InsectOR – webserver for sensitive identification of insect olfactory receptor genes from non-model genomes

Snehal D. Karpe[1,#a,#b], Vikas Tiwari[1] and Sowdhamini Ramanathan [1*]

[1] National Centre for Biological Sciences (NCBS), TIFR, Bengaluru, Karnataka, India

[#a] Current address: Laboratory of Experimental Hematology, Institut Jules Bordet, Université Libre de Bruxelles, Brussels, Belgium

[#b] Current address: Unit of Animal Genomics, GIGA, University of Liège, Liège, Belgium

[*] Corresponding author

Email : mini@ncbs.res.in (RS)

Abstract

Insect Olfactory Receptors (ORs) are diverse family of membrane protein receptors responsible for most of the insect olfactory perception and communication, and hence they are of utmost importance for developing repellents or pesticides. Hence, accurate gene prediction of insect ORs from newly sequenced genomes is an important but challenging task. We have developed a dedicated webserver, 'insectOR', to predict and validate insect OR genes using multiple gene prediction algorithms, accompanied by relevant validations. It is possible to employ this sever nearly automatically and perform rapid prediction of the OR gene loci from thousands of OR-protein-to-genome alignments, resolve gene boundaries for tandem OR genes and refine them further to provide more complete OR gene models. InsectOR outperformed the popular genome annotation pipelines (MAKER and NCBI eukaryotic genome annotation) in terms of overall sensitivity at base, exon and locus level, when tested on two distantly related insect genomes. It displayed more than 95% nucleotide level precision in both tests. Finally, given the same input data and parameters, InsectOR missed less than 2% gene loci, in contrast to 55% loci missed by MAKER for *Drosophila melanogaster*. The webserver is freely available on the web at http://caps.ncbs.res.in/insectOR/. All major browsers are supported. Website is implemented in Python with Jinja2 for templating and bootstrap framework which uses HTML, CSS and JavaScript/Ajax. The core pipeline is written in Perl.

Introduction

Insect biology has been studied extensively over the years for human benefit – to collect honey, pollinate crops, ward off pests, etc. Recently, these diverse species are also being used as model organisms for modern experiments to understand their (and in-turn our own) biology in intricate details. Advent of Next Generation Sequencing (NGS) technologies has given us powers to study this

vast diversity at genomic level [1]. Through projects like i5k, thousands of insect genomes and transcriptomes will be available soon and we need powerful bioinformatics tools to analyse the data[2,3].

Efforts are underway to exploit understanding of insect olfaction to manage pests and disease vectors [4–6]. Insect Olfaction is also an interesting system for study due to its commonalities and differences with the vertebrate olfactory system [7–9]. The discovery of insect olfactory receptors was itself largely dependent on the early bioinformatics analyses looking for novel protein coding regions with mammalian 'GPCR-like' properties in *Drosophila melanogaster* genome, which were further validated using antennae-specific expression [10–13]. Further OR discoveries in other genomes started to depend on their homology with the Drosophila ORs [14–16].

Later, vast differences in the average numbers and sub-families of ORs were observed across various insect orders (Hansson and Stensmyr, 2011, Montagné et al., 2015). Although OR repertoires from multiple species are available today, they still remain elusive in the genome due to this diversity. Insect ORs is a diverse family of proteins varying across insect orders [18]. In addition, the gene models of ORs also vary from one sub-family to another e.g. various OR subfamilies within the insect order Hymenoptera uniquely possess 4 to 9 exons [19]. This leads to lack of well-curated OR queries for use within general genome annotation pipelines. These automated genome annotations usually start with de novo gene predictions followed by homology-based corroborations. Probably, as these pipelines are trained on only one or few model organism annotations before use, they fail to capture the entire OR gene repertoire in an insect genome. Our previous work has shown that only 60-70% of the total OR gene content is recovered by the general gene annotation pipelines [19,20]. ORs are mostly selectively expressed only in antennae, differ from one insect order to another and undergo rapid births and deaths as per the requirements of each species, which causes missing and miss-annotations in the de novo and homology-based gene prediction of these genes. Hence, special

efforts (e.g. antennal transcriptome sequencing or extensive manual curation) are necessary to detect insect ORs with good sensitivity and precision [21].

Some of these problems could be alleviated by giving preference to homology-based gene predictions. In spite of that, we may find faulty gene predictions. ORs are usually present in tandem repeats in insect genomes and the alignments with OR protein queries may span two different gene regions and give erroneous gene predictions. This can also lead to miss-annotation of the gene and intron-exon boundaries. This problem could be addressed by transcriptome sequencing of the antenna, which is often costly and dependent on the availability of the antennae samples. It is also most likely to not cover the entire OR gene repertoire in cases of time-dependent/exposure-dependent expression of the OR genes [22]. Pipelines like OMIGA [23] are dedicated for insect genomes, but require transcriptome evidence to recognize OR genes. Hence most insect genome assembly and annotation projects are followed up by time-consuming, further experimental data or laborious homology dependent manual curation of ORs. To the best of our knowledge, currently there is only one recently developed, dedicated pipeline or webserver for prediction of genes from a single protein family as diverse as insect olfactory receptors, however it has been tested on the Niemann-Pick type C2 (NPC2) and insect gustatory receptor (GR) gene families and not olfactory receptors [24]. Hence a pipeline, with simplified and specific search for this OR family, without incorporating problems of general genome annotation pipelines, is of great value to the ever-growing insect genomics community.

We developed such a computational stand-alone pipeline during annotation of ORs from two solitary bees[20]. We have improved it further, added modules to assist automated refining and validation of genes and we are presenting it here in the form of a webserver, insectOR. Redundant hits are filtered, starting from alignment of multiple ORs to the genome of interest, to provide sensitive prediction of OR gene models.

