nature COMMUNICATIONS

ARTICLE

Received 17 Jul 2013 | Accepted 13 Mar 2014 | Published 20 May 2014

DOI: 10.1038/ncomms4636 OPEN

Molecular traces of alternative social organization in a termite genome

Nicolas Terrapon[1,*,†], Cai Li[2,3,*], Hugh M. Robertson[4], Lu Ji[2], Xuehong Meng[2], Warren Booth[5,†], Zhensheng Chen[2], Christopher P. Childers[6], Karl M. Glastad[7], Kaustubh Gokhale[8], Johannes Gowin[9,†], Wulfila Gronenberg[10], Russell A. Hermansen[11], Haofu Hu[2], Brendan G. Hunt[7,†], Ann Kathrin Huylmans[1,†], Sayed M.S. Khalil[5,12], Robert D. Mitchell[5], Monica C. Munoz-Torres[13], Julie A. Mustard[8], Hailin Pan[2], Justin T. Reese[6], Michael E. Scharf[14], Fengming Sun[2], Heiko Vogel[15], Jin Xiao[2], Wei Yang[2], Zhikai Yang[2], Zuoquan Yang[2], Jiajian Zhou[2], Jiwei Zhu[5], Colin S. Brent[16], Christine G. Elsik[6,17], Michael A. D. Goodisman[7], David A. Liberles[11], R. Michael Roe[5], Edward L. Vargo[5], Andreas Vilcinskas[18], Jun Wang[2,19,20,21,22], Erich Bornberg-Bauer[1], Judith Korb[9,†], Guojie Zhang[2,23] & Jürgen Liebig[8]

Although eusociality evolved independently within several orders of insects, research into the molecular underpinnings of the transition towards social complexity has been confined primarily to Hymenoptera (for example, ants and bees). Here we sequence the genome and stage-specific transcriptomes of the dampwood termite *Zootermopsis nevadensis* (Blattodea) and compare them with similar data for eusocial Hymenoptera, to better identify commonalities and differences in achieving this significant transition. We show an expansion of genes related to male fertility, with upregulated gene expression in male reproductive individuals reflecting the profound differences in mating biology relative to the Hymenoptera. For several chemoreceptor families, we show divergent numbers of genes, which may correspond to the more claustral lifestyle of these termites. We also show similarities in the number and expression of genes related to caste determination mechanisms. Finally, patterns of DNA methylation and alternative splicing support a hypothesized epigenetic regulation of caste differentiation.

[1] Institute for Evolution and Biodiversity, Westfälische Wilhelms-Universität, Münster D48149, Germany. [2] China National GeneBank, BGI-Shenzhen, Shenzhen 518083, China. [3] Centre for GeoGenetics, Natural History Museum of Denmark, University of Copenhagen, Øster Voldgade 5-7, Copenhagen 1350, Denmark. [4] Department of Entomology, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, USA. [5] Department of Entomology and W. M Keck Center for Behavioral Biology, North Carolina State University, Raleigh, North Carolina 27695, USA. [6] Division of Animal Sciences, University of Missouri, Columbia, Missouri 65211, USA. [7] School of Biology, Georgia Institute of Technology, Atlanta, Georgia 30332, USA. [8] School of Life Sciences, Arizona State University, Tempe, Arizona 85287, USA. [9] Behavioural Biology, University of Osnabrück, Osnabrück D49076, Germany. [10] Department of Neuroscience, University of Arizona, Tucson, Arizona 85721, USA. [11] Department of Molecular Biology, University of Wyoming, Laramie, Wyoming 82071, USA. [12] Department of Microbial Molecular Biology, Agricultural Genetic Engineering Research Institute, Giza 12619, Egypt. [13] Genomics Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA. [14] Department of Entomology, Purdue University, West Lafayette, Indiana 47907, USA. [15] Department of Entomology, Max Planck Institute for Chemical Ecology, Jena D-07745, Germany. [16] Arid Land Agricultural Research Center, United States Department of Agriculture, Maricopa, Arizona 85138, USA. [17] Division of Plant Sciences, University of Missouri, Columbia, Missouri 65211, USA. [18] Institut für Phytopathologie und Angewandte Zoologie, Justus-Liebig-Universität Giessen, Giessen D35390, Germany. [19] Department of Biology, University of Copenhagen, Copenhagen DK-1165, Denmark. [20] Princess Al Jawhara Center of Excellence in the Research of Hereditary Disorders, King Abdulaziz University, 21589 Jeddah, Saudi Arabia. [21] Macau University of Science and Technology, Avenida Wai long, Taipa, Macau 999078, China. [22] Department of Medicine, University of Hong Kong, Hong Kong. [23] Centre for Social Evolution, Department of Biology, University of Copenhagen, Universitetsparken 15, DK-2100 Copenhagen, Denmark. * These authors contributed equally to this work. † Present addresses: Architecture et Fonction des Macromolécules Biologiques, Aix-Marseille Université, 13288 Marseille, France (N.T.); Department of Biological Sciences, The University of Tulsa, Tulsa, Oklahoma 74104, USA (W.B.); Evolutionary Biology & Ecology, University of Freiburg, D 79117 Freiburg, Germany (J.G. or J.K.); Department of Entomology, University of Georgia, Griffin Campus, Griffin, Georgia 30223 USA (B.G.H.); Department Biology II, Ludwig-Maximilians-University Munich, Planegg-Martinsried D-82152, Germany (A.K.H.). Correspondence and requests for materials should be addressed to J.K. (email: judith.korb@biologie.uni-freiburg.de) or to G.Z. (email: zhanggj@genomics.org.cn) or to J.L. (email: jliebig@asu.edu).

© 2014 Macmillan Publishers Limited. All rights reserved.

Termites are major pests of human structures, with an annual worldwide cost in damage and control estimated at US$40 B[1]. However, in tropical habitats termites are pivotal for ecosystem function and maintaining biodiversity[2]. Their complex societies have enhanced their environmental adaptability, contributing to their success. Similar to eusocial Hymenoptera (ants, some bees and wasps), termites are characterized by a caste system in which a few individuals reproduce (queens and, in termites, kings) while the large majority (workers and soldiers) perform tasks such as foraging, brood care or defence[3]. Despite these similarities to eusocial Hymenoptera, termite societies have a phylogenetically distinct origin and divergent biological traits. They form a monophyletic clade nested within the Blattodea[4], indicating a single origin of termite eusociality, whereas eusociality independently evolved multiple times within the Hymenoptera[3]. Termites are hemimetabolous, having several immature stages that become more adult like with each transition, while Hymenoptera have a holometabolous development in which the final larval stage develops via a pupa into adulthood. Despite having distinct lineages with different phylogenetic constraints, termites and eusocial Hymenoptera have evolved similar social and physiological traits. Understanding the selective pressures and specific adaptations necessary to achieve comparable outcomes requires detailed comparisons, particularly at a genetic level. While annotated genomes have been published for seven ant species[5–10] and the honey bee *Apis mellifera*[11], sequence data are limited for termites and Blattodea in general. Most termite genetic studies have focused on the development of soldiers with few addressing differences between queens and workers[12]; therefore, it remains unclear whether eusocial Hymenoptera and termites convergently 'exploited' similar mechanisms to achieve similar ends.

Here we report the sequence and analysis of the first termite genome of the lower termite *Zootermopsis nevadensis nuttingi* (Termopsidae), together with genome-wide gene expression data of various caste and developmental stages. We compare these results with previous findings for eusocial Hymenoptera to identify common and divergent associations of traits linked to eusociality. We show significant expansion of gene families related to male reproduction and chemoperception, suggesting differences between *Z. nevadensis* and eusocial Hymenoptera in mating biology and communication, respectively. We also identify similarities and differences between orders in major gene families that are thought to play a role in the evolution and maintenance of eusociality. These include genes involved in endocrinology, immunity, reproductive development and caste differentiation. *Z. nevadensis* also exhibits a high level of DNA methylation, which may support a function in regulating phenotypic plasticity, as has been hypothesized for eusocial Hymenoptera. Collectively, the results are an important advance in our ability to elucidate the evolution and mechanistic basis of insect eusociality both within the termites and across taxa.

Results

Genome assembly and transcriptomes. Genome sequencing utilized a colony consisting of one naturally inbred family with complete homozygosity at four microsatellites that normally have two to three alleles in this population[13] (Supplementary Note 1; Supplementary Table 1). After sequencing, reads were strictly filtered leading to an average coverage depth of 98.4 × with an estimated genome size of 562 Mb (Methods; Supplementary Fig. 1; Supplementary Table 2). This is the smallest genome known among termites and roaches[14]. The assembly yielded 93,931 scaffolds, including 85,940 singleton contigs, with an N50 length of 740 kb and a coverage of 493.5 Mb, or 88% of the genome (Methods; Supplementary Table 3).

Expression analysis of 25 transcriptomes of different sex, developmental stage and caste (Fig. 1; Table 1; Supplementary Note 2; Supplementary Tables 4 and 5) was used to highlight the molecular basis of caste and life stage evolution of *Z. nevadensis*. We identified gene families that are specifically overexpressed in some castes or life stages (Fig. 2; Methods and Supplementary Note 2.6; Supplementary Tables 6 and 7). Five of these families were significantly expanded in the termite lineage (see below). Finally, we observed caste-specific expression of orphan genes, that is, genes without any identifiable orthologues, supporting the recently predicted lineage-specific function of orphans[15].

A total of 15,876 protein-coding genes are reported in OGSv2.2 (Methods; Supplementary Figs 2 and 3; Supplementary Table 8) with most (95.9%) of them supported by expression data from subsequent transcriptomes. We performed the annotation of the non-coding RNA (ncRNA) and repeated elements (Methods; Supplementary Tables 9–12) and predicted protein sequences were functionally annotated using Interpro domains, gene ontology (GO) terms and the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (Methods). Pfam annotation yielded 17,505 predicted domains (3,460 distinct families found in 52% of the termite proteins). Proteins of *Z. nevadensis*, eight reference insect species, the crustacean *Daphnia pulex* and a non-arthropod species *Caenorhabditis elegans* were clustered in orthologue groups, which were used along with functional annotation for genome quality assessment (Methods; Supplementary Note 3; Supplementary Tables 13 and 14). High orthologue and annotation coverage, despite the deep rooting of the termite lineage in the insect phylogeny (see following section), highlight the quality of the final assembly and gene models (Fig. 3).

Phylogenetic position of termites. Molecular approaches to reconstructing insect phylogeny have been hampered by

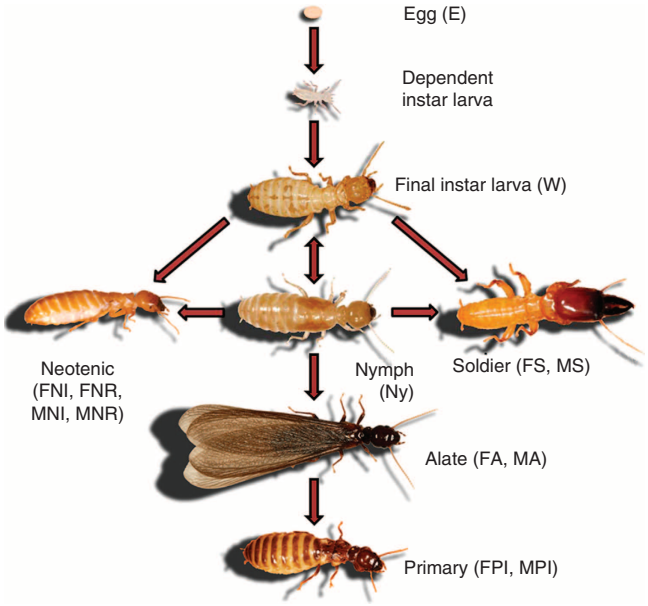


Figure 1 | Simplified developmental pathway of *Z. nevadensis* with sequenced castes and life stages. Abbreviations indicate (i) sex: female (F) or male (M); (ii) life stage: egg (E), final instar larva without wing buds (hereafter 'worker'—W), nymph with wing buds (Ny), soldier (S), alate (A), primary (P) or neotenic (N) reproductive; (iii) gonadal activity: currently inactive (I) or reproducing (R).

© 2014 Macmillan Publishers Limited. All rights reserved.

Table 1 | Expanded protein families in *Z. nevadensis* and related families showing significant differential expression.

				Reference arthropods counts									Over-/under -expression in samples						
	Protein family	Pfam IDs	Termite counts	D. mel	T. cas	N. vit	A. mel	C. flo	H. sal	P. hum	A. pis	D. pul	Egg	Juveniles	Soldiers	Alate ♂	Alate ♀	Male reprod.	Female reprod.
Chemo-perception	Ionotropic receptors	PF00060	134	18	29	14	7	10	12	14	13	114	-/-	-/-	-/-	-/-	-/-	1/1	5/3
	Ionotropic receptors	PF00060 PF10613	24	2	1	2	2	3	3	1	2	14	-/-	-/-	-/-	-/-	-/-	1/-	1/1
Regulation in female reproductives	Zinc finger C2H2	PF00096	215	77	96	81	88	87	31	94	210	108	1/-	-/-	-/-	-/-	-/-	1/2	**81/20**
	Zinc finger C2H2 + AD	PF00096 PF07776	32	38	40	14	11	4	2	7	6	1	-/-	-/-	-/-	-/-	-/-	-/-	**19/1**
	Histone	PF00125	20	93	17	71	16	19	17	25	17	105	2/-	-/-	-/-	-/-	-/-	-/8	**14/-**
Male mating biology	SINA	PF03145	33	3	16	2	1	1	1	4	3	3	-/-	-/-	1/-	1/-	-/-	**6/6**	**13/4**
	KELCH10	PF07707 PF00651 PF01344	37	9	9	10	8	6	4	10	78	7	-/-	-/-	-/-	-/-	-/-	**25/-**	3/7
	Kelch_1	PF01344	20	3	3	6	5	0	3	2	65	10	-/-	-/-	-/-	-/-	-/-	**14/-**	-/4
	PKD channel	PF08016	10	6	2	1	1	1	1	3	2	1	-/-	-/-	-/-	-/-	-/-	**7/-**	1/3
	Alpha tubulins	PS01161 PS01162	14	5	4	4	6	7	6	6	4	8	-/-	-/-	-/-	-/-	-/-	3/-	-/3
	BTB-Kelch	PF00651 PF01344	6	0	1	4	0	0	0	1	6	0	-/-	-/-	-/-	-/-	-/-	**4/-**	-/1
	BACK-Kelch	PF07707 PF01344	4	0	1	0	0	0	0	1	19	0	-/-	-/-	-/-	-/-	-/-	**4/-**	-/2
	ADAMTS	PF01562 PF01421 PF00090 PF05986	5	1	2	1	2	1	1	1	0	0	-/-	-/-	-/-	-/-	-/-	**5/-**	-/1

A. mel, Apis mellifera; A. pis, Acyrthosiphon pisum; C. flo, Camponotus floridanus; D. mel, Drosophila melanogaster; D. pul, Daphnia pulex; H. sal, Harpegnathos saltator; N. vit, Nasonia vitripennis; P. hum, Pediculus humanus; T. cas, Tribolium castaneum; .

Protein families, described by their domain architecture (domain names and Pfam IDs), are grouped by their putative biological roles. For each family, the number of proteins in *Z. nevadensis*, labelled by 'Termite counts' (blue background column), is compared with reference species, labelled 'Reference arthropods counts'. Significant expansions in *Z. nevadensis* are highlighted in yellow. Numbers of proteins with caste- or sex-specific differential expression patterns (over-/underexpression with false discovery rate <0.05 or 0.01 for types of samples having replicates (underlined) or unique samples without replicates, respectively) are given, with purple highlighting families for which the observed pattern is shared by a significant majority of members. Alpha tubulin domains originated from PRINTS.

numerous pitfalls; the rapid radiation of lineages early in the history of insects, compounded with widely varying rates of evolution (as measured by DNA substitutions), and a paucity of comparative genomes from basal insects have resulted in several controversies[16]. Using maximum likelihood and Bayesian approaches applied to sequences of both DNA and translated proteins, we were able to support the basal position of termites, being the outgroup to all other major insect taxa that possess representatives with draft genome sequences (Fig. 3; Methods; Supplementary Figs 4 and 5). Previously, only two genomes of hemimetabolous species were available as outgroups to the Endopterygota/holometabola group: *Acyrthosiphon pisum* (pea aphid) and *Pediculus humanus* (body louse). These two genomes exhibit features that might bias comparative analysis: substantial fragmentation of the genome sequence, with numerous paralogues and split gene models in *A. pisum* and a massive genome/proteome reduction owing to a parasitic lifestyle in *P. humanus*. Unlike these genomes, that of *Z. nevadensis* has minimal domain fragmentation and a protein number and orthologue/in-paralogue proportion that is in the range for the majority of insect genomes (Supplementary Note 3). These more standard characteristics enhance the value of *Z. nevadensis* as a third hemimetabolous genome and as a new outgroup to insect genomes.

The basal position of the termite genome thus allowed us to test recent hypotheses[17,18] regarding the evolution and synteny of the Osiris and Yellow-gene families (Supplementary Note 4; Supplementary Figs 6 and 7; Supplementary Table 15). These insect-specific families emerged and underwent multiple duplications in the insect ancestor, but orthologues and synteny have then been strongly conserved in all insect genomes. We found that the hemimetabolous *Z. nevadensis* had orthologues to the *Yellow-b* and *Osiris-1* and *-5* subfamilies, which were previously presumed to be specific to Endopterygota/holometabola. We were also able to identify microsyntenic regions that were larger than previously identified in the fragmented and reduced genomes of *A. pisum* and *P. humanus*.

Comparative genomics reveals gene family expansions. Our phylogeny provided the basis for all subsequent genomic analysis to identify species-specific features. Comparative analyses of gene families were conducted to identify major evolutionary changes in the termite genome (Supplementary Note 5; Supplementary Fig. 8; Supplementary Tables 16–20). When comparing single-copy genes across all sequenced insects, the most notable finding was the absence of several opsin orthologues, the photosensitive proteins used in vision. With only two opsin genes, *Z. nevadensis* has the smallest repertoire among insects, possibly as a result of principally living in the dark during most of their lifetime. We also found two instances of horizontal gene transfer from entomopathic viruses (Supplementary Note 5.2). Further, we tested gene families for lineage-specific expansion or contraction. Nine families exhibit expansion in *Z. nevadensis*, the majority being differentially expressed among developmental stages, castes and genders (Table 1). Four related families are not expanded but show similar differential expression across castes (Table 1). These proteins probably play key roles in *Z. nevadensis* life history, such as mating biology and communication, and are examined in the following sections.

Coexpansion of genes related to male fertility. Of the gene families that underwent significant expansion in *Z. nevadensis*, four exhibit a significant male-specific overexpression and have

© 2014 Macmillan Publishers Limited. All rights reserved.

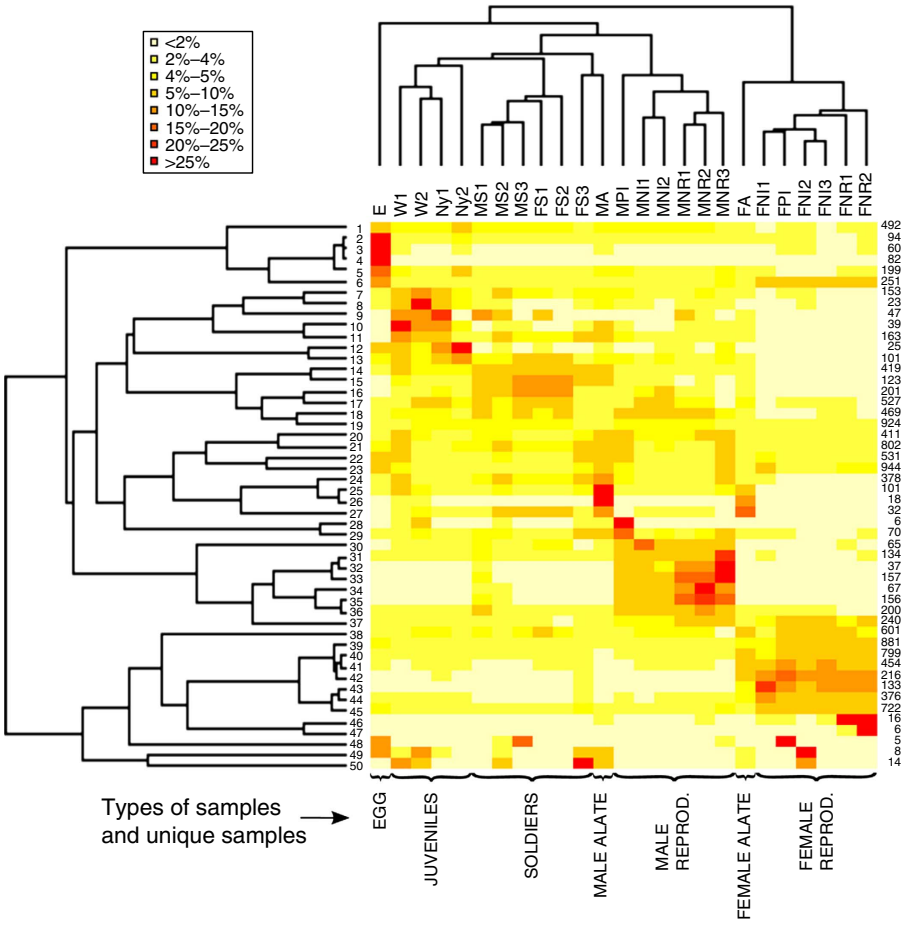


Figure 2 | Heatmap for the 50-class clustering of RNA-seq transcriptomes. Gene expression, normalized across samples to percentages, ranging from absent (pale yellow) to overexpressed (deep red). Cluster sizes are indicated on the right. Trees at the top and left correspond to hierarchical clusterings of the samples and of cluster-averaged expression, respectively. Type of samples with similar expression patterns (juveniles, soldiers, and male and female reproductives) and unique samples (egg, male and female alates), labelled at the bottom, indicate grouping for differential expression analysis.

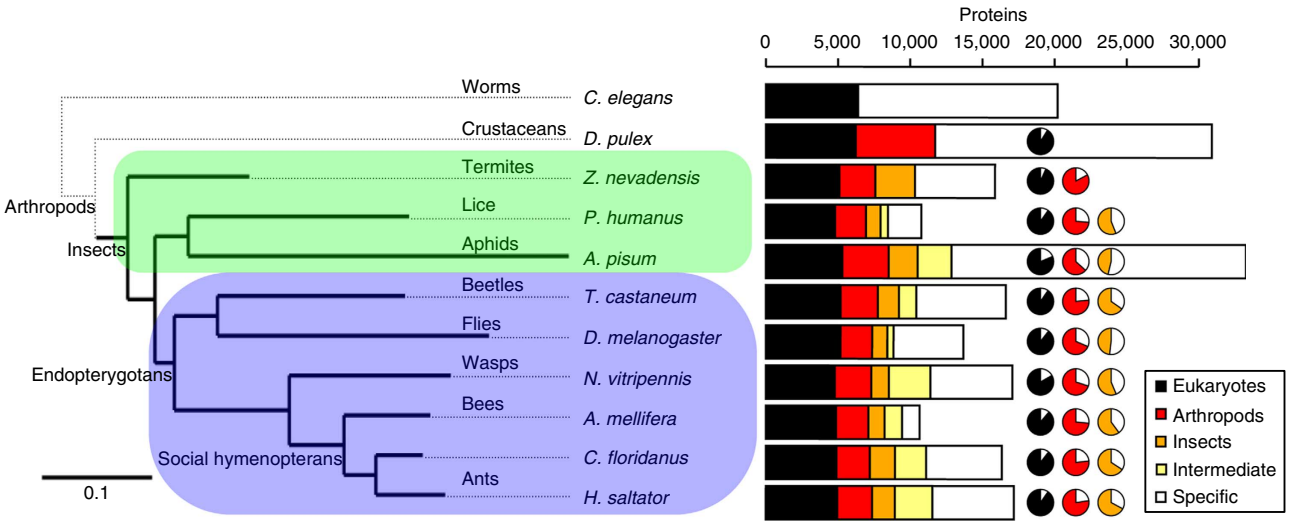


Figure 3 | Phylogeny and orthology relationships with reference genomes. Tree topology (filled bold line) was obtained by maximum likelihood analysis on proteins. Hemimetabolous and holometabolous species are depicted by a green and blue background, respectively. For each organism, proteins are represented by bars and classified according to orthoMCL analysis: 'Eukaryote' clusters with *C. elegans* and arthropods; 'Arthropod' clusters with *D. pulex* and insects; 'Insect' clusters with *Z. nevadensis* and another insect; 'Specific' groups species-specific paralogues and unclustered proteins; 'Intermediate' groups all others clustered proteins. Pie charts illustrate each species' representativeness of the total number of clusters in the categories eukaryote, arthropod and insect.

© 2014 Macmillan Publishers Limited. All rights reserved.

putative associations with male spermatogenesis or cell division (Table 1; Fig. 4; Supplementary Note 6; Supplementary Figs 9–13). Kelch-like protein 10 (KLHL10 with BTB-BACK-Kelch tri-domain) and the seven-in-absentia (SINA) proteins have been associated with E3 ubiquitin ligase complexes involved in protein degradation in spermatids[19,20]. Alpha-tubulins interact with SINA during cell division[21] while the human homologue to SINA binds a member of the polycystin (*PKD*) genes that are regularly expressed in testes[22,23]. In addition to their expansion, 19 members of the *KLHL10* family, and one *PKD* and one *alpha-tubulin* show signs of positive selection (Supplementary Table 16). Of note, we found that genes with male-specific overexpression mainly occurred within *Z. nevadensis*-specific subtrees in each family, with 100% (25 of 25) *KLHL10* genes, 67% (4 of 6) *SINA* genes, 75% (3 of 4) *alpha-tubulins* and 88% (7 of 8) *PKD* genes in these specific subtrees being overexpressed in male reproductives. A fifth gene family of extracellular proteases (*ADAMTS*, a disintegrin and metalloprotease with thrombospondin motifs), while not significantly expanded in *Z. nevadensis*, has the largest known copy number among insects suggesting a neofunctionalization in *Z. nevadensis*. Members of the *ADAMTS* family and the related *ADAM* gene family are involved in spermatogenesis[24,25]. Four of the five *ADAMTS* genes are significantly overexpressed in reproductive males. Another significantly expanded gene family, the monodomain *Kelch*, as well as two non-expanded but significantly differentially expressed families, *BTB-Kelch* and *BACK-Kelch*, share domains with *KLHL10* genes and may have similar functions. Collectively, the data suggest an expanded role for spermatogenesis regulation in termite evolution.

Chemoperception. Expansion was also observed in genes pertaining to chemical communication, a crucial component of insect societies[26]. Annotation of the four major gene families involved in insect chemoperception (Supplementary Note 7; Supplementary Figs 14–16; Supplementary Tables 21–24) identified 336 genes in *Z. nevadensis*, of which 280 were potentially functional. While this number is much higher than is typically observed in insects, it is intermediate to that of bees and ants[7,27,28], reflecting the central role of odorants in eusociality.

While the total gene numbers are comparable, their distribution within gene families diverged greatly from what has been observed in Hymenoptera. Odorant receptors (ORs), which confer most of the specificity and sensitivity of insect olfaction, are expanded in ants (344–400)[7,9,27] and honey bees (163)[28], but only 69 (63 intact) were found in *Z. nevadensis*. While the gustatory receptor (GR) repertoire in *Z. nevadensis* of 87 genes (80 intact) is comparable to that of other social insects (range 10–97 copies), *Z. nevadensis* shows lineage-specific expansions in different gene subfamilies compared with eusocial Hymenoptera such as the carbon dioxide receptors[7,27,28] (Supplementary Note 5.5). The ionotropic receptor (IR) family, implicated in gustation and olfaction in *Drosophila*[29], is expanded to its greatest known extent in *Z. nevadensis*, with 150 genes (137 intact). Only 10–32 copies have been observed in eusocial

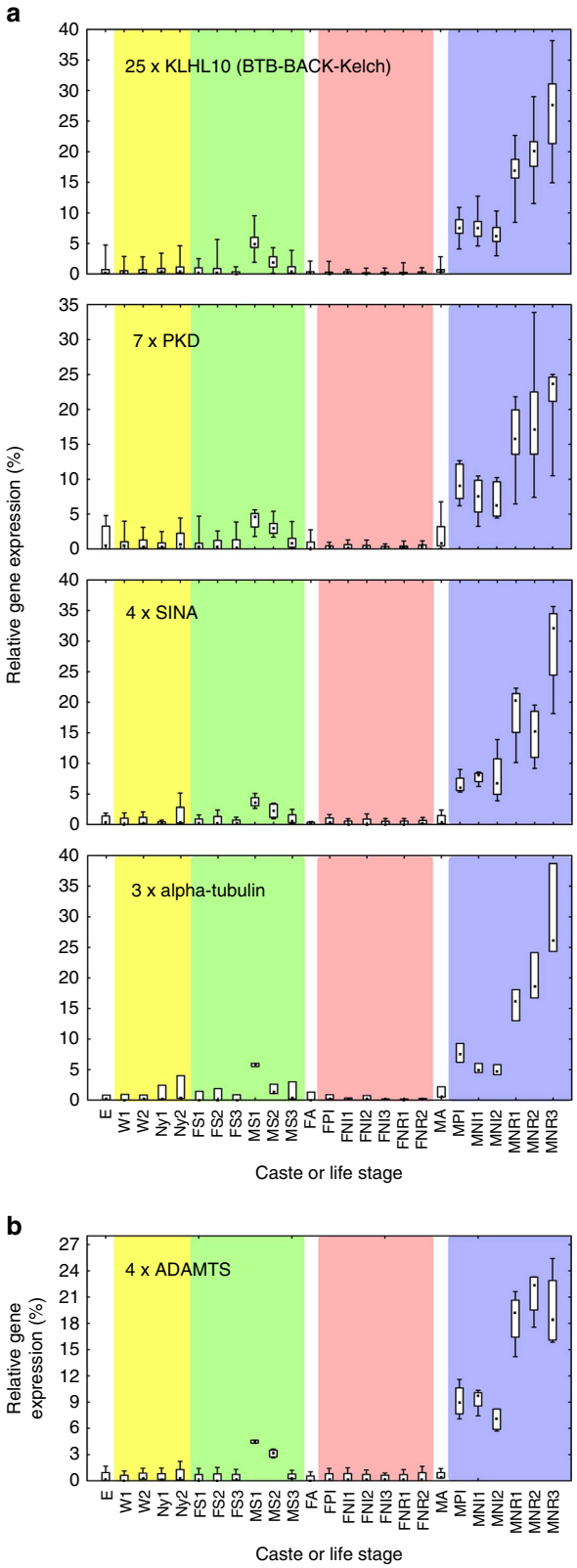


Figure 4 | Overexpressed genes in male reproductives from select gene families. Relative expression of genes that showed upregulated gene expression in male reproductives and (**a**) that occurred within *Z. nevadensis*-specific subtrees in each gene tree family and that belonged to significantly expanded gene families and (**b**) that belonged to a gene family with the largest known copy number among insects, but that is not significantly expanded. Expression of individual genes was normalized to 100% across castes and life stages. Medians, 50% intervals and minima/maxima are shown. The number of genes of the respective gene family included for each individual castes or life stages is indicated before the name of the gene family. The procedure to determine the significance of the differential expression of genes and families is detailed in Methods. Genes were significantly overexpressed at false discovery rate <0.05. Yellow bar, workers and nymphs; green bar, soldiers; pink bar, reproductive females; blue bar, reproductive males; for further abbreviations, see Fig. 1.

© 2014 Macmillan Publishers Limited. All rights reserved.

Hymenoptera species[7,27,28]. The termite IR repertoire includes 13 conserved members present throughout insects, and expansions in three subfamilies of 17, 48 and 66 genes, respectively (corresponding to two domain architectures in Table 1). The large difference in the numbers of ORs and IRs provides an opportunity to look at the organization of the olfactory lobe, the first centre for the processing of olfactory information in the insect brain. The antennal lobe is composed of densely packed glomeruli formed from axon terminals projected from receptor neurons in the antennae (Supplementary Fig. 17). Since sensory neurons expressing the same chemoreceptor extend their axons into the same glomerulus, the numbers of olfactory receptors and glomeruli in the insect antennal lobe usually match[30]. Of the 72 olfactory glomeruli of *Z. nevadensis* estimated based on histological sections (Supplementary Note 7.5), most are probably accounted for by the 63 functional ORs. As a result, only a small number of IRs and GRs can be involved in olfaction, and the remainder must be involved with gustation. The relatively low number of olfactory receptors may indicate that *Z. nevadensis* has a limited ability to discriminate odours compared with eusocial Hymenoptera.

Immunity. Parasites and pathogens are generally expected to be important drivers of social insect evolution as their colonial lifestyle creates genetically homogenous populations living in high density, that are ideal targets for infection[31]. Indeed, the decaying wood in which *Z. nevadensis* lives is also a pathogen-rich environment. To test for a proposed link between eusociality and disease resistance, we analysed immune genes in *Z. nevadensis* (Supplementary Note 8; Supplementary Fig. 18; Supplementary Table 25). We identified all of the immune-related pathways described for *Drosophila melanogaster* and other insects, including pattern recognition, signalling and gene regulation. We found six Gram-negative-binding proteins (GNBPs), which is more than in other insect genomes (maximum of 3 in *Nasonia vitripennis*) but fewer than in the crustacean *D. pulex* (10). Phylogenetic analysis revealed one general insect and five termite-specific GNBPs, supporting the hypothesis that these genes expanded early in isopteran evolution. The termite-specific group includes two GNBPs (GNBP1 and GNBP2) that seem to be under positive selection, at least in some species[32]. Five genes of the immune signalling pathway are significantly overexpressed in female reproductives (Fig. 5) probably indicating that they invest more in immune defence.

We found only three antimicrobial peptides (AMPs): attacin, diptericin and an orthologue of the termite defensin-like gene termicin. This was unexpected, given that an expansion of AMPs in the ant *Pogonomyrmex barbatus* has been proposed as a response to living in a pathogen-rich environment[7]. However, at least one of the AMPs, termicin, is under strong positive selection in several termite species[32]. These results imply that pathogens play important roles in eusocial insects but that mechanisms to fight these threats differ in a taxon-specific manner.

Reproductive division of labour. Caste differentiation and a reproductive division of labour is the hallmark of insect eusociality[26]. Proposed regulators of this division in eusocial Hymenoptera include vitellogenins (Vgs), juvenile hormone (JH), biogenic amines and modulators such as JH-binding protein, the insulin/insulin-like growth factor signalling pathway and yellow/major royal jelly protein-like genes. All of these factors appear to interact in complex ways to coordinate development with exogenous cues. We analysed these genes in detail to determine their roles in *Z. nevadensis* division of labour (Supplementary Note 9; Supplementary Figs 19–28; Supplementary Tables 26–32) and caste differentiation (Supplementary Note 10; Supplementary Fig. 29; Supplementary Table 33).

Vgs, precursors to the egg yolk protein vitellin, may also be used outside egg production, as they have been seemingly co-opted to help regulate caste determination in honey bees[33,34]. We identified four *Vgs* in *Z. nevadensis*, two of which seem to be recent duplicates that occur in tandem in the genome and are highly conserved (Supplementary Note 9.2; Supplementary

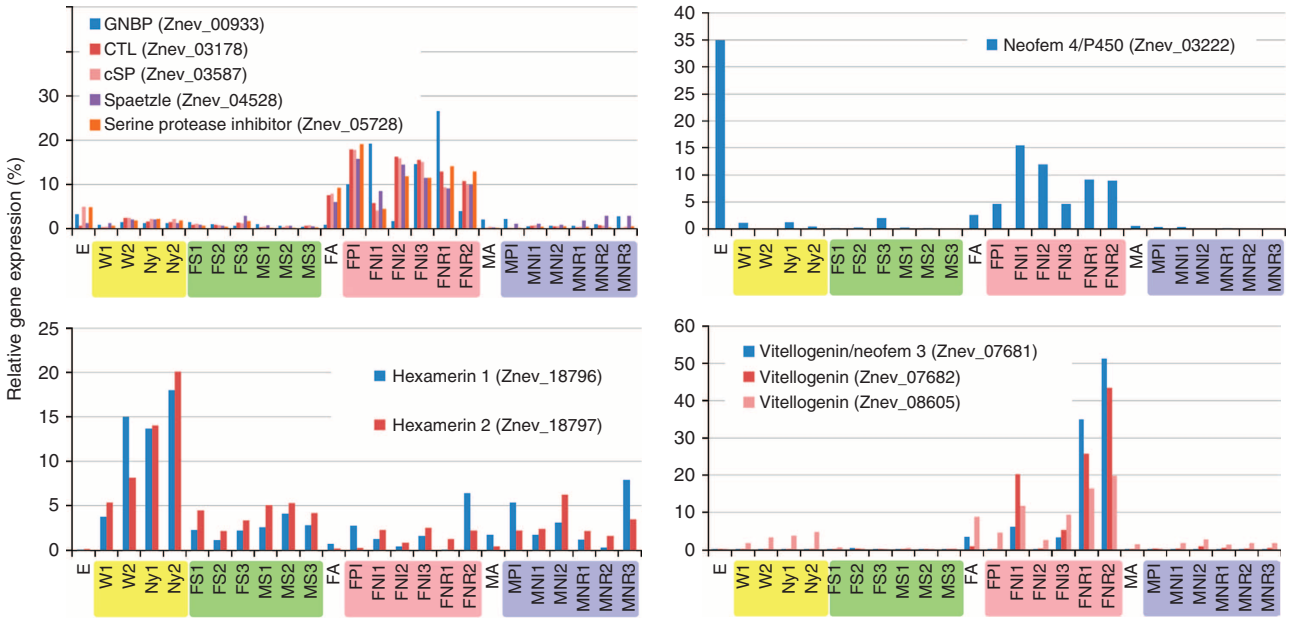


Figure 5 | Relative expression of several genes implicated in insect immunity, caste differentiation and division of labour. Percent expression is reported for immune-specific genes, hexamerins, P450 and Vgs. All genes are significantly differentially expressed (false discovery rate <0.05, see Methods). Yellow bar, workers and nymphs; green bar, soldiers; pink bar, reproductive females; blue bar, reproductive males. For further abbreviations, see Fig. 1.

