



**HAL**  
open science

## Genome sequence of the wheat stem sawfly, *Cephus cinctus*, representing an early-branching lineage of the Hymenoptera, illuminates evolution of hymenopteran chemoreceptors

Robert Robertson, Robert Waterhouse, Kimberly Walden, Livio Ruzzante, Maarten Reijnders, Brad Coates, Fabrice Legeai, Joanna Gress, Sezgi Biyiklioglu, David Weaver, et al.

### ► To cite this version:

Robert Robertson, Robert Waterhouse, Kimberly Walden, Livio Ruzzante, Maarten Reijnders, et al.. Genome sequence of the wheat stem sawfly, *Cephus cinctus*, representing an early-branching lineage of the Hymenoptera, illuminates evolution of hymenopteran chemoreceptors. *Genome Biology and Evolution*, 2018, 10 (11), pp.2997-3011. 10.1093/gbe/evy232 . hal-01926352

**HAL Id: hal-01926352**

**<https://inria.hal.science/hal-01926352>**

Submitted on 18 Jul 2019

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

# Genome Sequence of the Wheat Stem Sawfly, *Cephus cinctus*, Representing an Early-Branching Lineage of the Hymenoptera, Illuminates Evolution of Hymenopteran Chemoreceptors

Hugh M. Robertson<sup>1,\*†</sup>, Robert M. Waterhouse<sup>2,3,\*†</sup>, Kimberly K.O. Walden<sup>1</sup>, Livio Ruzzante<sup>2,3</sup>, Maarten J.M.F. Reijnders<sup>2,3</sup>, Brad S. Coates<sup>4</sup>, Fabrice Legeai<sup>5</sup>, Joanna C. Gress<sup>6,7</sup>, Sezgi Biyiklioglu<sup>7</sup>, David K. Weaver<sup>8</sup>, Kevin W. Wanner<sup>7</sup>, and Hikmet Budak<sup>7</sup>

<sup>1</sup>Department of Entomology, University of Illinois at Urbana-Champaign

<sup>2</sup>Department of Ecology and Evolution, University of Lausanne, Switzerland

<sup>3</sup>Swiss Institute of Bioinformatics, Lausanne, Switzerland

<sup>4</sup>United States Department of Agriculture, Agricultural Research Service (USDA ARS), Corn Insects & Crop Genetics Research Unit, Ames, Iowa

<sup>5</sup>Ecology and Genetics of Insects, INRA, UMR IGEPP 1349, Rennes, France

<sup>6</sup>Department of Science and Mathematics, Abraham Baldwin Agricultural College, Tifton, Georgia

<sup>7</sup>Department of Plant Sciences and Plant Pathology, Montana State University

<sup>8</sup>Department of Land Resources and Environmental Sciences, Montana State University

<sup>†</sup>These authors contributed equally to this work.

\*Corresponding authors: E-mails: hughrobe@uiuc.edu; robert.waterhouse@unil.ch.

**Accepted:** October 12, 2018

**Data deposition:** This project has been submitted to the National Center for Biotechnology Information (NCBI) as BioProjects PRJNA297591 and PRJNA168335 (NCBI GCF\_000341935.1), and all the Sequence Read Archive raw read accessions are listed in supplementary table S1, Supplementary Material online.

## Abstract

The wheat stem sawfly, *Cephus cinctus*, is a major pest of wheat and key ecological player in the grasslands of western North America. It also represents the distinctive Cephoidea superfamily of sawflies (Symphyta) that appeared early during the hymenopteran radiation, but after three early-branching eusymphytan superfamilies that form the base of the order Hymenoptera. We present a high-quality draft genome assembly of 162 Mb in 1,976 scaffolds with a scaffold N50 of 622 kb. Automated gene annotation identified 11,210 protein-coding gene models and 1,307 noncoding RNA models. Thirteen percent of the assembly consists of ~58,000 transposable elements partitioned equally between Class-I and Class-II elements. Orthology analysis reveals that 86% of *Cephus* proteins have identifiable orthologs in other insects. Phylogenomic analysis of conserved subsets of these proteins supports the placement of the Cephoidea between the Eusymphyta and the parasitic woodwasp superfamily Orussoidea. Manual annotation and phylogenetic analysis of families of odorant, gustatory, and ionotropic receptors, plus odorant-binding proteins, shows that *Cephus* has representatives for most conserved and expanded gene lineages in the Apocrita (wasps, ants, and bees). *Cephus* has also maintained several insect gene lineages that have been lost from the Apocrita, most prominently the carbon dioxide receptor subfamily. Furthermore, *Cephus* encodes a few small lineage-specific chemoreceptor gene family expansions that might be involved in adaptations to new grasses including wheat. These comparative analyses identify gene family members likely to have been present in the hymenopteran ancestor and provide a new perspective on the evolution of the chemosensory gene repertoire.

**Key words:** Cephidae, Cephoidea, odorant receptor, gustatory receptor, ionotropic receptor, odorant-binding protein.

© The Author(s) 2018. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

## Introduction

The wheat stem sawfly, *Cephus cinctus* Norton, is a major pest of wheat in western North America with a southward expanding geographic range likely driven by localized adaptation to cultivated crops from surrounding wildlands (Beres et al. 2011; Lesieur et al. 2016; Adhikari et al. 2018; Varella et al. 2018). It is a native species that also uses many other grass hosts and hence plays an important role in the ecology of grasslands (Cockrell et al. 2017). Worldwide, as insect burdens on wheat farmers, members of the genus *Cephus* rival the Hessian fly, *Mayetiola destructor*, a dipteran for which a genome sequence is available (Zhao et al. 2015). Female sawflies lay eggs in growing wheat stems and the larvae grow within the stem lumen (fig. 1). Once in the final instar, the sole survivor of larval cannibalism in each stem retreats to the base of the stem, which it girdles in preparation for pupation (Buteler et al. 2009, 2015). Economic loss is from both the larval presence and the subsequent collapse or lodging of the wheat stem (Bekkerman and Weaver 2018). As a result of this life history, wheat stem sawflies are protected from control using conventional sprayed insecticides. The development of alternative approaches relies on an improved understanding of *C. cinctus* chemical ecology (Cossé et al. 2002; Weaver et al. 2009) and associated behaviors on host plants (Buteler et al. 2009; Varella et al. 2017), but large-scale molecular biology studies to date remain limited to antennal transcriptomics that examined chemosensory genes (Gress et al. 2013) and analyses of larval and adult RNA sequencing (RNA-seq) data that focused on the identification of noncoding RNA transcripts (Cagirici et al. 2017). A key first step in this direction is to build genomic and additional transcriptomic resources for this sawfly that could, for example, facilitate the development of new resistant wheat strains using expression of RNA interference (RNAi) constructs against essential wheat stem sawfly genes.

Sawflies (Symphyta) have many distinctive aspects, from their distinguishing morphological features of the absence of a “wasp-waist” and a saw-like female ovipositor to their mostly herbivorous ecology including caterpillar-like larvae, however they are a paraphyletic assemblage of seven superfamilies with at least 8000 species. An additional significance to the wheat stem sawfly is that the genus *Cephus* represents the family Cephidae and superfamily Cephoidea, and phylogenetic analysis of the Hymenoptera shows that this is a distinctive lineage of sawflies that split after the earlier-branching sawflies of the superfamilies Xyeloidea, Tenthredinoidea, and Pamphloidea, now proposed to be called the Eusymphyta, and before the woodwasp superfamilies of Siricoidea and Xiphidriidea and finally the parasitic woodwasps or Orussoidea (Vilhelmsen 2001; Schulmeister et al. 2002; Sharkey et al. 2011; Branstetter et al. 2017; Peters et al. 2017). This sawfly therefore represents an interim step in the origin of the major monophyletic lineage of the

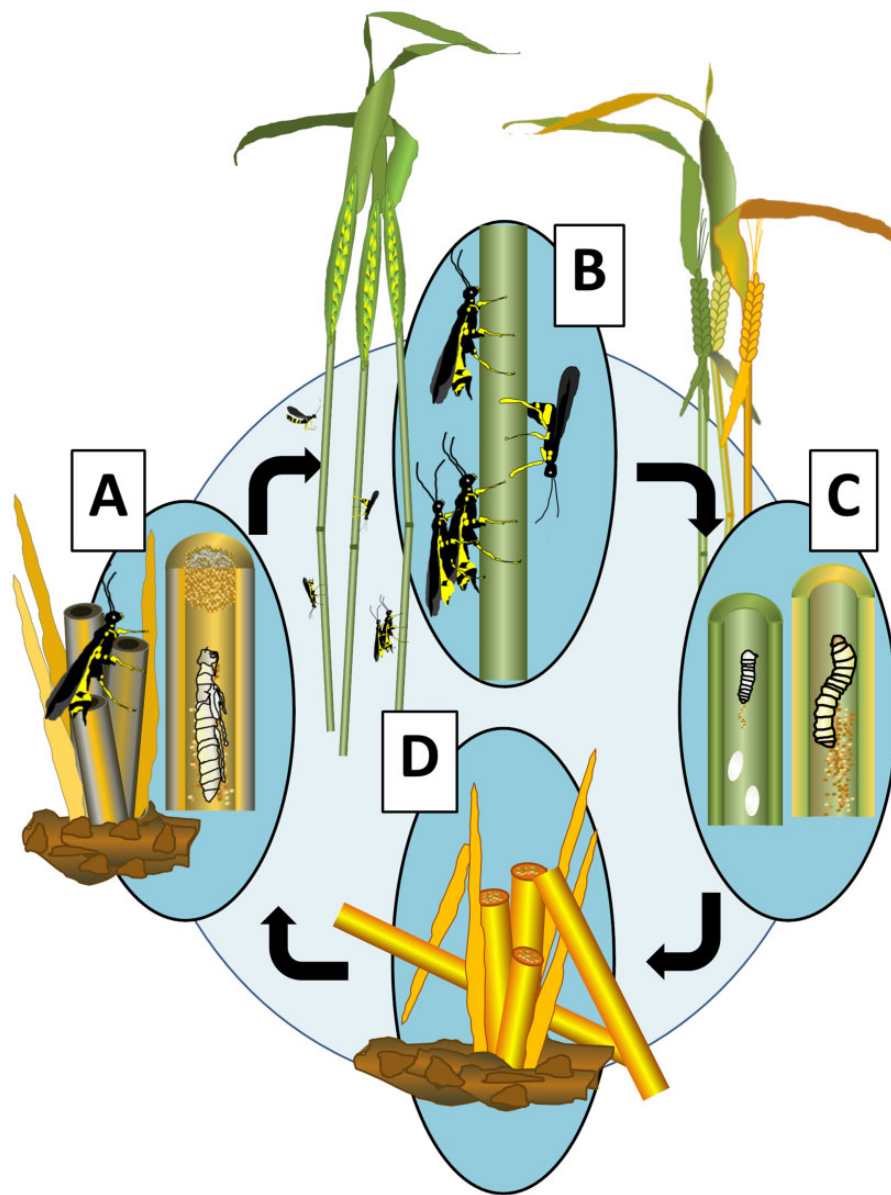
Apocrita, and so provides an important comparator and out-group for understanding some of the peculiarities of the genomes, gene contents, and biology of the apocritan wasps, ants, and bees. Genomic resources for *C. cinctus* therefore also contribute to augmenting the growing phylogenetic representation of sequenced hymenopteran genomes (Branstetter et al. 2018).