Methods

Input parameters

Exonerate alignment file with additional Generic Feature Format (GFF) annotations[25] generated from insect genome of interest and query OR sequences are mandatory inputs. The related FASTA files of genome and OR proteins are also necessary for better refinement of the roughly predicted gene models. The choice of the best protein queries for this search is a crucial step that can be better addressed by the user with the help of directions given on the 'About' page of the webserver and hence it is currently not automated. This also reduces the resources spent on performing Exonerate on the webserver. More directions on how to run exonerate can also be found at the 'About' page of insectOR.

Users can also choose to provide genome annotation from any other source (GFF format) for additional comparisons with insectOR predictions. One can additionally choose to perform validation of the predicted proteins using HMMSEARCH [26] against 7tm_6, the Pfam[27,28] protein family domain which is characteristic of insect ORs. The presence or absence of the 7tm_6 domain is recorded. Users may also choose one or more of the three trans-membrane prediction (TMH) methods – TMHMM2[29,30], HMMTOP2[31,32] and Phobius[33]. If all three methods are selected, additional Consensus TMH prediction is performed[34]. InsectOR provides an option to perform additional annotation using known motifs of the insect ORs with the help of MAST tool from the MEME motif suite [35,36]. Users can search for default set of 10 protein motifs predicted for *A. florea* ORs [19] or they may upload their own motifs of interest.

Output

Statistics on the total number of predicted genes/gene fragments, complete and partial genes, gene regions with and without putative start sites and pseudogenous/normal gene status are provided in the final summary of the output (Fig 1A). Additionally, details of the genes encoding proteins with 7tm_6 domains are provided. Novel OR gene regions annotated by insectOR that are absent in the user-provided gene annotations are also counted. The details of each predicted OR gene can be studied from the table available in the next tab (Fig 1B). If the genome sequence is provided by the user, these gene predictions are displayed in the Dalliance web-embedded genome viewer [37] (Fig 1C). In case annotations from any other source are provided they are also displayed in the genome viewer and trimmed version of GFF file overlapping with insectOR prediction is available for download. Dalliance displays results in a customizable manner for easy comparison with user-provided gene annotations. Fig 1C illustrates, user-provided genes from NCBI GFF file. Zooming in onto particular regions gives more information on the coding nucleotides and the protein sequence translated by them. For the predicted OR gene regions from insectOR, final gene structure is reported in GFF and BED12+1 format and the putative CDS/transcript and protein sequence are also provided, all of which are available for download. One may use the GFF/BED12+1 formatted output/s on one of the various genome annotation editing tools (like Artemis[38], Ugene[39], Web Apollo[40] etc.) for further manual curation and editing of these genes. The gene regions with the status of 'partial' or 'pseudogenous' or 'without start codon' can be particularly targeted for curation. If user chooses to perform TMH validation by any of the three third-party methods mentioned before, a bar-plot representing the distribution of number of helices predicted by each selected TMH prediction method is plotted (Fig 1D). If all the three are selected, consensus TMH [34] is predicted and insectOR provides details of the four TMH predictions in a new result tab (Fig 1E). In case motif search is selected, the results are available at the last tab (Fig 1F).

Fig 1. A sample output from insectOR. Section A-B and D-F are outputs derived for input Exonerate alignments for *Drosophila melanogaster* whereas section C displays information derived from *Habropoda laboriosa* alignments. Two or more of these sections are available in the output depending on the analysis chosen by the user.

Annotation algorithm

Core annotation algorithm is written natively in Perl. It also invokes several other tools as mentioned in the 'Input' section. This algorithm processes the Exonerate alignment data to sensitively predict the OR coding gene regions and also performs validations as discussed next (Fig 2). The problem of missing and mis-annotation of tandemly repeated OR genes is addressed using 'divide and conquer' policy as described below.

Initially, OR protein-to-genome alignments are identified on the genome as follows. The exonerate output is read for each alignment. For every new genomic scaffold (target in the alignment), a virtual scaffold with the similar length with score '0' at each nucleotide position is created. Subsequently for each alignment, the score at every corresponding nucleotide position is incremented by one. This leads to virtual subalignments of OR protein-to-genome alignments demarcated by islands of higher scores (rough OR loci) on the base string of repeated '0' scores (non-'OR' loci). As stringent cutoffs are advised for the allowed intron lengths while performing Exonerate alignment (e.g. 2000 nucleotides or less), this step helps to distinguish ('divide') between tandem OR genes in the form of closely situated but distinct alignment islands/clusters.

This is followed by the next step of selecting the best alignment/gene model for a set of sub-alignments. The sub-alignments may sometimes be too short to include full length gene alignments due to stringent intron length cutoffs. Such smaller alignment regions correspond to fragments of gene models. To resolve this, initially, the best alignment per set is selected based on the Exonerate

alignment score. Corresponding query proteins for each of these best alignments in each cluster are identified as the best query proteins for the related clusters. For example, query protein OR2, OR3 and OR1 are shown as the best scoring queries in the alignment clusters 1, 2 and 3 from left to right in Fig 2. For the best queries selected per cluster, all other alignments on the same genomic scaffold are retained. In this way, from multiple redundant alignments, insectOR retains the best scoring alignment and also their neighbouring alignments from the same best scoring query.