© 2014 Macmillan Publishers Limited. All rights reserved.

Fig. 19). One of these duplications is closely related to *Neofem3*, a reproductive-specific gene in three other termite species (*Cryptotermes secundus*, *C. cynocephalus* and *Reticulitermes flavipes*)[35–38]. Three of the four *Vg* genes, including the recent duplications, were significantly overexpressed in reproducing queens (Fig. 5). One of the *Vg* genes is also moderately expressed in non-reproductive workers and nymphs (Fig. 5), suggesting it has a function outside oogenesis. This is similar to the expanded functionality observed in duplicated Vgs in the ants *Solenopsis invicta* and *P. barbatus*[9,39]. In these cases, as with the honey bee, Vg appears to have acquired a role in regulating behavioural caste, and a similar function may have developed in *Z. nevadensis*.

JHs have crucial and diverse functions in insect development, reproduction, longevity and both solitary and social behaviours[40]. Among the known functions of JH in termites are modulation of caste differentiation[12] and adult gonadal activity[41]. JH has different functions at different life stages, and can have multiple functions during the same stage[42]. In *Z. nevadensis*, we found all crucial enzymes of the JH III biosynthetic pathway and major regulators such as JH-binding proteins (Supplementary Note 9.3; Supplementary Table 27). We also found key enzymes in the synthesis of ecdysteroids, another essential hormone group. Unexpectedly, we found neuropeptides normally associated with moulting expressed in reproductives, suggesting a novel function, given that these adults do not moult (Supplementary Note 9.5).

Reproductive division of labour is associated with increased longevity of reproductives[43] and various histone-modifying enzymes are implicated in lifespan regulation[44]. In reproductive females, we observed significantly increased expression in two histone deacetylases, sirtuin 6 and 7, and one histone demethylase and other histone-modifying enzymes (see Table 1 and Supplementary Note 9.7; Supplementary Figs 21–25). Although the lifespan effects of sirtuin 7 are unclear[45], increased expression of sirtuin 6 leads to prolonged lifespan in male mice[46]. Along with the expression pattern in termites, sirtuins 1 and 6 are more highly expressed in longer-lived reproductive females in the ant *Harpegnathos saltator*[5]. In honey bees, queen longevity has also been linked to Vg, a possible antioxidant[47], and the overexpression of this protein in female termite reproductives may play a similar role.

In several eusocial insects reproductive division of labour is regulated through cuticular hydrocarbons[48]. In *Z. nevadensis*, reproductive status is conveyed by an abundance of four long-chained polyunsaturated alkenes[49]. Therefore, we anticipated reproductive-specific expression of genes encoding elongases and desaturases, typically required for their synthesis. Of the 16 elongase and 10 desaturase genes present in *Z. nevadensis* (Supplementary Note 9.8, Supplementary Figs 26 and 27), one of each was most highly expressed in the reproductive morphs with highly correlated expression patterns across all samples (Pearson's correlation: $N=25$, $r^2=0.93$, $P<0.00001$; Supplementary Fig. 28). The reproductive-specific coexpression of these genes makes them candidate regulators of hydrocarbon signalling in *Z. nevadensis*. Although elongase and desaturase genes involved in hydrocarbon pheromone synthesis have been identified in Diptera[50], this is the first indication of a similar combined function within the respective pathway in a eusocial insect.

Substantial progress has recently been made in the study of the molecular underpinnings of termite caste differentiation[12]. Across various species, several genes, including Cytochrome P450s and hexamerins, have been implicated in caste differentiation.

P450s are multifunctional haeme-thiolate enzymes found in all eukaryotes and bacteria. In insects, they contribute to oxidizing endogenous substrates (for example, hormones) and xenobiotic compounds (for example, secondary plant compounds). Members of the *CYP4* and *CYP15* family have been linked to JH biosynthesis and degradation in insects and termites[12,51], making them promising candidates for caste differentiation regulators. P450s have been linked to JH-dependent termite worker-to-soldier differentiation[12], and in *C. secundus* a *CYP4* gene, *Neofem4*, is specifically upregulated in reproductive females[36]. We found 76 *P450* genes *in Z. nevadensis*, with 55 having at least one complete P450 domain (Supplementary Note 10.1; Supplementary Table 33). Members of the CYP4 and CYP6 families each represent about one-third of the total. Their gene number was less than in solitary Diptera (*D. melanogaster*: 83, *Anopheles gambiae*: 111) but intermediate compared with eusocial Hymenoptera (for example, *A. mellifera*: 46; invasive Argentine ant, *Linepithema humile*: 111). Of the 69 genes with expression support, 10 were significantly overexpressed in workers and several others also exhibited caste-specific expression patterns, supporting a possible role in caste differentiation. Strikingly, a *Neofem4* orthologue is highly expressed in active female reproductives and their eggs (Fig. 5).

Hexamerins in solitary insects usually act as storage proteins[52]. We found five hexamerin genes with expression support, four of which were adjacent and probably evolved by tandem duplication (Supplementary Note 10.2; Supplementary Fig. 29). The hexamerins sorted into three groups: one grouped with other insects, two were Blattodea-wide/specific and two grouped with hexamerins found in *R. flavipes*. These latter two are involved in modulating soldier differentiation, probably by controlling JH availability[53]. Specifically, it has been proposed that hexamerins reduce JH availability in workers and nymphal stages, thereby inhibiting soldier development[54]. High hexamerin expression during these relatively plastic stages (Fig. 5) supports this hypothesis. The duplicated genes appear to have been co-opted during termite evolution to function in caste regulation, possibly by acting as a link between nutritional and hormonal (JH) signalling[55].

DNA methylation and alternative splicing. We used empirical and computational methods to determine the patterns and roles of DNA methylation in *Z. nevadensis* (Supplementary Note 11; Supplementary Table 34–38; Supplementary Fig. 30 and 31). We first identified homologues of DNA methyltransferases 1 and 3, indicating the presence of functional DNA methylation machinery. We then examined depletion of normalized CpG content (CpG o/e), an indicator of DNA methylation within animal genomes[56]. Levels of CpG depletion were 1.5–3 times greater than in other insects[57,58]. DNA methylation also preferentially targeted exons and introns (Fig. 6a)[57,58]. However, depletion of CpG dinucleotides was greater in introns than exons (Fig. 6a,b). In addition, DNA methylation in *Z. nevadensis* targeted the entire length of the gene bodies, except for the first exon following the translation start site, which was subject to lower methylation levels (Fig. 6b). Finally, we found a strong correlation between alternative splice events and depletion of CpG dinucleotides. Genes with relatively high methylation levels tended to be alternatively spliced more frequently (Fig. 6c).

Discussion

Compared with the sequenced genomes of eusocial Hymenoptera, that of *Z. nevadensis* has both similarities and some profound differences that probably reflect the very different phylogenies and life histories of these species. Among the most pronounced differences was a large expansions in genes associated with spermatogenesis. While males of the eusocial Hymenoptera produce numerous sperm, they usually finish spermatogenesis before completing metamorphosis[59].

© 2014 Macmillan Publishers Limited. All rights reserved.

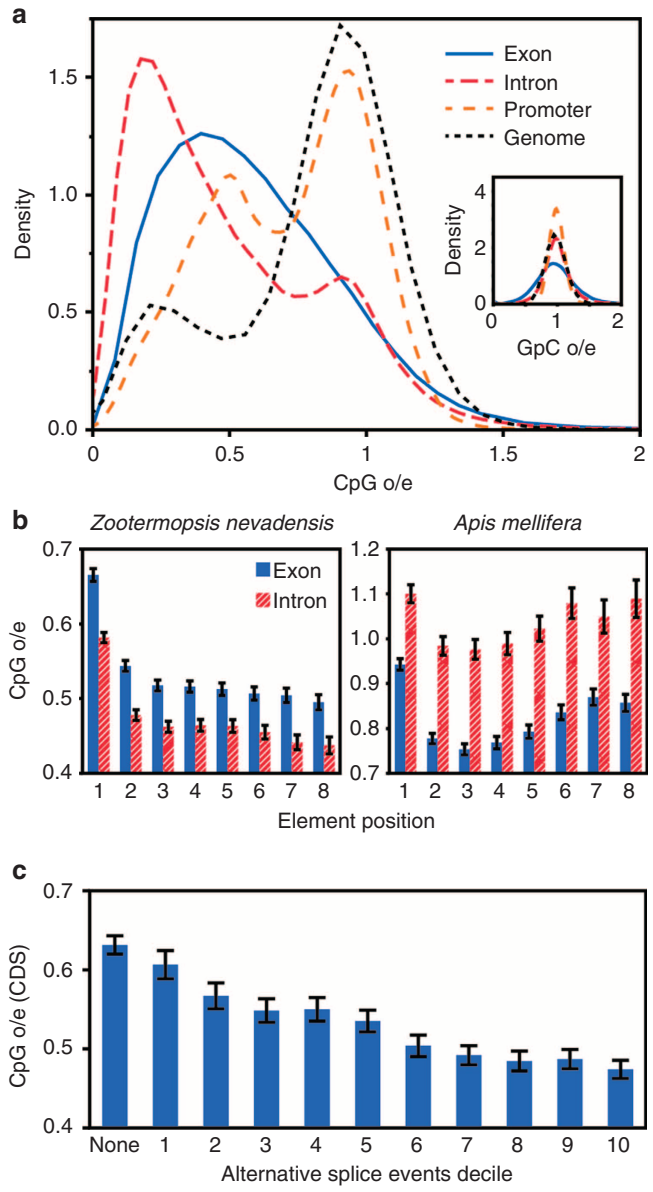


Figure 6 | Targets and regulatory correlates of DNA methylation in *Z. nevadensis*. (**a**) Distributions of CpG o/e for different genomic elements reveals substantial CpG depletion, and thus higher DNA methylation, in introns and exons relative to the genomic background ('genome', 1 kb windows) and the region 2 kb upstream of the translation start site (TSS) ('Promoter'). Inset depicts comparable distributions of GpC o/e, which control for effects unrelated to DNA methylation, such as GC content. (**b**) Positional CpG o/e of exons and introns in *Z. nevadensis* and *A. mellifera* relative to the first eight exons and introns in 5′–3′ orientation from the TSS. (Range in *n* for element: *Z. nevadensis* 4,349–12,140; *A. mellifera* 2,202–7,756.) (**c**) In *Z. nevadensis*, an increasing number of alternative splice events co-occurs with decreasing CpG o/e, suggesting that alternatively spliced genes are preferentially targeted by DNA methylation ($P<2.2\times10^{-16}$, the Kruskal–Wallis test, $n=2{,}366$ for 'none' and 904 for each decile 1–10). Means and 95% confidence intervals are plotted in **b** and **c**.

Furthermore, males generally die shortly after transferring their gametes to a receptive female. In contrast, termite males complete gamete maturation after their moult, mate repeatedly during their long lives and need to elevate sperm production throughout their lives to meet the growing requirements of an increasingly fecund queen or multiple queens. Such long-term pair bonding with remating is rare among insects[60]. Owing to changing seasonal needs, *Z. nevadensis* males cyclically activate and deactivate their testes, which may require additional adaptations. The coexpression of expanded genes of the KLHL10, SINA and alpha tubulin families in male reproductives, with potential roles in spermatogenesis, may reflect these added selective pressures on termite males.

Another striking difference is the low number of OR genes in *Z. nevadensis* compared with the eusocial Hymenoptera, which suggests differences in their ability to discriminate volatile substances and communicate with conspecifics. This difference may have evolved as a result of very different nesting behaviours. While Hymenoptera forage away from their nests, encountering a variety of odorants, including those from non-nestmate conspecifics, many of the basal termites, including *Z. nevadensis*, live their entire lives within a single log[61]. Most of the ants and the honey bee show sophisticated communication behaviour and nestmate recognition, and have an expanded number of ORs relative to *Z. nevadensis*. We would not expect that all termites have fewer ORs. The 'higher' termites, much like ants, have a more sophisticated division of labour, forage outside their nest and exhibit recruitment behaviour[61]. We predict that these species would show an increase in OR genes compared with *Z. nevadensis*, assuming the expansion of ORs is indicative of communication ability.

One area of similarity between the termite and ant genomes is an expansion of genes involved in the production of cuticular hydrocarbons used for communication. Relative to other insects, there is a greater number of desaturase genes in ants[62], although alkenes or polyunsaturated alkenes have not been found in all ants investigated[63]. *Z. nevadensis nuttingi* displays two alkadienes and two alkatrienes in the cuticular profile of reproductive individuals[49], but the number of desaturases is at the lower end of the number of putatively functional desaturase genes found in ants (10–23). If the desaturase genes are linked to more complex communication, we predict that an expansion of desaturase genes should be found in higher termites as we also predicted for genes associated with olfactory perception.

There is also evidence that immunity is an important factor in social evolution and that female reproductives invest specifically in immune defence. Compared with solitary insects there are expansions in number of immunity genes for both *Z. nevadensis* and ants[62]. However, the expansions occurred in different families, possibly as a result of different selective pressures. While the ant *P. barbatus* has a large number of AMP genes, AMPs are depleted in *Z. nevadensis*. AMPs may be counter selected to minimize deleterious effects on the microbial symbionts of the termite gut responsible for lignocellulose digestion. In addition, social hygiene and utilization of externalized antibacterial agents can reduce pathogen load in *Z. nevadensis*[64], further relaxing selection for more AMPs.

We also found evidence supporting the convergent co-option of storage proteins for regulating caste polyphenisms. Just as honey bees appear to utilize Vg to pace caste development via interactions with the endocrine system, termites may use hexamerins, P450s, and possibly Vg, as these families have gone through termite-specific gene duplication and are differentially expressed among castes. Evidence from other termites strongly indicates that hexamerins interact with JH and that a high expression of hexamerins, through interaction with JH, inhibits soldier development[53,54]. Vg may play a similar role, although this remains to be tested.

A final area of possible similarity between the eusocial insects is in their use of DNA methylation. In *Z. nevadensis*, the rate of methylation was found to be particularly high relative to other

© 2014 Macmillan Publishers Limited. All rights reserved.

insects[57,58]. DNA methylation is involved with gene regulation[65] and alternative splicing[66], and may be crucial to phenotypic plasticity such as caste differentiation. There is evidence that methylation plays a role in honey bee caste determination, specifically affecting the proportion of brood likely to develop into queens[67]. As has been observed in other insects[57,58], we found methylation primarily in the genic, rather than the intergenic regions of this termite's DNA. However, unlike the honey bee, methylation was greater in introns than exons (Fig. 6a,b). The relative similarity between levels of DNA methylation in introns and exons, as well as the lack of preferential 5′-targeting of DNA methylation (Fig. 6b) suggests that patterns of DNA methylation in *Z. nevadensis* may be more similar to those of basal invertebrate chordates[68], which exhibit relatively high levels of intragenic DNA methylation compared with those of holometabolous insects[68]. Regardless, the association between alternative splicing and DNA methylation we observed in this termite (Fig. 6c) supports the hypothesis that intragenic DNA methylation interacts with messenger RNA splicing to produce an array of phenotypes in eusocial insects[67,69].

Collectively, the results of our genome analyses substantially improve our understanding of the mechanisms that have allowed termites to develop and maintain a high degree of social complexity, providing a much needed comparative counterpoint to the wealth of genomic information available for eusocial Hymenoptera. These initial results highlight some of the commonalities and differences that arise from similar needs balanced with phylogenetic and environmental constraints. In addition, having this information for such a basal species greatly facilitates future endeavours to understand the evolution of insects in general.

Methods

Source of samples. Colonies of *Z. nevadensis nuttingi* were collected in their entirety within wood logs from Pebble Beach near Monterey, California, in November 2010. Species identity of each colony was confirmed by cuticular hydrocarbon analysis using gas chromatography–mass spectrometry. Colonies were transferred to artificial nests consisting of layered precavitated sheets of presoaked spruce (*Pinus glabra*). Nests were kept moist by periodic spraying with distilled water and were maintained in transparent plastic boxes under a 12L:12D light cycle at 20.5 °C. Colony 133 was used to provide all samples for genome sequencing (for additional details see Supplementary Note 1).

Genome sequencing. Sequencing reads were obtained by whole genome shotgun strategy using an Illumina Hiseq 2,000 at the BGI-Shenzhen. Seven paired-end libraries were constructed with insert sizes of 200, 500 and 800 bp, and 2, 5, 10 and 20 kb. DNA samples for genome sequences were derived from soldier heads to minimize contamination with gut content. Fifty heads were extracted for the construction of libraries up to 2 kb, while the DNA from another 150 heads was used for 5–20 kb libraries. In total, we obtained 68 Gb of raw reads. Before assembling, several filtering steps were applied to remove the following:

(1) reads with more than 10% of 'N' bases or polyA;
(2) low-quality reads, with more than 30 low-quality bases (Phred score $\leq$7);
(3) reads with adapter contamination (>10 bp adapter sequence, allowing a maximum of three mismatches);
(4) paired reads overlapping each other (>10 bp, allowing 10% mismatch); and
(5) PCR duplicates with identical reads between two paired-end reads.

The estimated genome size using the 17-nucleotide depth distribution was 562 Mb, which is similar to a previously published estimation. The averaged coverage depth is 98.4-fold and 92% of the bases have more than 20-fold coverage (Supplementary Fig. 1). Statistics are provided in Supplementary Table 2.

Genome annotation. The clean reads were assembled by SOAPdenovo into 7,049,535 preliminary contigs covering 396.3 Mbp, of which 613,353 were longer than 100 bp, covering 395.6 Mbp. After scaffolding, these contigs were assembled into 93,931 scaffolds (including 85,940 singleton contigs), yielding a 493-Mb assembly with 21.3 Mb of Ns. The result suggests that the assembly covers 88% of the *Z. nevadensis* genome, given the estimated genome size of 562 Mb. The scaffold N50 length of the assembly is 740 kb and the contig (continuous fragments extracted from the final assembly) N50 length is 20 kb. More details about contig/scaffold number and length are provided in Supplementary Table 3. Genome sequence and annotation data are available at http://www.termitegenome.org/?q=consortium_datasets. A genome browser is available at http://www.termitegenome.org/?q=browser.

Protein-coding prediction. As a first step, three different methods were used to predict gene models: homology-based, RNA-seq-based and *de novo*. For homology-based gene models, protein sequences from *A. mellifera*, *D. melanogaster*, *Homo sapiens* and two ants (*Camponotus floridanus* and *H. saltator*) were used. For each of these five species, the prediction pipeline included the following steps:

(1) use of TBLASTN (*E*-value < 1e − 5) for homology search;
(2) selection of the most similar gene loci when there were multiple candidates;
(3) exclusion of regions with identity <50%;
(4) use of GeneWise v2.0 to generate gene model structures; and
(5) for incomplete gene models, search of 30 bp in the upstream/downstream region to find start/stop codons (7,764 open reading frames (ORFs) completed).

Finally, the five homology-based gene predictions were merged into a union set of 20,005 genes (selecting the longest gene models when models overlapped). For RNA-seq data, we built transcriptomes from samples of different life stages or castes of *Z. nevadensis* (see Supplementary Note 2). TopHat v1.3.3 was used to align raw reads against the genome to identify exon–exon splice junctions, and then Cufflinks v0.8.2 was used to reconstruct 1,232,735 transcripts from the spliced alignments. Applying the same merging pipeline as for homology-based predictions to the obtained transcripts resulted in 38,123 gene models with intact ORFs. For *de novo* prediction, Augustus and SNAP programs were used. After masking the repeats in the genome, 500 genes from homologue-based prediction (with intact ORFs) were selected to train Augustus and SNAP. As a result, 21,224 gene models were predicted by Augustus and 43,140 by SNAP.

As a second step, gene models resulting from the three methods were merged into an integrated gene set through multiple filtering steps as follows:

(1) RNA-seq gene models were separated into two sets: multiple-exon and single-exon (resp. 11,900 and 26,223 gene models, respectively). As many single-exon genes tend to be incomplete transcripts, we kept the multiple-exon set as the basis of the integrated gene set and subsequent steps were performed to improve the basic set.
(2) If more than one gene model of the multiple-exon set overlapped a unique gene model from homology-based prediction, the homology-based model took precedence over the RNA-seq model.
(3) Gene models from homology-based prediction not supported by the RNA-seq multiple-exon set but with good homology evidence (Genewise scores $\geq$80 and CDS length $\geq$150 bp) were added to the integrated gene set.
(4) Single-exon genes from RNA-seq data supported by homology-based prediction, where the homology-based prediction was also a single-exon model, were added to the integrated gene set.
(5) Genes from *de novo* prediction, which did not overlap with any gene in the integrated gene set, were added to the gene set, if they obtain a significant hit (BLASTP *E*-value < 1e − 5) to a Swissprot protein.
(6) Genes containing transposon-related Interpro domains were removed.
(7) Manual curation for some genes of interest was performed using the Apollo Genome Annotation editor[70]. Manual annotations are available at http://www.termitegenome.org/?q=consortium_datasets.

We obtained an integrated gene set (named OGS v2.1) of 17,737 gene models (Supplementary Table 8). Most of them have expression support (73%) while just 2.7% are predicted by *de novo* approaches only (Supplementary Fig. 2). Subsequently, ~1,186 genes were identified as transposon-related genes through Interpro domain annotation and orthoMCL clustering (see below). These genes were not considered in some of the following analyses (for example, the transcriptomics analysis Supplementary Note 2), and they were removed from the next gene set release (OGS v2.2—consisting of 15,876 proteins). Finally, the OGSv2.2 contains additional gene models built through manual curation.

ncRNA annotation. Four types of ncRNAs were annotated in our analysis: transfer RNA (tRNA), ribosomal RNA (rRNA), microRNA and small nuclear RNA (Supplementary Table 4). tRNA genes were predicted by tRNAscan-SE with eukaryote parameters. rRNA fragments were identified by aligning the rRNA template sequences from invertebrate animals to the *Z. nevadensis* genome using BLASTN with an *E*-value cutoff of 1e − 5. MicroRNA and small nuclear RNA genes were inferred by the INFERNAL software, using release 9.1 of the Rfam database.

Repeat annotation. The presence of repeats in the genome was examined using two different approaches:

First, known transposable elements (TEs) were identified using RepeatMasker (version 3.2.6) against the Repbase. This step identified 13 Mb of known TEs, comprising 2.6% of the genome (Supplementary Table 5).

© 2014 Macmillan Publishers Limited. All rights reserved.

Second, a *de novo* repeat library was constructed using RepeatScout with default parameters. The generated consensus sequence for each repeat family was then used as reference in RepeatMasker to identify additional high and medium copy repeats (>10 copies) in the genome assembly. This allowed us to identify an additional 119 Mb of repetitive sequences spanning 24.3% of the genome (Supplementary Table 6), consisting primarily of unknown repeats. For non-interspersed repeat sequences, we ran RepeatMasker with the '-noint' option, which is specified for simple repeats, satellites and low complexity repeats. Tandem repeats were also predicted using Tandem Repeat Finder software, with parameters set to 'Match = 2, Mismatch = 7, Delta = 7, PM = 80, PI = 10, Minscore = 50 and MaxPeriod = 12'.

In total, repetitive sequences make up 26% of the assembled genome (Supplementary Table 7).

Protein functional annotation. The function of proteins was predicted using three methods: protein domains, GO and KEGG pathway.

Domains are the evolutionary units of protein-coding genes and their emergence and modular rearrangements are strongly associated with adaptive processes that are not always obvious at the gene level. Initial and principal domain annotation was performed using the Pfam database (release 24.0) and HMMER software (version 3.0). Only domains satisfying the Pfam-recommended thresholds (gathering cutoffs) were retained. Overlaps were resolved heuristically by selecting the domain with a strategy based on the best *E*-value. Additional domains were assigned using Interproscan (v4.3) including SUPERFAMILY, GENE3D, TIGRFAMs, SMART, PROSITE and PRINTS domain models. Proteins were further annotated by GO terms. We created GO annotations with two levels of confidence.

The first level is high confidence. Pfam domains and other domains from the previously mentioned databases are annotated by GO terms (pfam2go, smart2go and so on mappings), as well as the Interpro entries (interpro2go). The annotation of a given domain corresponds to the GO terms shared by all annotated proteins possessing this domain (or significantly over-represented for superfamily2go). Hence, when a domain is identified in a protein, the GO annotation of this domain can be safely transferred to that protein.

The second level is low confidence. Blast2GO, which is prone to 35% annotation errors, was applied with default values to transfer GO terms significantly over-represented in the best BLASTP hits of termite's proteins against the NCBI non-redundant database. Note that based on the BLASTP results against the NCBI, we created the list of orphan genes in *Z. nevadensis*, if no detectable homology (*E*-value < 10e − 3) was obtained.

The KEGG Automatic Annotation Server was used to assign proteins to KEGG orthology (KO) groups, using the recommended eukaryote set plus all other available arthropods as references. The KO system is the basis of the KEGG database, since it links annotated proteins to KEGG's metabolic pathways (PATHWAY) and functional ontology (BRITE). We produced two KO annotation sets: single-best hit was used for preliminary analysis of gene family expansion/contraction, while BBH (bi-directional best hit) was used for preliminary gain-and-loss analysis.

Transcriptome analysis. Transcriptomes samples were collected from several colonies. Raw reads were aligned on the genome using TopHat. We used edgeR to normalize libraries across samples (trimmed mean of *M*-values) and to identify differentially expressed genes. Significant over- or underexpression in unique samples without replicates (egg, male alate, female alate) was determined through pairwise comparisons to other types of samples with replicates (juveniles, soldiers, male reproductives, female reproductives), using the DEseq software, with a stringent threshold of 0.01 for the false discovery rate (that is, corrected *P* values with the Benjamini–Hochberg formula). Differentially expressed genes in types of samples including replicates were identified using the edgeR package with a false discovery rate threshold of 0.05 against all other types of samples and further exclusion of differentially expressed genes in unique samples. From these lists of individual genes, we identified gene families for which a significant number of members, as determined by Fisher's exact test, were differentially expressed. Expression values for each gene were calculated using the reads per kilo base per million formula. Genes with low expression levels (reads per kilo base per million <5 in all samples) were removed to reduce possible bias. Clustering, using K-means with Euclidean distance was determined with Cluster3.0 and visualized with R.

Comparative analysis. Domain architectures was used to define protein families and compare the *Z. nevadensis* protein repertoire with those of *D. melanogaster*, *T. castaneum*, *N. vitripennis*, *A. mellifera*, *C. floridanus*, *H. saltator*, *A. pisum*, *P. humanus* and the crustacean *D. pulex*. Families exhibiting expansion were detected using Fisher's exact tests (pairwise comparisons of *Z. nevadensis* with other genomes) on protein domain architecture counts using Pfam, or Interpro for uncovered families (Supplementary Note 5.5). In addition, the OrthoMCL procedure was run using standard parameters on these species and *C. elegans* to allow a finer perspective over subfamilies (Supplementary Note 3.2).

Phylogenetic analysis. We conducted phylogenetic analysis to ascertain the position of termites in the evolution of arthropods. Several data sets of orthologous proteins were tested (see below), but the procedure used for phylogenetic reconstructions was the same for each data set:

(1) protein sequences were aligned with MAFFT;
(2) alignments were cleaned with Gblocks;
(3) protein alignments were concatenated into a unique protein superalignment;
(4) the underlying DNA superalignment was deduced using only the first two nucleotides of each codon (custom scripts); and
(5) four topologies were computed using the two superalignments (nucleotides and amino acids) from two perspectives: a maximum likelihood approach (morePhyML script based on PhyML) and a Bayesian approach (Phylobayes software). For the evolutionary models, we use standard settings, that is the LG + Γ4 + I model for morePhyML with amino acids and the GTR + Γ4 + I model for nucleotide data, and the GTR + Γ4 + CAT model for Phylobayes.

Preliminary analyses followed the classical approach for phylogeny at the genome level, that is, we used 2,318 orthoMCL clusters with 1:1 orthologues in the nine arthropod reference species (see Supplementary Table 13). However, because of the previously mentioned limitations (the sparse genome sampling and the ancient rapid radiation of lower Neoptera, and the fast evolution of Paraneoptera representatives, especially the pea aphid *A. pisum*, causing 'long branch attractions' [LBA]), four different topologies were obtained regarding the relative branching of the termite *Z. nevadensis* and the Paraneoptera (*P. humanus* and *A. pisum*). Then, to limit the LBA, several filters were tested including evolution speed (as measured by lowest rates of substitution rates), reduced compositional bias and domain composition. However, none of the filter was successful to obtain an agreement of the four topologies.

For secondary analyses, we excluded the *D. pulex* genome since its position in the tree was already clear while it was likely a burden for the tree reconstruction for several reasons:

(1) it is known for its fast evolution (highly adaptive);
(2) it contains a large number of paralogues (see also above for *D pulex*-specific genome features). With likely differential loss of genes, this might produce incorrect 1:1 orthology relationships;
(3) we observed that *D. pulex* has many split gene models (see Supplementary Note 3.4), which may probably be misinterpreted as highly derived genes and thus introduce further artefacts;
(4) the high phylogenetic distance of this non-insect species from the common ancestor of termites and Paraneoptera might prevent clarification of the ancient rapid radiation of these taxa.

We then searched for clear 1:1 orthologues in the expressed sequence tag (EST) data sets of the NCBI non-redundant database. We used reciprocal BLAST best hits with the termite proteins as queries and required a threshold *E*-value of 10e − 50 and a match covering at least 60% of the proteins. We identified three taxa, Diplura, Archaeognatha and Thysanura, offering a common set of 16 orthologous proteins. These taxa are outgroups of Neoptera and belong to the Hexapoda clade (Supplementary Fig. 4). The topologies resulting from this second analysis agreed with the most recently published topologies and the fourth topology had a polytomy that did not contradict the other proposed grouping (Supplementary Fig. 5).

References

1. Rust, M. K. & Su, N. Y. Managing social insects of urban importance. *Annu. Rev. Entomol.* **57,** 355–375 (2012).
2. Abe, T., Bignell, D. E. & Higashi, M. *Termites: Evolution, Sociality, Symbioses, Ecology* (Springer, 2000).
3. Korb, J. Termites hemimetabolous diploid white ants? *Front. Zool.* **5,** 15 (2008).
4. Inward, D., Beccaloni, G. & Eggleton, P. Death of an order: a comprehensive molecular phylogenetic study confirms that termites are eusocial cockroaches. *Biol. Lett.* **3,** 331–335 (2007).
5. Bonasio, R. *et al.* Genomic comparison of the ants *Camponotus floridanus* and *Harpegnathos saltator*. *Science* **329,** 1068–1071 (2010).
6. Smith, C. R. *et al.* Draft genome of the red harvester ant *Pogonomyrmex barbatus*. *Proc. Natl Acad. Sci. USA* **108,** 5667–5672 (2011).
7. Smith, C. D. *et al.* Draft genome of the globally widespread and invasive Argentine ant (*Linepithema humile*). *Proc. Natl Acad. Sci. USA* **108,** 5673–5678 (2011).
8. Suen, G. *et al.* The genome sequence of the leaf-cutter ant *Atta cephalotes* reveals insights into its obligate symbiotic lifestyle. *PLoS Genet.* **7,** e1002007 (2011).
9. Wurm, Y. *et al.* The genome of the fire ant *Solenopsis invicta*. *Proc. Natl Acad. Sci. USA* **108,** 5679–5684 (2011).
10. Nygaard, S. *et al.* The genome of the leaf-cutting ant *Acromyrmex echinatior* suggests key adaptations to advanced social life and fungus farming. *Genome Res.* **21,** 1339–1348 (2011).
11. The Honeybee Genome Sequencing Consortium. Insights into social insects from the genome of the honeybee *Apis mellifera*. *Nature* **443,** 931–949 (2006).

© 2014 Macmillan Publishers Limited. All rights reserved.

12. Miura, T. & Scharf, M. E. in *Biology of Termites: A Modern Synthesis*. (eds Bignell, D. E., Roisin, Y. & Lo, N.) 211–253 (Springer, 2011).
13. Booth, W. *et al.* Population genetic structure and colony breeding system in dampwood termites (*Zootermopsis angusticollis* and *Z. nevadensis nuttingi*). *Insect. Soc.* **59,** 127–137 (2012).
14. Koshikawa, S., Miyazaki, S., Cornette, R., Matsumoto, T. & Miura, T. Genome size of termites (Insecta, Dictyoptera, Isoptera) and wood roaches (Insecta, Dictyoptera, Cryptocercidae). *Naturwissenschaften* **95,** 859–867 (2008).
15. Tautz, D. & Domazet-Lošo, T. The evolutionary origin of orphan genes. *Nat. Rev. Genet.* **12,** 692–702 (2011).
16. Whitfield, J. B. & Kjer, K. M. Ancient rapid radiations of insects: challenges for phylogenetic analysis. *Annu. Rev. Entomol.* **53,** 449–472 (2008).
17. Shah, N., Dorer, D. R., Moriyama, E. N. & Christensen, A. C. Evolution of a large, conserved, and syntenic gene family in insects. *G3 (Bethesda)* **2,** 313–319 (2012).
18. Ferguson, L. C., Green, J., Surridge, A. & Jiggins, C. D. Evolution of the insect yellow gene family. *Mol. Biol. Evol.* **28,** 257–272 (2011).
19. Kaplan, Y., Gibbs-Bar, L., Kalifa, Y., Feinstein-Rotkopf, Y. & Arama, E. Gradients of a ubiquitin E3 ligase inhibitor and a caspase inhibitor determine differentiation or death in spermatids. *Dev. Cell* **19,** 160–173 (2010).
20. Lorick, K. L. *et al.* RING fingers mediate ubiquitin-conjugating enzyme (E2)-dependent ubiquitination. *Proc. Natl Acad. Sci. USA* **96,** 11364–11369 (1999).
21. Germani, A. *et al.* SIAH-1 interacts with alpha-tubulin and degrades the kinesin Kid by the proteasome pathway during mitosis. *Oncogene* **19,** 5997–6006 (2000).
22. Kim, H., Jeong, W., Ahn, K., Ahn, C. & Kang, S. Siah-1 interacts with the intracellular region of polycystin-1 and affects its stability via the ubiquitin-proteasome pathway. *J. Am. Soc. Nephrol.* **15,** 2042–2049 (2004).
23. Zhou, J. Polycystins and primary cilia: primers for cell cycle progression. *Ann. Rev. Physiol.* **71,** 83–113 (2009).
24. Porter, S., Clark, I. M., Kevorkian, L. & Edwards, D. R. The ADAMTS metalloproteinases. *Biochem. J.* **386,** 15–27 (2005).
25. Edwards, D. R., Handsley, M. M. & Pennington, C. J. The ADAM metalloproteinases. *Mol. Aspects Med.* **29,** 258–289 (2008).
26. Wilson, E. O. *The Insect Societies* (Harvard Univ. Press, 1971).
27. Zhou, X. *et al.* Phylogenetic and transcriptomic analysis of chemosensory receptors in a pair of divergent ant species reveals caste-specific signatures of odor coding. *PLoS Genet.* **8,** e1002930 (2012).
28. Robertson, H. M. & Wanner, K. W. The chemoreceptor superfamily in the honey bee, *Apis mellifera*: expansion of the odorant, but not gustatory, receptor family. *Genome Res.* **16,** 1395–1403 (2006).
29. Benton, R., Vannice, K. S., Gomez-Diaz, C. & Vosshall, L. B. Variant ionotropic glutamate receptors as chemosensory receptors in *Drosophila*. *Cell* **136,** 149–162 (2009).
30. Vosshall, L. B., Wong, A. M. & Axel, R. An olfactory sensory map in the fly brain. *Cell* **102,** 147–159 (2000).
31. Schmid-Hempel, P. *Parasites in Social Insects* (Princeton Univ. Press, 1998).
32. Bulmer, M. S. Evolution of immune proteins in insects. eLS (2010).
33. Amdam, G. V., Norberg, K., Hagen, A. & Omholt, S. W. Social exploitation of vitellogenin. *Proc. Natl Acad. Sci. USA* **100,** 1799–1802 (2003).
34. Nelson, C. M., Ihle, K. E., Fondrk, M. K., Page, R. E. & Amdam, G. V. The gene vitellogenin has multiple coordinating effects on social organization. *PLoS Biol.* **5,** e62 (2007).
35. Korb, J., Weil, T., Hoffmann, K., Foster, K. R. & Rehli, M. A gene necessary for reproductive suppression in termites. *Science* **324,** 758–758 (2009).
36. Weil, T., Rehli, M. & Korb, J. Molecular basis for the reproductive division of labour in a lower termite. *BMC Genomics* **8,** 198 (2007).
37. Scharf, M. E., Wu-Scharf, D., Zhou, X., Pittendrigh, B. R. & Bennett, G. W. Gene expression profiles among immature and adult reproductive castes of the termite *Reticulitermes flavipes*. *Insect Mol. Biol.* **14,** 31–44 (2005).
38. Weil, T., Korb, J. & Rehli, M. Comparison of queen-specific gene expression in related lower termite species. *Mol. Biol. Evol.* **26,** 1841–1850 (2009).
39. Corona, M. *et al.* Vitellogenin underwent subfunctionalization to acquire caste and behavioral specific expression in the harvester ant *Pogonomyrmex barbatus*. *PLoS Genet.* **9,** e1003730 (2013).
40. Nijhout, H. F. *Insect Hormones* (Princeton Univ. Press, 1994).
41. Brent, C. S., Schal, C. & Vargo, E. L. Endocrine effects of social stimuli on maturing queens of the dampwood termite *Zootermopsis angusticollis*. *Physiol. Entomol.* **32,** 26–33 (2007).
42. Brent, C. S. in *Organization of Insect Societies*. (eds Gadau, J. & Fewell, J) 105-127 (Harvard Univ. Press, 2009).
43. Keller, L. & Genoud, M. Extraordinary lifespans in ants: a test of evolutionary theories of ageing. *Nature* **389,** 958–960 (1997).
44. Finkel, T., Deng, C. X. & Mostoslavsky, R. Recent progress in the biology and physiology of sirtuins. *Nature* **460,** 587–591 (2009).
45. Tsai, Y. C., Greco, T. M., Boonmee, A., Miteva, Y. & Cristea, I. M. Functional proteomics establishes the interaction of SIRT7 with chromatin remodeling complexes and expands its role in regulation of RNA polymerase I transcription. *Mol. Cell. Proteomics* **11,** 60–76 (2012).
46. Kanfi, Y. *et al.* The sirtuin SIRT6 regulates lifespan in male mice. *Nature* **483,** 218–221 (2012).
47. Amdam, G. V. *et al.* Hormonal control of the yolk precursor vitellogenin regulates immune function and longevity in honeybees. *Exp. Gerontol.* **39,** 767–773 (2004).
48. Liebig, J. in *Insect Hydrocarbons: Biology, Biochemistry, and Chemical Ecology*. (eds Blomquist, G. J. & Bagnères, A. G.) 254-281 (Cambridge Univ. Press, 2010).
49. Liebig, J., Eliyahu, D. & Brent, C. Cuticular hydrocarbon profiles indicate reproductive status in the termite *Zootermopsis nevadensis*. *Beh. Ecol. Sociobiol.* **63,** 1799–1807 (2009).
50. Wicker-Thomas, C. & Chertemps, T. in *Insect Hydrocarbons: Biology, Biochemistry, and Chemical Ecology* (eds Blomquist, G. J. & Bagnères, A. G.) 53-74 (Cambridge Univ. Press, 2010).
51. Feyereisen, R. in *Insect Molecular Biology and Biochemistry*. (ed Gilbert L. I.) 236-316 (Elsevier, 2012).
52. Burmester, T. & Scheller, K. Ligands and receptors: common theme in insect storage protein transport. *Naturwissenschaften* **86,** 468–474 (1999).
53. Zhou, X., Tarver, M., Bennett, G., Oi, F. & Scharf, M. Two hexamerin genes from the termite *Reticulitermes flavipes*: sequence, expression, and proposed functions in caste regulation. *Gene* **376,** 47–58 (2006).
54. Zhou, X., Oi, F. M. & Scharf, M. E. Social exploitation of hexamerin: RNAi reveals a major caste-regulatory factor in termites. *Proc. Natl Acad. Sci. USA* **103,** 4499–4504 (2006).
55. Scharf, M. E., Buckspan, C. E., Grzymala, T. L. & Zhou, X. Regulation of polyphenic caste differentiation in the termite *Reticulitermes flavipes* by interaction of intrinsic and extrinsic factors. *J. Exp. Biol.* **210,** 4390–4398 (2007).
56. Yi, S. V. & Goodisman, M. A. D. Computational approaches for understanding the evolution of DNA methylation in animals. *Epigenetics* **4,** 551–556 (2009).
57. Glastad, K. M., Hunt, B. G. & Goodisman, M. A. D. Evidence of a conserved functional role for DNA methylation in termites. *Insect Mol. Biol.* **22,** 143–154 (2013).
58. Glastad, K. M., Hunt, B. G., Yi, S. V. & Goodisman, M. A. D. DNA methylation in insects: on the brink of the epigenomic era. *Insect Mol. Biol.* **20,** 553–565 (2011).
59. Heinze, J. & Hölldobler, B. Fighting for a harem of queens—physiology of reproduction in *Cardiocondyla* male ants. *Proc. Natl Acad. Sci. USA* **90,** 8412–8414 (1993).
60. Boomsma, J. J. Beyond promiscuity: mate-choice commitments in social breeding. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **368 (2013).**
61. Korb, J., Buschmann, M., Schafberg, S., Liebig, J. & Bagnères, A. G. Brood care and social evolution in termites. *Proc. Biol. Sci.* **279,** 2662–2671 (2012).
62. Simola, D. F. *et al.* Social insect genomes exhibit dramatic evolution in gene composition and regulation while preserving regulatory features linked to sociality. *Genome Res.* **23,** 1235–1247 (2013).
63. Martin, S. & Drijfhout, F. A review of ant cuticular hydrocarbons. *J. Chem. Ecol.* **35,** 1151–1161 (2009).
64. Rosengaus, R. B., Traniello, J. F. & Bulmer, M. S. in *Biology of Termites: A Modern Synthesis*. (eds Bignell, D. E., Roisin, Y., Lo, N.) 165-191 (Springer, 2011).
65. Jones, P. A. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat. Rev. Genet.* **13,** 484–492 (2012).
66. Shukla, S. *et al.* CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. *Nature* **479,** 74–79 (2011).
67. Kucharski, R., Maleszka, J., Foret, S. & Maleszka, R. Nutritional control of reproductive status in honeybees via DNA methylation. *Science* **319,** 1827–1830 (2008).
68. Zemach, A., McDaniel, I. E., Silva, P. & Zilberman, D. Genome-wide evolutionary analysis of eukaryotic DNA methylation. *Science* **328,** 916–919 (2010).
69. Li-Byarlay, H. *et al.* RNA interference knockdown of DNA methyltransferase 3 affects gene alternative splicing in the honey bee. *Proc. Natl Acad. Sci. USA* **110,** 12750–12755 (2013).
70. Lewis, S. E. *et al.* Apollo: a sequence annotation editor. *Genome Biol.* **3,** research0082-0082.14 (2002).