Here, we report a high-quality draft genome assembly and annotation for *C. cinctus*, with genome-wide analyses of protein orthology and repeat content, and phylogenomic analysis that supports the placement of Cephoidea between the Tenthredinoidea and Orussoidea. Supported by extensive manual gene annotation efforts, our phylogenetic analyses focused on four major gene families encoding chemosensory proteins of ecological importance. As a member of an early radiating lineage, *Cephus* provides a new perspective on the likely ancestral repertoires and subsequent lineage-specific evolution of odorant-binding proteins (OBPs) and odorant, gustatory, and ionotropic receptors (ORs, GRs, and IRs), across the Hymenoptera.

## Materials and Methods

### DNA Isolation, Genomic Library Preparation, Sequencing, and Genome Assembly

Genomic DNA was isolated from a single male (~1 cm long weighing 4 mg) and separately from ten pooled females (supplementary table S1, Supplementary Material online). Briefly, insects were ground in liquid nitrogen before lysing in a SDS solution overnight with Proteinase K. The homogenate was treated with RNaseA, and proteins/debris were collected after high-salt precipitation and centrifugation. After ethanol precipitation, the DNA was resuspended in 10 mM Tris and evaluated on an agarose gel and by Qubit quantification. The following libraries were generated for sequencing: 500-bp and 1.5-kb insert shotgun libraries from the single male sawfly, and 3- and 5-kb insert mate-pair libraries from the pooled female DNA. The 500-bp and 1.5-kb insert shotgun libraries were prepared with Illumina TruSeq DNAseq Sample Prep kit. The 3- and 5-kb mate-pair libraries were prepared similarly except a custom linker was ligated between the fragment ends to facilitate mate-pair recovery. All libraries were sequenced individually in single lanes for at least 100 cycles on Illumina GAIIx or HiSeq2000 machines using Illumina TruSeq SBS Sequencing kit v3. Bases were called with Casava v1.6 or 1.82. The custom 3- and 5-kb mate-pair libraries were filtered for properly oriented reads of the appropriate insert size and uniqueness using custom pipeline scripts (available from <ftp://ccb.jhu.edu/pub/dpuiu/Bees/scripts/>). Raw Illumina reads were 5'- and 3'-trimmed for nucleotide-bias and low-quality bases using the FASTX Toolkit ([http://hannonlab.cshl.edu/fastx\\_toolkit/](http://hannonlab.cshl.edu/fastx_toolkit/)). Trimmed reads were error-corrected by library with



**FIG. 1.**—Life cycle of the wheat stem sawfly, *Cephus cinctus* Norton. (A) After overwintering in diapause in a senesced wheat stem, a prepupa metamorphoses in late spring or early summer. The pupal stage lasts 10–12 days, after which the newly eclosed adult chews a hole in a frass plug left by the larva to emerge from the stem. (B) The adults can mate immediately after emergence, either on the wheat stubble or on new host stems. The short-lived females oviposit 30–40 eggs before dying in 7–10 days. (C) Newly deposited eggs in the stem lumen, usually from multiple females, hatch in 5–10 days. The neonate larvae feed on the parenchyma lining the stem interior. As the larvae grow through four to five larval instars they forage on the stem lining throughout the stem by boring through the nodes. The duration of larval feeding can range from 3 to 8 weeks. (D) Within the stem, cannibalism leads to a single surviving larva that migrates to the base of the stem at plant senescence and ripening. The larva girdles the interior of the ripened stem wall, cutting a groove that encircles the entire stem. Wind or gravity causes the weakened stem to break and lodge, leaving the unharvested wheat head on the soil. Below this groove, the larva makes a frass plug before migrating to the base of the stem near the crown. Within a thin hibernaculum, the prepupa remains protected in obligate diapause from late summer through to warming the following spring. This period of inactivity can last >9 months. Image created by Megan L. Hofland and Norma J. Irish, Department of Land Resources and Environmental Sciences, Montana State University.

Quake (Kelley et al. 2010) counting 19-mers. SOAPdenovo v2.04 (Luo et al. 2012) was employed with  $K = 49$  to assemble the 500-bp insert shotgun library reads followed by scaffolding with iteratively longer-insert shotgun and

mate-pair libraries and use of GapCloser v1.12 to close gaps generated in the scaffolding (Luo et al. 2012). SOAPdenovo estimated average sequence coverage for the genome at 56 $\times$ .

### RNA Isolation, RNAseq Library Preparation, Sequencing, and Transcriptome Assembly

We undertook RNAseq studies of whole animals and tissues relevant to chemosensation such as antennae, heads without antennae, abdomen tips, and abdomens without tips (supplementary table S1, Supplementary Material online). Entire animals or body parts were ground in 1 ml Trizol in glass tissue grinders and filtered over a Qiagen Qiashreder column. The homogenate was extracted with chloroform and the RNA was precipitated with linear polyacrylamide (10 mg/mL) and isopropanol. RNA pellets were washed with 75% ethanol and resuspended in RNase-free water. RNA was quantified with a Qubit RNA Broad Range Assay Kit on a Qubit fluorometer (Life Technologies). RNA was visualized using ethidium bromide on a 1.0% agarose gel or on a BioAnalyzer (Agilent Genomics). The RNAseq libraries were prepared from an average cDNA fragment size of 250 bases using the Illumina TruSeq Stranded RNAseq Sample Prep kits. The libraries were individually barcoded and quantitated using qPCR before pooling and sequencing from both ends with TruSeq SBS Sequencing kits v2 or 3 for 100 cycles on a HiSeq2000 or with a HiSeq4000 Sequencing kit v1 for 150 cycles on a HiSeq4000 instrument. Bases were called with Casava v1.6 or 1.82 or bcl2fastq v2.17.1.14. Reads were trimmed as for DNA sequencing. The resulting trimmed reads from the male and female 2011 and the male and female antennal 2013 samples were assembled using Trinity (Release 2014-04-13) (<http://trinityrnaseq.github.io/>) (Haas et al. 2013).

### Automated Gene Modeling

Annotation of the *Cephus* genome assembly was performed by the NCBI using their Eukaryotic Genome Annotation Pipeline ([https://www.ncbi.nlm.nih.gov/genome/annotation\\_euk/process/](https://www.ncbi.nlm.nih.gov/genome/annotation_euk/process/)), with experimental support from the RNAseq and transcriptome. The results are available at [www.ncbi.nlm.nih.gov/genome/annotation\\_euk/Cephus\\_cinctus/101/](http://www.ncbi.nlm.nih.gov/genome/annotation_euk/Cephus_cinctus/101/), and constitute CcinOGSv1.0 (Official Gene Set).

### Manual Gene Modeling

Full descriptions of gene modeling methods are presented in the Supplementary Material online for each gene family. Briefly, TBLASTN searches of the genome assembly were performed using relevant proteins from other Hymenoptera and other insects, at stringencies relevant to the various families, at the i5k Workspace@NAL (Poelchau et al. 2018) where the genome assembly is presented. Relevant NCBI gene models were examined in the linked Apollo browser (Lee et al. 2013) and modified if necessary in light of gene model information from other species, as well as the extensive RNAseq, which was mapped to the genome and informed exon–intron boundaries. The 5′- and 3′-untranslated regions (UTRs) of the automated models were carefully adjusted based on the

RNAseq mapping, and any extensive overlaps with the 5′- or 3′-UTRs of highly expressed neighboring genes were noted to avoid problems with subsequent analysis of expression levels. Genes were named in the Apollo browser according to relevant approaches for each gene family (Supplementary Material online), and the manually modeled and annotated genes were merged with the NCBI models to generate CcinOGSv1.1, which is available from i5k Workspace@NAL. All new protein sequences are available in FASTA format in Auxiliary file 3. Phylogenetic methods are detailed in the Supplementary Material online and alignments are available from H.M.R.

### MicroRNA Annotation

MicroRNA (miRNA) annotation was performed as outlined previously (Cagirici et al. 2017). High-confidence miRNAs for Hexapoda were downloaded from miRBase.org (Release 21) and a query was formed for a homology-based screen of the genome assembly.