Next, these best neighbouring alignments arising from each query are concatenated into complete protein alignments, if they are arranged congruently in the correct orientation and sequence on the genomic scaffold. In some cases, the boundary region in the alignments may be extended and the same region from the query may be aligned to the two different successive locations that need to be merged (as shown in the Fig 2 for query protein OR1; Amino acids 45 to 50 are aligned at two different locations on the scaffold whereas the flanking regions are different – 5 to 50 and 45 to 150). These are the cases of wrong extensions of the alignment fragments into introns. For such overlapping regions of the query, the possible exon-intron splicing sites are predicted based on the presence of ‘gt’ towards the 3’ terminus of the previous exon (region where a protein fragment is aligned) and presence of ‘ag’ towards 5’ terminus of the next exon (region where next protein fragment is aligned). The remaining regions are trimmed. All the possible combinations of such fragments are generated keeping the length of the overlapping region constant (e.g. In the above case of protein query OR1, there are 6 amino acids overlapping – 45 to 50. All combination of the concatenated nucleotide fragments giving rise to 18 nucleotide regions with flanking splice sites are considered). Next, the combination of splicing sites and their scores are compared to each other. The concatenated region providing the best similarity-based score on the Exonerate alignment is retained. In this way, insectOR finds the best possible splicing sites in cases of the fragmented alignments and stitches them to generate more complete alignments/gene models. In some cases, genes may possess more

than one isoform that are formed by alternative splicing. In such cases, similar region of a query protein may be aligned at two consecutive locations (e.g. duplicated exons that are alternatively spliced to give different isoforms). If the overlap is less than 20% of the any of the two query regions, when aligned, the two hits are kept separate. In case of overlap, multiple parameters, such as completeness of the gene, higher protein length, non-pseudogenous nature and presence of START codon are examined (in that order).

Fig 2. Annotation algorithm part of the insectOR webserver. Steps in the annotation algorithm are displayed here in cartoon representation.

For further refinement of gene boundaries, each gene/genic fragment (referred as prediction-1 (P1) hits are used as input for another gene structure prediction tool called “GeneWise”. GeneWise is known to perform well for one-to-one protein-to-DNA alignments[41]. The genomic locus of each P1 hit is allowed to extend on either side depending upon the length of the hit and maximum boundary extension of 6000 bp. This empirical cut-off was provided based on the average intergenic region observed for multiple insect genomes. Along with the extended genomic locus se-quence, the best aligned query for that region (determined earlier) are given as input to GeneWise. For each P1 hit, corresponding predicted (P2) hits are generated by running GeneWise. Further, for each locus, both P1 and P2 hits are compared. If P1 and P2 hits are overlapping, then the best of two is retained and otherwise both the hits are retained. Final hit is modified by locating the START and STOP codons (20 amino acids) upstream or downstream of the current start and end of the alignment and it is finally assigned a name according to its genomic location. Also, the presence of ATG (start codon) at the N-terminus and pseudogenizing elements (frameshifts or stop codons with respect to the query pro-

tein) are noted and included in the gene name. Based on the user-provided completion cut-off (default: 300 amino acids), a genic region is either declared as complete or partial.

In the last step of the pipeline, various validations on the predicted protein sequences are performed. Although TMH prediction programs are not very accurate (and may predict less or more than 7 helices for an insect OR), the presence of at-least few TMHs (depending on the protein fragment length) is necessary for validation. More robust validation comes from the search for '7tm_6' domain. Users may also choose to scan for protein motifs of interest in the predicted proteins. With more ongoing research on insect ORs, presence or absence of certain OR protein motifs may provide affirmation of their specific insect order origin [42] and might also provide clues regarding the kind of the odorants they bind to and may even assist in deorphanization of few of these ORs [43]. Evidence of more precise gene boundaries of ORs of closely related genomes will certainly improve OR prediction through homology-based annotation.

Implementation

The core annotation pipeline, as described in the previous section, is invoked from the insectOR website. The webserver is written in Python with Jinja2 for templating and bootstrap framework which uses HTML, CSS and JavaScript/Ajax. Dalliance and its API is used for genome annotation visualization [37]. InsectOR also makes use of file conversion tools like faToTwobit [44], gff_to_bed.py (https://github.com/vipints/GFFtools-GX/blob/master/gff_to_bed.py) and bedToBigBed[45,46] for visualisation of the predictions.

Results and discussion

Evaluation

We discuss the number of ORs we find in two insect genomes through InsectOR webserver in detail. Although a comparable webserver/method is not available for OR gene prediction specifically, another general gene annotation pipeline (MAKER) was tested by providing comparable parameters. MAKER was tuned for OR detection by specifying the maximum intron length of 2000 and by providing the same input query proteins for its Exonerate runs as provided for the corresponding insectOR runs. OR search in *Drosophila melanogaster* demonstrated the performance of our method on a well-annotated species. The second example demonstrated how the search for ORs in a blueberry bee (H. labriosa) was made simpler and automatic, using the core pipeline that forms the basis of this webserver. Taking our own published final annotations of ORs from blueberry bee as a reference [20], the raw results from the current modified webserver and two other general annotation pipelines were compared. The general performance of insectOR was found to be better than the others as described below.

Case study 1: *Drosophila melanogaster* ORs

To test insectOR on a well-studied model organism, we chose fruit-fly *Drosophila melanogaster* genome (assembly Release 6 plus ISO1 MT) belonging to insect order Diptera.

The Ensembl reference gene annotations were taken as standard and only OR related information was retained. It possesses 61 OR gene loci encoding 65 OR mRNAs (including isoforms). For testing insectOR, the query protein dataset was built from well-curated 727 non-Drosophila OR protein sequences from NCBI non-redundant protein database belonging to the order Diptera.

Exonerate [25] alignment of these proteins against the Drosophila genome was performed and it was provided as an input to insectOR. For de novo gene prediction within MAKER [47], two methods - AUGUSTUS 2.5.5 [48] and SNAP [49] were implemented. HMM gene model of ‘aedes’ was used for training AUGUSTUS and that of ‘mosquito’ was used for training SNAP de novo gene

predictions as the gene models from the same non-'Drosophila' species were not available for the two methods. The predictions from insectOR and MAKER[47] were compared with those of the NCBI as reference using 'gffcompare' (http://ccb.jhu.edu/software/stringtie/gff.shtml). The results of the comparison are discussed in Table 1.

Table 1. *D. melanogaster* OR gene prediction assessment.