Acknowledgements

We thank the administrators of the Pebble Beach Company for permission to collect termites and Navdeep Mutti for initial help in RNA and DNA sampling. This work was supported by the Agriculture and Food Research Initiative Competitive Grant number 2007-35302-18172 from the USDA National Institute of Food and Agriculture to J.L. and C.S.B.; and a research grant from the Deutschen Forschungsgemeinschaft (DFG) to J.K. (KO1895/6) and LOEWE Research Focus 'Insect Biotechnology' to A.V. Mention of trade names or commercial products in this article is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the US Department of Agriculture. USDA is an equal opportunity provider and employer.

© 2014 Macmillan Publishers Limited. All rights reserved.

Author contributions

Project design: E.B.-B., J.K., G.Z. and J.L.; termite collection: C.S.B. and J.L.; population genetics: W.B. and E.L.V.; genome assembly: C.L., J.W. and Zu.Y.; functional annotation: C.L., H.P., N.T. and W.Y.; manual annotation: J.G., A.K.H., L.J., S.M.S.K., J.A.M., X.M., H.M.R., M.E.S., N.T., A.V. and J.X.; gene set annotation: H.H., C.L., J.W., Zh.Y., Zu.Y. and J.Zho., repeat annotation: F.S.; genome browser BGI: Z.C., Apollo browser: C.P.C., C.G.E., J.T.R. and M.C.M.T.; transcriptomics analysis: L.J., H.H., X.M., Zh.Y and N.T.; comparative analysis: A.K.H., L.J., C.L., D.A.L., R.A.H., X.M., N.T., J.W. and J.X.; phylogenetic analysis: J.G., C.L. and N.T.; biogenic amine receptors: J.A.M.; cytochrome P450: M.E.S., A.K.H. and J.G.; desaturases, elongases: J.L. and N.T.; expanded male-specific genes: J.L., C.L. and N.T.; histone-modifying enzymes: K.G.; immune genes: A.V., H.V., J.K. and N.T.; JH-pathway: R.M., R.M.R., N.T. and J.Z.; methylation prediction: K.M.G., B.G.H. and M.A.D.G.; neuropeptides: S.M.S.K. and R.M.R.; olfaction: H.M.R. and W.G.; vitellogenins: A.K.H., C.L. and N.T.; osiris genes: N.T. and A.K.H.; hexamerins, DNMT and yellow genes: N.T.; manuscript writing: N.T., L.C., C.S.B., E.B.B., J.K., G.Z. and J.L.

Additional information

Accession Codes: The Whole Genome Shotgun project for the dampwood termite *Z. nevadensis* has been deposited in the GenBank nucleotide core database under the accession code AUST00000000. RNA sequencing data have been deposited in the GenBank sequence read archive (SRA) under the accession code SRP022929.
Supplementary Information accompanies this paper at http://www.nature.com/naturecommunications

Competing financial interests: The authors declare no competing financial interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

How to cite this article: Terrapon, N. *et al.* Molecular traces of alternative social organization in a termite genome. *Nat. Commun.* 5:3636 doi: 10.1038/ncomms4636 (2014).

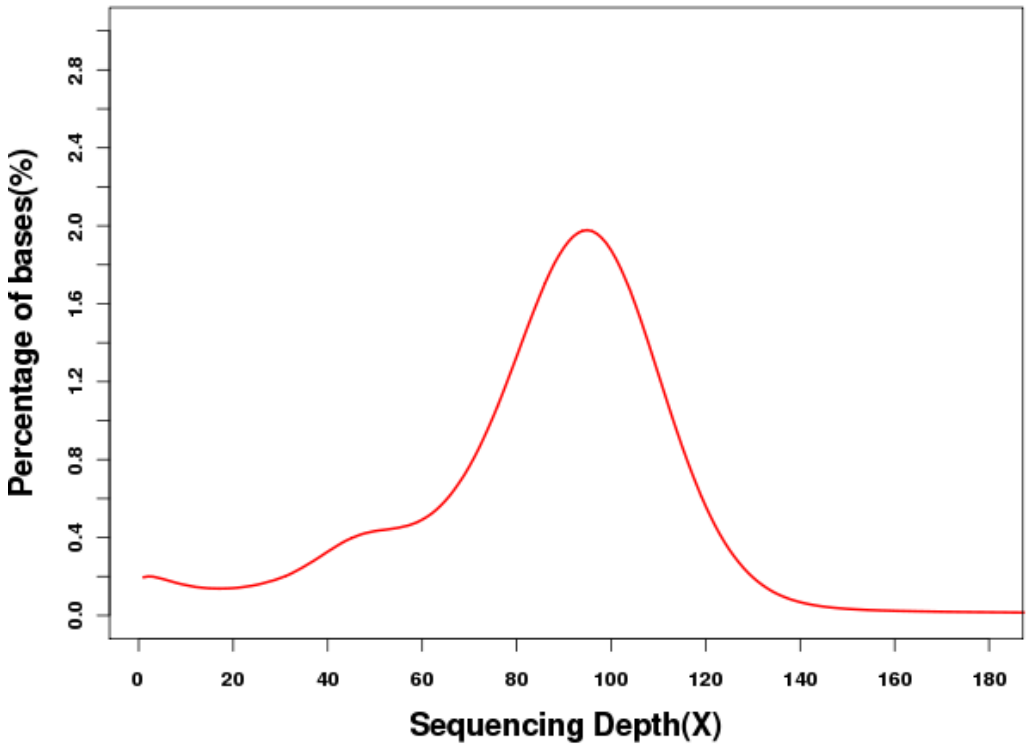
This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-sa/3.0/

© 2014 Macmillan Publishers Limited. All rights reserved.

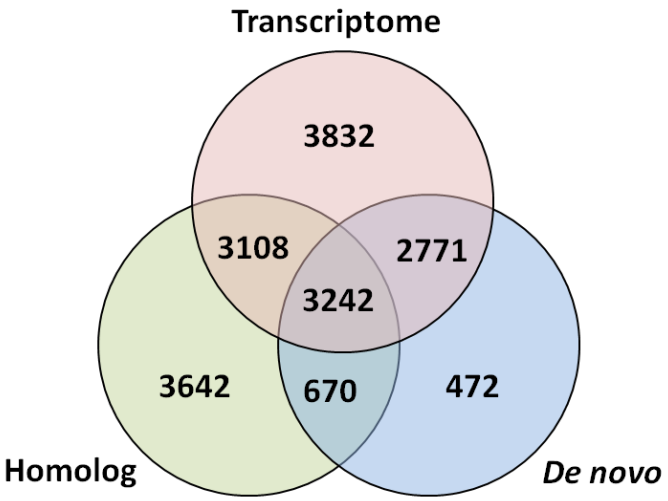
Supplementary Information

Molecular traces of alternative social organization in a termite genome

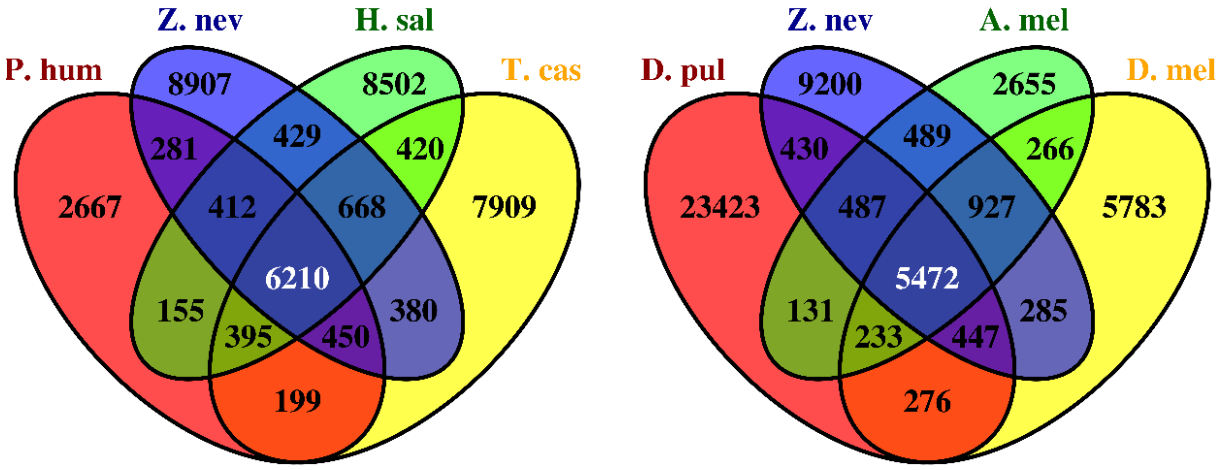
Supplementary Figures



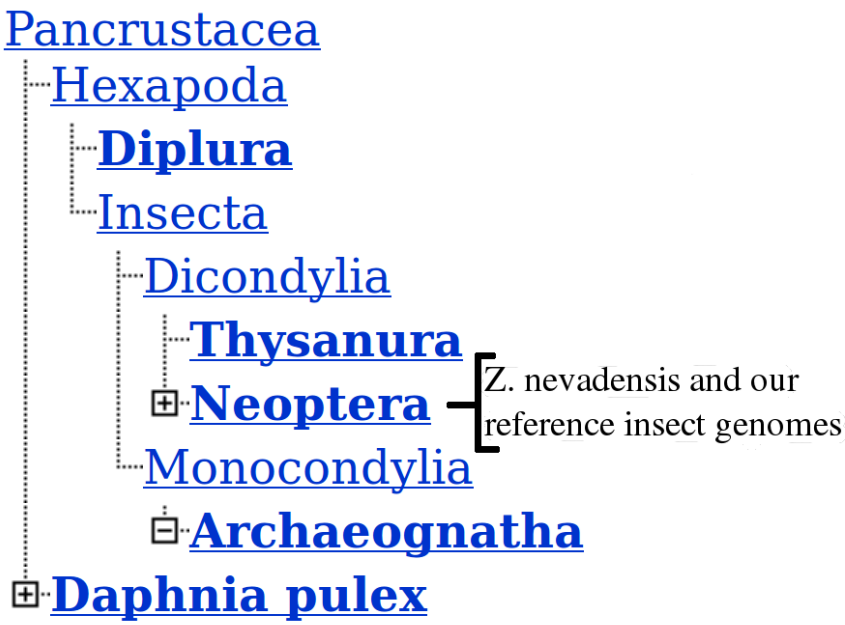
Supplementary Figure 1. Coverage depth of assembled genome



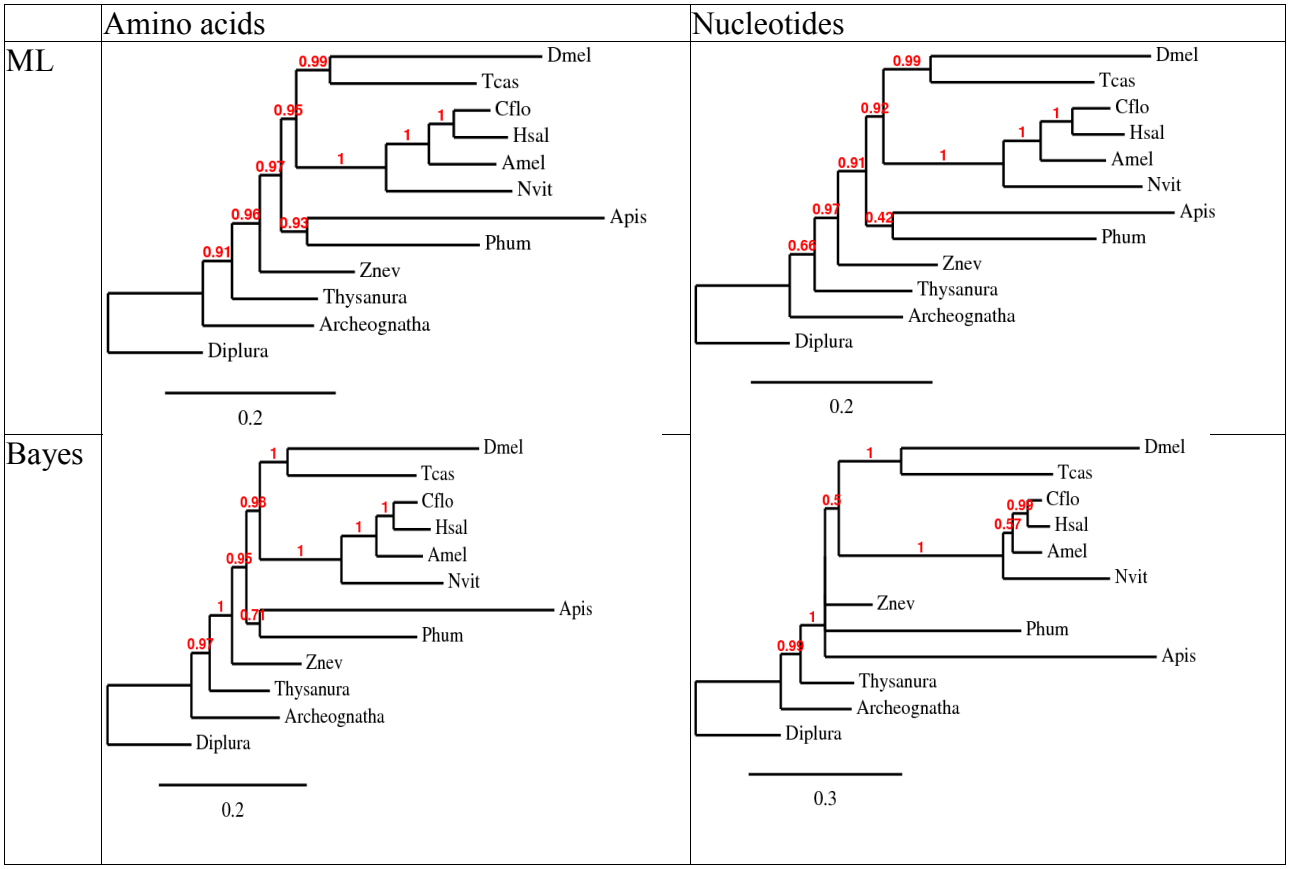
Supplementary Figure 2. Venn diagram of models predictions for *Z. nevadensis* protein coding genes.



Supplementary Figure 3. Venn diagrams of termite protein coding genes clustered with other arthropod proteins by orthoMCL procedure.

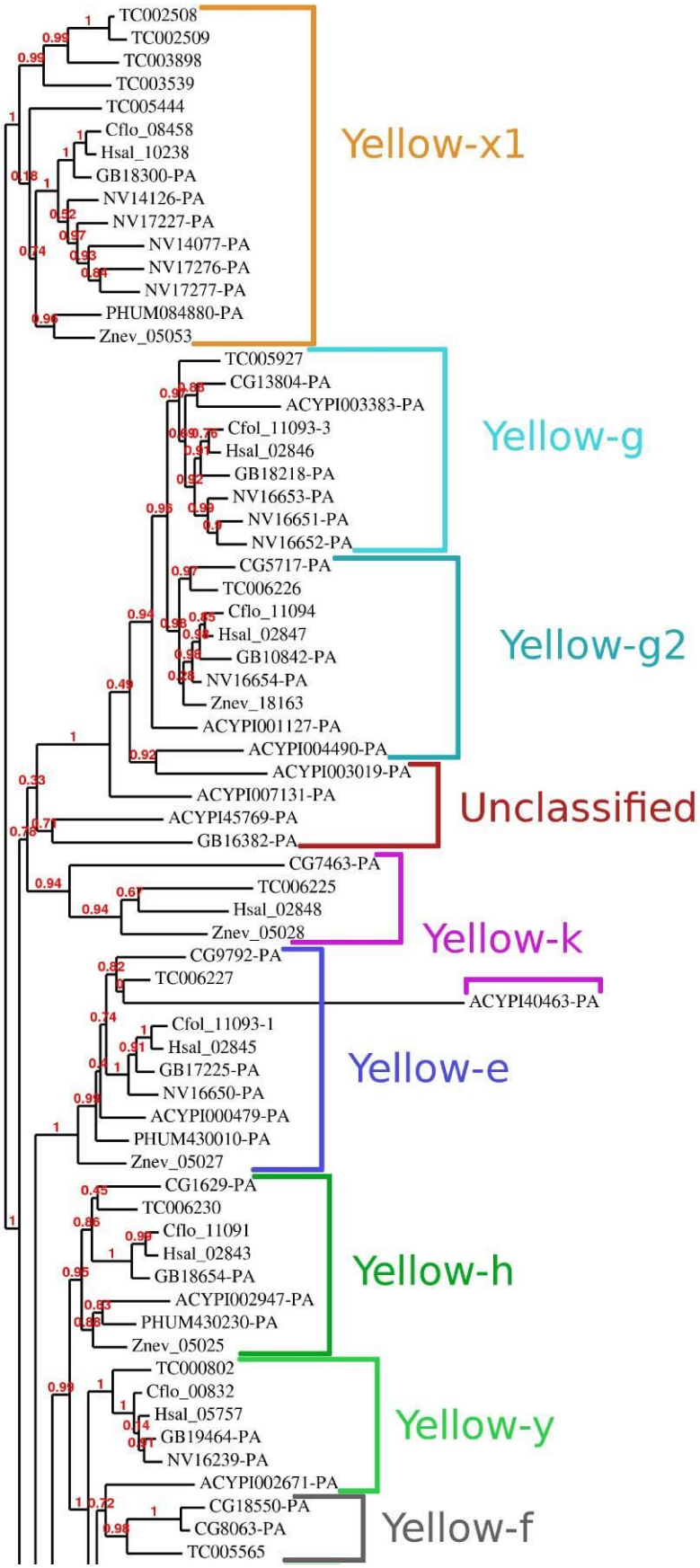


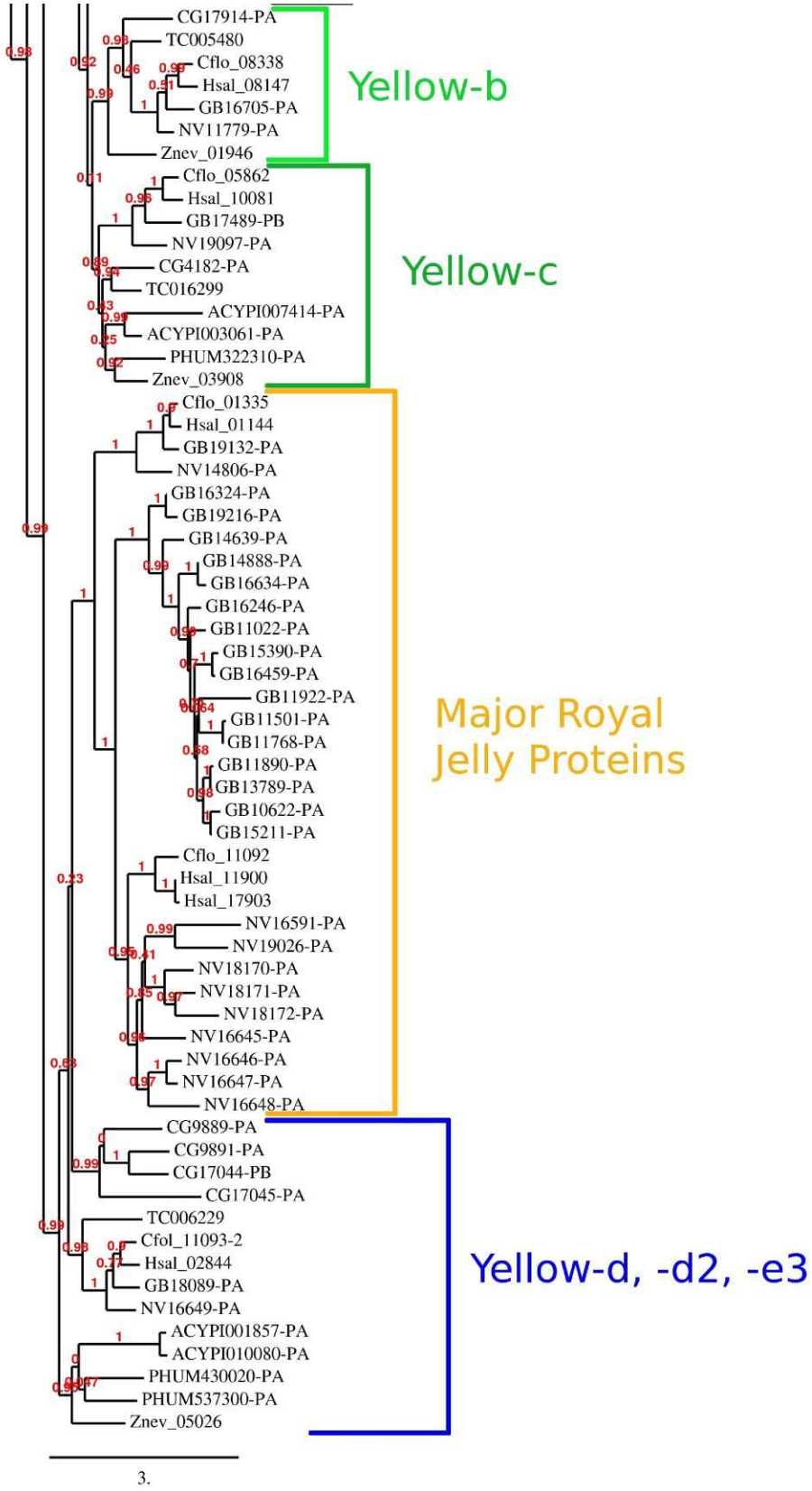
Supplementary Figure 4. NCBI taxonomy classification for candidate outgroups. This topology illustrates evolutionary relationships between Neoptera and the three lineages proposed as outgroups to replace the distant and fast evolving *D. pulex*.



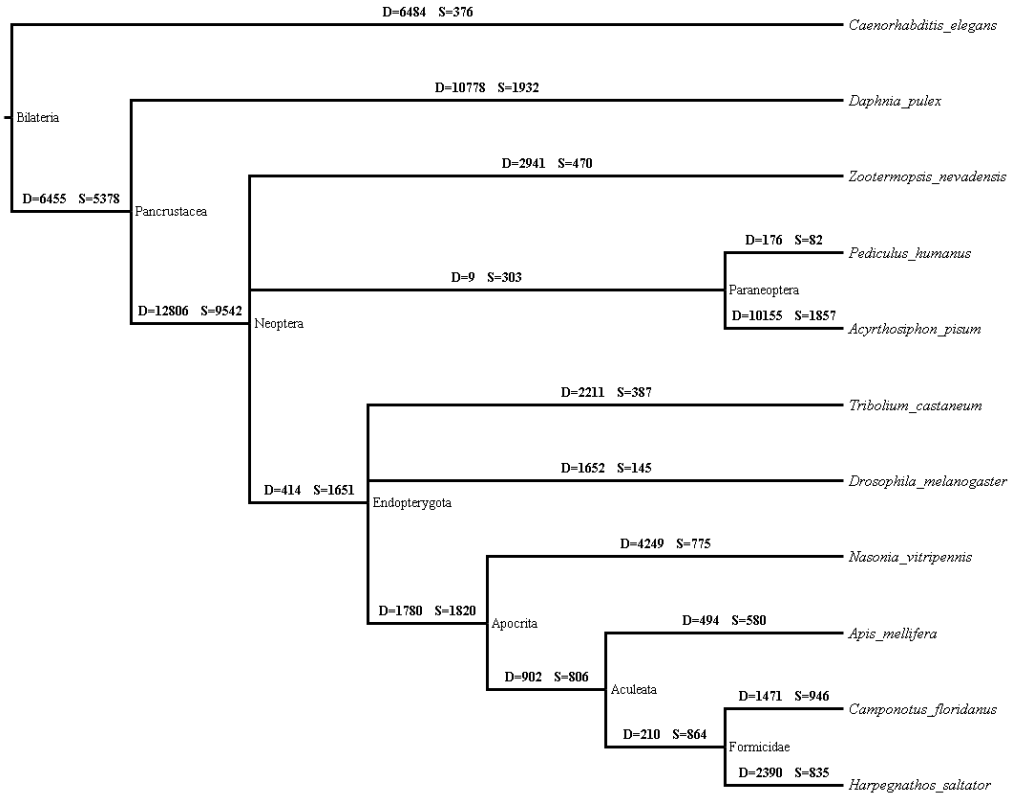
Supplementary Figure 5. Phylogenetic analyses showing agreement across the obtained topologies. Species names are abbreviated using the first letter of the genus in capital and the three first letters of the species (e.g. 'Apis' correspond to *A. pisum* and not to the bee *A. mellifera* which abbreviation is 'Amel').

Supplementary Figure 6. Osiris gene cluster in *Z. nevadensis* and five reference species. Genes are represented as squares with arrows indicating orientation. Gene names are indicated above while the subfamily classification and gene length are given below. A white cross in the square indicates non-detection of the corresponding domain with the Pfam recommended threshold. Intergenic space is also indicated, with a double large slash to notify genomic sequences larger than 1 Mb. Distinct scaffolds are separated by a vertical line.

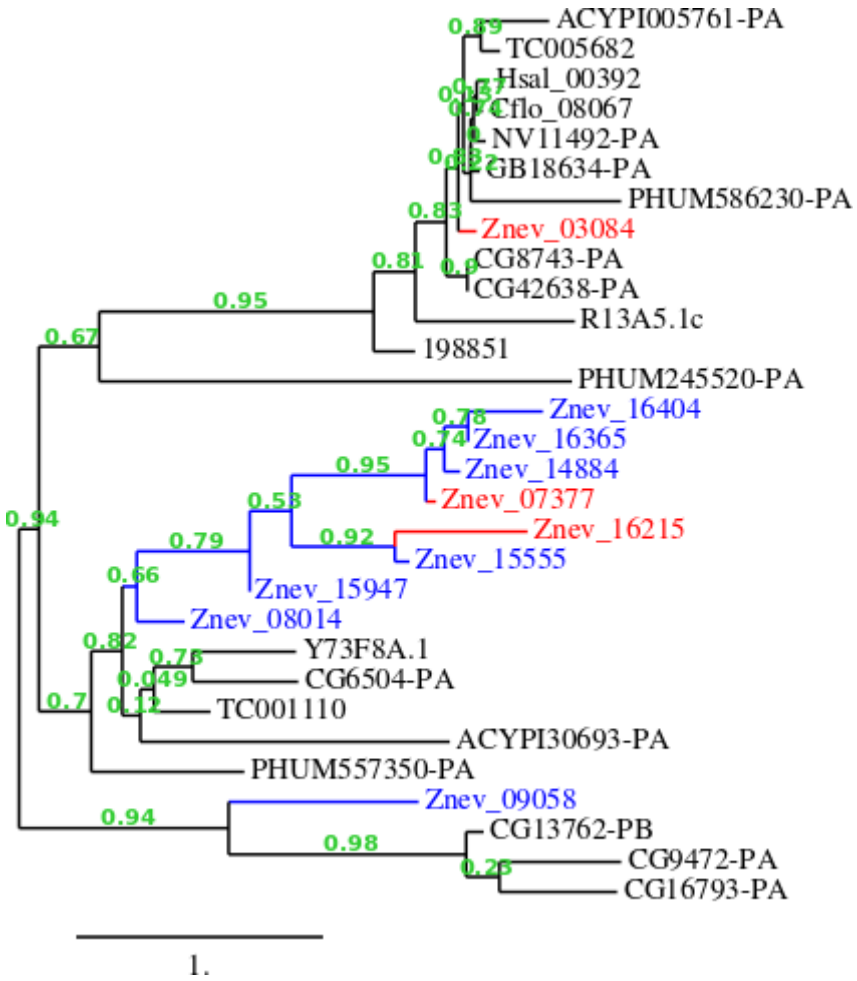

TC002508
TC002509
TC003898
TC003539
TC005444
Cflo_08458
Hsal_10238
GB18300-PA
NV14126-PA
NV17227-PA
NV14077-PA
NV17276-PA
NV17277-PA
PHUM084880-PA
Znev_05053
Yellow-x1
TC005927
CG13804-PA
ACYPI003383-PA
Cfol_11093-3
Hsal_02846
GB18218-PA
NV16653-PA
NV16651-PA
NV16652-PA
Yellow-g
CG5717-PA
TC006226
Cflo_11094
Hsal_02847
GB10842-PA
NV16654-PA
Znev_18163
ACYPI001127-PA
Yellow-g2
ACYPI004490-PA
ACYPI003019-PA
ACYPI007131-PA
ACYPI45769-PA
GB16382-PA
Unclassified
CG7463-PA
TC006225
Hsal_02848
Znev_05028
Yellow-k
CG9792-PA
TC006227
ACYPI40463-PA
Cfol_11093-1
Hsal_02845
GB17225-PA
NV16650-PA
ACYPI000479-PA
PHUM430010-PA
Znev_05027
Yellow-e
CG1629-PA
TC006230
Cflo_11091
Hsal_02843
GB18654-PA
ACYPI002947-PA
PHUM430230-PA
Znev_05025
Yellow-h
TC000802
Cflo_00832
Hsal_05757
GB19464-PA
NV16239-PA
ACYPI002671-PA
Yellow-y
CG18550-PA
CG8063-PA
TC005565
Yellow-f



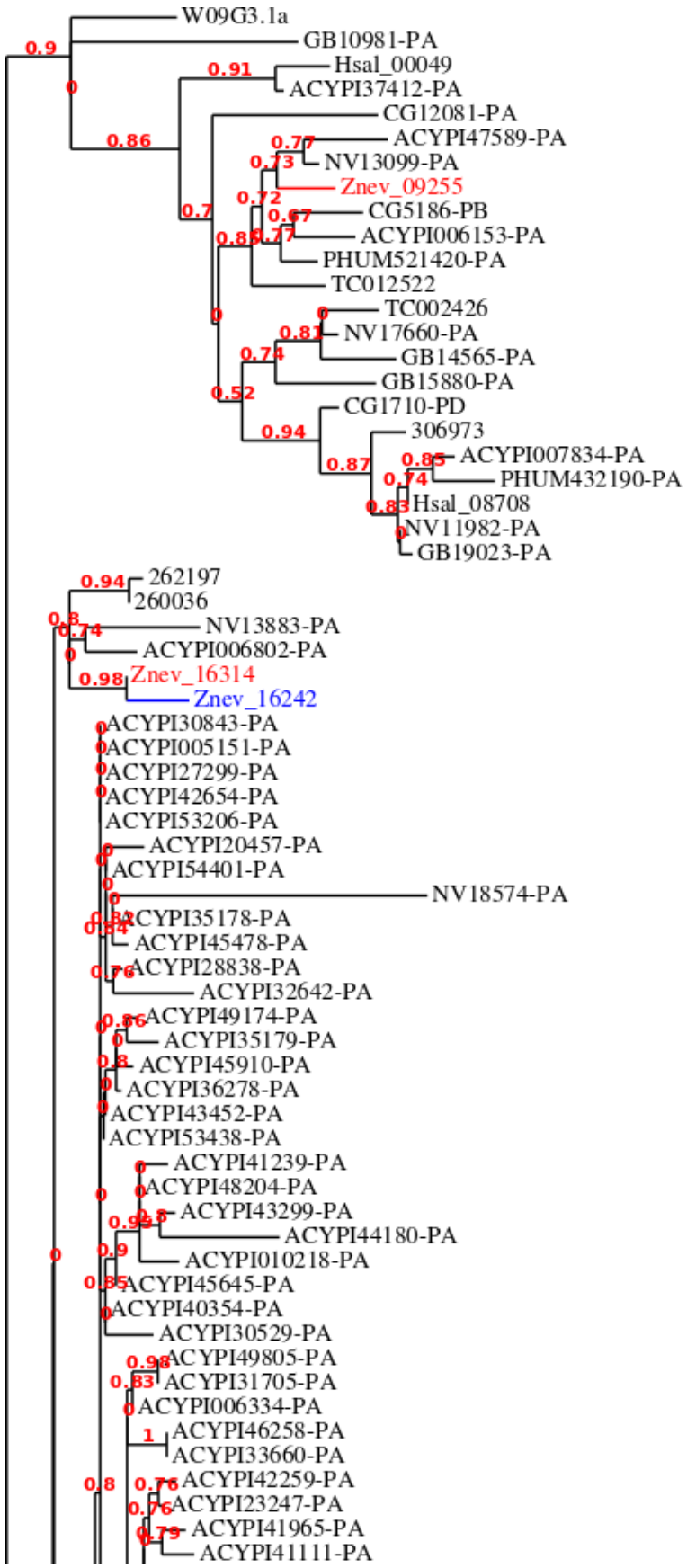
Supplementary Figure 7. Phylogeny of the Yellow gene family. For the protocol, refer to S5.3, for the mapping of protein IDs to species refer to Supplementary Table 13.