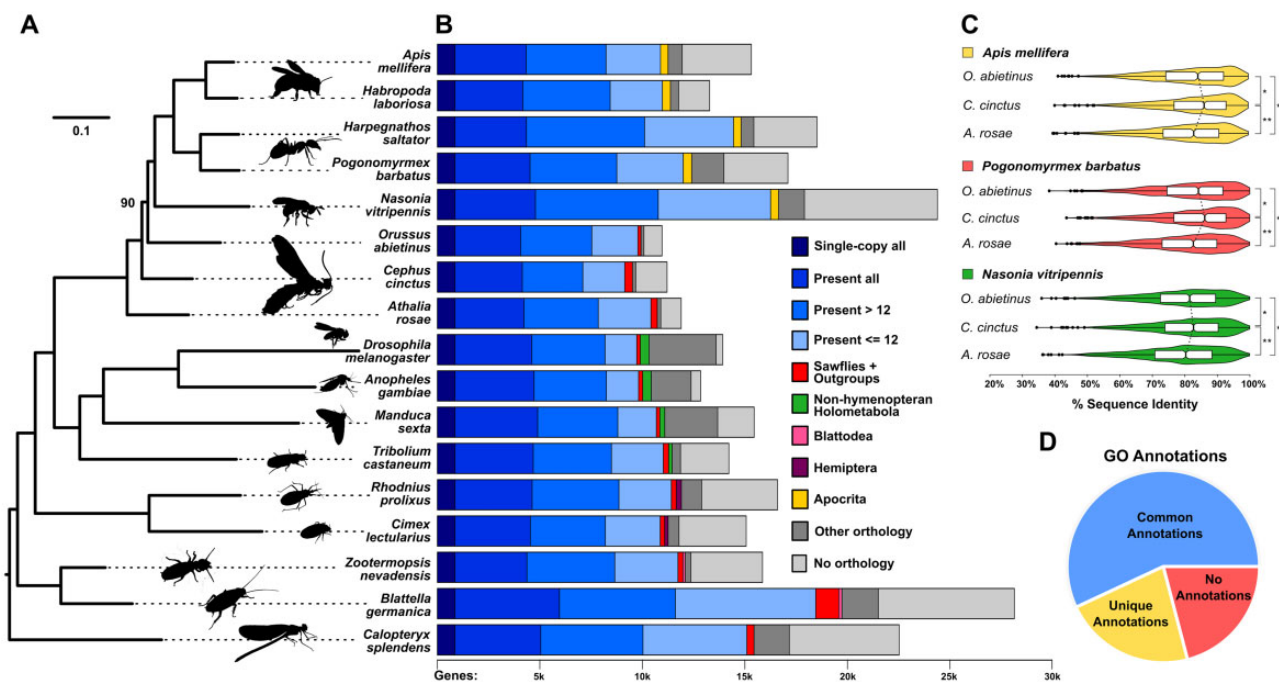
### Phylogenomics, Orthology, Gene Set, and Genome Completeness

The maximum likelihood molecular species phylogeny (fig. 2) was estimated from the concatenated protein sequence superalignment of 852 single-copy orthologs, using the automated gene models of CcinOGSv1.0 extracted from OrthoDB v9.1 (Zdobnov et al. 2017), using the 17 species shown in supplementary table S3, Supplementary Material online. The presence, absence, and copy-numbers of orthologs from 116 insect species were assessed to partition genes from each species into the distribution of categories shown in the bar chart (fig. 2). For each of the two sawflies and the woodwasp, percent amino acid identities with orthologs from *Nasonia vitripennis*, *Pogonomyrmex barbatus*, and *Apis mellifera*, were calculated from the trimmed multiple sequence alignments of 1,035 single-copy Hymenoptera orthologs (fig. 2, violin plots). Gene Ontology (GO) terms from *Drosophila melanogaster*, *A. mellifera*, *Tribolium castaneum*, and *Anopheles gambiae* were compared to identify *C. cinctus* genes with orthologs associated with GO terms assigned to at least two (i.e., annotations in common), just one (i.e., unique annotations), or none of the four species. Completeness in terms of expected gene content for the genomes and annotated gene sets of each of the eight analyzed hymenopteran species was assessed using both Hymenoptera ( $n = 4,415$ ) and Insecta ( $n = 1,658$ ) Benchmarking Universal Single-Copy Orthologs (Waterhouse et al. 2018). See the Supplementary Material online for full details of software versions and options/parameters used for each analysis.

### Transposable and Repeat Element Annotation

Although a repeat analysis for masking purposes is part of the NCBI gene annotation steps, we undertook a dedicated





**FIG. 2.**—Phylogenomics, orthology, sequence identity, and functional annotation of *Cephus cinctus* genes. (A) Maximum likelihood molecular phylogeny estimated from the aligned protein sequences of 852 single-copy orthologs from 17 selected insect species, including three sawflies, *Athalia rosae*, *C. cinctus*, and *Orussus abietinus*, with the banded demoiselle, *Calopteryx splendens* (Odonata), as the outgroup. Branch lengths represent substitutions per site, all nodes except for the placement of *O. abietinus* (labeled) achieved 100% bootstrap support. (B) Total gene counts per species partitioned into categories from single-copy orthologs in all 17 species, present (i.e., allowing gene duplications) in all or most (>12) or with patchy species distributions ( $\leq 12$ ), to lineage-restricted orthologs (sawflies and outgroups, non-hymenopteran Holometabola, Blattodea, Hemiptera, or Apocrita), genes with orthologs from other sequenced insect genomes, and those with currently no identifiable orthologs. (C) Distributions of percent amino acid identities of 1,035 single-copy orthologs from each of the three sawflies with *Apis mellifera*, *Pogonomyrmex barbatus*, and *Nasonia vitripennis*. Violin plots show the smoothed densities of each distribution, with boxplots indicating medians and upper and lower quartiles. Asterisks indicate statistically significant differences with *P* values  $< 1e-108$  (\*\*) and  $< 1e-39$  (\*). (D) Functional inferences for *C. cinctus* genes from Gene Ontology (GO) annotations of orthologs from *A. mellifera*, *Anopheles gambiae*, *Drosophila melanogaster*, or *Tribolium castaneum*. Of the 7,595 *C. cinctus* genes with orthologs from any of these four species, 57% were in orthologous groups with GO terms shared among at least two of the four species (common annotations), 22% with GO terms from only one species (unique annotations), and 21% where no GO terms were associated with any of the orthologs (no annotations).

analysis using the REPET package (Quesneville et al. 2005; Flutre et al. 2011). The computational pipeline to predict and classify transposable element (TE) integrations used both de novo- (TEdenovo) and homology-based methods (TEannot). A summary of the REPET pipeline can be found at <https://urgi.versailles.inra.fr/Tools/REPET> and detailed description of the applications and methods are in the [Supplementary Material](#) online.

### Expression Analysis

Levels of gene expression were determined by aligning the trimmed RNAseq reads from each library against the chemoreceptor and OBP transcripts from the i5k Workspace@NAL, occasionally truncated to avoid overlap with the UTRs of well-expressed neighboring genes, using the Burrows–Wheeler Aligner (BWA) (Li and Durbin 2009) and Read1 of each read pair. Samtools (Li et al. 2009) was used to sort, index, and

summarize the BWA. Read counts were standardized as Read counts Per Kilobase transcript per Million mapped reads in each library (RPKM) to facilitate comparisons across the 17 samples ([supplementary table S1, Supplementary Material](#) online).

## Results

### A High-Quality Genome Assembly and Annotation

Genome assembly in Hymenoptera is facilitated by generating most of the sequence from a single haploid male, thus avoiding complications from sequence and length differences between haplotypes. Partly as a result, our draft genome assembly is of relatively high contiguity for a short-read-only assembly. It consists of 10,707 contigs with a N50 of 45 kb that were connected into 1,976 scaffolds with a N50 scaffold size of 622 kb, for a total size of 162 Mb including 3 Mb of gaps between contigs within scaffolds. This assembly (v1) is

available from the National Center for Biotechnology Information (NCBI) as BioProjects PRJNA297591 and PRJNA168335 (NCBI GCF\_000341935.1). It is of a size and quality comparable to those of other Hymenoptera in GenBank (supplementary table S2, Supplementary Material online). We assessed the completeness of this assembly using BUSCO (Benchmarking Universal Single-Copy Orthologs) (Waterhouse et al. 2018), which revealed high completeness scores of 97.6–99.5%, with few duplicated (0.3–0.7%), fragmented (0.3–1.7%), or missing (0.2–0.7%) genes. These assembly completeness estimates are on a par with or marginally better than for the other assessed hymenopteran genomes (supplementary table S4, Supplementary Material online). In addition, 86–94% of the ILLUMINA RNAseq reads described below mapped to the assembly (NCBI *C. cinctus* Annotation Release 100—[https://www.ncbi.nlm.nih.gov/genome/annotation\\_euk/Cephus\\_cinctus/100/](https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Cephus_cinctus/100/)). These high levels of completeness and RNAseq read mapping indicate that this draft genome is a high-quality assembly that supports effective automated and manual gene modeling, as demonstrated below.

To support gene modeling and assess gene expression across lifestyles and tissues, we performed paired-end ILLUMINA RNAseq on three samples of larvae, two samples each of adult males and females, one sample each of male and female pupae, and one sample each of antennae, heads without antennae, abdomen tips, and abdomens without tips from adult males and females (supplementary table S1—Auxiliary file 1, Supplementary Material online). These libraries ranged in size from 3.5 to 12 Gb. An available RNAseq data set of 717,345,454-pyrosequencing reads from mixed-sex antennae (Gress et al. 2013) was also employed. Gene modeling was performed by NCBI using their GNOMON pipeline, and yielded 11,210 protein-coding genes and 41 pseudogenes, as well as 1,307 noncoding RNA genes. A total of 25,937 transcripts were modeled, with 25,189 of those having support from the RNAseq and/or homology with other arthropod proteins. The mean number of transcripts per gene is 2.3 (range 1–50) and the mean number of exons per gene is 8.9 (range 1–108). The protein-coding gene set is comparable in completeness (as estimated by alignment to the REFSEQ proteins from *D. melanogaster*) to that of the eusymphytan sawfly *Athalia rosae* and the woodwasp *Orussus abietinus* (NCBI *C. cinctus* Annotation Release 100—[https://www.ncbi.nlm.nih.gov/genome/annotation\\_euk/Cephus\\_cinctus/100/](https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Cephus_cinctus/100/)). BUSCO analysis of the quality of this automated annotation in terms of expected gene content reflected the high BUSCO completeness scores for the genome assembly, identifying 98.7–99.4% complete, and few duplicated (0.4–0.7%), fragmented (0.2–0.7%), or missing (0.4–0.6%) genes. The close match of BUSCO scores for the assembly and annotations indicate that the annotation strategy was successful. These annotation completeness estimates are on a par with or marginally better than for the other

assessed hymenopteran genomes (supplementary table S4, Supplementary Material online). This automated gene set therefore provides a confident base for comparative genomics, as well as subsequent manual gene annotation.

### Transposable and Repeat Element Content

Copies of transposable elements and other repeats comprise major portions of most eukaryotic genomes. Our repeat prediction pipeline identified 51,432 simple sequence repeats (SSRs) of unit length  $\geq 2$  that encompassed 1.53 Mb (~1% of the assembled genome), within which hexanucleotide repeats were most abundant (supplementary fig. S1, Supplementary Material online). A total of 64,215 TE fragments totaling 20.8 Mb (~13% of the assembled genome) were resolved into 57,948 predicted copies using the Long\_join method to connect disrupted portions of the same integration. These predictions are available as a track at the i5k Workspace@NAL browser for this genome (<https://i5k.nal.usda.gov/cephus-cinctus>). The PASTEC classification of predicted repeat elements from the REPET TEdenovo pipeline identified 840 unique elements placed into six Class-I families and five Class-II families (supplementary table S5, Supplementary Material online). Overall, although the copy numbers of predicted elements are approximately equal between Class-I ( $n = 21,429$ ) and Class-II ( $n = 21,473$ ) elements, the latter occupies ~2-fold greater portion of the genome (Class-I: 5.3 Mb vs. Class-II: 10.2 Mb). Among the Class-I retroelements, long-interspersed nuclear elements (LINEs) are the most abundant, but long terminal repeat (LTR) elements occupy a greater proportion of the genome. Incomplete copies of Class-II elements across all families are most abundant, including among the nonautonomous short interspersed nuclear element (SINE) and terminal-repeat retroelements in miniature (TRIM) elements (supplementary table S5, Supplementary Material online). Inverted repeat (TIR) family members are most abundant and encompass the greatest proportion of the genome among Class-II DNA elements. The miniature inverted repeat transposable elements (MITEs) are the second most abundant family, but likely due to being the only fully nonautonomous family of Class-II elements they occupy a smaller proportion of the genome than TIR or *Maverick* elements.