Reference mRNAs (Ensembl): 65				
OR prediction method	**insectOR**		**Maker**	
No of predicted genes/gene-fragments	62		25	
Proteins with one or more 7tm_6 predictions	56		24	
Proteins with multiple 7tm_6 predictions	0		9	
Missed exons	8.6%		56.4%	
Missed loci	1.60%		55.70%	
Matching loci	35		17	
	Sensitivity	**Precision**	**Sensitivity**	**Precision**
Base	87	99	43	84
Exon	74	76	37	78
Locus	57	61	28	68

Out of the total 62 OR gene/fragments predicted by insectOR, 56 can be validated using 7tm_6 and they also show 99% base level precision, which means that almost all the OR gene loci are identified at correct locations. Fifty-five of these had length more than 300 amino acids. InsectOR showed better sensitivity at base, exon and locus levels. Some genes containing ORs, predicted by

insectOR, were not complete at the boundaries and hence it showed less precision at the exon and locus level, as compared to MAKER[47]. At the exon and locus level precision calculation, gffcompare method searches for exact matches (with only 10 bp allowed deviation at the boundaries) to be qualified for a true positive hit [50]. However, this better precision at the exon and locus level for MAKER [47] was at the cost of sensitivity and it missed more than 50% of the OR gene loci completely. The output of gffcompare for *Drosophila melanogaster* is available in S1 File. This execution took around 3 hours to process Exonerate alignment file (9.1MB size containing 2099 alignments) on insectOR.

Case study 2: *H. laboriosa* ORs

We evaluated performance of insectOR for a species from another insect order – Hymenoptera (includes bees, ants and wasps). As discussed before, the basis of this pipeline was developed during annotation of ORs from two solitary bees – *Habropoda laboriosa* (Blueberry bee) and Dufourea novaeangliae, of which we have compared *H. laboriosa* predictions here [20]. Compared to our previous analysis on *A. florea* ORs, which required manual intervention, we found significant extent of automation for the complete annotation of *H. laboriosa* using insectOR. When the final set of genes (coming from our complete semi-automated annotation) were compared with those from NCBI eukaryotic genome annotation pipeline (https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Habropoda_laboriosa/100/) [51], significant improvement was observed in the coverage of the total number of OR genes and accuracy of gene models, as discussed. To summarise, after our complete semi-automated analysis, 42 completely new OR gene regions were found (27% of the total blueberry ORs found) as compared to the NCBI genome annotations. Eighty-two OR genes (54% of total blueberry ORs) already covered by NCBI gene annotations had serious problems with the gene and intron-exon boundaries that were corrected.

An example of this is shown in Fig 1C where middle panel of 'User-uploaded-genes' shows prediction of ORs by NCBI annotation pipeline and the last panel shows predictions from insectOR. In this case, the NCBI gene annotation has predicted one fused gene model for four distinct OR gene loci, as it has missed to predict the last exon in each of these genes. Also, it has missed the second gene region completely which is a pseudogene due to presence of an in-frame STOP codon TGA (as seen in the zoomed-in version – STOP codon is shown to translate into '*'). For more details on the number of novel and modified genes, please see the supplementary information in Karpe et. al., 2017.

Here we have compared raw OR gene predictions from insectOR (without further manual curation) with those from MAKER[47] and NCBI[51] (Table 2). The final manually curated gene predictions from the above mentioned paper were taken as the reference. These 1249 curated OR protein sequences (without self-OR sequences) were used as input for Exonerate within MAKER[47]. Similar to Drosophila, MAKER[47] annotations were carried out using de novo gene predictions from AUGUSTUS 2.5.5[48] and SNAP[49], both trained on gene models from *A. mellifera*. In the raw output of our current insectOR webserver, 151 OR gene/gene-fragments were predicted. Out of these, 103 were complete (>300 amino acids in length) and 134 displayed presence of 7tm_6 domain. We could find only 133 OR proteins predicted by MAKER and only 62 by NCBI. Out of these 133 ORs predicted using MAKER, 65 were complete. But, 23 of the probable complete ones were more than 500 amino acids in length and were fused protein predictions indicating that providing similar maximum intron length cut-off for Exonerate was not enough for fine-tuning for OR gene prediction within MAKER. Similar fused proteins were observed for NCBI gene predictions. This is reflected in the number of proteins with multiple 7tm_6 domains from MAKER and NCBI. As shown in the Table 2, for all the measures of performance of the prediction, insectOR performed better than MAKER and NCBI annotations. This example is provided for sample execution at insectOR. The output of gffcompare for *Habropoda laboriosa* is available in S2 File. The sample

execution took less than ten minutes to process Exonerate alignment file (45.9MB size containing 13180 alignments) on insectOR. Furthermore, we applied InsectOR on five other insect genomes and these results are organized in S1-S4 Tables.

Table 2. *H. laboriosa* OR gene prediction assessment.

Reference mRNAs [20]**: 151**						
OR prediction method	insectOR		Maker		NCBI	
No of predicted genes/gene-fragments	151		133		62	
Proteins with one or more 7tm_6 predictions	134		92		62	
Missed exons	13.9%		32.30%		49.30%	
Missed loci	0.7%		15.30%		14.00%	
Matching loci	57		6		14	
	Sensitivity	**Precision**	**Sensitivity**	**Precision**	**Sensitivity**	**Precision**
Base	87	95	73	85	54	80
Exon	65	68	31	40	33	55
Locus	38	39	4	5	9	23

Conclusion

InsectOR is a first-of-a-kind webserver for the prediction of ORs from newly sequenced genome of insect species. Insect OR genes are diverse across various taxonomical categories and hence these are hard to detect for general genome annotation pipelines, which also tend to wrongly predict

fused tandem OR gene models. InsectOR outperforms such general genome annotation methods in providing accurate gene boundaries, reducing the efforts spent on manual curation of this huge family of proteins. Overall, InsectOR performed well across two different insect orders and provided best sensitivity and good precision amongst the methods tested here for OR gene prediction.