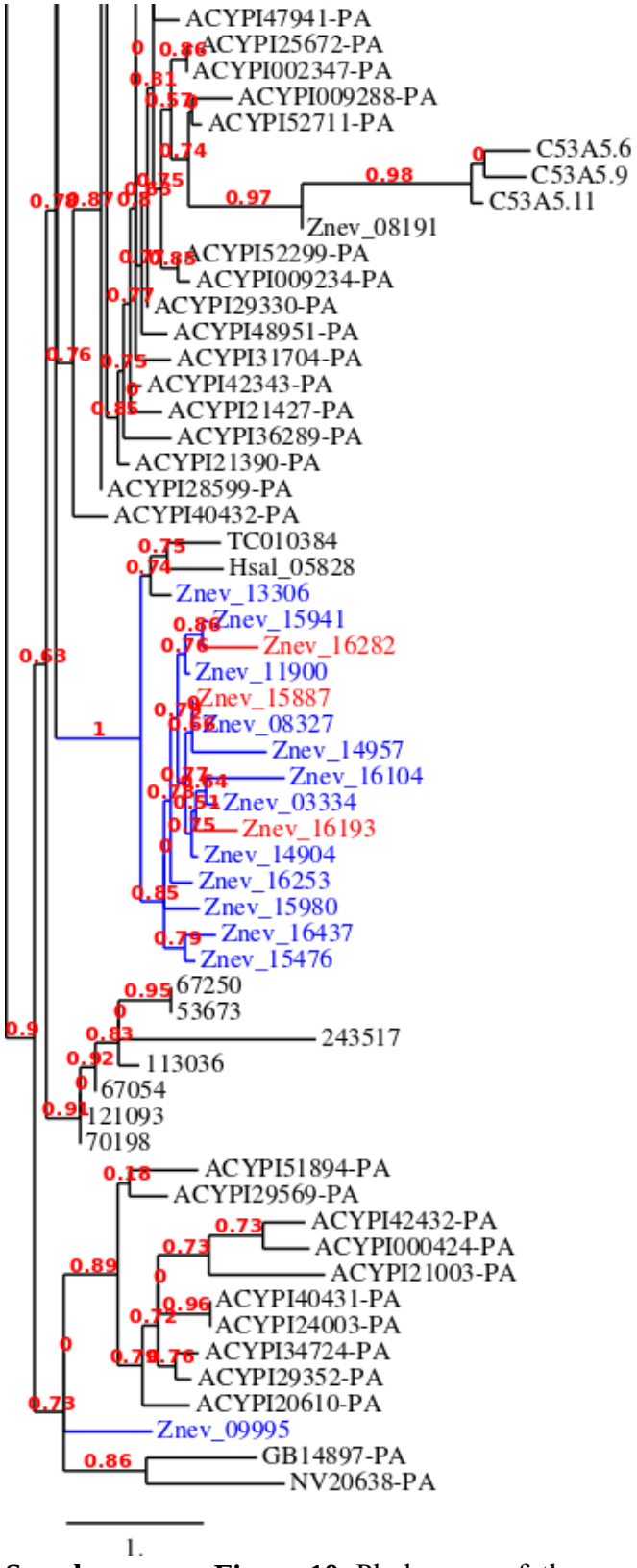


Supplementary Figure 8. Phylogenetic analysis of the *Zootermopsis nevadensis* species tree. Species tree for the eleven different taxa used in the phylogenetic analysis. The numbers above each branch are separated into duplications (D) and selective events (S).

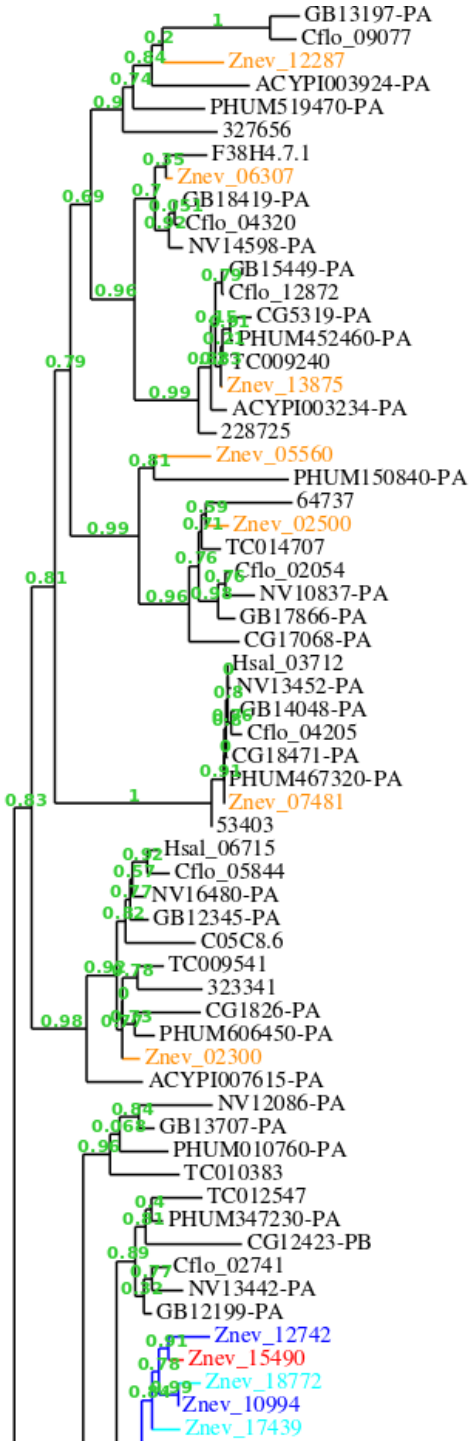


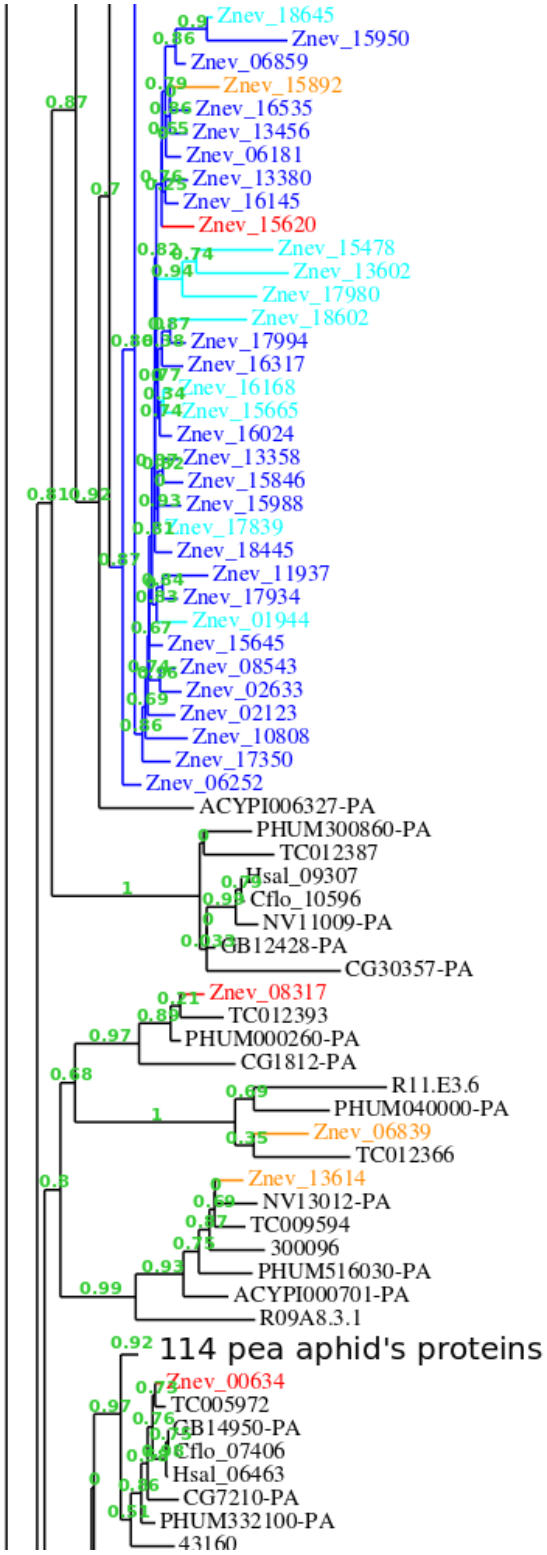
Supplementary Figure 9. Phylogeny of the PKD family. For the protocol, refer to Supplementary Notes 3.3, for the mapping of protein IDs to species refer to Supplementary Table 13. Termite proteins and lineage are coloured in blue if differentially expressed in males, in red if differentially expressed in other samples or if not differentially expressed.

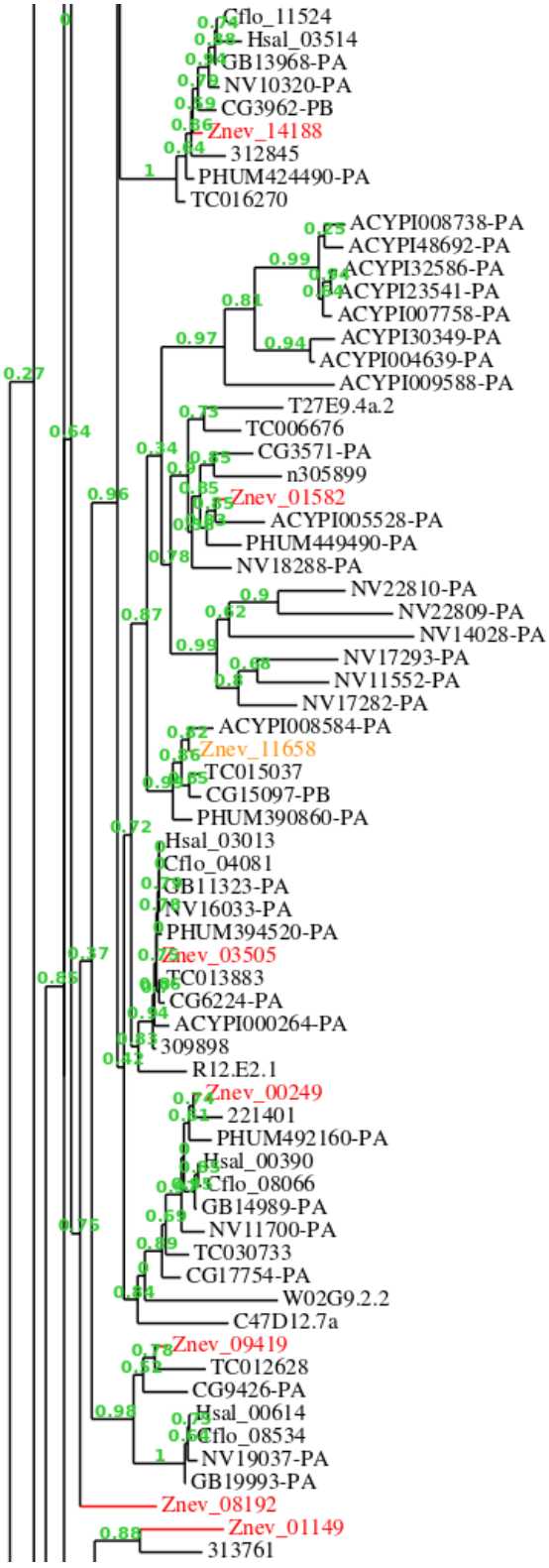

W09G3.1a
GB10981-PA
Hsal_00049
ACYPI37412-PA
CG12081-PA
ACYPI47589-PA
NV13099-PA
Znev_09255
CG5186-PB
ACYPI006153-PA
PHUM521420-PA
TC012522
TC002426
NV17660-PA
GB14565-PA
GB15880-PA
CG1710-PD
306973
ACYPI007834-PA
PHUM432190-PA
Hsal_08708
NV11982-PA
GB19023-PA
262197
260036
NV13883-PA
ACYPI006802-PA
Znev_16314
Znev_16242
ACYPI30843-PA
ACYPI005151-PA
ACYPI27299-PA
ACYPI42654-PA
ACYPI53206-PA
ACYPI20457-PA
ACYPI54401-PA
NV18574-PA
ACYPI35178-PA
ACYPI45478-PA
ACYPI28838-PA
ACYPI32642-PA
ACYPI49174-PA
ACYPI35179-PA
ACYPI45910-PA
ACYPI36278-PA
ACYPI43452-PA
ACYPI53438-PA
ACYPI41239-PA
ACYPI48204-PA
ACYPI43299-PA
ACYPI44180-PA
ACYPI010218-PA
ACYPI45645-PA
ACYPI40354-PA
ACYPI30529-PA
ACYPI49805-PA
ACYPI31705-PA
ACYPI006334-PA
ACYPI46258-PA
ACYPI33660-PA
ACYPI42259-PA
ACYPI23247-PA
ACYPI41965-PA
ACYPI41111-PA

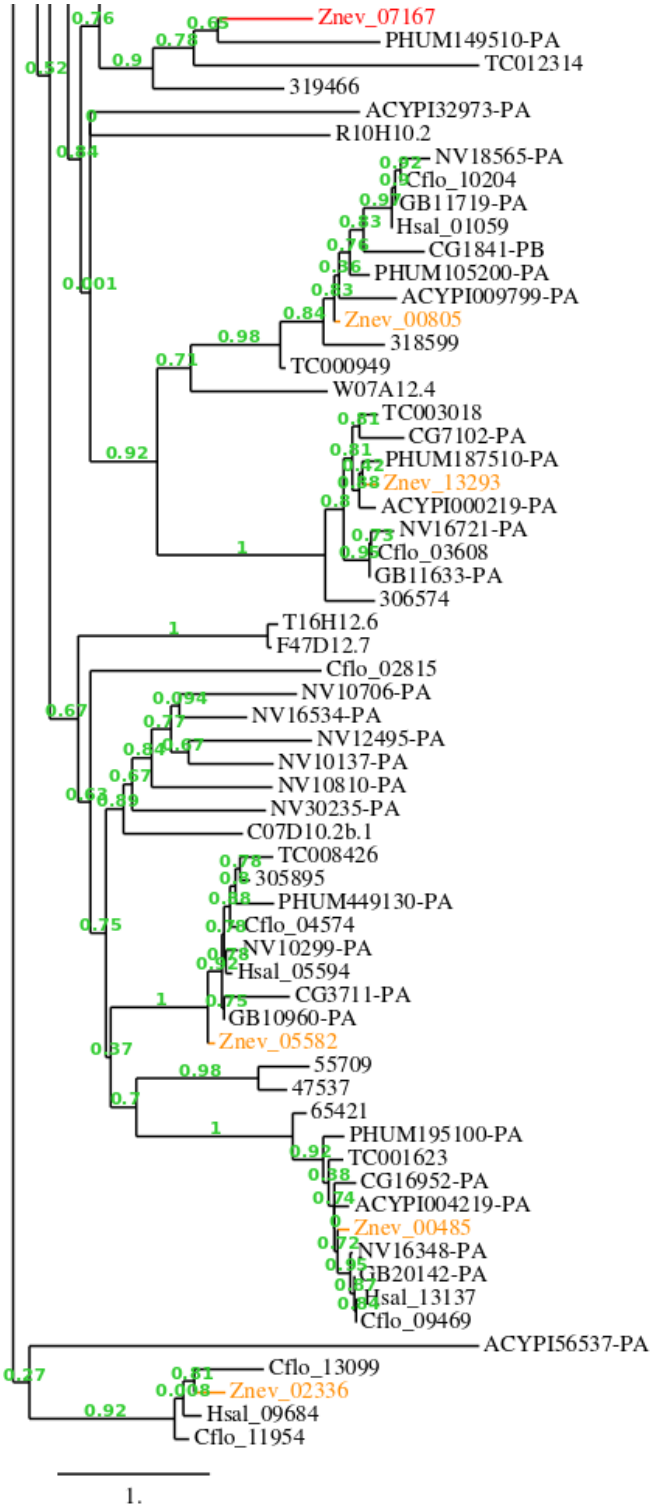


Supplementary Figure 10. Phylogeny of the monodomain Kelch family. For the protocol refer, to Supplementary Notes 3.3, for the mapping of protein IDs to species refer to Supplementary Table 13. Termite proteins and lineage are coloured in blue if differentially expressed in males, in red otherwise.

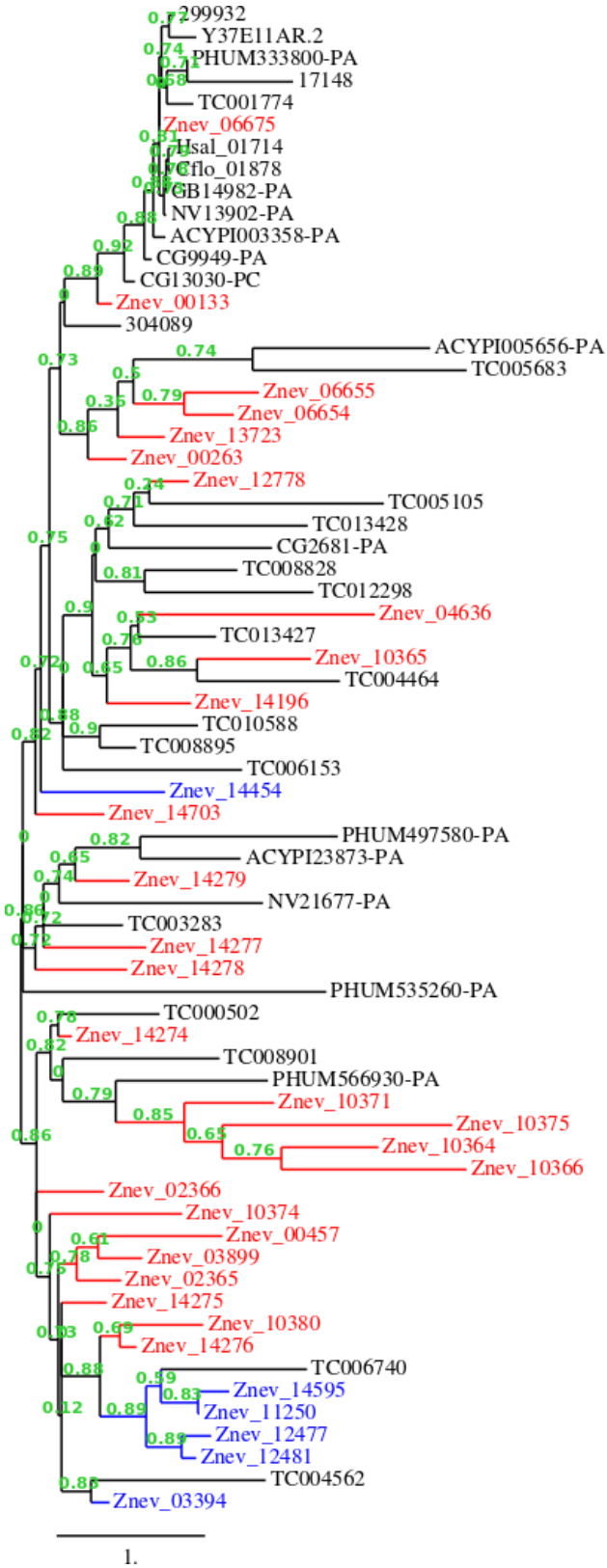

GB13197-PA
Cflo_09077
Znev_12287
ACYPI003924-PA
PHUM519470-PA
327656
F38H4.7.1
Znev_06307
GB18419-PA
Cflo_04320
NV14598-PA
GB15449-PA
Cflo_12872
CG5319-PA
PHUM452460-PA
TC009240
Znev_13875
ACYPI003234-PA
228725
Znev_05560
PHUM150840-PA
64737
Znev_02500
TC014707
Cflo_02054
NV10837-PA
GB17866-PA
CG17068-PA
Hsal_03712
NV13452-PA
GB14048-PA
Cflo_04205
CG18471-PA
PHUM467320-PA
Znev_07481
53403
Hsal_06715
Cflo_05844
NV16480-PA
GB12345-PA
C05C8.6
TC009541
323341
CG1826-PA
PHUM606450-PA
Znev_02300
ACYPI007615-PA
NV12086-PA
GB13707-PA
PHUM010760-PA
TC010383
TC012547
PHUM347230-PA
CG12423-PB
Cflo_02741
NV13442-PA
GB12199-PA
Znev_12742
Znev_15490
Znev_18772
Znev_10994
Znev_17439


Znev_18645
Znev_15950
Znev_06859
Znev_15892
Znev_16535
Znev_13456
Znev_06181
Znev_13380
Znev_16145
Znev_15620
Znev_15478
Znev_13602
Znev_17980
Znev_18602
Znev_17994
Znev_16317
Znev_16168
Znev_15665
Znev_16024
Znev_13358
Znev_15846
Znev_15988
Znev_17839
Znev_18445
Znev_11937
Znev_17934
Znev_01944
Znev_15645
Znev_08543
Znev_02633
Znev_02123
Znev_10808
Znev_17350
Znev_06252
ACYPI006327-PA
PHUM300860-PA
TC012387
Hsal_09307
Cflo_10596
NV11009-PA
GB12428-PA
CG30357-PA
Znev_08317
TC012393
PHUM000260-PA
CG1812-PA
R11.E3.6
PHUM040000-PA
Znev_06839
TC012366
Znev_13614
NV13012-PA
TC009594
300096
PHUM516030-PA
ACYPI000701-PA
R09A8.3.1
114 pea aphid's proteins
Znev_00634
TC005972
GB14950-PA
Cflo_07406
Hsal_06463
CG7210-PA
PHUM332100-PA
43160

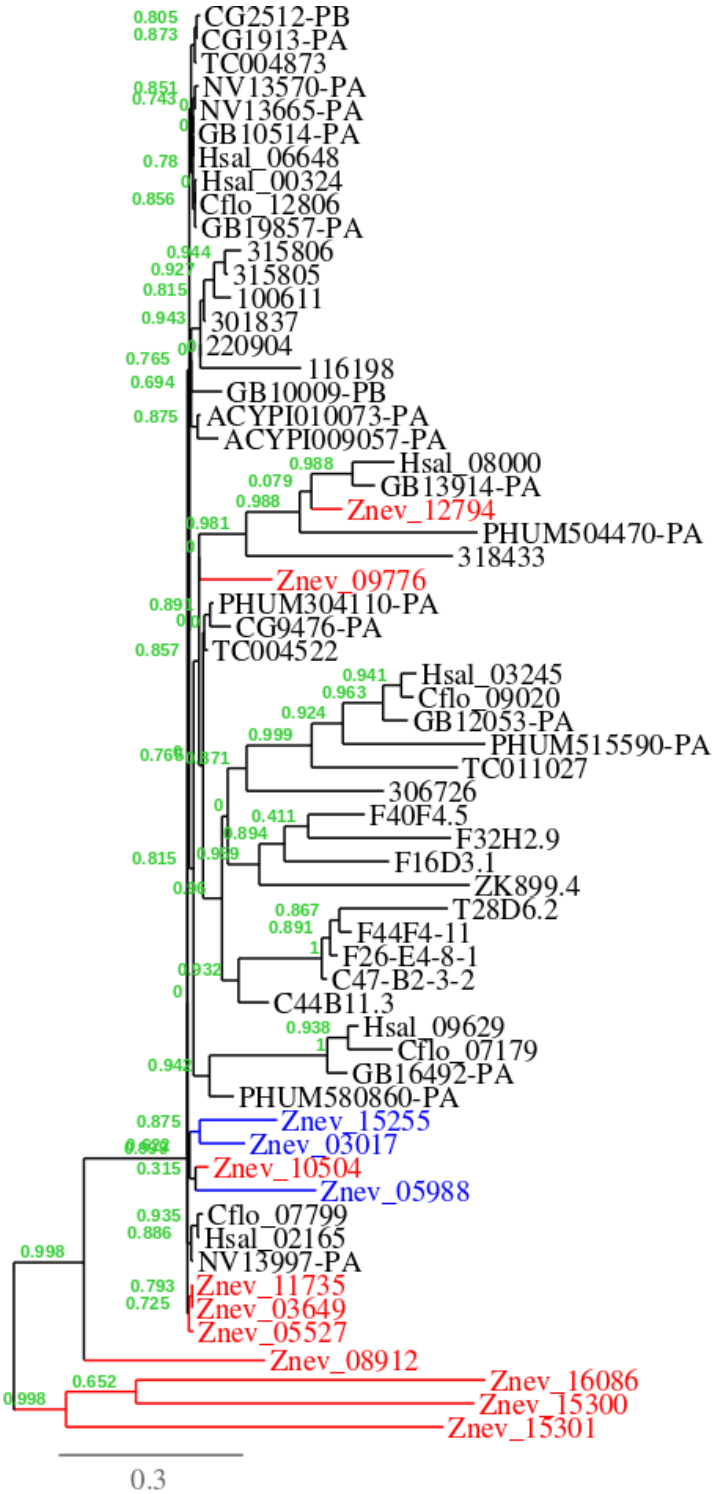

Cflo_11524
Hsal_03514
GB13968-PA
NV10320-PA
CG3962-PB
Znev_14188
312845
PHUM424490-PA
TC016270
ACYPI008738-PA
ACYPI48692-PA
ACYPI32586-PA
ACYPI23541-PA
ACYPI007758-PA
ACYPI30349-PA
ACYPI004639-PA
ACYPI009588-PA
T27E9.4a.2
TC006676
CG3571-PA
n305899
Znev_01582
ACYPI005528-PA
PHUM449490-PA
NV18288-PA
NV22810-PA
NV22809-PA
NV14028-PA
NV17293-PA
NV11552-PA
NV17282-PA
ACYPI008584-PA
Znev_11658
TC015037
CG15097-PB
PHUM390860-PA
Hsal_03013
Cflo_04081
GB11323-PA
NV16033-PA
PHUM394520-PA
Znev_03505
TC013883
CG6224-PA
ACYPI000264-PA
309898
R12.E2.1
Znev_00249
221401
PHUM492160-PA
Hsal_00390
Cflo_08066
GB14989-PA
NV11700-PA
TC030733
CG17754-PA
W02G9.2.2
C47D12.7a
Znev_09419
TC012628
CG9426-PA
Hsal_00614
Cflo_08534
NV19037-PA
GB19993-PA
Znev_08192
Znev_01149
313761



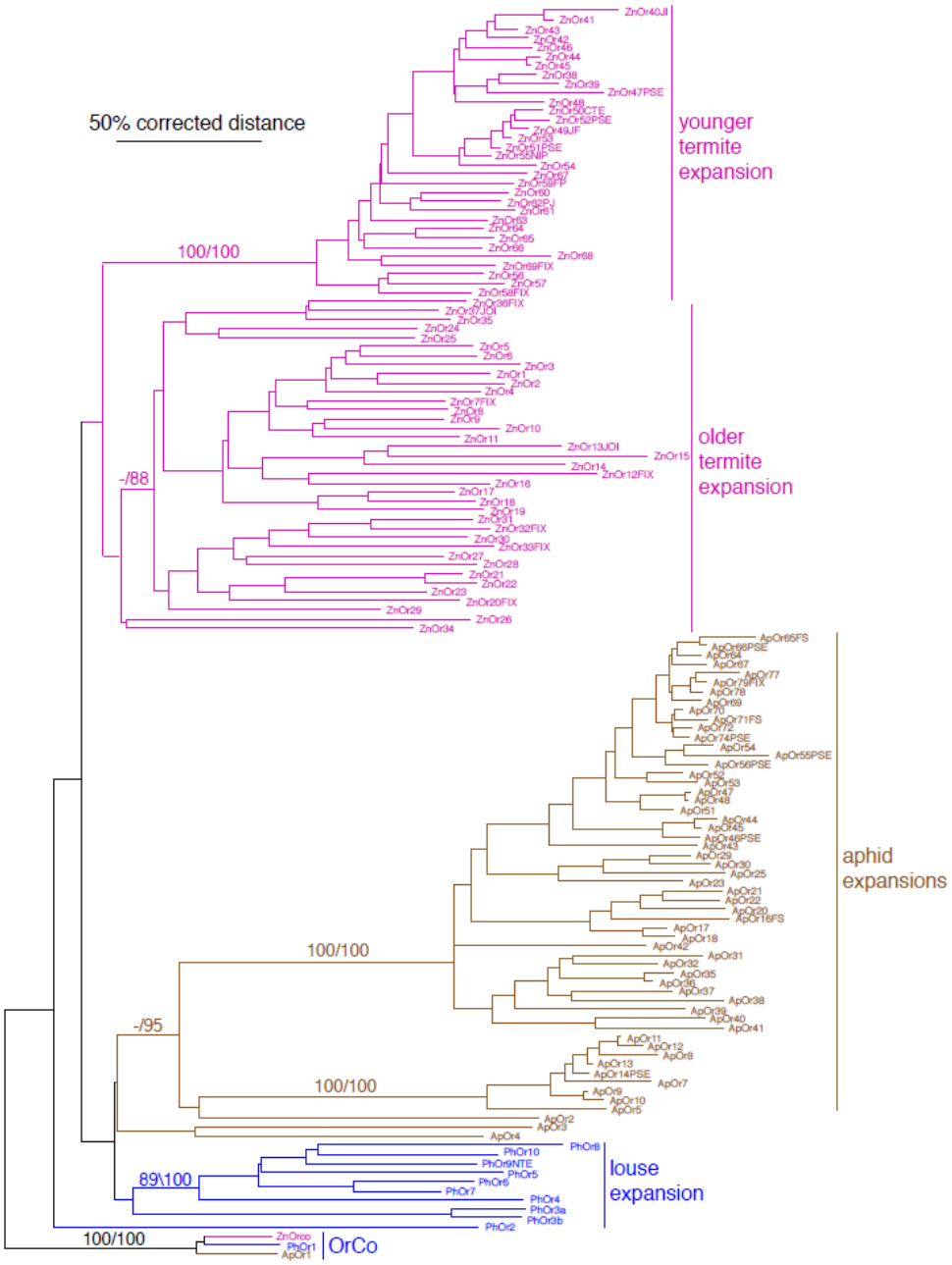
Supplementary Figure 11. Phylogeny of the multidomain BTB-BACK-Kelch family. For the protocol, refer to Supplementary Notes 3.3, for the mapping of protein IDs to species refer to Supplementary Table 13. Termite proteins and lineage are coloured in blue if differentially expressed in males, in red otherwise when these proteins exhibit the exact tridomain architecture BTB-BACK-Kelch. Light blue and orange are used for proteins having only two of the three required domains and/or additional domains, light blue for proteins differentially expressed in males, orange otherwise.



Supplementary Figure 12. Phylogeny of the SINA family. For the protocol, refer to Supplementary Notes 3.3, for the mapping of protein IDs to species refer to Supplementary Table 13. Termite proteins and lineage are coloured in blue if differentially expressed in males, in red otherwise.

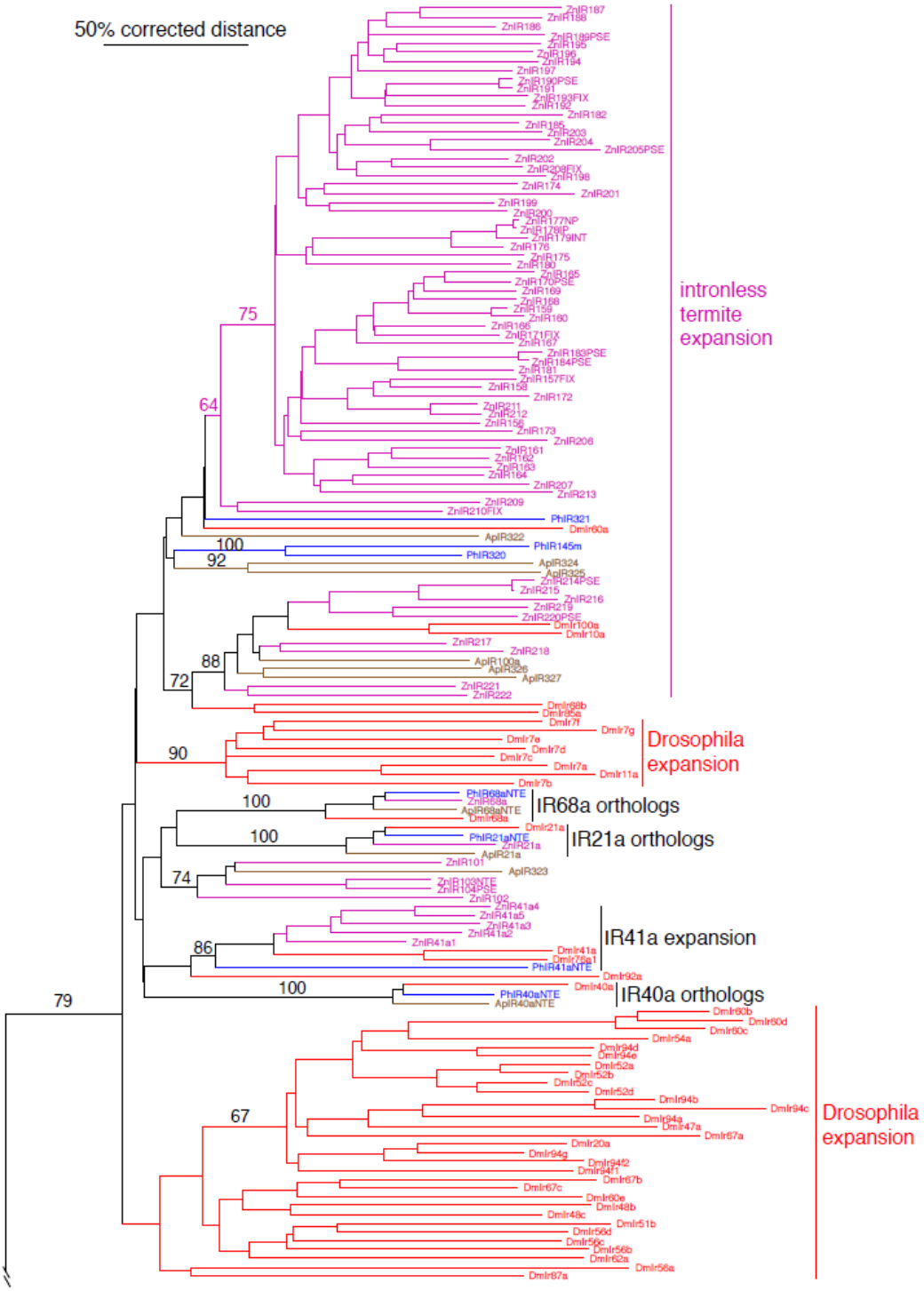


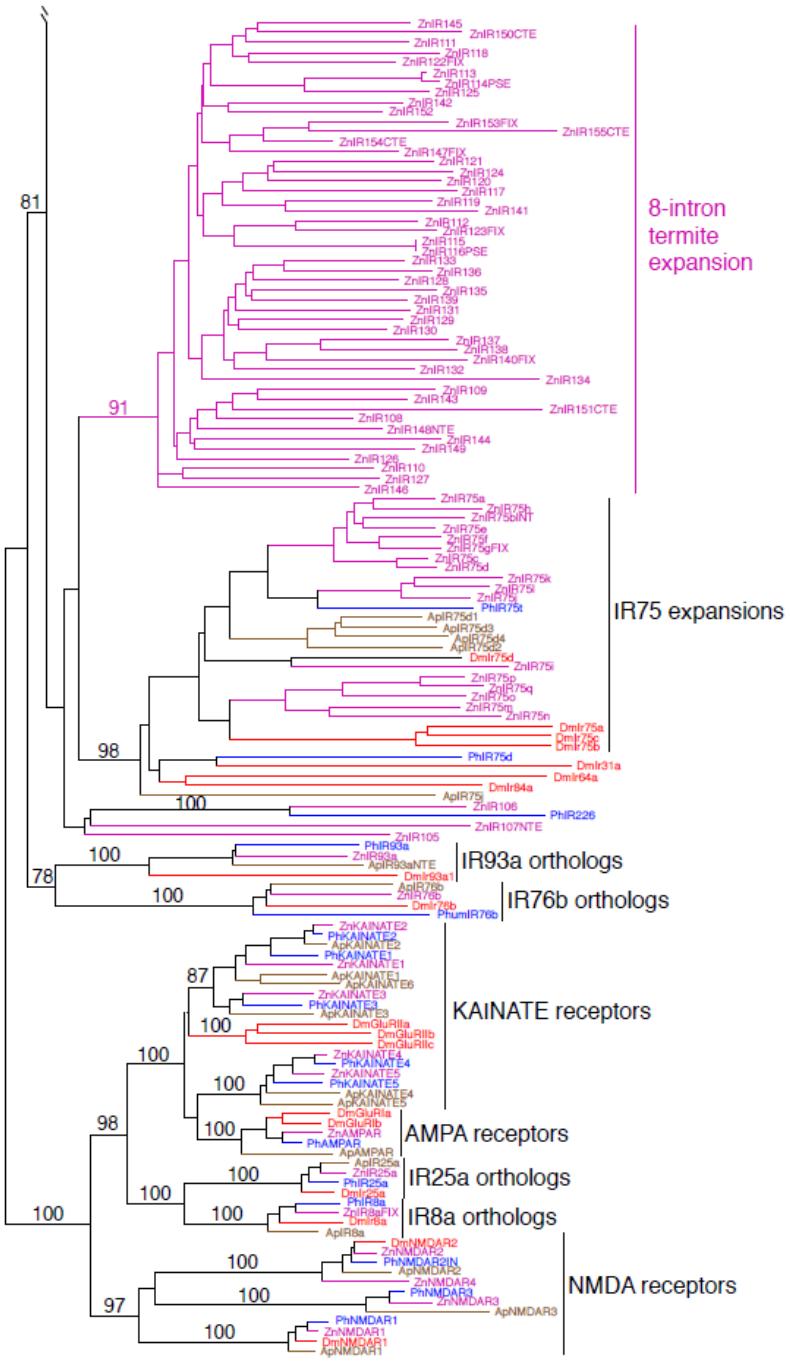
Supplementary Figure 13. Phylogeny of the alpha tubulin family. For the protocol, refer to Supplementary Notes 3.3, for the mapping of protein IDs to species refer to Supplementary Table 13. Termite proteins and lineage are coloured in blue if differentially expressed in males, in red otherwise.