### Phylogenomics and Protein Orthology

Phylogenomic analyses using large sets of conserved proteins are currently the most robust approach to resolving phylogenetic placement of species. The molecular species phylogeny (fig. 2A), estimated from concatenated protein sequence alignments of single-copy orthologs, supports the current view of sawfly radiations early during the evolution of the Hymenoptera (Peters et al. 2017), with Cephidae placed between the earlier-branching Tenthredinoidea (represented by the turnip sawfly, *A. rosae*) and the later-branching

Crussoidea (represented by the parasitic wood wasp, *O. abietinus*). Orthology delineation across insects identified orthologs for 86.4% of *C. cinctus* genes, almost three-quarters of which have orthologs in all or most of seven other hymenopterans and nine representative species from six other insect orders (fig. 2B). The two sawflies and the woodwasp have similarly low total gene counts and show similar proportions of genes in each orthology category. In contrast, the other representative hymenopterans have generally higher total gene counts including a fraction of seemingly apocritan-specific orthologs.

The small fraction of sawfly genes with orthologs in outgroup species, but not in other Hymenoptera, highlight potential gene losses in the apocritan ancestor. These include orthologs of *D. melanogaster* genes *TrpA1* (*Transient receptor potential cation channel A1*) involved in responses to heat and noxious chemicals (Luo et al. 2017), and the G-protein-coupled >receptor *Lgr1* (*Leucine-rich repeat-containing G protein-coupled receptor 1*) involved in development (Vandersmissen et al. 2014). This fraction also includes a gene encoding a CPCFC family cuticular protein, previously described by Vannini et al. (2015) as being present in *C. cinctus* but missing from apocritan species. Further examples of putative losses in the apocritan ancestor from our detailed analysis of chemoreceptor gene families are elaborated below.

Examining ortholog sequence conservation between sawfly and Apocrita species showed that *C. cinctus* proteins had significantly higher amino acid identities (Wilcoxon signed rank tests,  $P < 1e-39$ ) with the Apocrita than either *A. rosae* or *O. abietinus* (fig. 2C). This apparently slower rate of sequence divergence in *C. cinctus* may at least partially explain the uncertainty of the placement of *O. abietinus* in the species phylogeny (fig. 2A), which was alternatively placed as sister to *C. cinctus* in 10% of bootstrap samples. With little detailed functional characterization of *C. cinctus* genes to date, putative functions can instead be inferred by identifying orthologs from well-studied insects such as the honey bee, malaria mosquito, fruit fly, or flour beetle, to tentatively link Gene Ontology terms to the majority of *C. cinctus* genes (fig. 2D).

### MicroRNAs

miRNAs are a distinctive regulatory system in genomes, some of which are conserved across insects while others are unique to each lineage. Thirty-six high-confidence putative miRNAs from 22 different miRNA families were identified. Of these, 24 are located on the sense strand while 12 are on the antisense strand (table 1).

### The *C. cinctus* Chemosensory Gene Repertoire

Insects depend on members of three major gene families for most of the sensitivity and specificity of their senses of smell and taste (Leal 2013; Joseph and Carlson 2015). The first two

**Table 1**

miRNA ID and Number of miRNA Annotations

mirnaID	On Sense	On Antisense	Total
miR-10-5p	2	0	2
miR-137-3p	2	0	2
miR-1-3p	1	0	1
miR-14-3p	0	2	2
miR-1-5p	1	0	1
miR-184-3p	1	0	1
miR-190-5p	0	2	2
miR-210-3p	1	0	1
miR-210-5p	1	0	1
miR-275-3p	0	1	1
miR-277	0	1	1
miR-279-3p	0	1	1
miR-2796-3p	1	0	1
miR-281-3p	3	0	3
miR-281-5p	1	0	1
miR-2a-3p	0	1	1
miR-315-5p	0	1	1
miR-71-5p	0	1	1
miR-8-3p	2	0	2
miR-8-5p	1	0	1
miR-87a-3p	0	1	1
miR-927a-5p	1	0	1
miR-929-5p	1	0	1
miR-92b	1	0	1
miR-92b-3p	3	0	3
miR-981-3p	1	0	1
miR-iab-4-5p	0	1	1
<b>SUM</b>	<b>24</b>	<b>12</b>	<b>36</b>

are the odorant receptor (OR) and gustatory receptor (GR) families, which together form the insect chemoreceptor superfamily (Robertson et al. 2003) of seven-transmembrane ligand-gated ion channels. The third family are the Ionotropic Receptors (IRs) that are evolutionarily unrelated to the insect chemoreceptor superfamily, being a variant lineage of the ionotropic glutamate receptor superfamily (Benton et al. 2009; Rytz et al. 2013; van Giesen and Garrity 2017; Rimal and Lee 2018). Our manual annotation focused on these three gene families, as well as the OBP family of secreted small globular proteins, resulting in high-quality gene models from which to build robust inferences of their evolutionary histories.

The OR family consists of a highly conserved odorant receptor coreceptor (*Orco*) gene that has 1:1 orthologs throughout the pterygote insects (Ioannidis et al. 2017) and forms dimers with each of the many “specific” ORs (Benton et al. 2006). In *Cephus* there are 72 OR genes including the expected single highly conserved *Orco* (supplementary table S6, Supplementary Material online), a total considerably smaller than the large OR families of *A. mellifera* (176 proteins—Robertson and Wanner 2006), *Bombus terrestris* (166—Sadd et al. 2015), *P. barbatus* (400—Smith et al. 2011), and

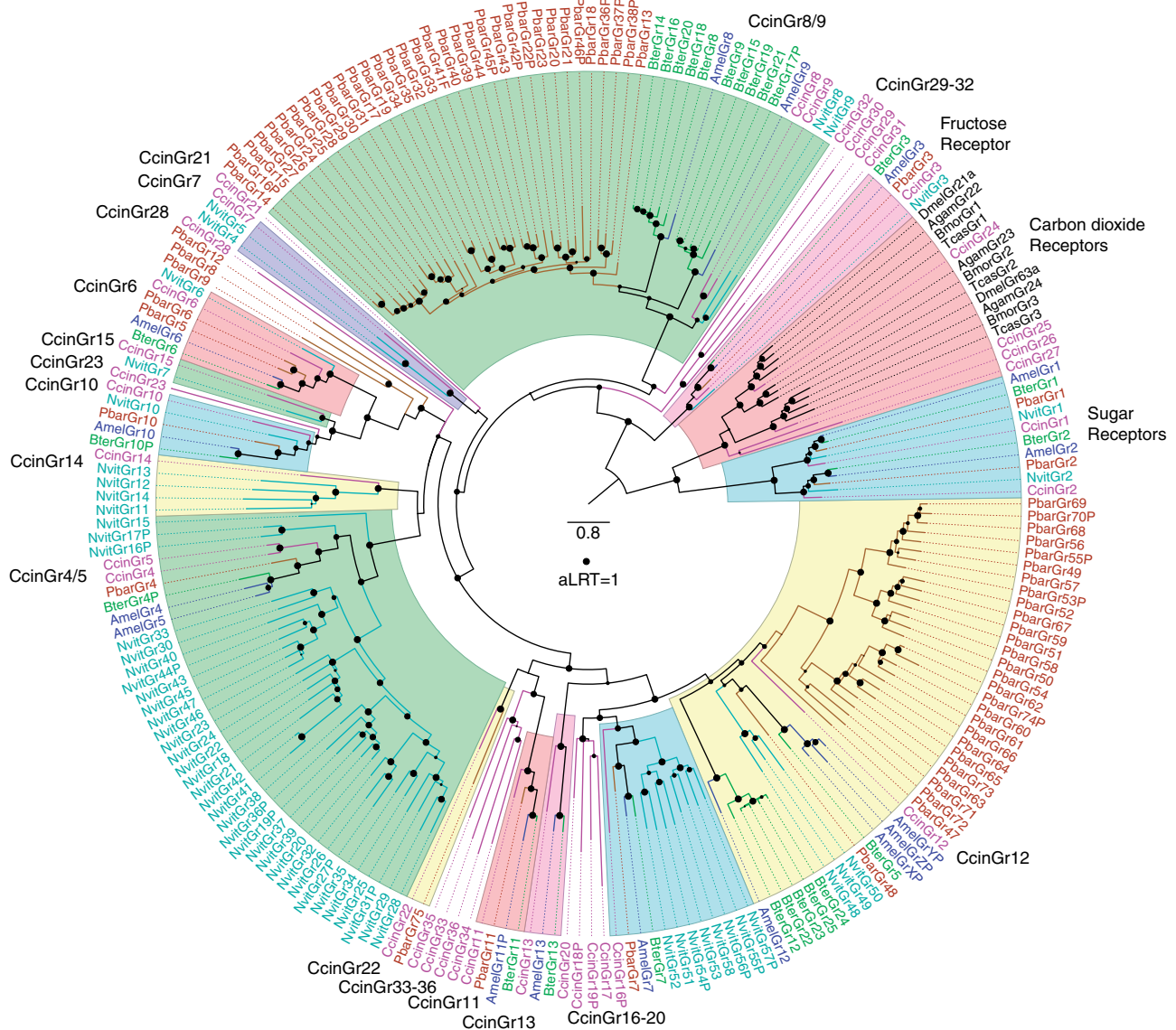


*N. vitripennis* (301—Robertson et al. 2010). Phylogenetic analysis using protein sequences from these species, employed because they have undergone similarly intensive manual annotation, reveals that the *Cephus* ORs are largely present as a single gene, or sometimes small clusters of genes, at the base of numerous small clades and sometimes highly expanded subfamilies in these other hymenopterans (supplementary fig. S2, Supplementary Material online). Most remarkably, a single gene (*CcinOr69*) is at the base of the extremely large 9-exon subfamily, which is particularly highly expanded in ants (Smith et al. 2011). Many receptors in this subfamily have been shown to mediate perception of ant cuticular hydrocarbons of importance to social interactions, although some might also perceive other chemicals and some cuticular hydrocarbons are detected by ORs outside this subfamily (Pask et al. 2017; Slone et al. 2017). Similarly, a small species-specific clade consisting of *CcinOr1-9* is at the base of the large-tandem-array subfamily of 61 genes in *A. mellifera* (Robertson and Wanner 2006) that includes a receptor for the primary queen pheromone substance, *AmelOr11* (Wanner et al. 2007), a tandem array that has persisted throughout hymenopteran evolution. In addition to this small *Cephus*-specific clade of tandem-array subfamily members, there are three more *Cephus*-specific clades of 4, 8, and 8 genes, which are extremely small compared with the large gene subfamily expansions observed in the apocritans. *Cephus* appears to have lost a few OR lineages present in the other hymenopterans, and has retained several lineages that are missing from the Apocrita but are commonly still present in other symphytans and/or the wood wasp (Supplementary Material online). This analysis suggests that ancestral hymenopterans had ~40 OR lineages.