InsectOR performance is dependent on the initial query set, hence there is a manual intervention of the right choice of queries. Where possible, it is best to employ query sequences which are evolutionarily close. Though InsectOR annotations are not yet complete for few genes near the gene-boundaries, it displays the relevant information showing whether each gene is incomplete or pseudogenous. Further measures (limited manual editing or expression analysis) can be performed by the user to ensure completeness of these models. With current ongoing projects of sequencing 1000s of insect genomes and transcriptomes, the webserver has potential to serve many entomologists all over the world. We believe, it will reduce the overall time taken for final manual curation of OR genes, to about one-fourth, of the usual from our previous experience. It is a first step towards annotation methods tuned for huge protein families like ORs and in future it could be adapted to other similar diverse protein families.

Acknowledgements

We would like to thank, Mr. Murugavel Pavalam for extensive help with improving the core algorithm of insectOR and also for helping to implement it on the web platform. We would like to thank National Centre for Biological Sciences (NCBS) for infrastructural facilities.

References

1. Grimmelikhuijzen CJ, Cazzamali G, Williamson M, Hauser F. The promise of insect

genomics. Pest Manag Sci. 2007;63: 413–416. doi:10.1002/ps.1352

2. Robinson GE, Hackett KJ, Purcell-Miramontes M, Brown SJ, Evans JD, Goldsmith MR, et al. Creating a buzz about insect genomes. Science. 2011;331: 1386. doi:10.1126/science.331.6023.1386

3. Pennisi E. Biologists propose to sequence the DNA of all life on Earth. Science (80-). 2017. doi:10.1126/science.aal0824

4. Davis EE. Insect Repellents: Concepts of their Mode of Action Relative to Potential Sensory Mechanisms in Mosquitoes (Diptera: Culicidae)1. J Med Entomol. 1985;22: 237–243. doi:10.1093/jmedent/22.3.237

5. Hallem EA, Nicole Fox A, Zwiebel LJ, Carlson JR. Olfaction: Mosquito receptor for human-sweat odorant. Nature. 2004;427: 212–213. doi:10.1038/427212a

6. Wang G, Carey AF, Carlson JR, Zwiebel LJ. Molecular basis of odor coding in the malaria vector mosquito Anopheles gambiae. Proc Natl Acad Sci U S A. 2010;107: 4418–23. doi:10.1073/pnas.0913392107

7. Pelosi P, Maida R. Odorant-binding proteins in vertebrates and insects: similarities and possible common function. Chem Senses. 1990;15: 205–215. doi:10.1093/chemse/15.2.205

8. Ache BW, Young JM. Olfaction: Diverse Species, Conserved Principles. Neuron. 2005;48: 417–430. doi:10.1016/J.NEURON.2005.10.022

9. Kaupp UB. Olfactory signalling in vertebrates and insects: differences and commonalities. Nat Rev Neurosci. 2010;11: 188–200. doi:10.1038/nrn2789

10. Clyne PJ, Warr CG, Freeman MR, Lessing D, Kim J, Carlson JR. A Novel Family of Divergent Seven-Transmembrane Proteins. Neuron. 1999;22: 327–338. doi:10.1016/S0896-6273(00)81093-4

11. Vosshall LB, Amrein H, Morozov PS, Rzhetsky A, Axel R. A Spatial Map of Olfactory

Receptor Expression in the Drosophila Antenna. Cell. 1999;96: 725–736. doi:10.1016/S0092-8674(00)80582-6

12. Kim J, Moriyama EN, Warr CG, Clyne PJ, Carlson JR. Identification of novel multi-transmembrane proteins from genomic databases using quasi-periodic structural properties. Bioinformatics. 2000;16: 767–775. doi:10.1093/bioinformatics/16.9.767

13. Kim J, Carlson JR. Gene discovery by e-genetics: Drosophila odor and taste receptors. J Cell Sci. 2002;115: 1107–12.

14. Krieger J, Raming K, Dewer YME, Bette S, Conzelmann S, Breer H. A divergent gene family encoding candidate olfactory receptors of the moth *Heliothis virescens*. Eur J Neurosci. 2002;16: 619–628. doi:10.1046/j.1460-9568.2002.02109.x

15. Robertson HM, Warr CG, Carlson JR. Molecular evolution of the insect chemoreceptor gene superfamily in Drosophila melanogaster. Proc Natl Acad Sci U S A. 2003;100 Suppl: 14537–42. doi:10.1073/pnas.2335847100

16. Robertson HM, Wanner KW. The chemoreceptor superfamily in the honey bee, Apis mellifera: Expansion of the odorant, but not gustatory, receptor family. Genome Res. 2006;16: 1395–1403. doi:10.1101/gr.5057506

17. Montagné N, Fouchier A De, Newcomb RD, Jacquin-joly E, Montagné N, Fouchier A De, et al. Advances in the Identification and Characterization of Olfactory Receptors in Insects. Progress in Molecular Biology and Translational Science. 2015. pp. 55–80.

18. Hansson BS, Stensmyr MC. Evolution of Insect Olfaction. Neuron. 2011;72: 698–711. doi:10.1016/j.neuron.2011.11.003

19. Karpe SD, Jain R, Brockmann A, Sowdhamini R. Identification of Complete Repertoire of Apis florea Odorant Receptors Reveals Complex Orthologous Relationships with Apis mellifera. Genome Biol Evol. 2016;8: 2879–2895. doi:10.1093/gbe/evw202

20. Karpe SD, Dhingra S, Brockmann A, Sowdhamini R. Computational genome-wide survey of odorant receptors from two solitary bees Dufourea novaeangliae (Hymenoptera: Halictidae) and Habropoda laboriosa (Hymenoptera: Apidae). Sci Rep. 2017;7: 10823. doi:10.1038/s41598-017-11098-z

21. Missbach C, Dweck HK, Vogel H, Vilcinskas A, Stensmyr MC, Hansson BS, et al. Evolution of insect olfactory receptors. Elife. 2014;3: e02115. doi:10.7554/eLife.02115

22. Engsontia P, Sanderson AP, Cobb M, Walden KKO, Robertson HM, Brown S. The red flour beetle's large nose: an expanded odorant receptor gene family in Tribolium castaneum. Insect Biochem Mol Biol. 2008;38: 387–397. doi:10.1016/j.ibmb.2007.10.005