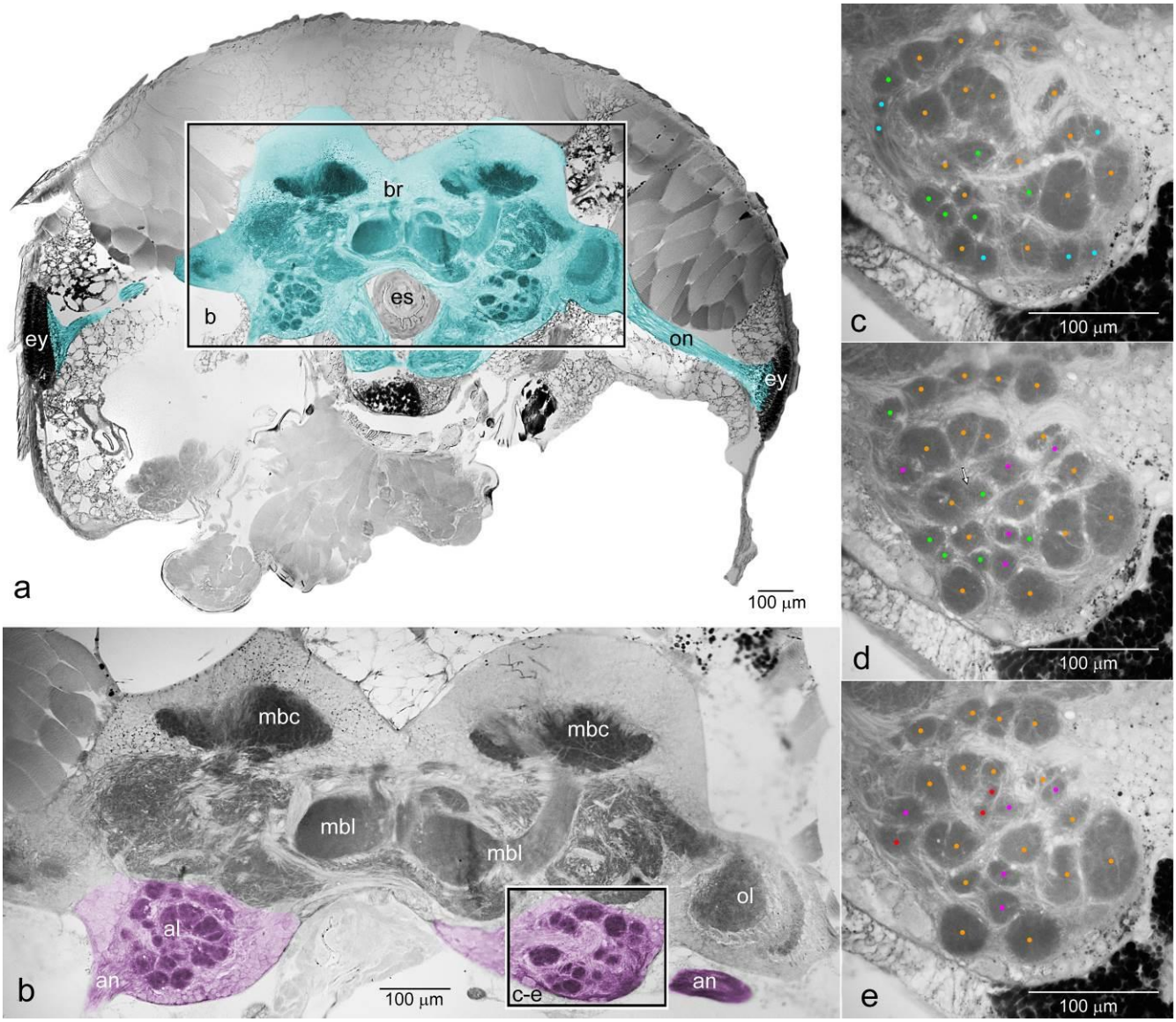
Supplementary Figure 14. Phylogenetic tree of the termite, aphid, and louse ORs. This is a corrected distance tree. The DmOr83b orthologs ZnOrCo, ApOr1, and PhOr1 were declared as the outgroup to root the tree, based on the basal position of this gene in the Or family in analysis of the entire chemoreceptor superfamily in *D. melanogaster*. The termite, aphid, and louse protein names are highlighted in purple, brown, and blue, respectively, as are the branches leading to them to emphasize gene lineages. Support for major branches is shown above each branch as the percentage of 10,000 bootstrap replications of uncorrected distances followed by Bayesian posterior probabilities. Comments on each gene lineage are on the right. Suffixes after the gene/protein names are: PSE – pseudogene; NTE – N-terminus missing; CTE – C-terminus missing; INT – internal sequence missing; FIX – sequence fixed with raw reads; JOI – gene model joined across scaffolds; multiple suffixes are abbreviated to single letters.

Supplementary Figure 15. Phylogenetic tree of the termite, aphid, louse, and select other insect GRs. This is a corrected distance tree and was rooted at the midpoint in the absence of a clear outgroup, an approach that clearly indicates the distinctiveness of the sugar and carbon dioxide receptors, as well as their putative distant relatedness. The termite, aphid, louse, and other insect gene/protein names are highlighted in purple, brown, blue, and red, respectively, as are the branches leading to them to emphasize gene lineages. Bootstrap support levels in percentage of 10,000 replications of uncorrected distance, followed by Bayesian posterior probabilities, are shown above major branches. Comments on major gene lineages are on the right. Suffixes after the gene/protein names are: PSE – pseudogene; NTE – N-terminus missing; CTE – C-terminus missing; INT – internal sequence missing; FIX – sequence fixed with raw reads; JOI – gene model joined across scaffolds; multiple suffixes are abbreviated to single letters. Gene models have been updated in OGSv2.2.

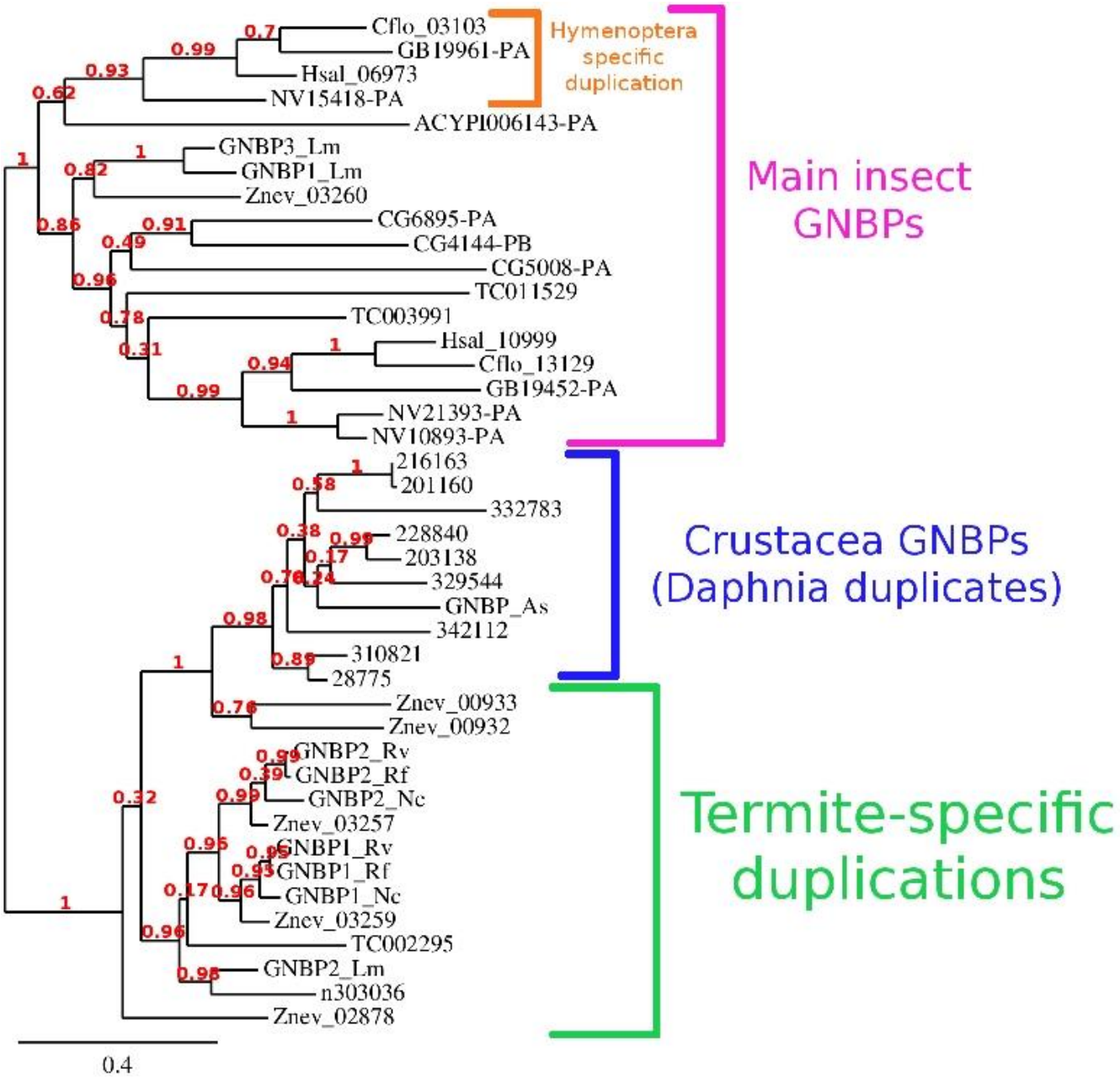

50% corrected distance
intronless termite expansion
Drosophila expansion
IR68a orthologs
IR21a orthologs
IR41a expansion
IR40a orthologs
Drosophila expansion
75
64
100
92
88
72
90
100
100
74
86
100
79
67



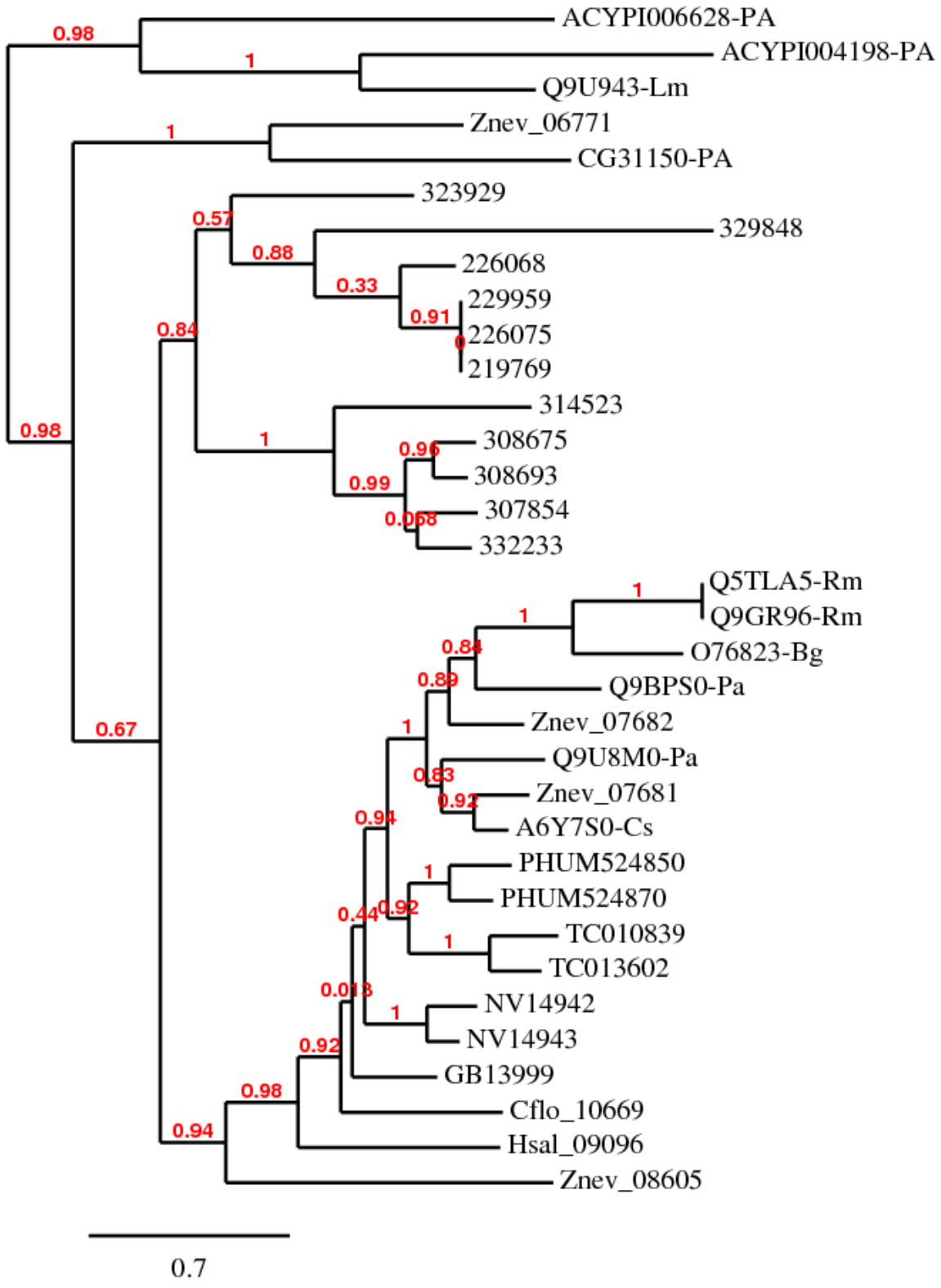
Supplementary Figure 16. Phylogenetic tree of the termite, aphid, louse, and *D. melanogaster* IRs**.** This is a corrected distance tree and was rooted with the iGluRs and IR25a/8a as the outgroup, based on their highly conserved sequences and ancestral position in the family. The termite, aphid, louse, and Drosophila gene/protein names are highlighted in purple, brown, blue, and red, respectively, as are the branches leading to them to emphasize gene lineages. Bootstrap support levels, expressed as the percentage of 10,000 replications of uncorrected distance analysis, are shown above major branches. Comments on gene lineages are on the right. Suffixes after the gene/protein names are: PSE – pseudogene; NTE – N-terminus missing; CTE – C-terminus missing; INT – internal sequence missing; FIX – sequence fixed with raw reads; multiple suffixes are abbreviated to single letters.



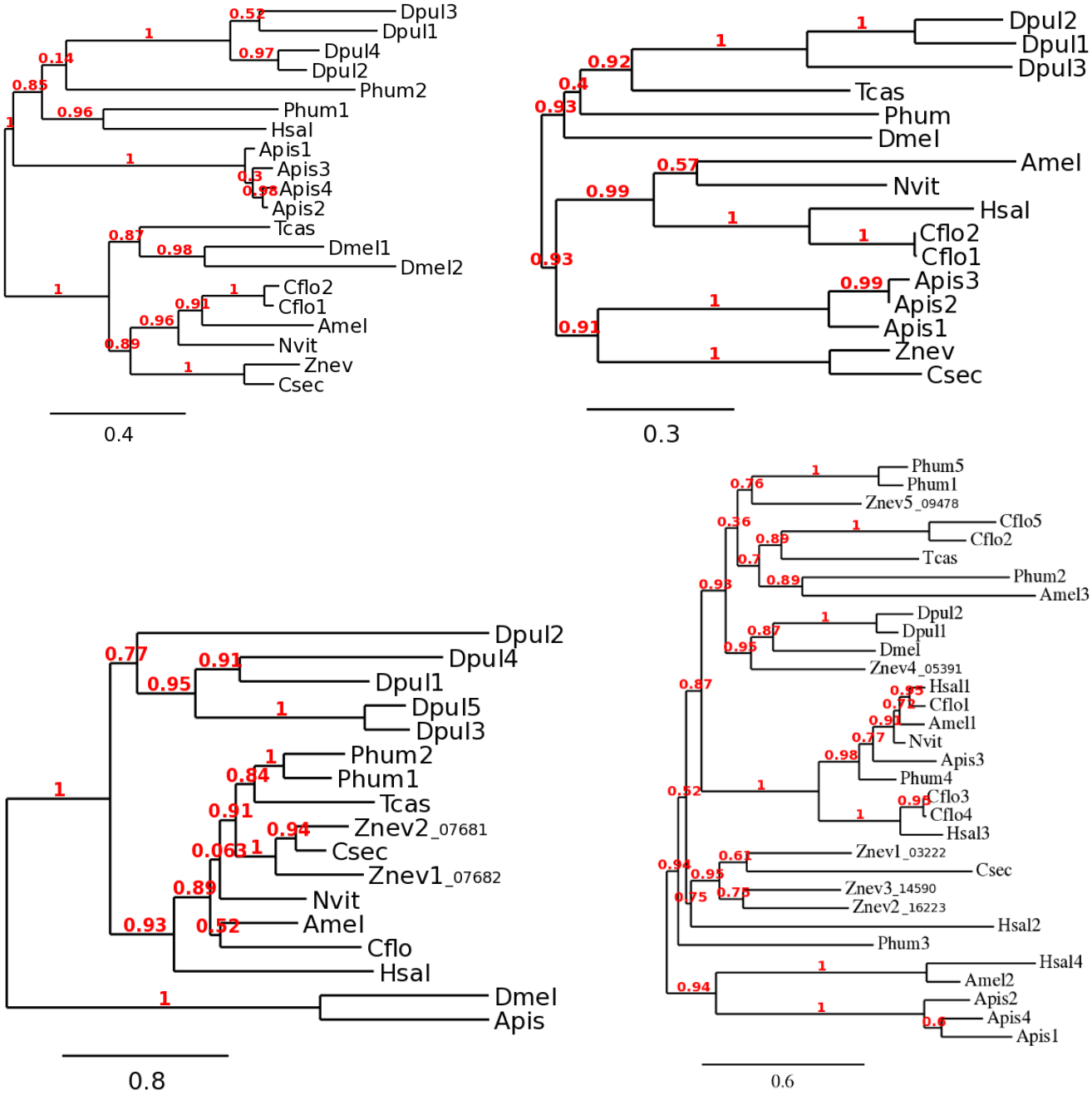
Supplementary Figure 17. Sections of olfactory glomeruli. Microphotographs of the dorsal part of the head (**a**; montage of three sections), the brain (**b**) and the right antennal lobe (**c-e**) of *Zootermopsis sp*. (frontal sections). Brain in (**a**) highlighted in cyan, antennal lobes in (b) highlighted in magenta; olfactory glomeruli in (**c-e**) colour coded as follows: cyan - glomeruli only present in upper section (**c**); green - glomeruli present in upper and middle section; magenta - glomeruli present in middle and lower section; red - glomeruli present only in lower section; orange - glomeruli present in all three sections. Arrow in (**d**) points at area where two adjacent glomeruli are difficult to discriminate; al - antennal lobe; an - antennal nerve; br - brain; es - esophagus; ey - eye; mbc - mushroom body calyx; mbl - mushroom body lobes; on - optic nerve; ol - optic lobe.



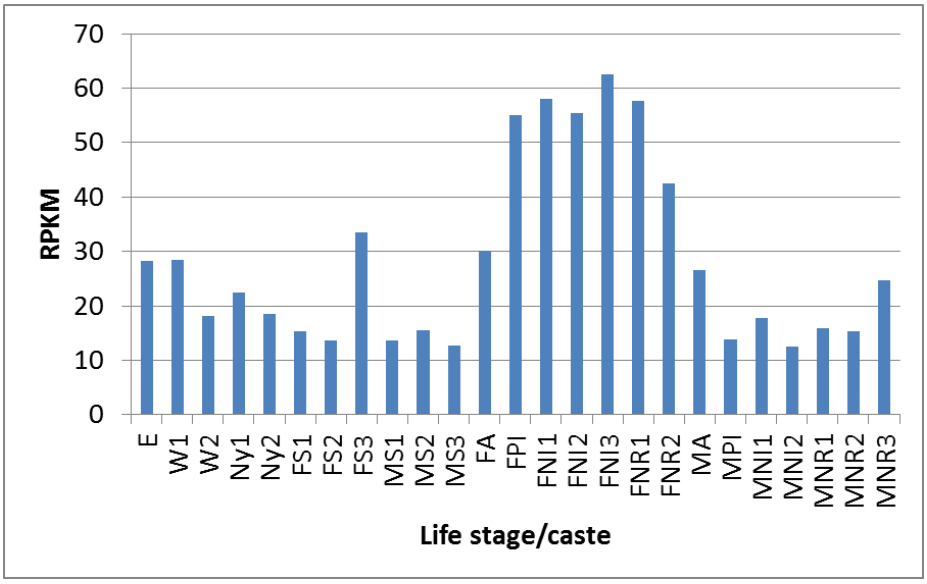
Supplementary Figure 18. Phylogeny of the GNBP family. For the protocol, refer to Supplementary Notes 3.3, for the mapping of protein IDs to species refer to Supplementary Table 13. Additional sequences of Isoptera species are labelled "*gene name_species initials*".



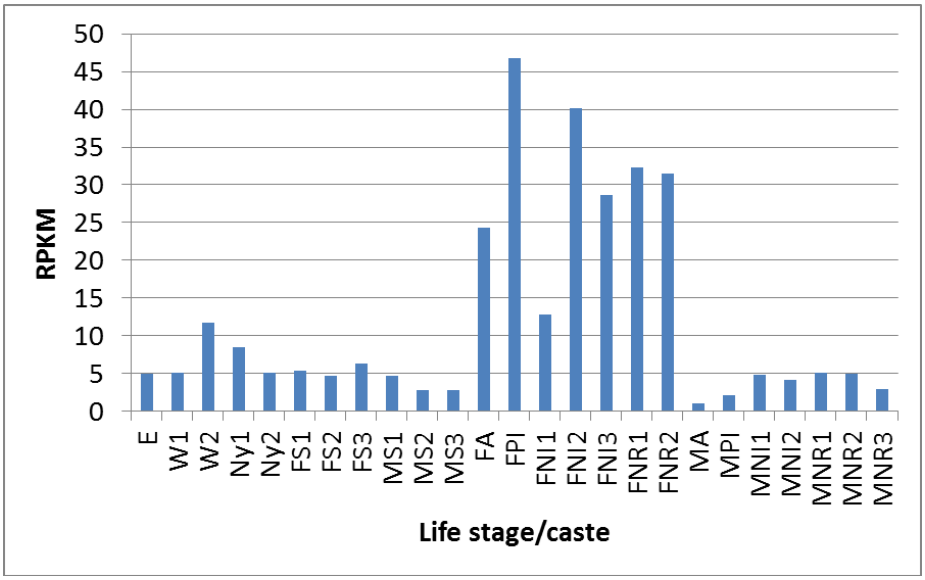
Supplementary Figure 19. Phylogeny of the Vitellogenin family in the reference species and *Z. nevadensis*. For the protocol, refer to Supplementary Notes 3.3, for the mapping of protein IDs to species refer to Supplementary Table 13. Additional sequences of Isoptera species are labelled "uniprot id-*species initials*".



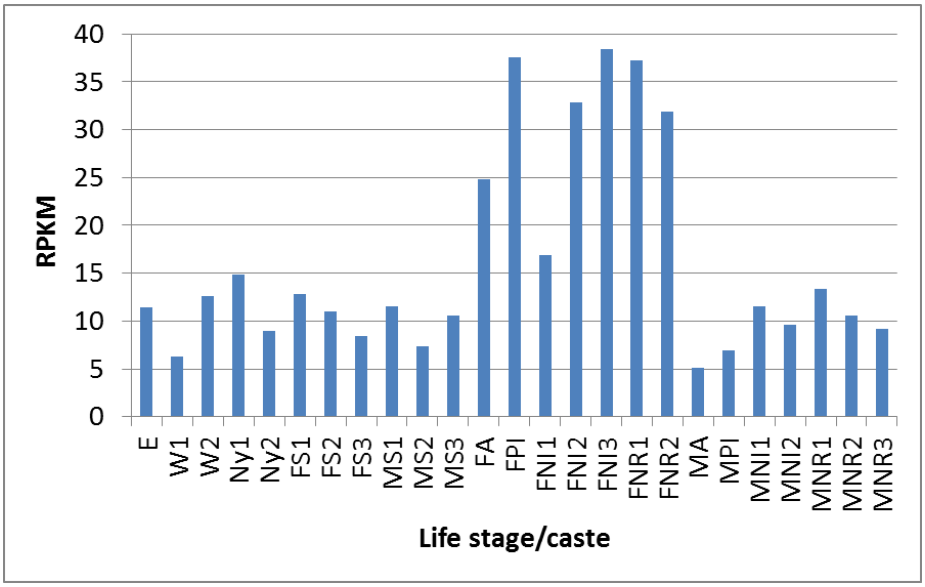
Supplementary Figure 20. Phylogenetic trees for four of the Neofem genes. Neofem1 (top-left), Neofem2 (top-right), Neofem3 (bottom-left) and Neofem4 (bottom-right). Genes for the different species are labelled with the initial of the genus, the three first letters of the species and a number.



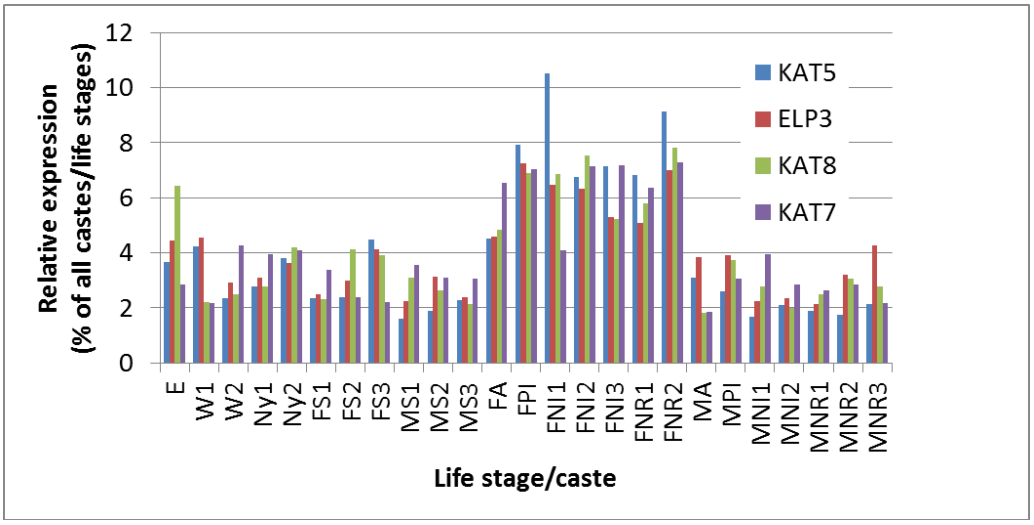
Supplementary Figure 21. Expression levels of sirtuin 6 across castes and life stages. Female reproductive castes show the highest gene expression. Normalized RPKM values have been standardized to 100% across the different life stages and castes. For abbreviations see Supplementary Table 4.



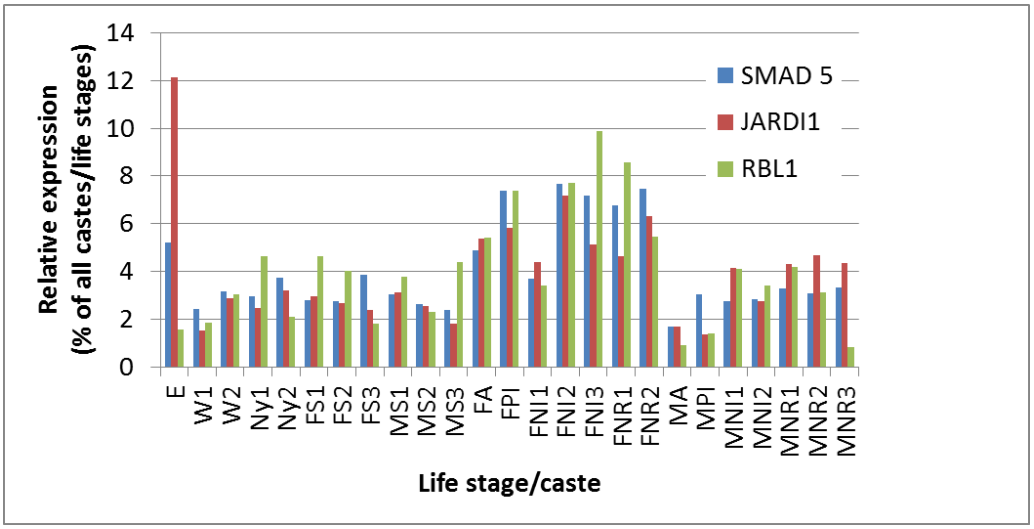
Supplementary Figure 22. Expression levels of sirtuin 7 castes and life stages. Female reproductive castes show the highest gene expression. Normalized RPKM values have been standardized to 100% across the different life stages and castes. For abbreviations see Supplementary Table 4.



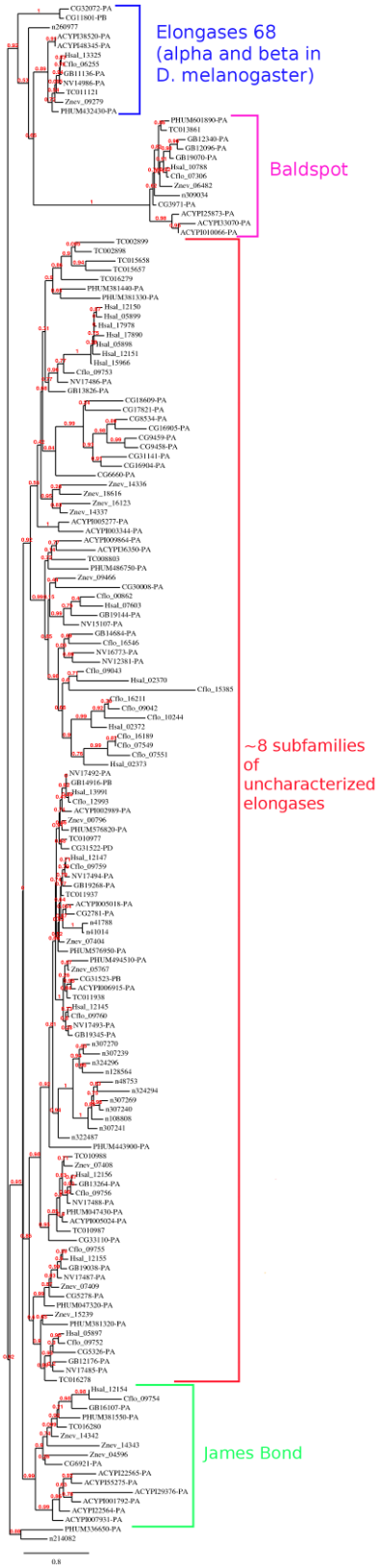
Supplementary Figure 23. Expression levels of KDM4C across castes and life stages. Female reproductive castes show the highest gene expression. Normalized RPKM values have been standardized to 100% across the different life stages and castes. For abbreviations see Supplementary Table 4.



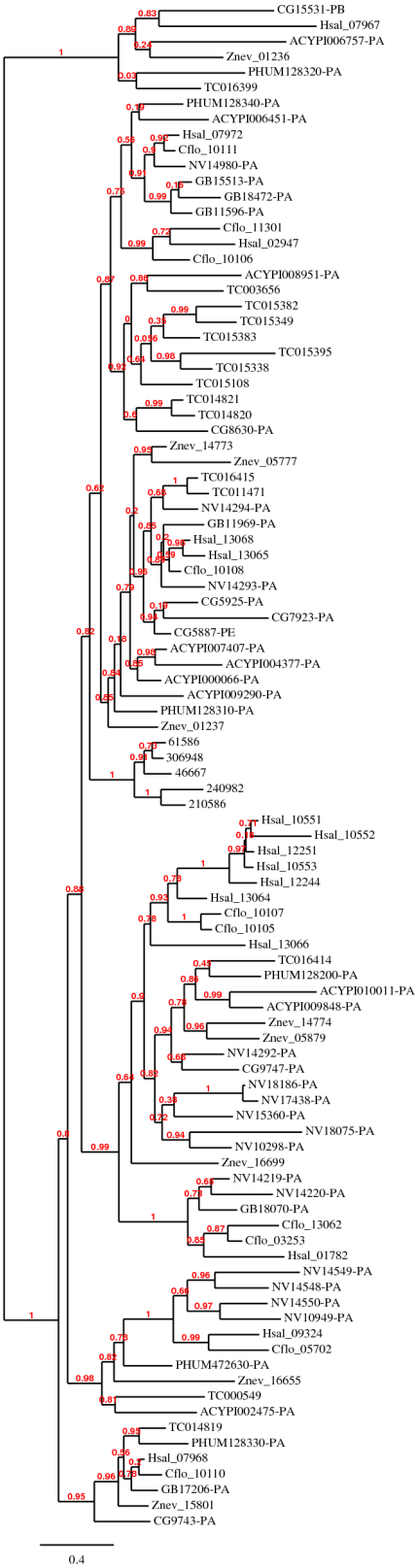
Supplementary Figure 24. Gene expression levels of histone acetyl transferases across castes and life stages. Female reproductive castes show the highest gene expression. Normalized RPKM values have been standardized to 100% across the different life stages and castes. For abbreviations see Supplementary Table 4.



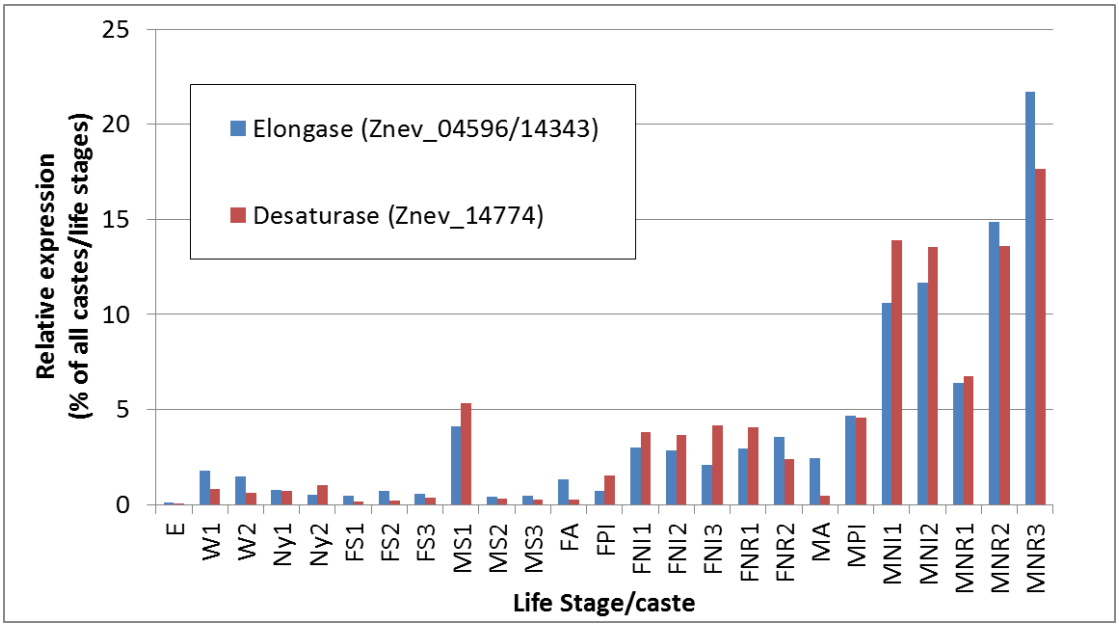
Supplementary Figure 25. Gene expression levels of histone methyl transferases across castes and life stages. Female reproductive castes show the highest gene expression. Normalized RPKM values have been standardized to 100% across the different life stages and castes. For abbreviations see Supplementary Table 4.



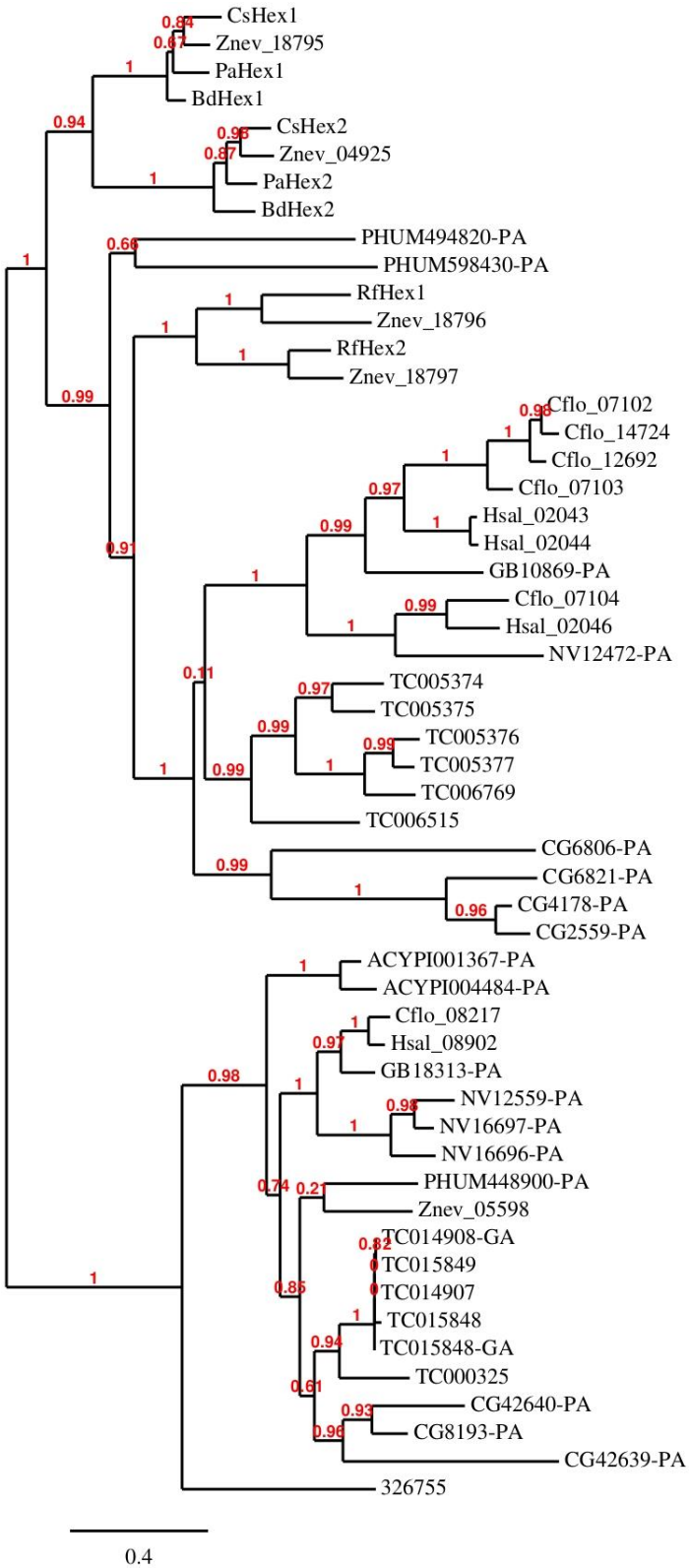
Supplementary Figure 26. Phylogeny of the elongase family. For the protocol, refer to Supplementary Notes 3.3, for the mapping of protein IDs to species refer to Supplementary Table 13.