The GR family is a highly diverse family of primarily taste receptors, but also some olfactory receptors such as the carbon dioxide receptors of flies and moths, including receptors for sugars and a vast array of “bitter” compounds, as well as light and heat. It is an ancient family within the Metazoa, dating back to early animals (Robertson 2015; Saina et al. 2015; Eyun et al. 2017), but was lost from vertebrates, which instead employ primarily G-protein-coupled receptors for both olfaction and taste. We identified a total of 35 GR genes in *C. cinctus* (supplementary table S7, Supplementary Material online), an intermediate number compared with the other species with 13 in *A. mellifera* (Robertson and Wanner 2006), 25 in *B. terrestris* (Sadd et al. 2015), 75 in *P. barbatus* (Smith et al. 2011), and 47 in *N. vitripennis* (Robertson et al. 2010). In addition to single orthologs of the two sugar receptors known for hymenopterans (*CcinGr1/2*) and a single candidate fructose receptor (*CcinGr3*), *Cephus* has two clear orthologs of the carbon dioxide receptors of other endopterygote insects (*CcinGr24/25*), although the third carbon dioxide receptor GR lineage is missing (fig. 3). In addition, there are two more genes clearly related to the carbon dioxide receptor clade (*CcinGr26/27*), which are part of an expanded

subfamily in exopterygote insects from which the carbon dioxide receptors evolved (Ioannidis et al. 2017). The remaining GRs, like the ORs, are placed in the phylogeny as single genes or small species-specific clades at the base of most GR lineages and subfamilies, including several large expansions in ants and wasps. There are again several highly divergent GR lineages in *Cephus* that are commonly present in other symphytans and the woodwasp (Supplementary Material online), but have been lost from the more derived Hymenoptera, specifically *CcinGr16–20*, 21, 23, 28, 29–32, 33–36 (fig. 3). Unlike for the ORs, there are no obvious losses of entire GR lineages from *Cephus*. These comparisons lead to an estimate of ~30 GR lineages in ancestral hymenopterans, maintained in *Cephus* but with some lineages lost from, and others greatly expanded in, the Apocrita.

The IRs have an extracellular ligand-binding domain supported by three transmembrane domains, and also function as ligand-gated ion channels. They are involved in both olfaction, primarily being expressed in olfactory sensory neurons in coeloconic sensilla that do not express ORs and sensing acids and amines, and gustation with large subfamilies expressed in gustatory organs, at least in *Drosophila melanogaster* (Rytz et al. 2013; Koh et al. 2014). A few are also involved in sensing temperature and humidity (Knecht et al. 2017; van Giesen and Garrity 2017; Rimal and Lee 2018). The IR family in *C. cinctus* is made up of 49 genes (supplementary table S8, Supplementary Material online), roughly double the number of *A. mellifera* (21—Croset et al. 2010), *B. terrestris* (22—Sadd et al. 2015), and *P. barbatus* (27—Smith et al. 2011 and Supplementary Material online), but considerably smaller than *N. vitripennis* for which our manual annotation efforts increased the gene count from just ten in Croset et al. (2010) to 153 genes, albeit 54 (35%) are pseudogenes (Supplementary Material online). *Cephus* has the expected single orthologs of the highly conserved *Ir8a*, *25a*, and *76b* genes, which encode proteins that function as coreceptors with other IRs (Rytz et al. 2013). It also has single orthologs for three of the four genes involved in thermo- and hygrosensation in *D. melanogaster* (*Ir21a*, *68a*, and *93a*), but *Ir40a* was lost at the base of the Hymenoptera. *Ir40a* cooperates with the other proteins to sense humidity in *D. melanogaster* (Knecht et al. 2017), so it would be interesting to discover how this gene loss, and idiosyncratic losses of *Ir21a* and *Ir68a* in various Hymenoptera, affect their abilities to perceive temperature and humidity. *Cephus* has only a single relative of the *Dmellr75a-c* genes that encode olfactory receptors for various acids (Prieto-Godino et al. 2017), a lineage that is expanded in wasps, but of comparable size to flies in ants and bees. There is a single hymenopteran set of orthologs for the *Dmellr41a* lineage of olfactory receptors that is commonly greatly expanded in other insects (Robertson et al. 2018). Like the ORs and GRs, *Cephus* has generally single representatives of other hymenopteran IR lineages, but also two *Cephus*-specific clades, one of several ancient gene lineages that appear to



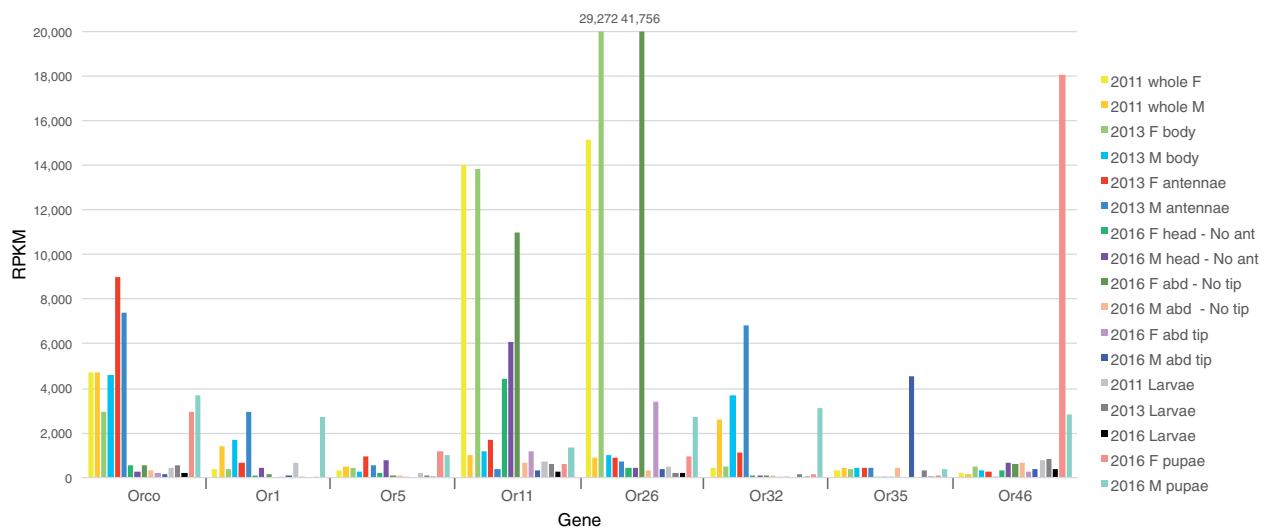
**FIG. 3.**—Phylogenetic relationships of the gustatory receptor family in Hymenoptera. The tree was rooted by declaring the sugar and carbon dioxide receptor lineages as the outgroup, based on their basal location within insect GRs in an analysis of the entire GR family in insects and other animals (Robertson 2015). GR names and the branches leading to them are colored by species, purple for *Cephus cinctus*, cyan for *Nasonia vitripennis*, brown for *Pogonomyrmex barbatus*, green for *Bombus terrestris*, blue for *Apis mellifera*, and black for other species. The CcinGr lineages are indicated outside the circle of names, and major lineages with CcinGr representatives are highlighted in colored segments. The scale bar indicates substitutions per site, and the filled circle indicates approximate likelihood ratio test (aLRT) support of 1.

have been lost from Apocrita, and one of very recent gene duplications (supplementary fig. S3, Supplementary Material online). These clades, along with a few bee and ant proteins and the majority of the *Nasonia* proteins, are related to the Dmellr7a-f/11a proteins and the large Dmellr20a clade implicated in gustation (Rytz et al. 2013; Koh et al. 2014). This analysis suggests that ancestral hymenopterans had ~34 IR gene lineages. Thus *Cephus* has the expected complement of conserved hymenopteran IRs, except for the *Ir40a* and *Ir41a* lineages, as well as both ancient lineages and very recent

species-specific clades, the latter of which, like those in the OR and GR families, might be involved in colonization of new grass species including wheat.

Finally, OBPs are small globular proteins commonly secreted by support cells at the base of chemosensory sensilla and thought to be involved in transport of commonly hydrophobic odorants across the chemosensillar lymph (Pelosi et al. 2014), although not all OBPs are expressed in chemosensory tissues (Forêt and Maleszka 2006; Pelosi et al. 2018). Furthermore, the only experimentally demonstrated biological

Downloaded from https://academic.oup.com/gbe/article-abstract/10/11/2997/5134337 by guest on 18 July 2019



**Fig. 4.**—Expression levels of selected ORs. Expression levels are shown as Reads Per Kilobase transcript per Million reads (RPKM) in each library for 8 of the 73 OR family genes and the 17 RNAseq libraries (detailed in [supplementary table S1, Supplementary Material](#) online) that sampled different life stages and tissues of male and female sawflies collected from 2011 to 2016. Abbreviations: abd, abdomen; ant, antennae; F, female; M, male.