23. Liu J, Xiao H, Huang S, Li F. OMIGA: Optimized Maker-Based Insect Genome Annotation. Mol Genet Genomics. 2014;289: 567–573. doi:10.1007/s00438-014-0831-7

24. Vizueta J, Sánchez-Gracia A, Rozas J. BITACORA: A comprehensive tool for the identification and annotation of gene families in genome assemblies. bioRxiv. 2019; 593889. doi:10.1101/593889

25. Slater GSC, Birney E. Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics. 2005;6: 31. doi:10.1186/1471-2105-6-31

26. Eddy SR, Crooks G, Green R, Brenner S, Altschul S. Accelerated Profile HMM Searches. Pearson WR, editor. PLoS Comput Biol. 2011;7: e1002195. doi:10.1371/journal.pcbi.1002195

27. Sonnhammer EL, Eddy SR, Durbin R. Pfam: a comprehensive database of protein domain families based on seed alignments. Proteins. 1997;28: 405–20.

28. Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, et al. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res. 2016;44: D279–D285. doi:10.1093/nar/gkv1344

29. Sonnhammer EL, von Heijne G, Krogh A. A hidden Markov model for predicting

411 transmembrane helices in protein sequences. Proc Int Conf Intell Syst Mol Biol. 1998;6: 175–
412 82.
413 30. Krogh A, Larsson B, von Heijne G, Sonnhammer EL. Predicting transmembrane protein
414 topology with a hidden Markov model: application to complete genomes. J Mol Biol.
415 2001;305: 567–80. doi:10.1006/jmbi.2000.4315
416 31. Tusnády GE, Simon I. Principles governing amino acid composition of integral membrane
417 proteins: application to topology prediction. J Mol Biol. 1998;283: 489–506.
418 doi:10.1006/jmbi.1998.2107
419 32. Tusnády GE, Simon I. The HMMTOP transmembrane topology prediction server.
420 Bioinformatics. 2001;17: 849–50.
421 33. Käll L, Krogh A, Sonnhammer EL. A Combined Transmembrane Topology and Signal
422 Peptide Prediction Method. J Mol Biol. 2004;338: 1027–1036. doi:10.1016/j.jmb.2004.03.016
423 34. Nagarathnam B, Karpe D, Harini K, Sankar K, Iftekhar M, Rajesh D, et al. DOR – a Database
424 of Olfactory Receptors – Integrated Repository for Sequence and Secondary Structural
425 Information of Olfactory Receptors in Selected Eukaryotic Genomes. Bioinform Biol Insights.
426 2014;8: 147–158. doi:10.4137/BBi.s14858
427 35. Bailey TL, Gribskov M. Combining evidence using p-values: application to sequence
428 homology searches. Bioinformatics. 1998;14: 48–54.
429 36. Bailey TL, Boden M, Buske F a., Frith M, Grant CE, Clementi L, et al. MEME Suite: Tools
430 for motif discovery and searching. Nucleic Acids Res. 2009;37: 202–208.
431 doi:10.1093/nar/gkp335
432 37. Down TA, Piipari M, Hubbard TJP. Dalliance: interactive genome viewing on the web.
433 Bioinformatics. 2011;27: 889–890. doi:10.1093/bioinformatics/btr020
434 38. Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, et al. Artemis:

435 sequence visualization and annotation. Bioinformatics. 2000;16: 944–5.
436 doi:10.1093/BIOINFORMATICS/16.10.944
437 39. Okonechnikov K, Golosova O, Fursov M. Unipro UGENE: a unified bioinformatics toolkit.
438 Bioinformatics. 2012;28: 1166–1167. doi:10.1093/bioinformatics/bts091
439 40. Lee E, Helt GA, Reese JT, Munoz-Torres MC, Childers CP, Buels RM, et al. Web Apollo: a
440 web-based genomic annotation editing platform. Genome Biol. 2013;14: R93. doi:10.1186/gb-
441 2013-14-8-r93
442 41. Birney E, Clamp M, Durbin R. GeneWise and Genomewise. Genome Res. 2004;14: 988–995.
443 doi:10.1101/gr.1865504
444 42. Miller R, Tu Z. Odorant Receptor C-Terminal Motifs in Divergent Insect Species. J Insect Sci.
445 2008;8: 1–10. doi:10.1673/031.008.5301
446 43. Ray A, van der Goes van Naters W, Carlson JR. Molecular determinants of odorant receptor
447 function in insects. J Biosci. 2014;39: 555–563. doi:10.1007/s12038-014-9447-7
448 44. Kent WJ. BLAT--the BLAST-like alignment tool. Genome Res. 2002;12: 656–64.
449 doi:10.1101/gr.229202. Article published online before March 2002
450 45. Kent WJ, Zweig AS, Barber G, Hinrichs AS, Karolchik D. BigWig and BigBed: enabling
451 browsing of large distributed datasets. Bioinformatics. 2010;26: 2204–2207.
452 doi:10.1093/bioinformatics/btq351
453 46. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features.
454 Bioinformatics. 2010;26: 841–2. doi:10.1093/bioinformatics/btq033
455 47. Cantarel BL, Korf I, Robb SMC, Parra G, Ross E, Moore B, et al. MAKER: an easy-to-use
456 annotation pipeline designed for emerging model organism genomes. Genome Res. 2008;18:
457 188–96. doi:10.1101/gr.6743907
458 48. Stanke M, Waack S. Gene prediction with a hidden Markov model and a new intron

submodel. Bioinformatics. 2003;19 Suppl 2: ii215-25.

49. Korf I. Gene finding in novel genomes. BMC Bioinformatics. 2004;5: 59. doi:10.1186/1471-2105-5-59

50. Burset M, Guigo R. Evaluation of Gene Structure Prediction Programs. Genomics. 1996;34: 353–367. doi:10.1006/GENO.1996.0298

51. Thibaud-Nissen F, Souvorov A, Murphy T, Dicuccio M, Kitts P. Eukaryotic Genome Annotation Pipeline. The NCBI Handbook [Internet] 2nd edition. Bethesda (MD): National Center for Biotechnology Information (US); 2013.