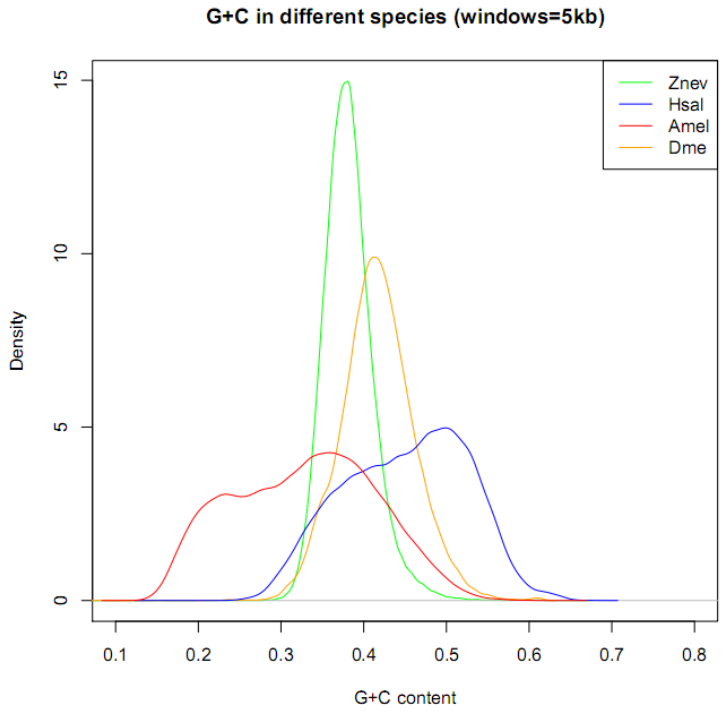
Supplementary Figure 27. Phylogeny of the desaturase family. For the protocol, refer to Supplementary Notes 3.3, for the mapping of protein IDs to species refer to Supplementary Table 13.



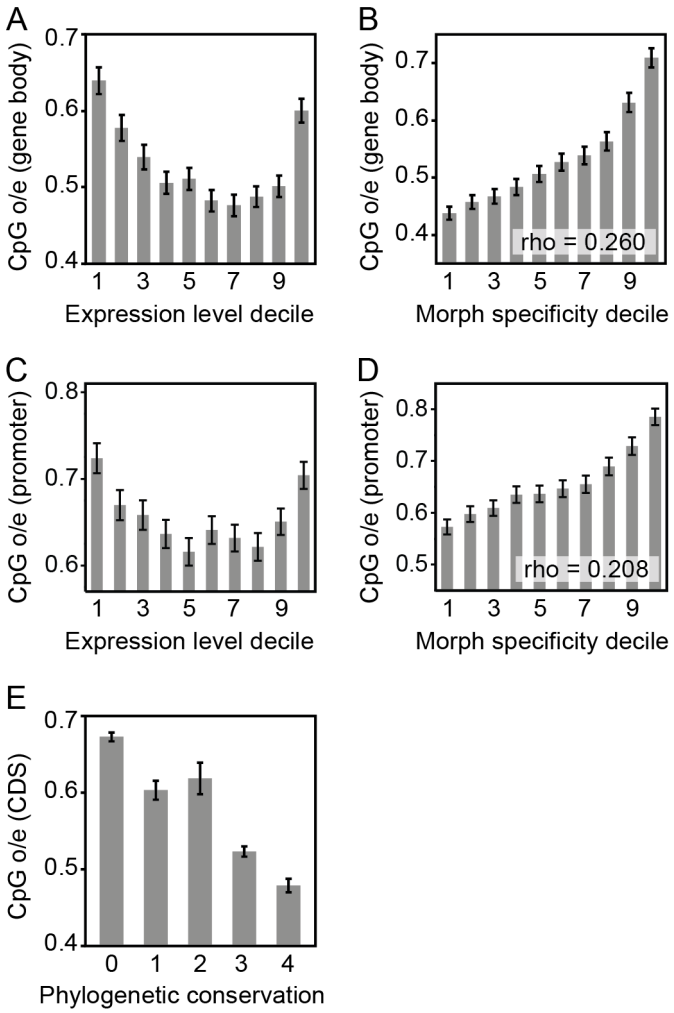
Supplementary Figure 28. Relative gene expression pattern of one elongase and one desaturase. The elongase consists of two gene models and RPKM values have been averaged accordingly. Normalized RPKM values have been standardized to 100% across the different life stages and castes. RPKM values of the two elongase fragments have been averaged. The desaturase is about ten times more highly expressed as the elongases. For abbreviations see Supplementary Table 4.



Supplementary Figure 29. Phylogeny of the Hexamerin family. For the protocol, refer to Supplementary Notes 3.3, for the mapping of protein IDs to species refer to Supplementary Table 13. Additional sequences of Isoptera species are labelled "*species initials*Hex1" or "*species initials*Hex2".



Supplementary Figure 30. Distribution of the G+C content in *Z. nevadensis, A. mellifera, D. melanogaster, and H. saltator*.



Supplementary Figure 31. Comparison of the evolutionary signature of DNA methylation (CpG o/e), gene expression, and phylogenetic conservation in *Z. nevadensis*. (A) The relationship between gene expression level and CpG o/e is parabolic, and (B) CpG o/e is positively correlated with morph specificity of expression, implying that ubiquitously expressed genes are preferentially methylated ($P < 2.2 \times 10^{-16}$, Spearman's rank correlation). Putative promoter regions 2kb upstream to the translation start site exhibit a (C) parabolic relationship with gene expression level and (D) a positive relationship with morph specificity ($P < 2.2 \times 10^{-16}$, Spearman's rank correlation), as observed for intragenic methylation. This suggests that promoter methylation is not widely associated with gene repression and indicates that the signal of promoter methylation could be driven in part by incomplete gene annotation. (E) Low CpG o/e (putatively methylated) genes are more phylogenetically conserved than genes with higher CpG o/e ($P < 2.2 \times 10^{-16}$, Kruskal-Wallis test). We assigned a score from zero to four to convey phylogenetic conservation, with 0 representing no detected orthologs, 1 indicating ortholog detection in insects only, 2 indicating detection in invertebrates only, 3 indicating detection in animals only, and 4 indicating detection across broad eukaryotic taxa (see methods). Means and 95% confidence intervals are plotted in all panels.

Supplementary Tables

Supplementary Table 1. Allelic diversity and Expected (H_E) and observed (H_O) heterozygosity

Locus	Colony alleles (1&2)	Pebble Beach Genotypes	Pebble Beach H_E	Pebble Beach H_O
Zoot-117	204	204	0.000	0.000
Zoot-73	229	226, 229	0.405	0.400
Zoot-254	200	200, 216	0.129	0.000
Za-18	257	255, 257, 259	0.551	0.267
Za-123	121	119, 121	0.066	0.067

Supplementary Table 2. Statistics of the read data before and after filtering

Insert Size(bp)	Average Reads Length	Raw Data Reads (M)	Raw Data Bases (Mb)	Filtered Data Reads (M)	Filtered Data Bases (Mb)	Coverage Depth (%)[1]
178	100	219.11	21,910.88	194.85	19,484.92	34.67
607	100	183.76	18,376.26	150.85	15,085.37	26.84
804	100	56.23	5,623.45	44.54	4,454.20	7.93
2,000	57.5	160.30	11,529.00	131.62	7,567.43	13.47
5,000	44	60.03	2,641.21	51.03	2,245.41	4.00
10,000	44	175.03	7,701.23	123.61	5,438.97	9.68
20,000	44	117.5	5759.41	23.64	1,040.22	1.85
Total		971.96	67782.03	720.14	55,316.52	98.43

[1] coverage depth is calculated based on the estimated genome size.

Supplementary Table 3. Statistics of the assembled genome

	Contig		Scaffold	
	Size (bp)	Number	Size (bp)	Number
N90	4,092	25,628	116,867	751
N80	8,217	17,756	302,910	509
N70	11,955	13,013	430,730	373
N60	15,756	9,578	584,779	274
N50	20,030	6,925	740,215	199
Longest	151,931		5,111,804	
Total Size	472,159,647		493,468,737	

Supplementary Table 4. Description of samples used for differential expression analysis

Sample[1]	**Abbrev**	**Colony**	**Number of individuals**	**Notes**[2]
Egg	E	127	700	Mixture of eggs of different developmental stages
Worker	W1	145	8	2 males and 2 females from both 3rd instar and 4th instar
Worker	W2	133	8	2 males and 2 females from both 3rd instar and 4th instar
Nymph	Ny1	145	2	1 male and 1 female
Nymph	Ny2	135	2	1 male and 1 female
Female soldier	FS1	138	1	
Female soldier	FS2	140	1	
Female soldier	FS3	-	1	
Female alate	FA	137	2	
Female primary with inactive ovaries	FPI	138	1	Ovaries with many ovarioles, corpora lutea, only one fully developed egg, hydrocarbon profile without reproductive peaks
Female neotenic with inactive ovaries	FNI1	125 + 140		Ovaries with many ovarioles, corpora lutea, but no eggs present, hydrocarbon profile of only one individual with reproductive peak (100%)
Female neotenic with inactive ovaries	FNI2	136-I	3	Ovaries with many ovarioles, corpora lutea, but no fully developed eggs, but one with signs of vittelogenesis, only one hydrocarbon profile with a small reproductive peak (13%) available
Female neotenic with inactive ovaries	FNI3	125 + 119		Ovaries with many ovarioles, corpora lutea, but no eggs present, hydrocarbon profile without reproductive peaks
Female neotenic reproductively active	FNR1	139		Ovaries with many ovarioles, corpora lutea, three fully developed eggs present, hydrocarbon profile with reproductive peak (94%)

Female neotenic reproductively active	FNR2	123	2	Both females with wing buds, well developed ovaries containing eggs, hydrocarbon profiles with reproductive peaks (27% and 28%)
Male soldier	MS1	122	1	
Male soldier	MS2	-	1	
Male soldier	MS3	116	1	
Male alate	MA	133	4	
Male primary with inactive testes	MPI	151B	1	Inconspicuous testes
Male neotenic with inactive testes	MNI1	140	1	Inconspicuous testes
Male neotenic with inactive testes	MNI2	13	1	Inconspicuous testes
Male neotenic with testes of intermediate development	MNR1	125	1	Testes enlarged, probably active, no reproductive-specific hydrocarbon peak present
Male neotenic reproductively active	MNR2	136-I	3	All males with large testes, hydrocarbon profile with reproductive peak (20%, only one available)
Male neotenic reproductively active	MNR3	123	2	Both males with large testes, hydrocarbon profiles with reproductive peaks (44% and 200%)

[1] Underlined samples belong to the first batch of RNA sequencing and were also used in gene prediction (see Methods in the main manuscript and, with MS2 and FS3, in the analysis of alternative splicing associated with DNA methylation (see Supplementary Notes 11.3)

[2] Some of the reproductive individuals show a reproductive-specific hydrocarbon (6,9,17-tritriacontatriene) within their cuticular hydrocarbon profile. The relative expression of this compound in the profile is given as percentage of the average peak area of heneicosane, tricosane, and pentacosane of this individual's hydrocarbon profile.

Supplementary Table 5. RNA-seq sample statistics.

Sample	# mapped reads	# total reads	Read mapped rate
E	61,187,583	88,584,594	69.1%
W1	74,352,244	100,262,480	74.2%
W2	77,112,183	96,920,562	79.6%
Ny1	60,431,068	72,240,428	83.7%
Ny2	56,497,718	72,629,630	77.8%
FS1	50,519,813	63,002,976	80.2%
FS2	50,542,226	72,082,622	70.1%
FS3	56,930,106	70,009,670	81.3%
FA	54,711,762	67,371,004	81.2%
FPI	67,385,188	77,803,592	86.6%
FNI1	63,143,716	79,202,218	79.7%
FNI2	55,597,964	56,399,559	98.6%
FNI3	76,248,768	97,389,526	78.3%
FNR1	59,568,834	69,247,868	86.0%
FNR2	57,152,829	70,806,064	80.7%
MS1	68,493,938	91,577,308	74.8%
MS2	56,729,960	65,878,948	86.1%
MS3	59,558,108	78,917,868	75.5%
MA	41,439,191	63,723,528	65.0%
MPI	90,071,507	116,281,472	77.5%
MNI1	69,478,502	90,846,806	76.5%
MNI2	68,039,463	84,632,764	80.4%
MNR1	72,224,783	89,376,970	80.8%
MNR2	44,823,513	52,973,217	84.6%
MNR3	53,010,498	74,815,512	70.9%

Supplementary Table 6. Differentially expressed genes in the *types of samples* and *unique samples*

	Over-expression	**Under-expression**	**Cumulated***
Juveniles	131	17	148
Soldier	146	9	155
Male reprod.	637	173	810
Female reprod.	2508	2657	5165
Egg	296	104	400
Male Alate	186	0	186
Female Alate	17	1	18
Cumulated*	3902	2960	6309

A blue background indicates the types of samples having replicates and while *unique samples* without replicates are depicted by in orange.

*Cumulated numbers of differentially expressed (DE) genes. This does not necessarily correspond to the sum since DE genes in *unique samples* are identified only against the four *types of samples* (with replicates) and not against other *unique samples* for statistical reasons, hence an under-expressed gene can be both listed for female alate and male alate for example. Note that this is not possible in the case of sample types.

Supplementary Table 7. Gene families differentially expressed (DE) across samples.

Group/ Putative function	Protein Family	Domain Architecture	Individuals with DE	DE expression	# of DE proteins	Expressed proteins	Fisher's p-value	Additional details
Male mating biology	BTB-BACK-Kelch	PF00651 PF07707 PF01344	*Male reprod.*	OVER	25	37	1.3E-24	Female reprod.: 3 over, 7 under
	Kelch	PF01344	*Male reprod.*	OVER	14	20	5.0E-16	Female reprod.: 4 under
	PKD	PF08016	*Male reprod.*	OVER	7	10	7.0E-08	Female reprod.: 1 over, 3 under
	BACK-Kelch	PF01344 PF07707	*Male reprod.*	OVER	4	4	5.8E-06	Female reprod.: 2 under
	BTB-Kelch	PF00651 PF01344	*Male reprod.*	OVER	4	6	8.0E-05	Female reprod.: 1 under
	Disintegrins & metalloproteinases (ADAMTS)	PF01562 PF01421 PF00090 PF05986	*Male reprod.*	OVER	4	5	2.8E-05	Female reprod.: 1 under
Queen regulation	Zinc finger C2H2	PF00096	*Female reprod.*	OVER	81	186	2.6E-14	Female reprod.: 20 under; Juveniles: 2 over; Egg 1 over; Male reprod.: 1 over, 2 under
	Zinc finger C2H2 and associate domain	PF00096 PF07776	*Female reprod.*	OVER	19	32	6.7E-07	Female reprod.: 1 under
	Histone	PF00125	*Female reprod.*	OVER	14	17	3.6E-08	Egg: 2 over; Male reprod.: 8 under
Egg development	Osiris DUF1676	PF07898	*Egg*	OVER	12	12	1.6E-20	
Cuticular protein	Chitin Binding	PF00379	*Egg, Juveniles /*	OVER / UNDER	20, 7 / 10	54	1.2E-19, 1.1E-06 /	Egg: 20 over, 1 under; Juveniles: 7 over; Male reprod.: 1 over, 10 under;

			Male reprod.				2.0E-09	Female reprod.: 4 over, 5 under
Pheromone binding proteins	PBP/GOBP	PF01395	*Female reprod.*	UNDER	18	24	1.4E−08	Male alate: 3 over; Juveniles: 2 over; Egg: 1 over; Female alate: 1 over
Esterase	Carboxylesterase family	PF00135	*Female reprod.*	UNDER	22	36	1.2E-07	Egg: 2 under; Solder: 1 over; Female reprod.: 3 over; Juveniles: 2 over; Male alate: 1 over
Reproduction biology	Seven-in-Absentia	PF03145	*Male & female reprod. +soldier&male alate (ALL)*	OVER	21	29	3.1E-06	Male reprod.: 6 over, 6 under; Female reprod.: 13 over, 4 under; Soldier: 1 over Male alate: 1 over
Protein kinase	Pkinases	PF00069	*Male & female reprod.+ Soldier (ALL)*	OVER	65	134	5.2E-06	Female reprod.: 40 over, 21 under; Male reprod.: 23 over, 1 under; Soldier: 2 over
Caste differentiation	P450	PF00067	*Egg, female reprod. / Juveniles*	UNDER / OVER	11, 29 / 6	67	4.5E-12, 1.9E-05 / 5.7E-05	Egg: 4 over, 11 under; Female reprod. 4 over, 29 under; Juveniles: 6 over; Male reprod.: 1 over, 2 under; Male alate: 1 over
Unknown	Leucine rich repeat LRR_1	PF00560	*ALL*	UNDER	27	58	5.7E-05	Male reprod.: 2 over, 1 under; Female reprod.: 7 over, 23 under; Egg: 1 over, 3 under; Male alate: 2 over; Soldier: 1 over

Supplementary Table 8. General statistics of predicted protein-coding genes.

	Gene set	Number	Average transcript length (bp)	Average CDS length (bp)	Average exons per gene	Average exon length (bp)	Average intron length (bp)
De novo	Augustus	21,224	7,526	1,113	5.06	219	1,577
	SNAP	43,140	10,095	775	4.74	163	2,490
	Merged[1]	13,706	16,266	1,429	7.64	187	2,233
Homolog	*A.mellifera*	8,106	10,052	1,320	6.64	198	1,548
	C.floridanus	14,957	5,151	826	3.95	209	1,464
	D.melanogaster	5,451	10,989	1,444	7.44	194	1,482
	H.sapiens	5,397	10,177	1,366	7.41	184	1,374
	H.saltator	13,834	5,545	844	4.09	206	1,523
	Merged[2]	20,005	4,995	793	3.81	208	1,493
RNA-seq		38,123	2,583	567	2.79	203	1,127
Final gene set		17,737	8,520	1177	5.60	210	1,595

[1] intersection set of Augustus and SNAP predictions, selecting the longest for overlapped gene models
[2] union set of all the homology-based predictions, selecting the longest for overlapped gene models

Supplementary Table 9. Non-coding RNA genes in the genome.

Type		Copy	Average length (bp)	Total length (bp)	% of genome
miRNA		96	114.59[1]	11,001	0.002229
tRNA		1860	73.29	136,319	0.027625
rRNA	18S	31	134.68	4,175	0.000846
	28S	6	76.17	457	0.000093
	5.8S	1	44.00	44	0.000009
	5S	15	78.73	1,181	0.000239
	Total	53	110.51	5,857	0.001187
snRNA	CD-box	9	95.56	860	0.000174
	HACA-box	0	0.00	0	0
	splicing	37	141.62	5,240	0.001062
	Total	46	132.61	6,100	0.001236

[1] average length of miRNA is average length of the predicted precursor miRNAs.

Supplementary Table 10. Results of homology-based prediction.

	RepeatMasker		Protein Masker		Combined TEs[1]	
	Length (bp)	% in genome	Length (bp)	% in genome	Length (bp)	% in genome
DNA	1,533,628	0.31	4,259,200	0.86	5,144,391	1.0
LINE	2,094,073	0.42	6,425,534	1.3	6,903,460	1.4
SINE	412,493	$8.4e^{-2}$	0	0	412,493	$8.4e^{-2}$
LTR	229,928	$4.7e^{-2}$	431,787	$8.8e^{-2}$	601,790	0.12
Others[2]	1,282	$2.6e^{-4}$	0	0	1,282	$2.6e^{-4}$
Unknown[3]	15,313	$3.1e^{-3}$	0	0	15,313	$3.1e^{-3}$
Total	4,216,826	0.85	11,113,336	2.3	13,001,826	2.6

[1] "Combined TEs" represent the non-redundant set of results of RepeatMasker and Protein Masker.
[2] "Others" refers to the repeats that can be classified by RepeatMasker, but not included by the classes above.
[3] "Unknown" refers to the repeats that can't be classified by RepeatMasker.

Supplementary Table 11. Results of *de novo* prediction.

Type	Length (bp)	% of genome
DNA	10,652,427	2.2
LINE	20,025,930	4.1
SINE	9,739,380	2.0
LTR	186,361	$3.8e^{-2}$
Simple repeat	1,385,258	0.28
Satellite	-	-
Others[1]	-	-
Unknown[2]	81,304,307	16.5
Total	119,786,086	24.3

[1] "Others" refers to the repeats that can be classified by RepeatMasker, but not included by the classes above.
[2] "Unknown" refers to the repeats that cannot be classified by RepeatMasker.

Supplementary Table 12. Summary of repeats in *Z. nevadensis* genome.

Type	Length (Bp)	% of genome
DNA	12782141	2.6
LINE	22129963	4.5
SINE	9781258	2.0
LTR	722363	0.15
Satellite	7015595	1.4
Minisatellite	5176160	1.1
Microsatellite	866499	0.18
Others	1282	$2.6e^{-4}$
Unknown	81316561	16.5
Total	128412165	26.0

Supplementary Table 13. Set of reference species.

Species	Abbrev.	Common name	Genome publication	OGS release	Protein IDs pattern
Drosophila melanogaster	D.mel	Fruit fly	(Adams *et al.,* 2000) [14]	5.27	CG…
Tribolium castaneum	T.cas	Flour beetle	(Richards *et al.,* 2008) [15]	3.0	TC…
Nasonia vitripennis	N.vit	Jewel wasp	(Werren *et al.,* 2010) [16]	1.2	NV…
Apis mellifera	A.mel	Honey bee	(THGSC, 2006) [17]	2.0	GB…
Camponotus floridanus	C.flo	Florida carpenter ant	(Bonasio *et al.,* 2010) [18]	3.3	Cflo_...
Harpegnathos saltator	H.sal	Jerdon's jumping ant	(Bonasio *et al.,* 2010) [18]	3.3	Hsal_...
Pediculus humanus	P.hum	Body louse	(Kirkness *et al.,* 2010) [19]	1.2	PHUM…
Acyrthosiphon pisum	A.pis	Pea aphid	(TIAGC *et al.,* 2010) [20]	2.0	ACYP…
Daphnia pulex	D.pul	Water flea	(Colbourne *et al.,* 2011) [21]	1.1	Numbers only
Caenorhabditis elegans	C.ele	Roundworm	(Coulson *et al.,* 1996) [22]	WS210	1 letter + combines letters & numbers

Supplementary Table 14. Indicators of genome and gene model quality for *Z. nevadensis* and reference genomes

Species	*C. elegans*	*D. pulex*	***Z. nevadensis***	*P. humanus*	*A. pisum*	*T. castaneum*	*D. melanogaster*	*N. vitripennis*	*A. mellifera*	*C. floridanus*	*H. saltator*
Tot. proteins	20212	30899	**17737**	10769	33267	16631	13689	17084	10660	16356	17191
CEGMA complete genes(%)	244 (98.4%)	245(98.8%)	**243 (98.0%)**	239 (96.37%)	236 (95.2%)	228 (91.9%)	248 (100%)	237 (95.6%)	234 (94.4%)	242 (97.6%)	243 (98.0%)
CEGMA partial genes(%)	248 (100%)	246 (99.2%)	**247 (99.6%)**	246 (99.2%)	248 (100%)	248 (100%)	248 (100%)	246 (99.2%)	247 (99.6%)	248 (100%)	247 (99.6%)
Clustered by OrthoMCL (%)	14168 (70.1%)	22686 (73.4%)	**12872 (72.6%)**	8562 (79.5%)	22947 (69.0%)	12037 (72.4%)	10352 (75.6%)	14546 (85.1%)	9533 (89.4%)	12391 (75.8%)	13625 (79.3%)
Clustered with termite (%)	5097 (29%)	7180 (40%)	–	7930 (45%)	7374 (42%)	8472 (48%)	7607 (43%)	7623 (43%)	7929 (45%)	8431 (48%)	8646 (49%)
KO SBH (%)	4,201 (20.8%)	8,503 (27.5%)	**5,913 (33.3%)**	4,621 (42.9%)	6,193 (18.6%)	5,623 (33.8%)	4,642 (33.9%)	4,160 (24.4%)	3,983 (37.4%)	4,821 (29.5%)	4,670 (27.2%)
PATHWAY SBH (%)	2,387 (11.8%)	5,344 (17.3%)	**3,699 (20.9%)**	2,839 (26.4%)	3,653 (11.0%)	3,429 (20.6%)	2,893 (21.1%)	2,637 (15.4%)	2,539 (23.8%)	3,046 (18.6%)	2,997 (17.4%)
BRITE SBH (%)	3,730 (18.5%)	7,282 (23.6%)	**5,149 (29.0%)**	3,959 (36.8%)	5,339 (16.0%)	4,781 (28.7%)	3,999 (29.2%)	3,599 (21.1%)	3,453 (32.4%)	4,170 (25.5%)	4,022 (23.4%)
KO BBH (%)	4,136 (20.5%)	4,913 (15.9%)	**4,268 (24.1%)**	3,920 (36.4%)	4,128 (12.4%)	4,345 (26.1%)	4,574 (33.4%)	3,550 (20.8%)	3,640 (34.1%)	3,728 (22.8%)	3,398 (19.8%)
Unique KO BBH (%)	2,736	3,828	**3,670**	3,448	3,315	3,550	4,574	4,345	3,174	3,255	3,280
Unique KO in PATHWAY BBH (%)	2,392	2,059	**2,285**	2,086	2,148	2,222	2,091	1,940	2,013	2,053	2,086
Unique PATH-WAY BBH (%)	279	289	**283**	282	286	284	282	278	283	280	281
Unique BRITE BBH (%)	2,344	3,232	**3,118**	2,939	2,791	2,996	2,781	2,594	2,710	2,786	2,798
With Pfam (%)	13374 (66.2%)	13651 (44.2%)	**9201 (51.9%)**	7727 (71.8%)	13028 (39.2%)	10397 (62.5%)	9948 (72.7%)	10156 (59.4%)	8288 (77.7%)	9072 (55.5%)	8927 (51.9%)
Amino-acid coverage	37.9%	28.0%	**31.8%**	33.8%	24.8%	34.0%	33.3%	28.7%	35.3%	30.9%	31.2%
Tot. occurrences	21197	21759	**17505**	14654	20809	19015	17633	17588	15747	16267	15700
Unique Pfam	3156	3462	**3460**	3310	3308	3395	3388	3255	3432	3395	3409

Monodom. (%)	10547 (78.9%)	10950 (80.2%)	**6768 (73.6%)**	5599 (72.5%)	10392 (79.8%)	7651 (73.6%)	7491 (75.3%)	7576 (74.6%)	5976 (72.1%)	6811 (75.1%)	6783 (76.0%)
Distinct Pfam in monodom.	2143	2493	**2274**	2128	2275	2225	2191	2134	2241	2231	2281
Multidom.	1264	1553	**1578**	1486	1467	1697	1423	1581	1602	1591	1476
Distinct architectures	2827	2701	**2433**	2128	2636	2746	2457	2580	2312	2261	2144
Distinct Pfam in multidom.	1554	1700	**1785**	1737	1688	1817	1706	1719	1820	1808	1733
Proteins with fragments (%)	–	1089	**472**	271	1021	333	118	493	384	683	474
Tot. fragmented occurrences	–	1110	**478**	271	1021	344	118	504	393	692	481
Distinct Pfam fragmented	–	410	**249**	204	346	220	88	215	243	263	257
Most frequent fragment (ID-#)	–	PF00069 – 50	**PF00067 – 24**	PF00001 – 13	PF00078 – 123	PF00078 – 16	PF00561 – 8	PF00078 – 71	PF02949 – 22	PF02949 – 94	PF00067 – 39
2nd most frequent fragment (ID-#)	–	PF00183 – 37	**PF00069 – 18**	PF00069 – 8	PF05970 – 47	PF00561 – 12	PF00069 – 4	F07727 – 29	PF00069 – 9	PF00067 – 47	F02949 – 24
3rd most frequent fragment (ID-#)	–	PF05970 – 30	**PF00001 – 14**	PF00071 – 5	PF04827 – 42	PF00135 – 10	PF00722 – 3	PF02949 – 19	PF00067 – 7	PF00026 – 27	PF00069 – 13
4th most frequent fragment (ID-#)	–	PF07714 – 28	**PF00858 – 12**	PF04547 – 5	PF00001 – 28	PF00856 – 6	PF00400 – 3	PF00075 – 18	PF02932 – 7	PF01433 – 26	PF00089 – 13

Note: Domain fragmentation statistics being based on the averaged domain length in arthropod genomes, such values have not been investigated for *C. elegans*

Supplementary Table 15. Repartition of the twelve Yellow sub-families in *Z. nevadensis* and reference species.

	-x (-x1)	-c	-y	-b	-e*	-d* (d2, e2, e3)	-h*	-g*	-g2‡*	-k*	-f†	MRJP†	Total
Z. nev	1	1	0	1	1	1	1	0	1	1	0	0	8
P. hum	1	1	0	0	1	2	1	0	0	0	0	0	6
A. pis	0	3	1	0	1	2	1	1	3-4	1	0	0	14
D. mel	1	1	1	1	1	3	1	1	1	1	2	0	12
T. cas	5	1	1	1	1	1	1	1	1	1	1	0	15
N. vit	4	1	1	1	1	1	1	3	1	0	0	10	24
A. mel	1	1	1	1	1	1	1	1	1-2	0	0	17	27
C. flo	1	1	1	1	1	1	1	1	1	0	0	2	11
H. sal	1	1	1	1	1	1	1	1	1	0	0	2	12

* Green background correspond to the largest syntenic observed in most species.
† Pale-yellow background indicates clade-specific subfamilies.
‡ Uncertainty between phylogenies and orthoMCL classification in *A. pisum* and *A. mellifera* are indicated by ranges for -g2 subfamily

Supplementary Table 16. Positively selected genes along the *Z. nevadensis* lineage following a duplication event.

orthoMCL Cluster	***Z. nevadensis* Gene**	**dN/dS Ratio**
OG2_00101	Znev_13509	1.3235
OG2_00101	Znev_03899	5.5283
OG2_00101	Znev_15827	12.3562
OG2_00101	Znev_01140	12.3562
OG2_00101	Znev_00469	12.3562
OG2_00101	Znev_05882	12.3562
OG2_00101	Znev_16056	12.3562
OG2_00101	Znev_14296	12.3562
OG2_00101	Znev_11516	12.3562
OG2_00101	Znev_03038	12.3562
OG2_00101	Znev_12281	12.3562
OG2_00101	Znev_16182	12.3562
OG2_00122	Znev_05988	1.2557
OG2_00122	Znev_03017	1.2557
OG2_00146	Znev_12008	2.3457
OG2_00287	Znev_06181	1.543
OG2_00287	Znev_02633	1.596
OG2_00287	Znev_11937	1.0198
OG2_00287	Znev_16317	1.0108
OG2_00287	Znev_15645	1.063
OG2_00287	Znev_15846	1.6358
OG2_00287	Znev_13456	1.0612
OG2_00287	Znev_16145	1.2565
OG2_00287	Znev_06859	2.2649
OG2_00287	Znev_13380	2.2649
OG2_00287	Znev_13358	2.2649
OG2_00287	Znev_08543	2.2649
OG2_00287	Znev_15950	1.4177
OG2_00287	Znev_18602	1.4177
OG2_00287	Znev_17350	1.4177
OG2_00287	Znev_17994	1.4177
OG2_00287	Znev_15988	1.4177
OG2_00287	Znev_16024	1.4177
OG2_00287	Znev_18445	1.325
OG2_00287	Znev_16535	1.325

OG2_00428	Znev_12777	1.8993
OG2_00428	Znev_14406	2.052
OG2_00428	Znev_08082	1.7274
OG2_00428	Znev_14923	1.002
OG2_00428	Znev_15151	1.282
OG2_00428	Znev_13057	1.0739
OG2_00428	Znev_14887	1.2755
OG2_00428	Znev_08781	2.2609
OG2_00428	Znev_14231	1.2895
OG2_00428	Znev_16127	1.5071
OG2_00428	Znev_14443	1.5071
OG2_00428	Znev_15491	1.5071
OG2_00428	Znev_14888	1.5071
OG2_00428	Znev_15995	1.5071
OG2_00428	Znev_14445	1.5071
OG2_00428	Znev_09978	1.5071
OG2_00428	Znev_16141	1.5071
OG2_00428	Znev_03938	1.5071
OG2_00428	Znev_16047	1.5071
OG2_00770	Znev_16011	1.1432
OG2_00770	Znev_14685	1.5544
OG2_00770	Znev_10969	1.7934
OG2_00770	Znev_02660	1.7934
OG2_00770	Znev_02661	1.7934
OG2_00770	Znev_14696	1.7934
OG2_00770	Znev_14691	1.7934
OG2_00770	Znev_14695	1.7934
OG2_01076	Znev_13807	2.0117
OG2_01076	Znev_16492	1.5589
OG2_01076	Znev_05731	1.1313
OG2_01076	Znev_11210	1.1816
OG2_01076	Znev_16107	1.4838
OG2_01116	Znev_08447	5.3352
OG2_01450	Znev_08719	1.011
OG2_01450	Znev_16378	1.011
OG2_02698	Znev_06417	1.331
OG2_02698	Znev_06725	1.8806
OG2_02698	Znev_16394	1.2044

OG2_02698	Znev_15193	1.2044
OG2_05213	Znev_14811	1.3433
OG2_05213	Znev_18329	1.3433
OG2_07911	Znev_04216	1.8735
OG2_07911	Znev_15430	1.8735
OG2_11134	Znev_10001	1.1265
OG2_11973	Znev_10407	1.0177
OG2_11973	Znev_10424	3.1535
OG2_12511	Znev_15941	6.9028
OG2_13378	Znev_04494	1.0258
OG2_13378	Znev_15500	1.0258
OG2_13998	Znev_00347	1.2275
OG2_14001	Znev_09199	1.0349
OG2_14001	Znev_12572	1.0349

Supplementary Table 17. Gene families with significant number of members having accelerated evolution in the termite *Z. nevadensis*.

Domain names	Domain IDs	Positively Selected proteins	Family size	Fisher's p-value
BTB BACK Kelch_1	PF00651 PF01344 PF07707	19	37	2.7 e-35
Pkinase	PF00069	7	23	5.9 e-6

Supplementary Table 18. Pfam domains uniquely found in *Z. nevadensis*, or shared with one reference species.

Pfam Name	**Pfam ID**	**Z. nevadensis protein**	**Shared with one of the 9 reference sp.**	**Hits to NCBI ESTs of other arthropods**	**Taxonomic distribution in Pfam**
Periviscerokinin family	PF08259	Znev_14145	Ø	*Dictyoptera*	Dictyoptera
Fungal protease inhibitor	PF12190	Znev_12190	Ø	*Lepidoptera*	Lepidoptera
Hypoxia-inducible factor-1	PF11413	Znev_11422	Ø	*Anopheles*	Metazoa
D-arabinono-1,4-lactone oxidase	PF04030	Znev_02059	Ø	*Ixodes*	Bacteria, Eukaryota
Cyclin-dependent kinase inhibitor 3	PF05706	Znev_01901	Ø	*Ixodes*	Bacteria, Eukaryota
GTP cyclohydrolase I feedback regulatory protein	PF06399	Znev_08813	Ø	*Ixodes*	Metazoa
Ubiquinol-cytochrome c reductase, N-terminal	PF09165	Znev_13839	Ø	*Ixodes, Hemitera, Lepidoptera*	Metazoa
Domain of Unknown Function 2678	PF10856	Znev_02895	Ø	*Ixodes*	Metazoa
Domain of Unknown Function 2781	PF10914	Znev_06292	Ø	*Ixodes*	Eukaryota
Domain of Unknown Function 1113	PF06541	Znev_13127	Ø	*Ixodes*	Bacteria, few Metazoa
Sensors of blue-light using FAD	PF04940	Znev_02454	Ø	Ø	Bacteria, few Euglenida and Heterolobosea
Telomere-length maintenance and DNA damage repair	PF11640	Znev_00687	Ø	Ø	Fungi, Metazoa
Cenp-O kinetochore centromere component	PF09496	Znev_09942	Ø	Ø	Eukaryota
Cor1/Xlr/Xmr conserved region	PF04803	Znev_11027	Ø	Ø	Metazoa
Rab geranylgeranyl transferase alpha-subunit, insert domain	PF07711	Znev_00417	Ø	Ø	Chordata
NUC202 domain	PF08166	Znev_01843	Ø	Ø	Chordata
APOBEC-like N-terminal domain	PF08210	Znev_11331	Ø	Ø	Metazoa

Peroxisome biogenesis factor 1, N-terminal	PF09263	Znev_06267	Ø	Ø	Chordata
Domain of Unknown Funtion 1725	PF08333	Znev_18672	Ø	Ø	Mammalia, few Bacteria, Plasmodium
Domain of Unknown Funtion 3250	PF11618	Znev_14588	Ø	Ø	Metazoa, some protists (Trypanosma, Paramecium, Trichomonas)
Cathepsin C exclusion domain	PF08773	Znev_18359	*D. pulex*	*Ixodes, Crustacea*	Metazoa, Apicomplexa, Giardia, Paramecium
Amidohydrolase family	PF07969	Znev_03104	*D. pulex*	Ø	Bacteria, few Archea and Eukaryota
Formiminotransferase domain	PF02971	Znev_13949	*D. pulex*	*Ixodes*	Bacteria, few Archea and Eukaryota
Formiminotransferase-cyclodeaminase	PF04961	Znev_13949	*D. pulex*	*Ixodes*	Bacteria, few Archea and Eukaryota
Formiminotransferase domain, N-terminal subdomain	PF07837	Znev_13949	*D. pulex*	*Ixodes*	Bacteria, few Archea and Eukaryota
Uncharacterized ACR, YdiU	PF02696	Znev_13943	*D. pulex*	*Ixodes*	Bacteria, few Eukaryota
Centromere protein B dimerisation domain	PF09026	Znev_05595	*A. pisum*	Ø	Metazoa
Sclerostin	PF05463	Znev_11239	*A. pisum*	Ø	Metazoa
Folate receptor	PF03024	Znev_04146	*P. humanus*	*Ixodes*	Eukaryota
Islet cell autoantigen ICA69, C-terminal domain	PF04629	Znev_12836	*P. humanus*	*Ixodes, Anopheles, Aedes*	Metazoa
Agenet domain	PF05641	Znev_17745	*P. humanus*	*Ixodes, Culex, Anopheles, Gryllus*	Metazoa, Viridiplantae
Methyltransferase TYW3	PF02676	Znev_10188	*P. humanus*	*Ixodes, Caligus*	Archea, Eukaryota

IspD	PF01128	Znev_00757	*P. humanus*	Ø	Bacteria, few Eukaryota
CW-type Zinc Finger	PF07496	Znev_01396	*A. mellifera*	*Solenopsis, Ixodes*	Metazoa, Viridiplantae, protists
DNA ligase IV	PF11411	Znev_05994	*C. floridanus*	*Solenopsis, Anopheles*	Metazoa
Stage III sporulation protein D	PF12116	Znev_07634	*H. saltator*	Ø	Bacteria
Immediate early response protein	PF05760	Znev_03512	*T. castaneum*	*Ixodes*	Metazoa
Domain of Unknown Function 1989	PF09347	Znev_09556	*D. melanogaster*	*Diptera*	Bacteria, few Eukaryota

Supplementary Table 19. Pfam domains lost in *Z. nevadensis* but present in all or at least eight reference arthropods.