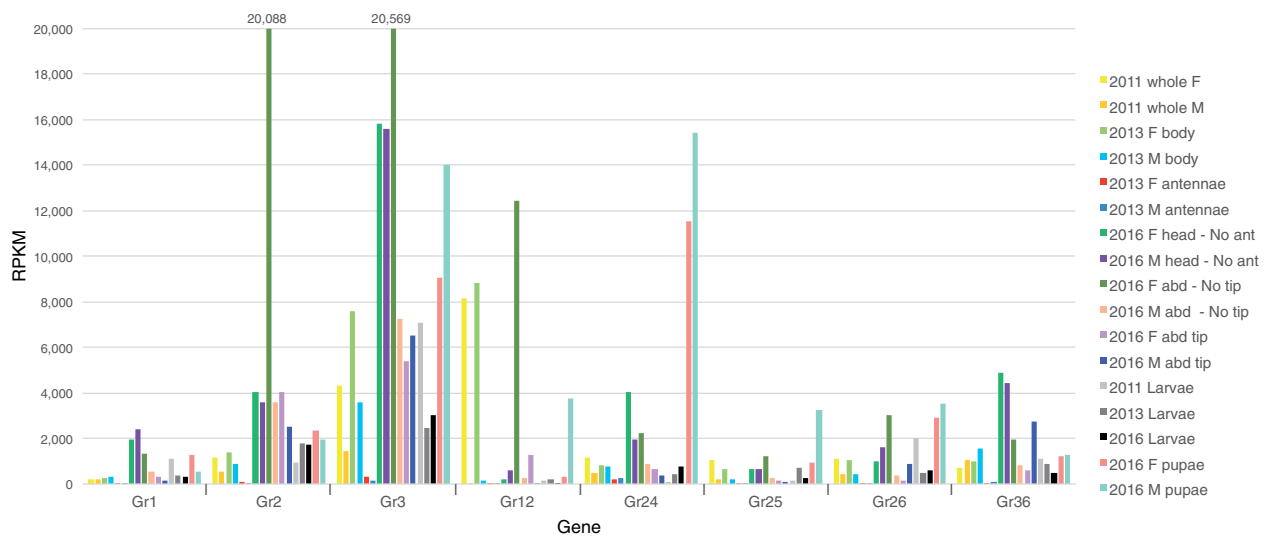
role for an OBP is that of quenching the signal of an odorant (Larter et al. 2016). Like the chemoreceptor families, hymenopterans have a wide range of numbers of OBPs, from 18 in the fire ant *Solenopsis invicta* (Gotzek et al. 2011) and 21 in *A. mellifera* (Forêt and Maleszka 2006) to 90 in *N. vitripennis* (Vieira et al. 2012). We identified 15 OBP genes and their encoded proteins in *C. cinctus* ([supplementary table S9, Supplementary Material](#) online), which like the chemoreceptors are frequently at the base of a clade of single orthologs or expansions in these other hymenopterans ([supplementary fig. S4, Supplementary Material](#) online).

#### Expression Levels of Chemoreceptors and OBPs

We examined the levels of expression of the chemoreceptors and OBPs in our 17 diverse RNAseq data sets from larvae, pupae, adults, and adult body parts ([supplementary table S1—Auxiliary file 1, Supplementary Material](#) online), using the manually annotated transcripts, truncated if necessary to avoid conflation with overlapping untranslated regions (UTRs) of highly expressed neighboring genes. The complete results are presented in [supplementary tables S10–S13 and figures S5–S8](#) (Auxiliary file 2), [Supplementary Material](#) online, with selected genes shown in [figures 4–7](#). For the ORs, *Orco* has the expected high levels of expression in male and female antennae, roughly twice the level seen in entire adult bodies or pupae, where chemoreceptor expression is initiated. As expected, heads without antennae, abdomen tips, and abdomens without tips had low levels of expression of *Orco*. Larvae also have low levels, perhaps because larvae only have a few olfactory sensilla given their feeding mode within wheat stems. These expected results for *Orco* provide confidence that our other results are reliable. Most OR genes have

appropriately low expression, primarily in antennae ([supplementary fig. S5, Supplementary Material](#) online), represented by *Or5* in [figure 4](#), while some of the divergent genes like *Or60* and *Or72* are barely expressed. There are a few instances of consistent sex bias. For example, *Or1* is consistently expressed  $\sim 4\times$  higher in males than females, while *Or30–32* exhibit  $5–7\times$  higher expression in males, although these three might be conflated as they are very similar genes in a tandem array (represented by *Or32* in [fig. 4](#)). All four are among the mostly highly expressed *Ors*, a result confirmed by prior quantitative PCR that also found *Or32* expression  $15\times$  higher in male antennae (Gress et al. 2013; [supplementary table S14, Supplementary Material](#) online), making these candidate receptors for female pheromone. The expression of *Ors 18, 63, 65, and 68* was higher in female antennae, in this study and in prior quantitative PCR experiments ( $5–15\times$  female bias, Gress et al. 2013; [supplementary table S14, Supplementary Material](#) online), making these candidate receptors for host plant volatiles or male-produced pheromones.

Some ORs exhibit unusual expression patterns, for example, the tandemly arrayed *Or11–13* are expressed at levels comparable to *Orco* in heads without antennae, and *Or11* is also extremely highly expressed in female abdomens. It would be interesting to see if the relatives of this small clade of *CcinOr11–14* in other Hymenoptera (*NvitOr11–33*, *PbarOr70–75*, *BterOr87–95*, and *AmelOr114/5*) are similarly expressed outside the obvious olfactory organs. A few other unusual results are high expression of *Or35* and *Or65* in male abdominal tips, and of *Or46* in pupae. An enigmatic result is the extraordinarily high expression of *Or26* in females, specifically their abdomens. The 3'-UTR of this gene overlaps that of a neighboring highly expressed DNA polymerase gene,



**Fig. 5.**—Expression levels of selected GRs. Expression levels in RPKM for 8 of the 36 GR family genes. Other details as for figure 4.

hence the analyzed transcript was shortened at the 3' end, so this result has to be treated with caution, nevertheless the reads that map to *Or26* are mostly correctly spliced, implying they are from this locus rather than the 3'-UTR of the polymerase.

Among the GRs, as expected the sugar receptors *Gr1/2* are expressed primarily in heads at levels comparable to or higher than *Orco*, however *Gr2* is also expressed in abdomens and abdominal tips, and unusually highly in female abdomens. The candidate fructose receptor *Gr3* is the mostly highly expressed GR, and as expected from the expression of its *Drosophila* ortholog *DmelGr43a* (Miyamoto et al. 2012), primarily in heads but also in abdomens, larvae, and pupae, and appropriately low in antennae. Most of the other GRs show the expected pattern of low levels of expression primarily in heads (supplementary fig. S6, Supplementary Material online), rather than antennae, a pattern exemplified by *Gr36* (fig. 5). Exceptions include *Gr12*, which is consistently highly expressed in females, primarily in abdomens. The two candidate carbon dioxide receptors, *Gr24/25*, are also not well expressed in antennae, but rather in heads, and *Gr24* is unusually highly expressed in pupae. The related carbon-dioxide receptor subfamily gene *Gr26* shows a similar pattern (unfortunately *Gr27* entirely overlaps the 3'-UTR of a highly expressed gene, so could not be evaluated).

Among the IRs, *Ir25a* and *76b*, two coreceptors, are the most highly expressed, primarily in heads, as are a few others like *Ir41a*, however *Ir76b* is extremely highly expressed in male abdominal tips, a clear result with almost entirely spliced reads. The third coreceptor, *Ir8a*, is expressed at a relatively low level. *Ir93a* is the most highly expressed of the three genes implicated in perception of temperature and humidity, with *Ir68a* the lowest of these three. It is noteworthy that these are the more conserved receptors whose orthologs in

*D. melanogaster* are primarily expressed in antennae (Rytz et al. 2013). Among the more divergent IRs, the most highly expressed are *Ir108* and *Ir120*, again primarily in heads, while *Ir140* is shown as an example of the generally low expression of the IRs with likely gustatory roles (fig. 6).

Finally, as expected the OBPs commonly exhibited far higher levels of expression than the chemoreceptors (supplementary fig. S8, Supplementary Material online). Perhaps surprisingly, only *Obp1*, *Obp4*, *Obp6*, *Obp11*, and *Obp12* are well expressed in antennae (and male abdominal tips for *Obp12*) (represented by *Obp6* and *Obp12* in fig. 7). *Obp2* and *Obp9* are primarily expressed in heads (and larvae for *Obp9*), while *Obp3*, *Obp5*, *Obp8*, and *Obp15* have unusually low levels of expression for OBPs, and mostly in larvae. *Obp7* and *Obp10* are primarily expressed in male abdomens, especially the tips, while *Obp13* is almost exclusively expressed in larvae. Finally, *Obp14* is extraordinarily highly expressed in multiple samples, most highly in female abdominal tips. The observation of high expression of several OBPs in abdominal tips might imply a nonolfactory role for these proteins, although some chemoreceptors are also unusually highly expressed there, so a chemosensory role there is also possible.

## Discussion

Genome sequencing offers the opportunity to explore the putative genomic or genetic basis of organismal biology that may be widely shared among species or highly lineage-specific. The informed interpretation of observed patterns of conservation and divergence requires a robust understanding of the evolutionary relationships among the considered species, which in the case of sawflies has remained unclear until relatively recently. While the separation of Cephidae from the other symphytan sawflies had been proposed for some time,



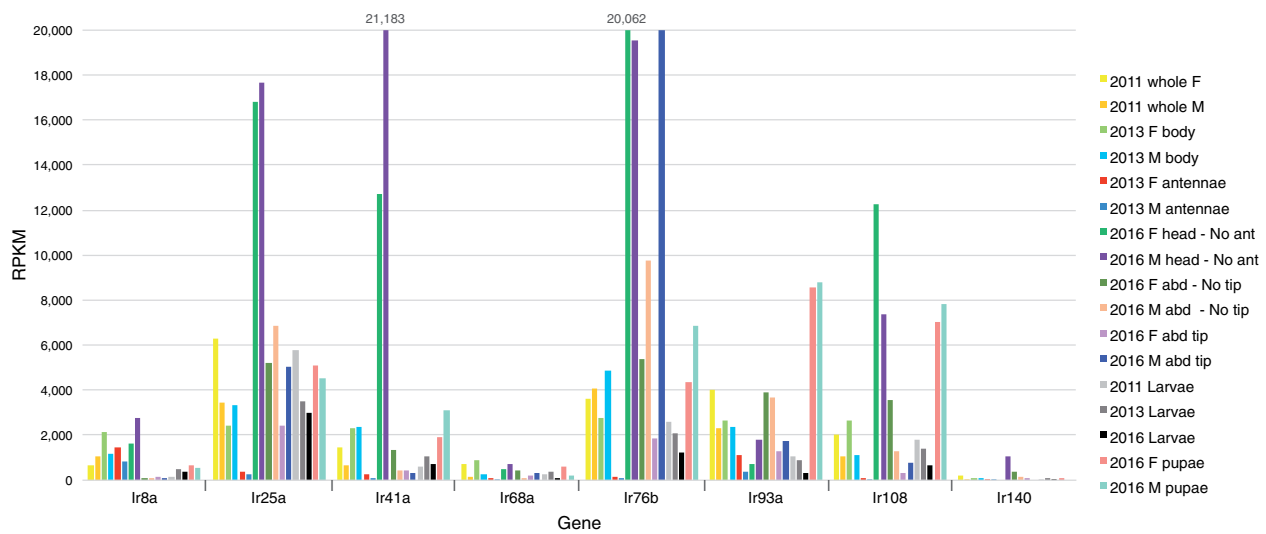


FIG. 6.—Expression levels of selected IRs. Expression levels in RPKM for 8 of the 49 IR family genes. Other details as for figure 4.

