional ionotropic receptors of *Drosophila melanogaster*. *Insect Mol Biol* 27:1-7.

Robertson HM, Warr CG, Carlson JR. 2003. Molecular evolution of the insect chemoreceptor gene superfamily in *Drosophila melanogaster*. *Proc Natl Acad Sci USA* 100 (suppl 2):14537-14542.

Robertson HM. 2015. The insect chemoreceptor superfamily is ancient in animals. *Chem Senses* 40:609-614.

Rytz R, Croset V, Benton R. 2013. Ionotropic receptors (IRs): chemosensory ionotropic glutamate receptors in *Drosophila* and beyond. *Insect Biochem Mol Biol* 43:888-897.

Saina M, Busengdal H, Sinigaglia C, Petrone L, Oliveri P, Rentzsch F, Benton R. 2015. A cnidarian homologue of an insect gustatory receptor functions in developmental body patterning. *Nature Comm* 6:e6243.

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

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

Vosshall LB, Wong AM, Axel R. 2000. An olfactory sensory map in the fly brain. *Cell* 102:147-159.

Vosshall LB, Hansson BS. 2011. A unified nomenclature system for the insect olfactory coreceptor. *Chem Senses* 36:497-498.

Wu C, Jordan MD, Newcomb RD, Gemmell NJ, Bank S, Meusemann K, Dearden PK, Duncan EJ, Grosser S, Rutherford K, et al. 2017. Analysis of the genome of the New Zealand giant collembolan (*Holacanthella duospinosa*) sheds light on hexapod evolution. *BMC Genomics* 18:e795.

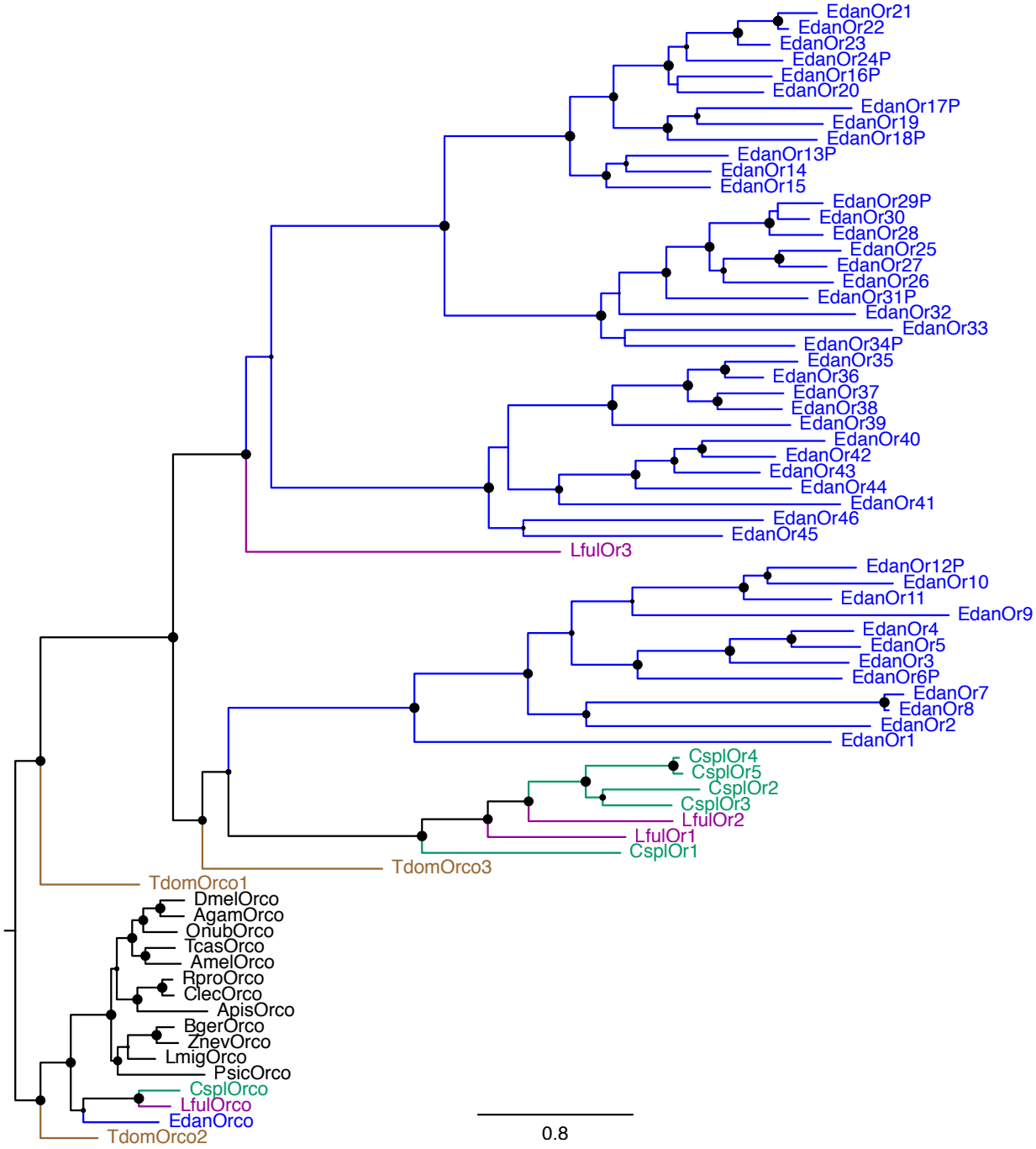

Figure 1. Phylogenetic tree of the OR family in basal insects. The tree was rooted by declaring the Orco proteins from neopterans, along with TdomOrco2, as the outgroup. Protein names are colored by species, *T. domestica*, *C. splendens*, *L. fulva*, and *E. danica* in brown, green, teal, and blue, respectively, as are the branches to them. Orco proteins from neopteran insects are in black. The suffix P after a name indicates a pseudogene. Support levels at nodes are indicated as filled circles representing approximate Likelihood Ratio Tests (aLRT) from PHyML v3.0 ranging from 0-1. The scale bar is substitutions per site.