Supporting information captions

S1 File. The output of gffcompare for *Drosophila melanogaster.* Detailed result of comparison of gene annotations by MAKER and insectOR to the NCBI annotations as reference for the *Drosophila melanogaster* genome.

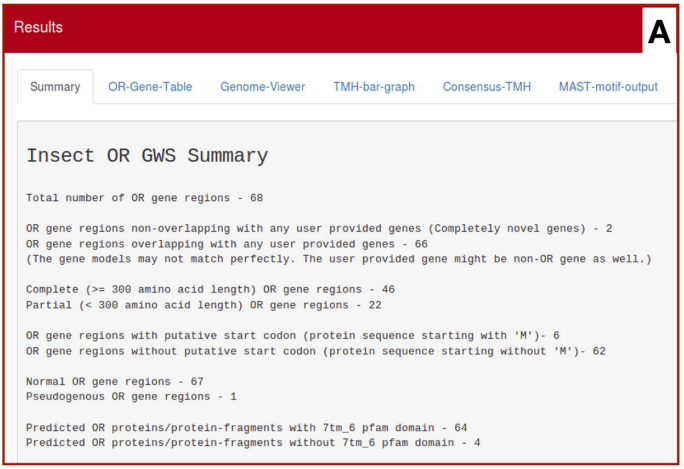
S2 File. The output of gffcompare for *Habropoda laboriosa.* Detailed result of comparison of gene annotations by MAKER, NCBI and insectOR to the curated annotations as reference for the *Habropoda laboriosa* genome.

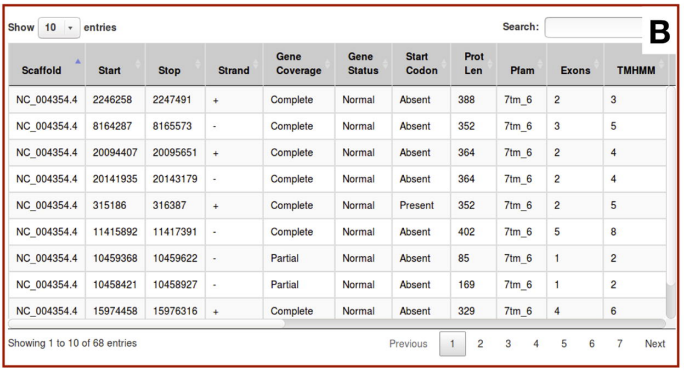
S1 Table. InsectOR prediction of ORs in *Dufourea novaeangliae.*

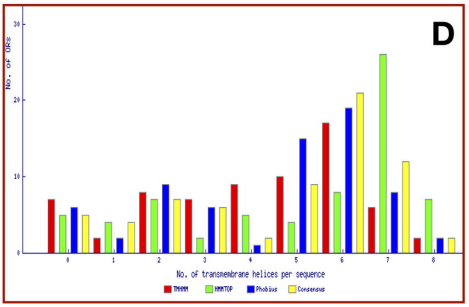
S2 Table. InsectOR prediction of ORs in *Apis florea.*

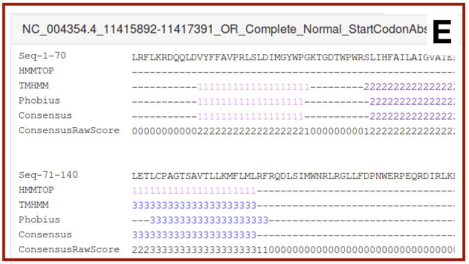
S3 Table. InsectOR prediction of ORs in *Anopheles gambiae.*

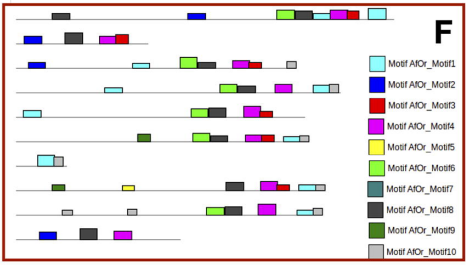
S4 Table. InsectOR prediction of ORs in *Leptinotarsa decemlineata.*