Pfam Domain Name	**Pfam Domain ID**	**Also lost in 1 reference sp.**	**Taxonomic distribution in Pfam**
MAP7 (E-MAP-115)	PF05672	Ø	Metazoa
Translocon-associated protein beta	PF05753	Ø	Eukaryota, Archea
DUF1075	PF06388	Ø	Metazoa
Janus/Ocnus family	PF05005	Ø	Metazoa, few protists and plants
Sox developmental protein N terminal	PF12444	Ø	Metazoa
Radial spoke protein 3	PF06098	Ø	Eukaryota
Translocase of the outer mitochondrial membrane	PF08038	D. pulex	Eukaryota
Replication protein A, C terminal	PF08784	P. humanus	Eukaryota
Asparaginase	PF00710	P. humanus	Eukaryota, Bacteria, Archea
Retrotransposon gag protein	PF03732	P. humanus	Eukaryota
BtpA family	PF03437	N. vitripennis	Eukaryota, Bacteria, Archea
Interferon-related protein conserved region	PF04836	N. vitripennis	Metazoa, few plants
Ubiquitin related modifier 1	PF09138	T. castaneum	Eukaryota, few Bacteria, Archea
DUF543	PF04418	T. castaneum	Eukaryota

Supplementary Table 20. Expanded gene families in *Z. nevadensis*.

Domain names	Domain IDs	Z. nev	D. mel	T. cas	N. vit	A. mel	C. flo	H. sal	P. hum	A. pis	D. pul
Zinc finger, C2H2 type	PF00096	215	77	96	81	88	87	31	94	210	108
Seven-in-absentia	PF03145	33	3	16	2	1	1	1	4	3	3
Ligand-gated ion channel	PF00060	134	18	29	14	7	12	10	13	14	114
Ligand-gated ion channel L-glutamate- and glycine-binding site	PF00060 PF10613	24	2	1	2	2	3	3	2	1	14
BTB BACK Kelch_1	PF00651 PF07707 PF01344	37	9	9	10	8	6	4	10	78*	7
Kelch_1	PF01344	20	3	3	6	5	0	3	2	65*	10
Zinc finger C2H2 type and Associated-Domain	PF00096 PF07776	32	38*	40*	14	11	4	2	7	6	1
PKD	PF08016	10	6*	2	1	1	1	1	3*	2	1
Alpha tubulin	PR01161 PR01162	14	5	4	4	6	7*	6	6	4	8*

* denotes cases of species without significance or similar signal.

Supplementary Table 21. Details of ZnOBP family genes and proteins

Gene[1]	Protein ID[2]	Scaffold[3]	Coordinates[4]	Strand[5]	Introns[6]	AAs[7]	ESTs[8]	Comments[9]
OBP1	Znev_08872	1026	383677-389680	+	7	150	2/1	Fine as is
OBP2	Znev_08873	1026	392149-396978	+	7	151	12/4	Fine as is
OBP3	Znev_08874	1026	402994-406257	+	6	139	12/4	Fine as is
OBP4FIX	Znev_08875/6 → Znev_19046†	1026	408887-413420	+	6	139	12/4	Assembly gap
OBP5	Znev_13321	584	39389-44529	-	6	145	12/4	Fine as is
OBP6	Znev_13320	584	30255-36412	-	6	145	12/4	Fine as is
OBP7	Znev_13319	584	19160-25932	-	6	150	12/4	Fine as is
OBP8	Znev_13318	584	7432-12232	-	6	146	12/4	Fine as is
OBP9JOI	-	584	<1-3227	-	6	144	10/4	Join across two scaffolds
	Znev_12542	874	<1-3481	+				
OBP10	Znev_19029†	874	5884-10361	+	6	167	10/4	New gene model
OBP11	Znev_12543	874	15844-24000	+	6	159	11/4	Fine as is
OBP12	Znev_02786	436	104642-108477	+	6	148	12/4	Changed final exon
OBP13	Znev_02787	436	110390-114161	+	5	155	6/0	Fine as is
OBP14	Znev_12612	672	74478-79283	+	5	173	12/4	Fine as is
OBP15	Znev_01423	204	1317512-1324337	-	6	143	12/4	Fine as is
OBP16	Znev_14610	511	134422-139115	-	5	145	12/4	Fine as is
OBP17	Znev_14609	511	119493-124936	-	6	151	12/4	Fine as is
OBP18	Znev_18969	511	111456-116777	-	6	151	12/4	New gene model
OBP19	Znev_14608†	511	106532-111221	-	6	146	12/4	Fine as is
OBP20	Znev_08639	695	330906-338315	+	6	156	12/4	Fine as is
OBP21FIX	Znev_08640	695	341351-346407	+	5	148	3/4	Assembly gap
OBP22	Znev_08641	695	350020-361640	+	6	178	12/4	Fine as is
OBP23NTE	Znev_14744	1046	48214-55125>	-	6	128	4/3	N-terminal exons missing
OBP24	Znev_14742	1046	37817-46288	-	8	239	12/4	Fine as is
OBP25	Znev_14741	1046	16019-30332	-	8	230	12/4	Fine as is
OBP26	Znev_14740 → Znev_19047†	1046	1632-9687	-	8	212	12/4	Add two N-term exons
OBP27	Znev_16171	631	4715-24658	+	8	218	10/4	Fine as is
OBP28	Znev_19002†	631	34244-42335	+	8	244	6/4	New gene model
OBP29	Znev_07853	385	339623-344616	+	8	311	12/4	Fine as is

[1] "Gene" – proposed gene name. Temporary suffixes specify: NTE – N-terminus missing; FIX – assembly has to be repaired; JOI – gene model spans scaffolds.
[2] "Protein ID" – official gene identifier in OGSv2_1, or in OGSv2.2 for new or repaired gene models (indicated by a †)
[3] "Scaffold" – the v1 genome assembly scaffold
[4] "Coordinates" – the nucleotide range from the first position of the start codon to the last position of the stop codon in the scaffold

[5] “Strand” – coding strand, + being forward and - meaning reverse;
[6] “Introns” – number of introns in coding region
[7] “AA” – length of encoded amino acid sequence
[8] “ESTs” – numbers of EST contigs spliced for at least one intron amongst twelve initial transcriptomes (with mixed sex soldiers instead of later differentiated) and number of spliced antennal transcriptome contigs out of 4
[9] “Comments” – comments on the OGSv2.1 gene model and repairs to the genome assembly

Supplementary Table 22. Details of ZnOR family genes and proteins

Gene[1]	Protein ID[2]	Scaffold[3]	Coordinates[4]	Strand[5]	Introns[6]	AAs[7]	ESTs[8]	Comments[9]
OrCo	Znev_11756	792	136055-147860	-	6	472	12/4	Fine as is
Or1	Znev_03993 → Znev_18978†	555	5699-11078	+	4	495	3/4	Multiple changes
Or2	Znev_03994/5 → Znev_18979†	555	13946-17296	+	4	488	8/4	Merge gene models
Or3	Znev_18973†	532	693940-697911	-	4	447	0/4	New gene model
Or4	Znev_08194 → Znev_18894†	292	74236-77814	-	4	510	0/0	N-terminus extension
Or5	Znev_08193 → Znev_18893†	292	68817-72044	-	4	499	1/4	Multiple changes
Or6	Znev_18892†	292	60688-64581	-	4	523	0/4	New gene model
Or7FIX → Or7CTE†	- → Znev_18891†	292	52131-56794	-	4	472	1/4	Assembly gap → Still missing C-term
Or8	Znev_18890†	292	45512-50062	-	4	494	0/1	New gene model
Or9	Znev_11216/5 → Znev_19033†	885	80832-83666	-	4	438	8/4	Merge gene models
Or10	Znev_13116 → Znev_18846†	163	228378-232840	+	4	466	10/4	Multiple changes
Or11	Znev_09049 → Znev_18824†	69	22768-28018	+	5	435	11/4	Multiple changes
Or12FIX	Znev_07294	360	387063-391376	-	4	528	3/4	Assembly gap
Or13JOI	-	1311	545025-546329>	+	4	517	3/0	Join across scaffolds
	Znev_05616	740	<1-3693	+				
Or14	Znev_19012†	740	5429-10612	+	4	478	1/4	New gene model
Or15	Znev_05617 → Znev_19013†	740	16586-20509	+	4	496	0/4	Multiple changes
Or16	Znev_18852†	172	1102630-1107716	+	4	495	4/4	New gene model
Or17	Znev_10718 → Znev_18908†	357	45378-50684	-	4	465	12/4	Needs N-term exon
Or18	Znev_18907†	357	30377-34056	-	4	475	1/4	New gene model
Or19	Znev_10716 → Znev_18906†	357	22477-26121	-	5	459	4/4	Multiple changes
Or20FIX	-	570	202435-205436	-	4	491	0/4	Assembly gap
Or21	Znev_07263 → Znev_18983†	570	197056-200498	-	4	484	1/4	Needs N-term exon
Or22	Znev_18982†	570	191320-195517	-	4	491	3/0	New gene model

Or23	Znev_18981†	570	182539-186439	-	4	454	6/4	New gene model
Or24	Znev_07749 → Znev_18948†	438	405680-411158	-	4	437	0/0	Multiple changes
Or25	Znev_07748 → Znev_18947†	438	399801-403544	-	4	493	5/4	Multiple changes
Or26	Znev_18988†	600	460342-462171	+	4	397	8/0	New gene model
Or27	Znev_03221 → Znev_18841†	147	1815868-1822504	+	5	422	8/4	Multiple changes
Or28	Znev_18823†	68	84056-88701	+	5	470	4/4	New gene model
Or29	Znev_19044†	987	362734-367669	+	5	507	5/0	New gene model
Or30	Znev_19008†	709	781196-786266	+	6	467	1/2	New gene model
Or31	Znev_19009†	709	790139-797469	+	6	479	0/0	New gene model
Or32FIX → Or32CTE†	- → Znev_19010†	709	803085-807630	-	6	497	7/4	Assembly gap → Still missing C-term
Or33FIX → Or33CTE†	- → Znev_18839†	131	241429-245950	+	6	470	0/4	Assembly gap → Still missing C-term
Or34	Znev_13573 → Znev_19038†	893	38252-45472	+	5	484	10/4	N-terminus extension
Or35	Znev_07564 → Znev_19021†	786	348517-354638	-	7	473	0/4	Multiple changes
Or36FIX → Or36CTE†	Znev_01504 → Znev_19027†	864	228780-235559	-	5	479	1/2	Assembly gap → Still missing C-term
Or37JOI → Or37CTE†	- → Znev_18975†	545	151066-154851>	+	4	438	0/4	Join across two → Still missing C-term
	-	C13664523	<1-542>	-				scaffolds and a contig
	-	650	388579-395437>	-				
Or38	Znev_07577 → Znev_18972†	512	11197-16103	-	6	406	12/4	Multiple changes
Or39	Znev_18971†	512	5991-9067	-	6	416	10/4	New gene model
Or40JI → Or40CTE†	- → Znev_18970†	512	1-3275	-	4	306	0/4	Join across scaffolds → Still missing C-term
	-	279	1-492	+				Exons missing in gap
Or41	Znev_18870†	279	5896-10249	+	6	422	0/4	New gene model
Or42	Znev_01292 → Znev_18871†	279	12877-17621	+	6	417	2/0	Multiple changes
Or43	Znev_18872†	279	19336-23255	+	6	420	1/4	New gene model
Or44	Znev_18873†	279	27824-32205	+	6	415	0/0	New gene model
Or45	Znev_18874†	279	33668-37793	+	6	408	0/4	New gene model

Or46	Znev_18875†	279	41234-45523	+	6	419	0/4	New gene model
Or47PSE	-	279	47454-51926	+	6	331	0/4	Pseudogene (1)
Or48	Znev_18838†	129	372236-376529	-	6	406	0/4	New gene model
Or49JF	-	6256	640-888>	+	6	400	3/4	Join across scaffolds
→ Or49NTE†	Znev_16112 → Znev_18836†	126	<1-11655	+				Assembly gap → Still missing N-term
Or50CTE	-	126	14915-21500>	+	6	372	5/4	Exon missing in gap
Or51PSE → Or51CTE†	- → Znev_18837†	126	24891-30440	+	6	309	2/4	Pseudogene (2) → Still missing C-term
Or52PSE	-	126	33191-39563	+	6	379	9/4	Pseudogene (1)
Or53	Znev_05538 → Znev_18855†	186	240-13420	+	6	398	9/4	Added final exon
Or54	Znev_19048†	1049	194454-199593	+	6	400	0/4	New gene model
Or55NIP	-	1049	205223-208760	+	6	236	1/3	Pseudogene (1)
Or56	Znev_19036†	892	466827-471918	-	6	412	1/4	New gene model
Or57	Znev_19035†	892	457189-463970	-	6	395	1/4	New gene model
Or58FIX → Or58CTE†	- → Znev_19034†	892	446027-453333	-	6	390	12/4	Assembly gap → Still missing C-term
Or59FP	-	3089	2671-8925	+	6	442	6/4	Assembly gap Pseudogene (1)
Or60	Znev_19068†	3089	11672-16852	+	6	416	0/4	New gene model
Or61	Znev_19069†	3089	23573-29605	+	6	410	2/4	New gene model
Or62FJP	-	3089	31472-36644>	+	6	410	0/4	Join across two
	-	C13704661	<1-608>	-				scaffolds and a contig
	-	1852	<1-571	+				Pseudogene (1)
Or63	Znev_19061†	1852	2421-9953	+	6	415	0/4	New gene model
Or64	Znev_11965 → Znev_19062†	1852	18783-23243	+	6	424	9/4	Needs first exon
Or65	Znev_19063†	1852	23965-29344	+	6	414	0/4	New gene model
Or66	Znev_19064†	1852	30878-35810	+	6	412	7/4	New gene model
Or67	Znev_11966 → Znev_19065†	1852	37646-42949	+	6	432	3/4	Multiple changes
Or68	Znev_19056†	1259	175515-182588	+	6	383	1/4	New gene model
Or69FIX → Or69NTE†	- → Znev_19070†	3390	76998-78926	-	5	376	0/0	Assembly gap

[1] "Gene" – proposed gene name. Temporary suffixes specify: PSE – pseudogene, NTE – N-terminus missing in gap, CTE – C-terminus missing in gap, INT – internal exon missing in gap; FIX – assembly has to be repaired; JOI – gene model spans scaffolds; multiple suffixes are abbreviated to single letters

[2] "Protein ID" – official gene identifier in OGSv2_1, or in OGSv2.2 for new or repaired gene models (indicated by a †)

[3] “Scaffold” – the v1 genome assembly scaffold
[4] “Coordinates” – the nucleotide range from the first position of the start codon to the last position of the stop codon in the scaffold
[5] “Strand” – coding strand, + being forward and - meaning reverse;
[6] “Introns” – number of introns in coding region
[7] “AA” – length of encoded amino acid sequence
[8] “ESTs” – numbers of EST contigs spliced for at least one intron amongst twelve initial transcriptomes (with mixed sex soldiers instead of later differentiated) and number of spliced antennal transcriptome contigs out of 4
[9] “Comments” – comments on the OGSv2.1 gene model, repairs to the genome assembly, and pseudogene status (numbers in parentheses are the number of obvious pseudogenizing mutations)

Supplementary Table 23. Details of ZnGR family genes and proteins

Gene[1]	**Protein ID**[2]	**Scaffold**[3]	**Coordinates**[4]	**Strand**[5]	**Introns**[6]	**AAs**[7]	**ESTs**[8]	**Comments**[9]
Gr1	Znev_02094 → Znev_19016†	777	814564-821236	-	7	423	Many*	Add N-terminal exon
Gr2	Znev_19015†	777	803928-809967	-	7	447	Few	New gene model
Gr3	Znev_09959/8 → Znev_19040†	919	162188-168763	-	6	405	Many*	Fuse and extend models
Gr4	Znev_09879 → Znev_19039†	912	410716-415455	-	7	403	Many*	Add N-terminal exon
Gr5NTE	Znev_18821†	68	<32808-41036	+	7	458	Few	N-terminus incomplete
Gr6	Znev_14192 → Znev_18822†	68	48136-55097	-	8	469	Many	Add N-terminal exon
Gr7	Znev_01299 → Znev_18876†	279	215979-219439	-	6	430	Some	Multiple changes
Gr8	Znev_01300 → Znev_18877†	279	221444-224539	+	6	420	Many*	Extended N-terminal exon
Gr9	Znev_01301 → Znev_18878†	279	228411-232381	+	6	444	None	Multiple changes
Gr10	Znev_01302 → Znev_18879†	279	234522-238790	+	6	430	Many*	Part of long model
Gr11	Znev_01302 → Znev_18880†	279	241130-245213	+	6	436	Few*	Part of long model
Gr12	Znev_01302 → Znev_18881†	279	247218-250802	+	6	433	None*	Part of long model
Gr13	Znev_01303 → Znev_18882†	279	252456-256566	+	6	422	None*	Multiple changes
Gr14	Znev_07457 → Znev_18818†	52	93263-99803	+	6	431	Few*	Multiple changes
Gr15	Znev_04375 → Znev_18964†	485	109158-115361	+	6	440	Few*	Multiple changes
Gr16	Znev_10879 → Znev_19022†	799	214486-220144	+	6	406	Many*	Fine as is
Gr17JOI → Gr17CTE†	Znev_08755	372	<1-12197	-	7	433	Few	Part of long model → Still missing C-term
	-	1437	17038-17448>	-				Join across two scaffolds
Gr18	Znev_08755 → Znev_18910†	372	16028-32690	+	7	432	None	Part of long model
Gr19	Znev_08756	372	34111-47954	+	7	451	None	Multiple changes

	→ Znev_18911†							
Gr20	Znev_18912†	372	48993-62813	+	7	444	None	New gene model
Gr21FIX	Znev_08758	372	67824-74438	+	6	465	None*	Assembly gap
Gr22	Znev_08759 → Znev_18913†	372	83792-100063	+	6	421	Few*	Different final exon
Gr23FIX	Znev_14427	37	7795-16468	+	7	411	Some	Assembly gap
Gr24PSE	-	37	17449-22022	-	7	300	None	Pseudogene (9)
Gr25	Znev_14429 → Znev_18816†	37	25380-39998	+	7	433	Many	Multiple changes
Gr26INT	Znev_14430	37	55346-64686	+	5	327	None	Two exons missing in gap
Gr27NC	-	37	74816-82235	+	6	329	Some	Two exons missing in gaps
Gr28INT	Znev_15974	1630	11254-22732	+	7	365	Some	Exon missing in gap
Gr29	Znev_18848†	168	2625-15167	+	7	415	Some	New gene model
Gr30	Znev_15114 → Znev_18849†	168	18240-28097	+	7	411	Some	Multiple changes
Gr31	Znev_18850†	168	29237-44180	-	7	433	None	New gene model
Gr32	Znev_15116 → Znev_18851†	168	45851-55526	-	7	433	Some	Multiple changes
Gr33	Znev_13534 → Znev_19000†	630	64803-70136	+	3	432	Some*	Extend first exon
Gr34	Znev_19001†	630	72659-76325	+	3	414	Few*	New gene model
Gr35a	Znev_02883 → Znev_18869†	273	1093908-1103091	-	3	435	Few*	Alternatively-spliced
Gr35b	-	273	1093908-1099845	-	3	422	Few*	Alternatively-spliced
Gr36	Znev_19043†	942	78567-83784	-	3	439	Many	New gene model
Gr37	Znev_18805†	20	20467-23626	-	3	427	Few	New gene model
Gr38	Znev_18804†	20	16280-18724	-	3	417	None	New gene model
Gr39FIX → Gr39NTE†	- → Znev_18803†	20	12738-15000	+	3	435	None	Assembly gap → Still missing N-term
Gr40	Znev_18802†	20	7195-11310	-	3	422	Few*	New gene model
Gr41	Znev_18801†	20	3095-5558	+	3	380	None	New gene model
Gr42	Znev_18924†	392	57928-62303	+	3	469	None	New gene model
Gr43aFIX	-	36	<44078-53808	+	3	467	Few	Assembly gap
Gr43bPSE	-	36	46106-53808	+	3	417	None	Pseudogene (1)
Gr43c	- → Znev_18815†	36	49308-53808	+	3	446	None	Alternatively-spliced
Gr44	Znev_18914†	383	1652-2977	+	0	441	None	New gene model
Gr45	-	383	4067-5392	+	0	441	Few	New gene model

Gr46	Znev_18864†	257	663402-664709	-	0	435	None	New gene model
Gr47	Znev_18865†	257	743566-744963	+	0	465	Few	New gene model
Gr48	Znev_18927†	392	371655-373019	-	0	454	None	New gene model
Gr49	Znev_09636 → Znev_18926†	392	369678-371054	-	0	458	Many	Extend N-terminus
Gr50	Znev_18925†	392	366550-367872	-	0	440	Few	New gene model
Gr51	Znev_09653 → Znev_18928†	392	889320-890615	+	0	431	None	Fine as is
Gr52	Znev_09654 → Znev_18929†	392	892668-893969	+	0	433	Few	Extend N-terminus
Gr53	Znev_03015 → Znev_18932†	399	567233-568528	-	0	431	Some*	Extend N-terminus
Gr54FIX → Gr54CTE†	Znev_03014 → Znev_18931†	399	561865-563125	-	0	429	Many*	Assembly gap → Still missing C-term
Gr55	Znev_18842†	151	6213-7487	-	0	424	None	New gene model
Gr56FIX	-	151	44-1460	-	0	435	None	Assembly gap
Gr57	Znev_18991†	623	184651-185997	+	0	448	None	New gene model
Gr58PSE	-	623	188139-189398	+	0	419	None	Pseudogene (1)
Gr59	Znev_18992†	623	190545-191861	+	0	438	None	New gene model
Gr60	Znev_18993†	623	194807-196135	+	0	442	None	New gene model
Gr61	Znev_18994†	623	198969-200291	-	0	440	None	New gene model
Gr62	Znev_18995†	623	202550-203872	-	0	440	None	New gene model
Gr63	Znev_18996†	623	204898-206211	+	0	437	None	New gene model
Gr64	Znev_18997†	623	207683-208975	-	0	430	None	New gene model
Gr65	Znev_18998†	623	210551-211846	+	0	431	None	New gene model
Gr66	Znev_19026†	863	284939-286165	+	0	408	None	New gene model
Gr67	Znev_18921†	385	156759-158048	-	0	429	None	New gene model
Gr68	Znev_18916†	385	58940-60247	-	0	435	Few*	New gene model
Gr69	Znev_18966†	487	26997-28175	-	0	392	Few	New gene model
Gr70PSE	-	487	22094-23433	+	0	374	Few*	Pseudogene (1)
Gr71	Znev_18868†	264	450415-451761	+	0	448	None	New gene model
Gr72	Znev_18896†	311	45925-47175	+	0	416	Few	New gene model
Gr73	Znev_18897†	311	52379-53620	+	0	413	None	New gene model
Gr74	Znev_18898†	311	54599-55831	+	0	411	Few*	New gene model
Gr75	Znev_18899†	311	56730-57959	+	0	409	Few	New gene model
Gr76PSE	-	311	61438-62738	-	0	432	Few*	Pseudogene (1)
Gr77PSE	-	311	62994-63996	-	0	333	None*	Pseudogene (1)
Gr78	Znev_18900†	311	68503-69738	+	0	411	None*	New gene model

Gr79PSE	-	311	71483-72930	+	0	368	None*	Pseudogene (2)
Gr80	Znev_18901†	311	75660-76910	+	0	416	None	New gene model
Gr81FIX → Gr81NTE†	- → Znev_18902†	311	<77989-79215	+	0	416	Few*	Assembly gap → Still missing N-term
Gr82	Znev_18903†	311	81210-82460	+	0	416	None*	New gene model
Gr83	Znev_18840†	141	11010-12224	-	0	404	Some	New gene model
Gr84FIX	-	937	113805-115050	-	0	399	Few	Assembly gap
Gr85	Znev_19023†	827	760340-761554	+	0	404	Few	New gene model
Gr86	Znev_19024†	827	762205-763443	+	0	412	Few	New gene model
Gr87	Znev_18798†	3	1059582-1060847	+	0	421	Some	New gene model

[1] "Gene" – proposed gene name. Temporary suffixes specify: PSE – pseudogene, NTE – N-terminus missing in gap, CTE – C-terminus missing in gap, INT – internal exon missing in gap; FIX – assembly has to be repaired; JOI – gene model spans scaffolds; multiple suffixes are abbreviated to single letters

[2] "Protein ID" – official gene identifier in OGSv2_1, or in OGSv2.2 for new or repaired gene models (indicated by a †)

[3] "Scaffold" – the v1 genome assembly scaffold

[4] "Coordinates" – the nucleotide range from the first position of the start codon to the last position of the stop codon in the scaffold

[5] "Strand" – coding strand, + being forward and - meaning reverse;

[6] "Introns" – number of introns in coding region

[7] "AA" – length of encoded amino acid sequence

[8] "ESTs" – numbers of EST contigs spliced for at least one intron amongst twelve initial transcriptomes (with mixed sex soldiers instead of later differentiated) and number of spliced antennal transcriptome contigs out of 4. An asterix (*) indicates at least one EST contig spliced for at least one intron amongst four RNA-seq assemblies from antennae, except for Gr44-87, which are intronless genes for the coding region in which case only gene-long EST contigs are considered

[9] "Comments" – comments on the OGSv2.1 gene model, repairs to the genome assembly, and pseudogene status (numbers in parentheses are the number of obvious pseudogenizing mutations)

Supplementary Table 24. Details of ZnIR family genes and proteins

Gene[1]	Protein ID[2]	Scaffold[3]	Coordinates[4]	Strand[5]	Introns[6]	AAs[7]	ESTs[8]	Comments[9]
NMDAR1	Znev_00196 → Znev_18808†	24	312743-337196	+	17/0/0	962	12/0	Fine as is
NMDAR2	Znev_07796 → Znev_18965†	486	582664-678895	+	20/-/-	1058	12/0	Extend both ends
NMDAR3	Znev_05405-7	476	609022-721610	+	12/-/-	1845	12/0	Fuse and extend 3 models
NMDAR4	Znev_16944 → Znev_18980†	555	33715-106041	+	19/-/-	1055	10/0	Add exons to both ends
AMPAR	Znev_09358/9 → Znev_18860†	232	187856-355836	+	19/-/0	938	6/0	Fuse and extend gene models
KAINATE1	Znev_06788 → Znev_19051†	1158	229835-235464	-	16/0/0	885	12/2	N-terminus remains unclear
KAINATE2	Znev_06791	1158	238032-246102	+	14/1/1	890	12/4	Split gene model
KAINATE3	Znev_06791 → Znev_19052†	1158	249609-259650	+	15/0/0	898	12/4	Split gene model
KAINATE4	Znev_07758 → Znev_18949†	438	598281-650500	-	15/-/1	854	12/1	Fine as is
KAINATE5	Znev_05107	464	<1-45953	-	16/-/0	902	12/4	Assembly gap
	Znev_05185	1199	<1-5852	+				Join across scaffolds
IR25a	Znev_05203 → Znev_19054†	1199	350096-361142	+	16/2/0	935	12/4	Fine as is
IR8aFIX	Znev_03604 → Znev_18858†	200	221818-240151	+	14/1/1	869	12/4	Assembly gap
IR93a	Znev_01116 → Znev_18953†	457	1080724-1092791	-	18/0/0	864	9/3	Greatly expand 2-exon model
IR76b	Znev_08444 → Znev_18807†	22	1244248-1266674	-	8/1/0	532	12/4	Fine as is
IR68a	Znev_07531 → Znev_18905†	346	822176-828724	-	5/2/0	681	12/3	Fine as is
IR21a	Znev_00358 → Znev_18951†	453	1251330-1257797	+	8/0/0	806	12/4	Fine as is
IR41a1	Znev_02124 → Znev_19017†	777	1770311-1776126	-	4/-/-	744	0/0	Extend N-terminus
IR41a2	Znev_01449 → Znev_18835†	123	457353-462688	+	5/1/0	694	10/4	Add 1 N-terminal exon
IR41a3	Znev_08399 → Znev_18867†	264	121243-125621	-	5/1/1	720	9/4	Multiple changes

IR41a4	Znev_06514/5 → Znev_18862†	236	1344008-1350992	-	7/1/0	670	3/4	Fuse and extend gene models
IR41a5	Znev_06513 → Znev_18861†	236	1334074-1340855	-	7/1/0	662	4/4	Add 5 N-terminal exons
IR75a	Znev_05253 → Znev_18960†	468	846124-856855	-	8/1/0	577	12/4	Add 1 N-terminal exon
IR75bINT	Znev_05246 → Znev_18959†	468	642211-656214	-	10/-/0	675	7/3	Unfixable assembly gap
IR75c	Znev_18958†	468	633964-640357	-	8/-/0	675	2/4	New gene model
IR75d	Znev_05245 → Znev_18957†	468	625001-631127	-	8/0/0	632	6/0	Add 1 N-terminal exon All ESTs female?
IR75e	Znev_05244 →Znev_18956†	468	614472-622735	-	8/1/0	630	3/4	Add 4 N-terminal exons
IR75f	Znev_18955†	468	603215-612967	-	8/-/0	640	8/4	New gene model
IR75gFIX	Znev_05243 → Znev_18954†	468	593221-601222	-	8/1/0	601	4/3	Assembly gap
IR75h	Znev_10656 → Znev_19067†	2486	58002-63014	-	8/0/0	620	12/4	Fine as is
IR75i	Znev_18989†	602	204919-220707	-	9/-/-	637	0/0	New gene model
IR75j	Znev_10420 → Znev_19005†	671	135096144887	-	10/1/0	625	0/4	Add exons to both ends.
IR75k	Znev_18967†	495	722597-729545	+	11/0/0	626	5/4	New gene model
IR75l	Znev_01172 → Znev_18968†	495	730179-736802	+	10/-/0	647	2/4	Add exons to both ends
IR75m	Znev_00202 → Znev_18809†	24	427018-435155	+	9/-/0	627	10/0	Add 3 N-terminal exons
IR75n	Znev_00203 → Znev_18810†	24	439397-448255	+	9/-/0	650	11/4	Add 3 N-terminal exons
IR75o	Znev_00204 → Znev_18811†	24	455432-467919	+	10/1/0	627	7/4	Fine as is
IR75p	Znev_00205 → Znev_18812†	24	470390-480647	+	9/-/0	625	8/4	Add 2 N-terminal exons
IR75q	Znev_00206 → Znev_18813†	24	483018-492467	+	9/0/0	653	5/4	Add 8 N-terminal exons
IR101	Znev_09248 → Znev_18847†	167	292400-297621	+	5/1/0	571	12/0	Fine as is
IR102	Znev_01041/2 → Znev_19025†	827	1148795-1151901	-	1/0/2	572	12/4	Fuse and extend gene models

IR103NTE	Znev_12348 → Znev_19011†	715	272201-278908	-	5/-/-	627	0/0	N-terminus unidentified
IR104PSE	-	715	264379-271711	-	4/-/-	628	3/0	Pseudogene (1)
IR105	Znev_09919/20 → Znev_19059†	1256	327676-334279	-	8/1/0	664	12/0	Fuse gene models
IR106	Znev_02312 → Znev_18923†	386	12320-17768	+	7/-/-	729	8/1	Add 5 N-terminal exons
IR107NTE	Znev_07813 → Znev_19049†	1092	493262-501314	+	6/-/-	390	8/2	N-terminal 3 exons missing
IR108	Znev_00610 → Znev_18799†	4	2918460-2932528	+	8/0/0	618	10/3	Add 3 N-terminal exons
IR109	Znev_18819†	66	1613158-1617569	+	8/-/-	648	0/0	New gene model
IR110	Znev_00106 → Znev_18820†	66	1931874-1937127	-	8/0/0	648	6/2	Add exons to both ends
IR111	Znev_07840 → Znev_18922†	385	167985-173628	+	8/-/0	648	11/4	Add 3 N-terminal exons
IR112	Znev_07839 → Znev_18920†	385	146396-154811	-	8/0/0	605	12/4	Add 4 N-terminal exons
IR113	Znev_07838 → Znev_18919†	385	139404-145671	+	8/0/0	672	5/4	Add 7 N-terminal exons
IR114PSE	-	385	129124-135005	+	8/-/-	641	3/2	Pseudogene (1)
IR115	Znev_07837 → Znev_18918†	385	112727-124973	-	8/-/-	630	1/0	Multiple changes
IR116PSE	Znev_07836	385	103570-117692	-	6/-/-	471	0/0	Pseudogene (1)
IR117	Znev_07835 → Znev_18917†	385	91604-98286	-	8/-/-	620	2/0	Add exons to both ends
IR118	Znev_03021 → Znev_18933†	399	873670-881403	+	8/-/-	614	3/3	Multiple changes
IR119	Znev_03022 → Znev_18934†	399	883823-891727	-	8/-/-	620	0/0	Multiple changes
IR120	Znev_03023 → Znev_18935†	399	895133-902978	-	8/-/-	601	0/0	Second half of model
IR121	Znev_03023 → Znev_18936†	399	904939-914957	-	8/-/-	615	3/0	First half of model
IR122FIX	Znev_18937†	399	918050-926232	-	8/-/-	625	0/0	Assembly gap
IR123FIX	Znev_03024	399	929974-938393	-	8/-/-	646	3/0	Assembly gap
IR124	Znev_18938†	399	940994-947477	+	8/-/-	606	5/0	New gene model
IR125	Znev_03027	399	980109-987459	+	8/-/0	626	8/0	Add 3 N-terminal exons

	→ Znev_18939†							
IR126	Znev_18940†	399	992609-1001421	+	8/-/-	640	8/2	New gene model
IR127	Znev_03028 → Znev_18941†	399	1002503-1009402	+	8/-/-	616	9/0	Add 3 N-terminal exons
IR128	Znev_01303 → Znev_18883†	279	257528-262198	+	8/-/-	614	1/0	Part of long model
IR129	Znev_01303 → Znev_18884†	279	263797-269607	+	8/-/-	603	0/0	Part of long model
IR130	Znev_01303 → Znev_18885†	279	275960-282134	+	8/-/-	645	0/0	Part of long model
IR131	Znev_18886†	279	283992-288556	+	8/-/-	623	0/0	New gene model
IR132	Znev_01306 → Znev_18887†	279	289109-293358	+	8/-/-	617	4/0	Add exons to both ends
IR133	Znev_18888†	279	293830-298771	-	8/-/-	628	4/0	New gene model
IR134	Znev_18830†	92	1138590-1145915	-	8/-/-	629	1/0	New gene model
IR135	Znev_02183 → Znev_18829†	92	1129252-1135725	-	8/-/-	638	0/0	Part of long model
IR136	Znev_02183 → Znev_18828†	92	1121296-1128540	-	8/-/-	611	1/0	Part of long model
IR137	Znev_19018†	786	98773-109771	+	8/-/-	620	0/0	New gene model
IR138	Znev_19019†	786	111332-120196	+	8/-/-	615	3/1	New gene model
IR139	Znev_07549 → Znev_19020†	786	137356-148261	+	8/-/-	608	2/0	Part of long model
IR140FIX	Znev_07549	786	156259-164690	+	8/-/-	620	0/0	Assembly gap
IR141	Znev_11451/2 → Znev_18943†	415	230472-242092	+	8/-/-	634	7/0	Fuse and extend models
IR142	Znev_15343 → Znev_18915†	383	9433-21323	+	8/-/-	611	5/0	Add 6 N-terminal exons
IR143	Znev_00861 → Znev_18825†	87	1273157-1282825	+	8/-/-	653	8/2	Add 7 N-terminal exons
IR144	Znev_00863 → Znev_18826†	87	1289416-1304821	+	9/1/0	667	8/2	Add 7 N-terminal exons
IR145	Znev_18859†	204	1138060-1160362	-	8/0/0	639	0/4	New gene model
IR146	Znev_15081 → Znev_18834†	117	251211-260976	+	8/0/0	620	8/0	Add 3 N-terminal exons
IR147FIX	Znev_10307	689	142024-155874	+	8/0/0	618	3/0	Assembly gap
IR148NTE	Znev_10552 → Znev_19050†	1119	240677-244209	-	7/-/0	507	3/4	First two exons missing