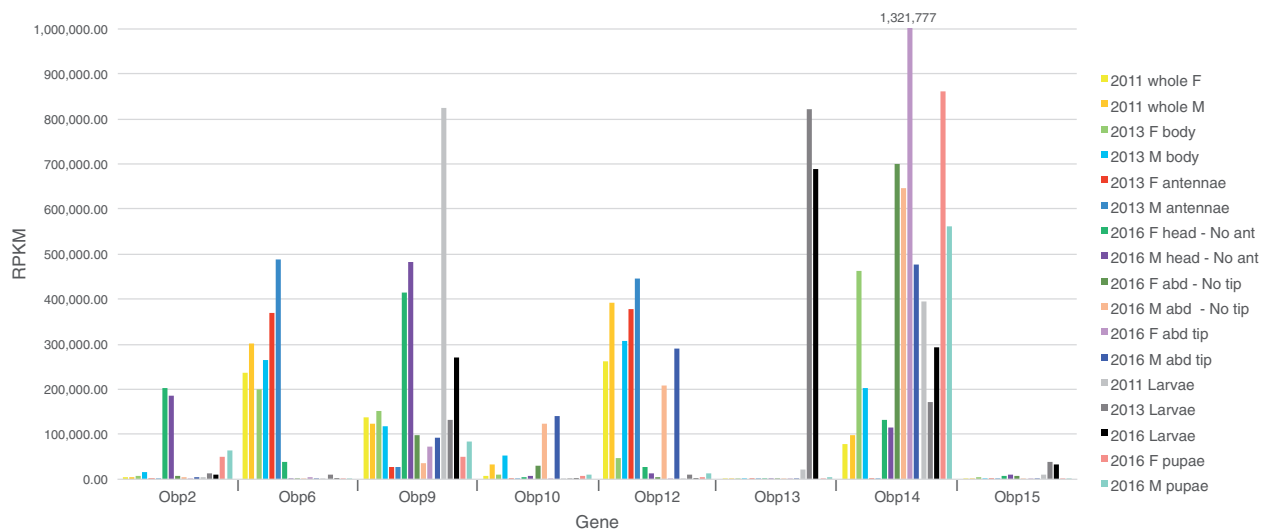


FIG. 7.—Expression levels of selected OBPs. Expression levels RPKM for 8 of the 15 Obp family genes. Other details as for figure 4.

the phylogenomics study of Peters et al. (2017), based on protein sequences deduced from transcriptomes, put this placement on more solid footing. Our genome-based phylogenomics analysis further supports placement of the Cephidae as a secondary radiation after the Eusymphyta, and with considerable confidence agrees with Peters et al. (2017) in the branching of the Cephidae before the orussid woodwasps. The apparently constrained rates of molecular evolution of *Cephus* genes relative to other sawflies and woodwasps, resulting in higher amino acid identities with their Apocrita orthologs, may partially explain previous difficulties in attempting to resolve these early hymenopteran radiations.

The availability of many other hymenopteran and other insect genomes means that the vast majority (86%) of *Cephus* proteins can be confidently assessed as having

orthologs in other species, despite the deep branching of this lineage. Functional information from orthologs in *D. melanogaster* and other insects allows assignment of putative biological roles for more than half of the annotated *C. cinctus* protein-coding genes. These associations reflect the ever-increasing knowledge of insect molecular biology and provide a strong foundation for future functional work on this sawfly.

Our repeat analysis from this 162-Mb genome assembly revealed that 1% consists of simple sequence repeats, among which are just 27 copies of the TTAGG telomeric repeat found to constitute the telomeres of the honey bee (Robertson and Gordon 2006). None of these TTAGG repeats are obviously at the end of long scaffolds where they might constitute telomeres, indicating that despite the high quality of our assembly, telomeric regions, and surely centromeric regions, remain unassembled, as is the case for most eukaryotic draft genome

sequence assemblies based on short-read technologies. Long-read technologies should be able to improve the assembly and recover at least the telomeres and perhaps centromeres, as has now been done for honey bee (NCBI GCF\_003254395.2). A further 13% of the assembly consists of transposons, partitioned roughly equally between Class-I and Class-II elements. This TE profile is in contrast to that seen in the honey bee where simple repeats make up 4% of the 250-Mb genome assembly and transposons comprise only 5% (Elsik et al. 2014). We still have only a rudimentary understanding of how genomes come to have such disparate transposon compositions.

Our analysis of the three major chemoreceptor families and the OBPs reveals *Cephus* genes at the base of most apocritan gene lineages, many as single-copy orthologs but also several with remarkably large expansions in wasps, ants, and/or bees. While only a few gene lineages appear to have been lost from *Cephus*, several are absent in the apocritans, most prominently the entire carbon dioxide receptor subfamily. This analysis pinpoints this loss to the apocritan ancestor, and leaves open the question of how extant apocritans are able to sense carbon dioxide concentrations. *Cephus* nevertheless does have a few small species-specific expansions of chemoreceptors that might mediate some relatively recent adaptations to new grasses including wheat, which could be important for future analysis of phylogeography of this recently adapted crop pest (Lesieur et al. 2016). Our identification of candidate pheromone and host volatile receptors, as well as ORs, GRs, and OBPs expressed on the ovipositor, will inform future studies of behavioral and chemical ecology (Bartelt et al. 2002; Cossé et al. 2002; Varella et al. 2017). Examination of gene expression levels reveals generally expected results, such as high expression of *Orco* in antennae and the sugar and fructose GRs and the IR coreceptors in heads, with low expression of most other receptors. Several exceptions are noted including sex-biased and tissue-specific expression. This complete cataloging of these chemosensory gene families now makes detailed functional studies of their involvement in the ecology of this sawfly feasible.

In summary, this high-quality draft genome assembly and automated gene annotation will facilitate further molecular biological studies on this sawfly, and offers opportunities to identify genes for engineering of sawfly-resistant wheat strains with RNAi constructs. It also augments the diversity of hymenopterans with sequenced genomes, particularly for the poorly sampled sawflies (Branstetter et al. 2018), making this insect order second only to Diptera for the number of species with genome sequences.

## Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

## Acknowledgments

We thank Alvaro Hernandez and the W.M. Keck Center for Comparative and Functional Genomics at the University of Illinois at Urbana-Champaign for genomic and RNA library construction and sequencing; Daniel Ence and Mark Yandell (University of Utah) for annotation of an earlier version of the genome assembly; Terence Murphy for assistance with the NCBI annotation; Monica Poelchau and Chris Childers for assistance with the i5k Workspace@NAL browser and official gene set v1.1; Chris Elsik for an Apollo browser at NasoniaBase; Masatsugu Hatakeyama, Bernhard Misof, Oliver Niehuis, and Jan Philip Oeyen for granting prepublication access to gene annotations of *Athalia rosae* and *Orussus abietinus*; and Megan Hofland and Norma Irish for preparing figure 1. This work was supported by funding from the United States Department of Agriculture (Grant Number AG2008-35302-188815 to H.M.R. and K.W.W.), Swiss National Science Foundation (Grant Number PP00P3\_170664 to R.M.W.), the Winifred-Asbjornson Plant Science Endowment to H.B., the Montana Wheat and Barley Committee and the Montana Grains Foundation to D.K.W. and K.W.W., who also received funds supporting this research from the AES allocation in HB 645 of the 61st Legislature of the State of Montana, and a contribution from Gene Robinson (University of Illinois at Urbana-Champaign). Part of this research was the result of a joint contribution from the United States Department of Agriculture (USDA), Agricultural Research Service (ARS) (CRIS Project 5030-22000-018-00D), and the Iowa Agriculture and Home Economics Experiment Station, Ames, IA (Project 3543). USDA is an equal employment opportunity provider. This article reports the results of research only.

## Literature Cited

- Adhikari S, Seipel T, Menalled FD, Weaver DK. 2018. Farming system and wheat cultivar affect infestation of and parasitism on *Cephus cinctus* in the Northern Great Plains. *Pest Manag Sci.* 74 (Early View):1–8.
- Bartelt RJ, Cossé AA, Petroski RJ, Weaver DK. 2002. Cuticular hydrocarbons and novel alkenediol acetates from wheat stem sawfly (*Cephus cinctus*): natural oxidation to pheromone components. *J Chem Ecol.* 28(2):385–405.
- Bekkerman A, Weaver DK. 2018. Modeling joint dependence of managed ecosystems pests: the case of the wheat stem sawfly. *J Agric Res Econ.* 43:172–194.
- Benton R, Sachse S, Michnick SW, Vosshall LB. 2006. Atypical membrane topology and heteromeric function of *Drosophila* odorant receptors in vivo. *PLoS Biol.* 4(2):e20.
- Benton R, Vannice KS, Gomez-Diaz C, Vosshall LB. 2009. Variant ionotropic glutamate receptors as chemosensory receptors in *Drosophila*. *Cell* 136(1):149–162.
- Beres BL, Dosdall LM, Weaver DK, Cárcamo HA, Spaner DM. 2011. Biology and integrated management of wheat stem sawfly and the need for continuing research. *Can Entomol.* 143(2):105–125.
- Branstetter M, et al. 2017. Phylogenomic insights into the evolution of stinging wasps and the origins of ants and bees. *Curr Biol.* 27(7):1019–1025.
- Branstetter M, et al. 2018. Genomes of the Hymenoptera. *Curr Opin Insect Sci.* 25:65–75.