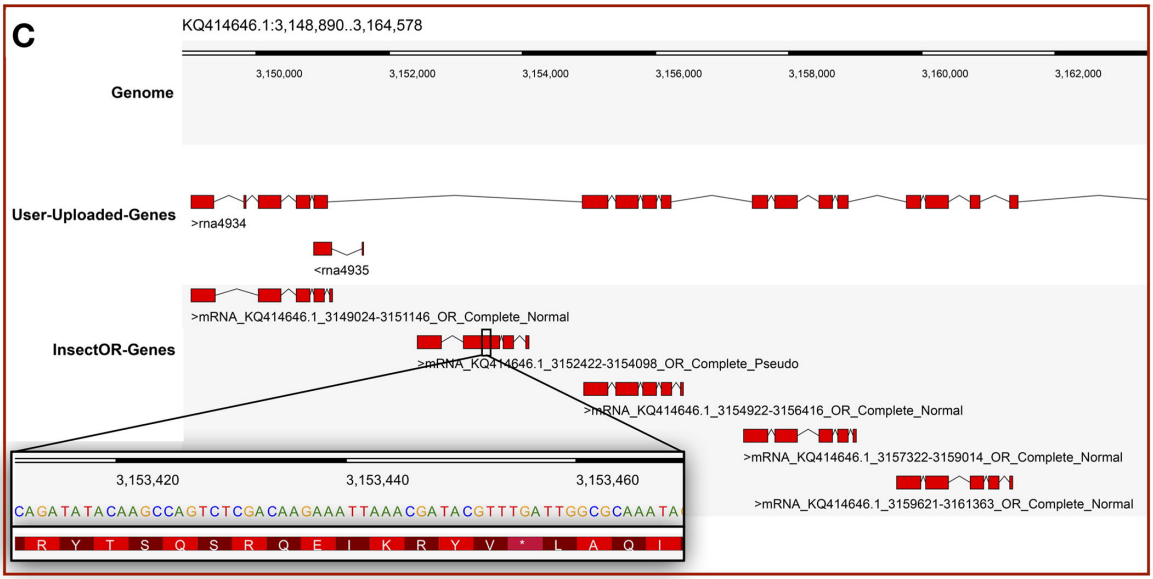

Results
A
Summary
OR-Gene-Table
Genome-Viewer
TMH-bar-graph
Consensus-TMH
MAST-motif-output
Insect OR GWS Summary
Total number of OR gene regions - 68
OR gene regions non-overlapping with any user provided genes (Completely novel genes) - 2
OR gene regions overlapping with any user provided genes - 66
(The gene models may not match perfectly. The user provided gene might be non-OR gene as well.)
Complete (>= 300 amino acid length) OR gene regions - 46
Partial (< 300 amino acid length) OR gene regions - 22
OR gene regions with putative start codon (protein sequence starting with 'M')- 6
OR gene regions without putative start codon (protein sequence starting without 'M')- 62
Normal OR gene regions - 67
Pseudogenous OR gene regions - 1
Predicted OR proteins/protein-fragments with 7tm_6 pfam domain - 64
Predicted OR proteins/protein-fragments without 7tm_6 pfam domain - 4


B
Show 10 entries
Search:
Scaffold Start Stop Strand Gene Coverage Gene Status Start Codon Prot Len Pfam Exons TMHMM
NC_004354.4 2246258 2247491 + Complete Normal Absent 388 7tm_6 2 3
NC_004354.4 8164287 8165573 - Complete Normal Absent 352 7tm_6 3 5
NC_004354.4 20094407 20095651 + Complete Normal Absent 364 7tm_6 2 4
NC_004354.4 20141935 20143179 - Complete Normal Absent 364 7tm_6 2 4
NC_004354.4 315186 316387 + Complete Normal Present 352 7tm_6 2 5
NC_004354.4 11415892 11417391 - Complete Normal Absent 402 7tm_6 5 8
NC_004354.4 10459368 10459622 - Partial Normal Absent 85 7tm_6 1 2
NC_004354.4 10458421 10458927 - Partial Normal Absent 169 7tm_6 1 2
NC_004354.4 15974458 15976316 + Complete Normal Absent 329 7tm_6 4 6
Showing 1 to 10 of 68 entries
Previous 1 2 3 4 5 6 7 Next


D
No. of ORs
No. of transmembrane helices per sequence
TMHMM
HMMTOP
Phobius
Consensus


E
NC_004354.4_11415892-11417391_OR_Complete_Normal_StartCodonAbs
Seq-1-70
HMMTOP
TMHMM
Phobius
Consensus
ConsensusRawScore
Seq-71-140
HMMTOP
TMHMM
Phobius
Consensus
ConsensusRawScore


F
Motif AfOr_Motif1
Motif AfOr_Motif2
Motif AfOr_Motif3
Motif AfOr_Motif4
Motif AfOr_Motif5
Motif AfOr_Motif6
Motif AfOr_Motif7
Motif AfOr_Motif8
Motif AfOr_Motif9
Motif AfOr_Motif10


C
KQ414646.1:3,148,890..3,164,578
Genome
3,150,000
3,152,000
3,154,000
3,156,000
3,158,000
3,160,000
3,162,000
User-Uploaded-Genes
>rna4934
<rna4935
InsectOR-Genes
>mRNA_KQ414646.1_3149024-3151146_OR_Complete_Normal
>mRNA_KQ414646.1_3152422-3154098_OR_Complete_Pseudo
>mRNA_KQ414646.1_3154922-3156416_OR_Complete_Normal
>mRNA_KQ414646.1_3157322-3159014_OR_Complete_Normal
>mRNA_KQ414646.1_3159621-3161363_OR_Complete_Normal
3,153,420
3,153,440
3,153,460
CAGATATACAAGCCAGTCTCGACAAGAAATTAAACGATACGTTTGATTGGCGCAAATA
R Y T S Q S R Q E I K R Y V * L A Q I


Initialization of virtual scaffold with zero at each position
Increment the score for each aligned query
OR 1
OR 2
OR 4
OR 3
OR 4
OR 1
OR 3
OR 1
Alignment cluster and best scoring query
OR 2
OR 3
OR 1
Identification of all alignments for best queries
OR 1
OR 2
OR 3
OR 1
OR 1
Stitch the aligned region based on overlapping regions
OR 1: 5-50 AA
OR 2
OR 3
OR 1 : 45-150 AA
OR 1: 148-190 AA
Identification of best alignment
Predicted Protein length = 320
Predicted Protein length = 100
Predicted Protein length = 100
Remove overlapping region and predict OR gene sequence and structure
Predicted Protein length = 320
Predicted OR protein (Prediction-1 (P1))
Extension of P1 hit boundaries based on length of hit
OR 1-P1
OR X-P1
Max 6k
OR Y-P1
OR 1
OR X
OR Y
Run GeneWise with modified locus and query OR for Prediction-2 (P2)
OR 1-P2
OR X-P2
OR Y-P2
OR 1
OR X
OR Y
Compare OR-P1 and OR-P2 and retain the best hit. For non-overlapping hits, retain both.
OR 1-P1
OR X-P2
OR X-P1
OR Y-P2
OR 1
OR X
OR X
OR Y
Modify the ends for start and stop codons
M
OR 1-P1
M
OR X-P2
M
OR X-P1
M
OR Y-P2
Validation and analysis of predicted ORs
HMMTOP
TMHMM
Phobius
7tm_6 Pfam domain prediction
Consensus Transmembrane Helix prediction
Protein motif Search