- Buteler M, Peterson RKD, Hofland ML, Weaver DK. 2015. A multiple decrement life table reveals that host plant resistance and parasitism are major causes of mortality for the wheat stem sawfly. *Environ Entomol.* 44(6):1571–1580.
- Buteler M, Weaver DK, Peterson RKD. 2009. Exploring the oviposition behavior of the wheat stem sawfly when encountering plants infested with cryptic conspecifics. *Environ Entomol.* 38(6):1707–1715.
- Cagirici HB, Biyiklioglu S, Budak H. 2017. Assembly and annotation of transcriptome provided evidence of miRNA mobility between wheat and wheat stem sawfly. *Front Plant Sci.* 8:e1653.
- Cockrell DM, et al. 2017. Host plants of the wheat stem sawfly (Hymenoptera: cephidae). *Environ Entomol.* 46(4):847–854.
- Cossé AA, Bartelt RJ, Weaver DK, Zilkowski BW. 2002. Pheromone components of the wheat stem sawfly: identification, electrophysiology, and field bioassay. *J Chem Ecol.* 28(2):407–423.
- Croset V, et al. 2010. Ancient protostome origin of chemosensory ionotropic glutamate receptors and the evolution of insect taste and olfaction. *PLoS Genet.* 6(8):e1001064.
- Elsik PG, et al. 2014. Finding the missing honey bee genes: lessons learned from a genome upgrade. *BMC Genomics* 15:e86.
- Eyun SJ, et al. 2017. Evolutionary history of chemosensory-related gene families across the Arthropoda. *Mol Biol Evol.* 34(8):1838–1862.
- Flutre T, Duprat E, Feuillet C, Quesneville H. 2011. Considering transposable element diversification in *de novo* annotation approaches. *PLoS One* 6(1):e16526.
- Forêt S, Maleszka R. 2006. Function and evolution of a gene family encoding odorant binding-like proteins in a social insect, the honey bee (*Apis mellifera*). *Genome Res.* 16(11):1404–1413.
- Gotzke D, Robertson HM, Wurm Y, Shoemaker D. 2011. Odorant binding proteins of the red imported fire ant, *Solenopsis invicta*: an example of the problems facing the analysis of widely divergent proteins. *PLoS One* 6(1):e16289.
- Gress JC, Robertson HM, Weaver DK, Dlakić M, Wanner KW. 2013. Odorant receptors of a primitive hymenopteran pest, the wheat stem sawfly. *Insect Mol Biol.* 22(6):659–667.
- Haas BJ, et al. 2013. *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc.* 8(8):1494.
- Ioannidis P, et al. 2017. Genomic features of the damselfly *Calopteryx splendens* representing a sister clade to most insect orders. *Genome Biol Evol.* 9(2):415–430.
- Joseph RM, Carlson JR. 2015. *Drosophila* chemoreceptors: a molecular interface between the chemical world and the brain. *Trends Genet.* 31(12):683–695.
- Kelley DR, Schatz MC, Salzberg SL. 2010. Quake: quality-aware detection and correction of sequencing errors. *Genome Biol.* 11(11):R116.
- Knecht ZA, et al. 2017. Ionotropic Receptor-dependent moist and dry cells control hygrosensation in *Drosophila*. *Elife* 6:e26654.
- Koh TW, et al. 2014. The *Drosophila* IR20a clade of ionotropic receptors are candidate taste and pheromone receptors. *Neuron* 83(4):850–865.
- Larter NK, Sun JS, Carlson JR. 2016. Organization and function of *Drosophila* odorant binding proteins. *Elife* 5:e20242.
- Leal WS. 2013. Odorant reception in insects: roles of receptors, binding proteins, and degrading enzymes. *Annu Rev Entomol.* 58:373–391.
- Lee E, et al. 2013. Web Apollo: a web-based genomic annotation editing platform. *Genome Biol.* 14(8):R93.
- Lesieur V, et al. 2016. Phylogeography of the wheat stem sawfly, *Cephus cinctus* Norton (Hymenoptera: cephidae): implications for pest management. *PLoS One* 11(12):e0168370.
- Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* 25(14):1754–1760.
- Li H, et al. 2009. The sequence alignment/map format and SAMtools. *Bioinformatics* 25(16):2078–2079.
- Luo J, Shen WL, Montell C. 2017. TRPA1 mediates sensation of the rate of temperature change in *Drosophila* larvae. *Nat Neurosci.* 20(1):34–41.
- Luo R, et al. 2012. SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. *Gigascience* 1:e18.
- Miyamoto T, Slone J, Song X, Amrein H. 2012. A fructose receptor functions as a nutrient sensor in the *Drosophila* brain. *Cell* 151(5):1113–1125.
- Pask GM, et al. 2017. Specialized odorant receptors in social insects that detect cuticular hydrocarbon cues and candidate pheromones. *Nat Commun.* 8:e297.
- Pelosi P, Iovinella I, Felicioli A, Dani FR. 2014. Soluble proteins of chemical communication: an overview across arthropods. *Front Physiol.* 5:e320.
- Pelosi P, Iovinella I, Zhu J, Wang G, Dani FR. 2018. Beyond chemoreception: diverse tasks of soluble olfactory proteins in insects. *Biol Rev Camb Philos Soc.* 93(1):184–200.
- Peters RS, et al. 2017. Evolutionary history of the Hymenoptera. *Curr Biol.* 27(7):1013–1018.
- Poelchau M, et al. 2018. Navigating the i5k Workspace@NAL: a resource for arthropod genomes. *Methods Mol Biol.* 1757:557–577.
- Prieto-Godino LL, et al. 2017. Evolution of acid-sensing olfactory circuits in drosophilids. *Neuron* 93(3):661–676.
- Quesneville H, et al. 2005. Combined evidence annotation of transposable elements in genome sequences. *PLoS Comput Biol* 1(2):166–175.
- Rimal S, Lee Y. 2018. The multidimensional ionotropic receptors of *Drosophila melanogaster*. *Insect Mol Biol.* 27(1):1–7.
- Robertson HM. 2015. The insect chemoreceptor superfamily is ancient in animals. *Chem Senses* 40(9):609–614.
- Robertson HM, et al. 2018. Enormous expansion of the chemosensory gene repertoire in the omnivorous German cockroach *Blattella germanica*. *J Exp Zool B Mol Dev Evol.* 330(5):265–278.
- Robertson HM, Gadau J, Wanner KW. 2010. The insect chemoreceptor superfamily of the parasitoid jewel wasp *Nasonia vitripennis*. *Insect Mol Biol.* 19:121–136.
- Robertson HM, Gordon KH. 2006. Canonical TTAGG-repeat telomeres and telomerase in the honey bee, *Apis mellifera*. *Genome Res.* 16(11):1345–1351.
- Robertson HM, Wanner KW. 2006. The chemoreceptor superfamily in the honey bee, *Apis mellifera*: expansion of the odorant, but not gustatory, receptor family. *Genome Res.* 16(11):1395–1403.
- Robertson HM, Warr CG, Carlson JR. 2003. Molecular evolution of the insect chemoreceptor gene superfamily in *Drosophila melanogaster*. *Proc Natl Acad Sci U S A.* 100(Suppl 2):14537–14542.
- Rytz R, Croset V, Benton R. 2013. Ionotropic receptors (IRs): chemosensory ionotropic glutamate receptors in *Drosophila* and beyond. *Insect Biochem Mol Biol.* 43(9):888–897.
- Sadd BM, et al. 2015. The genomes of two key bumblebee species with primitive eusocial organization. *Genome Biol.* 16:76.
- Saina M, et al. 2015. A cnidarian homologue of an insect gustatory receptor functions in developmental body patterning. *Nat Commun.* 6:e6243.
- Schulmeister S, Wheeler WC, Carpenter JM. 2002. Simultaneous analysis of the basal lineages of Hymenoptera (Insects) using sensitivity analysis. *Cladistics* 18(5):455–484.
- Sharkey MJ, et al. 2011. Phylogenetic relationships among superfamilies of Hymenoptera. *Cladistics* 27:1–33.
- Slone JD, et al. 2017. Functional characterization of odorant receptors in the ponerine ant, *Harpegnathos saltator*. *Proc Natl Acad Sci U S A.* 114(32):8586–8591.
- Smith CR, et al. 2011. Draft genome of the red harvester ant *Pogonomyrmex barbatus*. *Proc Natl Acad Sci U S A.* 108(14):5667–5672.
- Vandersmissen HP, Van Hiel MB, Van Loy T, Vleugels R, Broeck JV. 2014. Silencing *D. melanogaster Igr1* impairs transition from larval to pupal stage. *Gen Comp Endocrinol.* 209:135–147.

- van Giesen L, Garrity PA. 2017. More than meets the IR: the expanding roles of variant Ionotropic Glutamate Receptors in sensing odor, taste, temperature and moisture. *F1000Res* 6:1753.
- Vannini L, Bowen JH, Reed TV, Willis JH. 2015. The CPCFC cuticular protein family: anatomical and cuticular locations in *Anopheles gambiae* and distribution throughout Pancrustacea. *Insect Biochem Mol Biol*. 65:57–67.
- Varella AC, et al. 2017. Host plant quantitative trait loci affect specific behavioral sequences in oviposition by a stem-mining insect. *Theor Appl Genet*. 130(1):187–197.
- Varella AC, et al. 2018. Characterization of resistance to *Cephus cinctus* Norton (Hymenoptera: cephidae) in barley germplasm. *J Econ Entomol*. 111(2):923–930.
- Vieira FG, et al. 2012. Unique features of odorant-binding proteins of the parasitoid wasp *Nasonia vitripennis* revealed by genome annotation and comparative analyses. *PLoS One* 7(8):e43034.
- Vilhelmsen L. 2001. Phylogeny and classification of the extant basal lineages of the Hymenoptera. *Zool J Linnean Soc*. 131(4):393–442.
- Wanner KW, et al. 2007. A honey bee odorant receptor for the queen substance 9-oxo-2-decenoic acid. *Proc Natl Acad Sci U S A*. 104(36):14383–14388.
- Waterhouse RM, et al. 2018. BUSCO applications from quality assessments to gene prediction and phylogenomics. *Mol Biol Evol*. 35(3):543–548.
- Weaver DK, et al. 2009. Cultivar preferences of ovipositing wheat stem sawflies as influenced by the amount of volatile attractant. *J Econ Entomol*. 102(3):1009–1017.
- Zdobnov EM, et al. 2017. OrthoDB v9. 1: cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs. *Nucleic Acids Res*. 45(D1):D744–D749.
- Zhao C, et al. 2015. A massive expansion of effector genes underlies gall-formation in the wheat pest *Mayetiola destructor*. *Curr Biol*. 25(5):613–620.

Associate editor: Rebecca Zufall