IR149	Znev_07950 → Znev_18946†	420	499578-509564	-	8/1/0	669	12/4	Add exons to both ends
IR150CTE	Znev_12699/70 → Znev_19006†	686	3450-13564	-	8/-/-	529	8/1	Fuse and extend models; Last exon missing
IR151CTE	Znev_19007†	686	20126-26591	+	8/-/-	543	0/0	Missing last exon
IR152	Znev_09127 → Znev_18990†	623	139326-147106	-	8/-/-	686	0/0	Add first and last exons
IR153FIX	Znev_13812/3 → Znev_19030†	879	58921-74379	+	8/-/-	621	0/0	Assembly gap
IR154CTE	Znev_19031†	879	78299-85345	+	7/-/-	447	0/0	Last two exons missing
IR155CTE	Znev_19032†	879	89571-94547	+	8/-/-	510	0/0	Last exon missing
IR156	Znev_11873 → Znev_19045†	995	77196-79256	+	0/-/-	686	0/0	Convert to single exon model
IR157FIX	-	1353	1028339-104849	+	0/0/0	678	8/0	Fix polymorphic insertion
IR158	Znev_05865 → Znev_18817†	48	533821-535695	-	0/-/-	624	0/0	Fine as is
IR159	Znev_05304 → Znev_18986†	583	1612-3603	+	0/0/0	663	0/4	Fine as is
IR160	Znev_18987†	583	6915-8921	+	0/-/-	668	0/0	New gene model
IR161	Znev_01505 → Znev_19028†	864	251215-253353	+	0/0/0	712	12/4	Fine as is
IR162	Znev_03590 → Znev_18800†	18	1897383-1899356	-	0/-/-	657	0/0	Extend single exon model
IR163	Znev_08443 → Znev_18806†	22	1232989-1234914	+	0/-/-	641	0/0	Fine as is
IR164	Znev_10025 → Znev_18930†	395	91207-93051	+	0/1/0	614	10/2	Fine as is
IR165	Znev_18961†	483	1201-3201	-	0/0/0	667	0/4	New gene model
IR166	Znev_18962†	483	17252-19276	+	0/-/-	674	0/0	New gene model
IR167	Znev_14498 → Znev_18963†	483	25978-28785	+	1/-/-	664	0/0	Add N-terminal exon; Novel intron
IR168	Znev_18833†	115	748499-750508	+	0/0/0	669	0/4	New gene model
IR169	Znev_18977†	554	312176-314218	-	0/0/0	680	4/4	New gene model
IR170PSE	-	263	2529-4591	+	0/0/0	686	0/4	Pseudogene (1)
IR171FIX →IR171PAR†	Znev_01694 → Znev_18866†	263	6930-8763	+	0/0/0	680	6/1	Fix insertion in assembly
IR172	Znev_19058†	1517	410799-412637	+	0/-/-	662	0/0	New gene model
IR173	Znev_18952†	454	687742-689760	-	0/-/-	672	0/0	New gene model

IR174	Znev_09128 → Znev_18999†	623	215728-217812	+	0/0/0	694	8/0	Extend single exon model
IR175	Znev_18827†	92	785545-787542	+	0/-/-	665	0/0	New gene model
IR176	-	1477	28019-30130	+	0/0/0	703	8/0	New gene model
IR177NP	Znev_15584/5	416	<1-1945	+	0/-/-	648	7/4	Pseudogene (4)
IR178IP	Znev_15586	416	5403-7603	-	0/-/-	551	2/3	Pseudogene (4)
IR179INT	Znev_16376	1537	1045-2999	+	0/0/0	597	8/0	Has an assembly gap
IR180	Znev_13560 → Znev_18976†	552	12153-13907	+	0/-/-	584	6/0	Remove last two exons
IR181	Znev_09379 → Znev_18831†	100	546173-548137	+	0/-/-	654	5/0	Extend single exon model
IR182	Znev_18832†	100	603379-605407	+	0/-/-	675	0/0	New gene model
IR183PSE	Znev_16401	634	1574-3192	-	0/0/0	538	10/4	Pseudogene (5)
IR184PSE	-	205	5706-7364	+	0/0/0	553	12/4	Pseudogene (5)
IR185	Znev_09042 → Znev_19037†	892	675959-677938	-	0/0/0	659	10/0	Extend single exon model
IR186	Znev_18944†	417	76616-78592	+	0/0/0	658	0/4	New gene model
IR187	Znev_14829 → Znev_18945†	417	227203-229206	+	0/0/0	667	8/0	Extend single exon model
IR188	Znev_19004†	668	165329-167308	+	0/0/0	659	5/0	New gene model
IR189PSE	Znev_04642/3	668	754453-756536	-	0/0/0	693	6/0	Pseudogene (2)
IR190PSE	-	1355	13880-15790	-	0/-/-	629	4/4	Pseudogene (3)
IR191	-	1355	8716-10728	-	0/0/0	670	8/4	New gene model
IR192	Znev_19053†	1165	9628-11745	+	0/-/-	705	4/0	New gene model
IR193FIX	-	1165	<16933-18190	+	0/-/-	661	10/4	Assembly gap and frameshift
IR194	Znev_14932 → Znev_19057†	1319	171681-173705	+	0/0/0	674	11/4	Extend single exon model
IR195	Znev_18984†	577	137367-139340	+	0/-/-	657	6/4	New gene model
IR196	Znev_18904†	324	262851-264773	-	0/-/-	640	0/0	New gene model
IR197	Znev_19003†	658	253496-255514	+	0/-/-	672	5/0	New gene model
IR198	Znev_10970 → Znev_18950†	445	18924-21038	-	0/0/0	704	10/0	Extend single exon model
IR199	Znev_18856†	188	55897-58032	+	0/1/0	711	0/4	New gene model
IR200	Znev_18857†	188	59464-61972	+	1/0/0	694	0/4	New gene model; novel intron
IR201	Znev_18854†	185	329294-331312	+	0/-/-	672	0/0	New gene model
IR202	Znev_18942†	411	43923-45734	+	0/-/-	603	1/1	New gene model
IR203	Znev_13821 → Znev_19066†	2001	42748-44475	-	0/0/0	575	10/3	Fine as is

IR204	Znev_18853†	183	439510-441432	-	0/-/-	640	0/0	New gene model
IR205PSE	-	165	64456-66567	+	0/-/-	703	0/0	Pseudogene (1)
IR206	Znev_19014†	744	437941-439956	+	0/0/0	671	5/1	New gene model
IR207	Znev_18974†	537	119595-121907	-	1/0/0	663	8/0	New gene model; novel intron
IR208FIX →IR208PAR†	Znev_13524 → Znev_19055†	1259	103734-105574	-	0/1/0	601	10/4	Assembly gap
IR209	Znev_18863†	245	65950-67905	+	0/-/-	651	2/0	New gene model
IR210FIX	Znev_03026	399	974204-976116	+	0/0/0	605	12/4	Assembly gap
IR211	Znev_15827 → Znev_19060†	1844	33761-36548	-	1/-/-	663	0/0	Remove C-terminus from model
IR212	Znev_10077 → Znev_18985†	577	242505-245268	+	2/-/-	695	3/0	Two novel introns
IR213	Znev_18895†	293	1028668-1030791	+	0/0/0	707	10/3	New gene model
IR214PSE	-	927	477973-479837	-	0/-/-	593	0/0	Pseudogene (4)
IR215	Znev_19041†	927	1082529-1084442	+	0/-/-	637	0/0	New gene model
IR216	Znev_19042†	927	1086253-1088199	+	0/-/-	648	4/0	New gene model
IR217	Znev_18845†	155	634264-636366	+	0/-/-	700	3/0	New gene model
IR218	Znev_18844†	155	627530-629515	-	0/-/-	661	0/0	New gene model
IR219	Znev_18814†	24	3264563-3266470	+	0/-/-	635	3/0	New gene model
IR220PSE	-	24	3271165-3273141	+	0/-/-	658	1/0	Pseudogene (1)
IR221	Znev_18843†	151	52402-54390	-	0/-/-	662	0/0	New gene model
IR222	Znev_04891 →Znev_18889†	281	1075374-1077398	+	0/0/0	674	12/0	Extend single exon model

[1] "Gene" – proposed gene name. Temporary suffixes specify: PSE – pseudogene, NTE – N-terminus missing in gap, CTE – C-terminus missing in gap, INT – internal exon missing in gap; FIX – assembly has to be repaired; JOI – gene model spans scaffolds; PAR – partial gene models despite correction; multiple suffixes are abbreviated to single letters

[2] "Protein ID" – official gene identifier in OGSv2_1, or in OGSv2.2 for new or repaired gene models (indicated by a †)

[3] "Scaffold" – the v1 genome assembly scaffold

[4] "Coordinates" – the nucleotide range from the first position of the start codon to the last position of the stop codon in the scaffold

[5] "Strand" – coding strand, + being forward and - meaning reverse;

[6] "Introns" – number of introns in coding region / number of 5' UTR introns / number of 3' UTR introns (- indicates no EST evidence available)

[7] "AA" – length of encoded amino acid sequence

[8] "ESTs" – numbers of EST contigs spliced for at least one intron amongst twelve initial transcriptomes (with mixed sex soldiers instead of later differentiated) and number of spliced antennal transcriptome contigs out of 4, except for IR156-222, which are mostly intronless genes for the coding region in which case only nearly gene-long EST contigs are considered

[9] "Comments" – comments on the OGSv2.1 gene model, repairs to the genome assembly, and pseudogene status (numbers in parentheses are the number of obvious pseudogenizing mutations)

Supplementary Table 25. *Z. nevadensis* immune genes

Gene name	Gene function/pathway	*Z.nevadensis* Gene ID
Attacin	Effector/AMP	Znev_02297
Diptericin		Znev_09129
Termicin	Effector/AMP	
Lysozyme C-type	Lysozyme	Znev_03738
	Lysozyme	Znev_17477
	Lysozyme	Znev_11184
Lysozyme I-type	Lysozyme	Znev_07418
	Lysozyme	Znev_01125
	Lysozyme	Znev_11158
	Lysozyme	Znev_07419
GNBP	Pattern Recognition	Znev_03260
	Pattern Recognition	Znev_02878
	Pattern Recognition	Znev_03257
	Pattern Recognition	Znev_03259
	Pattern Recognition	Znev_00932
	Pattern Recognition	Znev_00933
PGRP	Pattern Recognition	Znev_07518
	Pattern Recognition	Znev_09910
	Pattern Recognition	Znev_08618
	Pattern Recognition	Znev_09909
	Pattern Recognition	Znev_07984
	Pattern Recognition	Znev_01249
18wheeler	Toll receptor	Znev_10053
Toll/Tollo	Toll receptor	Znev_00888
	Toll receptor	Znev_10041
	Toll receptor	Znev_10044
Toll	Toll receptor	Znev_13969
Toll or LRR	Toll receptor	Znev_05966
Toll-9	Toll receptor	Znev_10923
Toll-like receptor	Toll receptor	Znev_01370
Tollip (Toll-interacting protein)	TOLL pathway	Znev_07741
ECSIT (signal intermediate in Toll pathway)	TOLL pathway	Znev_07957
Pelle	TOLL pathway	Znev_09453
Tube	TOLL pathway	Znev_05846
TRAF	TOLL pathway	Znev_17913
	TOLL pathway	Znev_11985
	TOLL pathway	Znev_15253
Spaetzle	TOLL pathway	Znev_00635
	TOLL pathway	Znev_04528
Spaetzle-like	TOLL pathway	Znev_11366

	TOLL pathway	Znev_10323
	TOLL pathway	Znev_10324
DIF/DORSAL	TOLL pathway	Znev_07478
Easter (Spaetzle-Processing enzyme)	TOLL pathway	Znev_10162
JNK-interacting SapK	TOLL pathway	Znev_15407
Relish (NF-Kappa-B)	NF-K-B-related	Znev_11193
C-Jun/JNK	NF-K-B-related	Znev_00650
Mpk2	NF-K-B-related	Znev_02213
Cactus		Znev_07020
NF-kappa-B inhibitor alpha	NF-K-B-related	Znev_09660
NF-kappa-B inhibitor-like	NF-K-B-related	Znev_03760
Kappa-B-ras (NF-kappa-B inhibitor alpha-interacting)	NF-K-B-related	Znev_06522
NF-kappa-B-repressing factor	NF-K-B-related	Znev_05897
STAT	JAK-STAT pathway	Znev_16675
Cytokine receptor	JAK-STAT pathway	Znev_09344
JAK pathway STAM	JAK-STAT pathway	Znev_00434
JAK/hopscotch	JAK-STAT pathway	Znev_10639
IKB (I-Kappa-B)	IMD pathway	Znev_09963
IMD (immune deficiency)	IMD pathway	Znev_02405
FAS-assocaited factor (TNFRSF6)	IMD pathway	Znev_08114
Optineurin (NF-K-B modulator)	IMD pathway	Znev_04648
MAPKKK (TAK1)	IMD pathway	Znev_09904
MAPKKK	IMD pathway	Znev_12168
MYLIP (defense repressor)	IMD pathway	Znev_02112
NIK + IKBKB-BP (TRAF-like)	IMD pathway	Znev_04156
prophenoloxidase	PO-related	Znev_05598
Hemocyanin	PO-related	Znev_04925
	PO-related	Znev_04926
Coagulation factor XI	PO-related	Znev_16656
Prophenoloxidase-activating enzyme 2	PO-related	Znev_18221
TEP1	Thioester-containing protein	Znev_02879
TEP2	Thioester-containing protein	Znev_18586
TEP3	Thioester-containing	Znev_13964

	protein	
TEP4	Thioester-containing protein	Znev_05513
Dual oxidase/heme peroxidase	Peroxidase	Znev_17480
Dual oxidase/heme peroxidase	Peroxidase	Znev_00592
Peroxidasin/Chorion peroxidase	Peroxidase	Znev_16904
Peroxidasin/Chorion peroxidase	Peroxidase	Znev_01987
Peroxidasin/Chorion peroxidase	Peroxidase	Znev_16752
Peroxidasin/Chorion peroxidase	Peroxidase	Znev_09888
Peroxidasin/Chorion peroxidase	Peroxidase	Znev_02993
Peroxidasin/Chorion peroxidase	Peroxidase	Znev_16996
Peroxidase	Peroxidase	Znev_03574
Peroxidase	Peroxidase	Znev_03575
ATG2 (Autophagy-related protein 2)	Autophagy	Znev_14004
ATG2	Autophagy	Znev_11462
ATG3	Autophagy	Znev_06264
ATG4B	Autophagy	Znev_05555
ATG4D	Autophagy	Znev_10705
ATG5	Autophagy	Znev_12390
ATG6 (Beclin)	Autophagy	Znev_07723
ATG6 (Beclin)	Autophagy	Znev_17966
ATG7	Autophagy	Znev_17566
ATG7	Autophagy	Znev_04460
ATG8 (Gabarap)	Autophagy	Znev_13859
ATG9	Autophagy	Znev_01174
ATG10	Autophagy	Znev_06925
ATG12	Autophagy	Znev_07412
ATG16L1	Autophagy	Znev_10009
ATG16L1	Autophagy	Znev_10012
RB1-inducible coiled-coil	Autophagy	Znev_02531
ULK2 (unc-51-like kinase 2)	Autophagy	Znev_10659
ULK3	Autophagy	Znev_04866
Wipi1 (WD repeat domain phosphoinositide-interacting protein 1)	Autophagy	Znev_09191

Wipi3	Autophagy	Znev_13104
Wipi4	Autophagy	Znev_10905
WD repeat-containing protein 65	Autophagy	Znev_02990
APAF-1 (apoptotic protease activating factor 1)	Apoptosis	Znev_00248
APAF-2	Apoptosis	Znev_00246
BAX (Apoptosis inhibitor)	Apoptosis	Znev_15412
Ice (effector caspase-1)	Apoptosis	Znev_13497
Effector caspase	Apoptosis	Znev_14304
Effector caspase	Apoptosis	Znev_17676
Effector caspase	Apoptosis	Znev_18362
Effector caspase	Apoptosis	Znev_08436
DREDD/Nedd2	Apoptosis	Znev_13406
DCN1-like protein	Apoptosis	Znev_00522
Fas (TNFRSF6)-associated via death domain	Apoptosis	Znev_06474
Ankyrin repeat and death domain-containing protein	Apoptosis	Znev_09041
Galectin	Lectin	Znev_07311
Galectin	Lectin	Znev_04392
CTL	C-Lectin	Znev_00043
CTL	C-Lectin	Znev_01827
CTL	C-Lectin	Znev_05559
CTL	C-Lectin	Znev_16861
CTL	C-Lectin	Znev_05556
CTL	C-Lectin	Znev_00446
CTL (Macrophage mannose receptor 1)	C-Lectin	Znev_14909
CTL (sushi, von Willebrand factor type A, EGF and pentraxin domain-containing protein)	C-Lectin	Znev_08675
CTL (sushi, von Willebrand factor type A, EGF and pentraxin domain-containing protein)	C-Lectin	Znev_03178
NPC2-like	ML superfamily (MD-2-related lipid-recognition)	Znev_00122
MPA2 allergen	ML superfamily	Znev_11387

MPA2 allergen	ML superfamily	Znev_11388
MDL1	ML superfamily	Znev_17997
MDL2	ML superfamily	Znev_11389
SCARA (Scavenger receptor class A-like)	Scavenger Receptor A	Znev_09392
SCARA	Scavenger Receptor A	Znev_09907
SCARA	Scavenger Receptor A	Znev_03412
SCARA	Scavenger Receptor A	Znev_12289
SRCR (scavenger receptor class A-like, cysteine-rich)	Scavenger Receptor A	Znev_12950
SRCR	Scavenger Receptor A	Znev_12911
SCARB (scavenger receptor class B), croquemort type	Scavenger Receptor B	Znev_01894
SCARB, croquemort type	Scavenger Receptor B	Znev_01929
SCARB, croquemort type	Scavenger Receptor B	Znev_17151
SCARB	Scavenger Receptor B	Znev_09663
SCARB	Scavenger Receptor B	Znev_09665
SCARB	Scavenger Receptor B	Znev_09668
SCARB	Scavenger Receptor B	Znev_17970
SCARB	Scavenger Receptor B	Znev_05148
SCARB	Scavenger Receptor B	Znev_14227
SCARB	Scavenger Receptor B	Znev_05147
LDL receptor	LDL	Znev_11465
LDL receptor	LDL	Znev_11466
LDL receptor	LDL	Znev_11467
T-cell immunomodulatory protein		Znev_04939
IG-binding protein	TOR pathway	Znev_11186
IG-domain containing serine protease inhibitor	IG superfamily	Znev_15173
	Serpin	Znev_11231
	Serpin	Znev_03219
	Serpin	Znev_11099
	Serpin	Znev_01881
	Serpin	Znev_11098
	Serpin	Znev_11095
	Serpin	Znev_07264
	Serpin	Znev_02826
	Serpin	Znev_06875
	Serpin	Znev_11096
	Serpin	Znev_10043
	Serpin	Znev_02802

	Serpin	Znev_05728
Leukocyte elastase inhibitor	Serpin	Znev_06230
cSP (serine protease stubble)	cSP	Znev_16873
	cSP	Znev_17166
	cSP	Znev_12938
	cSP	Znev_07739
	cSP	Znev_05138
	cSP	Znev_03586
	cSP	Znev_03587
	cSP	Znev_03593
	cSP	Znev_06101
	cSP	Znev_09866
	cSP	Znev_12298
	cSP	Znev_05193
Superoxide dismutase	SOD	Znev_15634
	SOD	Znev_12756
	SOD	Znev_11634

Supplementary Table 26. Genes of the IIS and TOR pathways.

D. melanogaster protein name	*Z.nevadensis* protein ID
ILP4	Znev_07418
ILP7	Znev_12419
InR	Znev_16736
Ras	Znev_14471
akt/PKB	Znev_04339
S6K	Znev_16688
Atg1	Znev_10659
FOXO	Znev_14322
PI3K92E	Znev_01400
MAP4K3	Znev_03645
RagC	Znev_07460
TOR-C1	Znev_11128

Supplementary Table 27. Enzymes involved in the juvenile hormone III biosynthetic pathway

Isoprenoid pathway		JH branch	
Substrate ↓ Enzyme	*Z. nevadensis* protein ID	Substrate ↓ Enzyme	*Z. nevadensis* protein ID
Acetyl-CoA		Farnesyl-PP	
↓ acetoacetyl-CoA thiolase	Znev_06783	↓ farnesyl diphosphate pyrophosphatase*	Znev_13079
Acetoacetyl-CoA		Farnesol	
↓ HMG-S[1]	Znev_08283 Znev_14140	↓ Farnesol dehydrogenase*	Znev_12423†
HMG-CoA		Farnesal	
↓ HMG-R[2]	Znev_02974	↓ Farnesal dehydrogenase*	Znev_01798 Znev_03358
Mevalonate		Farnesoic acid	
↓ mevalonate kinase	Znev_14029	↓ JH methyltransferase	Znev_12145
Mevalonate-5-P		Methyl farnesoate	
↓ phosphomevalonate kinase	Znev_12655	↓ JH epoxidase	Znev_14299
Mevalonate-5-PP		Juvenile hormone III	
↓ Diphosphomevalonate decarboxylase	Znev_08133		
Isopentenyl-PP ← isopentnyl-diphosphate δ-isomerase → Dimethylally-PP	Znev_09747		
+Geranyl-PP ↓ farnesyl diphosphate synthase	Znev_02161		
Farnesyl-PP			

[1] HMG-S stands for 3-hydroxy-3-methylglutaryl-CoA synthase
[2] HMG-R – 3-hydroxy-3-methylglutaryl-CoA reductase
* No structural data are available to associate them specifically to JH III biosynthesis
† We are reporting here only the protein with highest E-value (best BLAST hit), while several paralogs are candidates and the true participant is unknown at this time. This was confirmed by the orthoDB database that clusters ~9 on average paralogs for all insect species, involving nine proteins (4 being adjacent) in *Z. nevadensis* genome: Znev_05770, Znev_05797, Znev_07337, Znev_12420, Znev_12421, Znev_12422, Znev_12423, Znev_17844 and Znev_17938.

Supplementary Table 28. Neuropeptide sequences in *Z. nevadensis*. Signal peptide are marked in green, putative active peptide in blue, basic cleavage sites in yellow, and glycine, used for amidation, in gray. Dots indicate missing sequences while dash indicts the termination code location. Signal peptids were predicted by SignalP while putative neuropeptides were determined based on similarity to other insect neuropeptides.

Name	Scaffold	Exons	Sequence
Adipokinetic hormone (AKH1)	13	870428-870264, 869344-86297	MSCIAKTIFVMVALIFVFCEAQVNFTPNWGKRSGLQDAPCKASTEAAMYIYKLIQNEAQKLLDCEKF GSN-
Adipokinetic hormone (AKH2)	13	463232-463222, 462105-461934, 461075-460929	MTSRRLCGRALLLVAVLNCLHFRTWGQVTFSRDWNAGKRSPADLQCSAIIKSADEFCRVLIEEFRQL AACETKSLLRFLKDYDDSQADIFMESQNGRQTPTNDLHQRNF-
Allatostatin A	171	135980-134841	MLGLQSSLGSLKMTLFSVLLLHLTVLVLGTASAPSETHETAEESSPVSAGGMGLVPQLEDSSSAENA ELDFVKRLYDFGLGKRAYSYVSEYKRLPVYNFGLGKRSKMYSFGLGKRSGTEGRLYSFGLGKRDYDD YAEENEDEDQTNGDEEFEDSDLDLMEKRERLYSFGLGKRARPYSFGLGKRSPSSGIQRLYGFGLGKR GGSLYSFGLGKRADGRLYSFGLGKRPVNSGRQSGSRFNFGLGKRSDIDYNEFDDELGEEAKGFPQGH RYYLGLGKREVAPSELDAIRNEEREKINYRDESRKNETAEGHHSGERVKRSLHYAFGLGKRAYDLES STIDTDEDDEARNDFARLIRRPFNFGLGKRIPLYDFGIGKRSER-
Allatostatin C	656	280604-280440	…LVDDDGSIETALINYLFAKQVVNRLRSQMDVSDLQRKRSYWKQCAFNAVSCFGK-
Allatostatin CC	1044	383367-383453	...MDLQRRGQQKGRVYWRCYFNAVTCFKRK-
Allatotropin	174	583932-583697, 581315-581183	MRASLSVNCMIAATVLVVLVLCDCVSSGPSYQNARNKPRTIRGFKNVALSTARGFGKRDGALSYLAD NANTASEPTLESLPVEWFVEELRTNPELARIIVHKFVDADQDGELSAEELLRPMY-
Bursicon alpha	619	76990-77023, 78944-79107, 80399-80524, 81678-81812	MACQQPSIQQIAVSAVLLLSLGYVVLVDAKDECQVTPVIHVLQYPGCVPKPIPSFACTGRCSSYLQV SGSKIWQMERSCMCCQESGEREANVSLFCPKAKAGERKFRKVSTKAPLECMCRPCSTVEESAVIPQE IAGYADEGPLSNHFRKSL-
Bursicon beta	619	69512-69647, 71345-71469, 71914-72084	MVSEVAWWFRSLLLFILFFVAVPSTVQQGEDVACETLPSEIHIIKEEFDDLGRLQRTCSSEVSVNKC EGACTSQVQPSVITPTGFLKECYCCRESFLRERIISLTHCYDPDGIRLTQEGQASLDVKIREPADCK CFKCGDFSR-
CCHamide	1501	271745-272095	…ANTSSLSNLASYEVRVGVVVAIRAVYVLFVAGGCSAFGHSCFGGHGKRTDGNAVLIPGPDSDQQPL LVFRPEEEDADDAMVQQGALPSAWSATAGISPRQPPPLPAKYNLTPFLRQW…
Corazonin	206	140903-140616, 135668-135546	MHLNSTASSSRCRSRMTGLLLIFCCLTGSILAQTFQYSRGWTNGRKRSGPSPQMLIPSSASGERLFQ NTDESSAISNPCSQLQRIRFLLGARNPQQFYFPCETWTDIFETPSEEVSERFRRKAHQDFAEGNNIE GN-
Crustacean Cardioactive Peptide (CCAP)	1199	504619-504720, 505675-505747, 506856-507016, 509125-509259, 510163-510192	MQMCHVVIGCSVAVLLMIIGLPLASCDSVIIQKRQIDPADVDRILDPKRKRPFCNAFTGCGKKRSDE SMGTLVELNSEPAVEDLSRKILSETKLWEAIQEARAELLRRRQEQLQLQEGQYATAVERPIPLSITG YRKKRSVIPEGTGNSLLTTSEPQDQSTKTWSR-
Diuretic hormone (DH31-like)	24	1021602-1021505, 1019148-1019038, 1008245-1008076, 1004735-1004663, 1004452-1004314	MNSCVVLLTSALLVGAVLMISVVNAVESAPLSSHRNNFISDQDSEPDSEYVLEMLARLGQSIIRAND LEKGLIALMMEAARTSETSVNFYQTTRLNNPEDNHLQTHRRENLKSYTNIVLSNKLDLQASQERFCS MKSIINPTGGLLISAIFSKRGLDLGLSRGFSGSQAAKHLMGLAAANYAGGPGRRRRSVDESS-

Diuretic hormone (DH37-like)	546	12402-12175, 10307-10033, 8396-8321,	MLAAVLTLLLSALVSCGTASIEPPLLEALAAPSADHETTSYLLPRLSAKFRPHGDWDSAPDPRFYVL TELQRESSQAARRMKRTGAVPSLSIVNPLDVLRQRLLLEIARRRMRQSQDQIQANREMLQTIGKRDA DQSQQRSSDDDDEDDEMDSEFMVSSTDEKSGSNRPRDTPDWSPSSGPRWDDEFASQHH-
Eclosion Hormone (EH)	495	1521789-1521529	MEQRKISEAVVLAMSVAFLAATVVVPSGATSYSIGVCIRNCAQCKKMFGPYFEGQLCADACVKFKGK IIPDCEDLASIAPFLNKFE-
Ecdysis triggering Hormone (ETH1 and ETH2)	83	193633-193422, 192636-192540, 192184-192052, 191678-191479	MVAAILVVLTACPSIISADETGTNFFLKSSKSVPRIGRRSEYDFLKASKEIPRIGRRREMSPLVSDT NYGVLRDANRPCGLGSEPPTQFQVRRGGPTTTYMRKGFAVLPSKTTEADYGLLGYDLRVILSGYRRF GRSISSPSSVWNDGKPLSWTSVEKTMEEAPELWKPDLWRKNSETFPLRDDFDVEQVVRRSPGRFGST KEIDEGRNQVEV-
Glycoprotein hormone-alpha2	1159	32615-32738, 33731-33906, 35348-35440	MFPVSWRLQCCYLLFIFVTILSVVSRTRARDAWERPGCHKVGHTRKISIPDCVEFHITTNACRGYCE SWAVPSAIDTLRVNPHQAITSVGQCCNIMDTEDVEVQVMCLDGTRDLVFKSAKSCSCYHCKKD-
Glycoprotein hormone-beta5	1159	69290-69192, 64784-64542	…MYAYKVTKTDSAGRVCWDVINVMSCWGRCDSNEISDWRFPYKRSFHPVCLHDTRAVSSATLQNCEE GVEPGTEVYEYLEALTCRCMVCKSSEASCEGLRYRGQRSGPFLVGGR-
Insulin like peptide (ILP)	420	83020-82828, 82510-82305	MWRLYLRLVAIAALCLCTLAQAQSDLFQLGDKRNTNKYCGRNLANMLRYVCNGNYYPMFKKASQDVE DVNDSGIWIQPLPIEEPQLQYPFHSRSNAATLVPGSLRRHTRGVYDECCRKSCTIQEMVSYCGSR-
Insulin like peptide (ILP)	420	97394-97163, 96172-95982	MWRACFRIVVVVALCLCSLAQSQSDIFPFPDKRPETKRYCGSNLVDILQLVCNGKYYSNINNSNNYS PHVGRKKSMPEADEDFWQQLQPVEEQMKFPFRSRSSVSTFAHRIFKRHTVGVAYECCINKGCTVYEL RSYCAP-
Ion Transport peptide	1070	534235-533997, 531875-531755, 531121-531120	MQHPHLTRILACSLLVSMIITSLLTSRTSGLAVGHSLHKRSFFEIQCKGVYDKSIFARLDRICEDCY NLFREPQLHSLCRKNCFTTDYFKGCLDVLLLQDEMEKYQTWIKQLHGAEPGV-
Myosuppressin	87	813942-813843, 812573-812482	MKHVCVVLICFLAALLAFSPLRVSAVPPPQCSPIARDSSNPLVDAGVKRQDVDHVFLRFGRRR-
Orcokinin	656	21204-21041, 18481-18262	…HSLQRSIVRFTSNRTVTVAGREVDGLAPFPRKTRSGLDSLSGVTFGWNKRLDSLRGITFGNQKRNF DEIDRSGFNSFVKKNFDEIDRSGFDGFVKKNFDEIDRVGFGSFVKRNAPFLLARSYEKENH-
Pigment Dispersing Factor (PDF)	204	1044447-1044548, 1045679-1045840	MKQLGAVILFFYLLTTEFTSAAIQLEDNRYLDKEFQTNAVNVRELATWIMQLLLHKGQQTICTHKRN SELINTLLGLPKILNEAGRK-
SIFamide	83	206943-206814, 204360-204176	MQNRVVATCVLLLAVLLLAEFATAAFRKPPFNGSIFGKRGSPTDVGGLRLPQPRYHLEFSKQNENGN VQLFPVVPTDSISGPTDVFRNVTTDGSFVDLNYPTPQ-
Sulfakinin	1210	1425995-1425630	MVATLILTLGVYLVLQYQHHAAVDAAPSSSDVVAAGGSNLEGPGQRGRSRSFLQTPPRSPQYMRARL VPVEPAADILNDFIIDDESMDFNKRQSDDYGHMRFGKREQFDDYGHMRFGRSLD-
Tachykinin	324	256762-257696	…RVRCRACAVLVVTLSLVAVVLCAPEESPKRAPSGFLGVRGKKDSAFVSEEAYNDVMEKRAPAMGFQ GVRGKKDDKRGPSMGFHGMRGKKDADSRAEFLQELLQDKRAPSMGFMGMRGKKEALDFDYFDKRAPS LGFQGMRGRRDGEYLSANRLGLIGVRVENGVNLEGDDYAEMSSDEDLEAGLQDAEEFSKRAPAANGF FGTRGKKVPANGFFGTRGKKGPSAGFFAMRGKKAPSAGFMEYQGPPVDLDTLLNYLGTAYQHGRDKR NGGRLPGSKKAPIGFLGTRGKKDWPTQQGNRRSRPLSGSGADP-
Vasotocin-neurophysin VT	412	22921-23055, 23756-23969, 24910-15055	MKMQLGTATLLAVFISLCTACLITNCPKGGKRAGTHSQELHTIRQCARCGPAKLGHCYGPAICCGPQ IGCLVATPDTARCLSEAASPVPCTAPTGAQCGEGKFAGRCTANGVCCTHESCHIDITCQLTTSDAPE LIDVSADQTNPLYSLYSSYQQENPGLGLSE-

Supplementary Table 29. Biogenic Amines Receptors in the *Z. nevadensis* genome

Dmel ortholog	**Dmel gene ID**	**Znev gene**	**Scaf-fold**	**Coding range (nt)**	**Apollo gene model**	**BGI gene model**	**Note**	**BLASTp Dmel E-value**	**Note**
DmDOP1	CG9652	ZnevDop1	363	137878-143350	Znev:06465845	Znev_10284	BGI model miss N-term	0.0	
DmDOPR2 Var A	CG18741	ZnevDop2	27	552846-639929	Znev:06465885	Znev_17077		0.0	
DmDDR2	CG33517	ZnevDOP3	1345	52101-33595	Znev:06465896	Znev_15682	C-term	0.0 (for whole protein)	Split across 2 scaffolds
			909	91306-91770	Znev:06466218		N-term		
DmDopEcR	CG18314	ZnevDopEcR	2025	11203-13064	Znev:06465919	Znev_16402		2.10^{-141}	Some amino acids missing around nt 12000 due to poor sequence quality
DmTYR1	CG7485	ZnevTyr1A	449	28767-30209	Znev:06465932	Znev_15639	No introns	6.10^{-157}	
DmTYR1	CG7485	ZnevTyr1B	449	281805-283247	Znev:06465941	Znev_15640	No introns	8.10^{-171}	
DmTYR2	CG7431	ZnevTyr2	263	906250-909735	Znev:06466229	Znev_01738		3.10^{-93}	
DmOAMB	CG3856	ZnevOctA1	113	114209-163721	Znev:06465950	Znev_10802		1.10^{-99}	
DmOCTB1R/ DmOA2	CG6919	ZnevOctB1R	311	203891-264810	Znev:06465991	Znev_13827	Splice variants	3.10^{-160}	
DmOCTB2R	CG33976	ZnevOctB2R	542	1172853-1174232	Znev:06465967	Znev_10946	No introns	0.0	
DmOCTB3R	CG42244	ZnevOctB3R	311	98777-155203	Znev:06465976	Znev_13825	Splice variants	3.10^{-145}	
Dm5HTR1A	CG16720	Znev5HT1A	461	353763-447380	Znev:06466039	Znev_09818		6.10^{-111}	

Dm5HTR1B	CG15113	Znev5HT1B	722	1107533-1133185	Znev:06466024	Znev_12811		6.10^{-71}	Some amino acids may be missing
Dm5HT7	CG12073	Znev5HT7	66	2803168-2844088	Znev:06466004	Znev_00165		0.0	
Dm5HT2	CG1056	Znev5HT2A	356	258154-220687	Znev:06466071	Znev_14559		4.10^{-117}	
DmOrphan	CG42796	Znev5HT2B	356	170064-139614	Znev:06466054	Znev_14558	Predicted 5HT receptor	6.10^{-114}	
DmOrphan	CG18208	Znev18208	144	309867-308728	Znev:06466101	Znev_17804	N-term	0.0 (for whole protein)	Split across 2 scaffolds
			702	1531623-1532291	Znev:06466090	Znev_13273	C-term		
DmOrphan	CG7918	Znev7918	267	270975-274214	Znev:06466110	Znev_11238	predicted acetylcholine receptor	7.10^{-115}	
DmMAcR-60C	CG4356	ZnevmAcR1	385	409081-417033	Znev:06466201	Znev_07856		3.10^{-165}	
DmOrphan	CG13579	Znev13579	672	146341-97055	Znev:06466163	Znev_12614	BGI model miss N-term	6.10^{-54}	
DmOrphan	CG12796	Znev12796	709	475408-477959	Znev:06466132	Znev_17808		3.10^{-88}	
DmAdoR	CG9753	ZnevAdoR1	1163	414004-385779	Znev:06466186	Znev_09501	BGI model split	6.10^{-147}	
						Znev_09502	BGI model miss C-term		
DmsNPFR	CG7395	ZnevsNPFR	144	1009067-1018584	Znev:06466015	Znev_08555	no introns	5.10^{-130}	
?	CG30106	ZnevPepOrph1	709	823700-825232	Znev:06466154	None	no introns	2.10^{-43}	
?	CG30106	ZnevPepOrph2	709	832920-834452	Znev:06466145	None	no introns	2.10^{-43}	

Supplementary Table 30. Best matching Swissprot proteins for histone deacetylases

Z. nevadensis gene ID	Human gene name	*Drosophila* gene name
Znev_03795	HDAC1	Rpd3
Znev_18002	HDAC3	HDAC3
Znev_00349	HDAC4	HDAC4
Znev_02211	HDAC6	HDAC6
Znev_12928	HDAC8	--
Znev_10901	HDAC11	--
Znev_11203	SIRT1	--
Znev_11971	SIRT2	Sir2
Znev_01239	SIRT3	--
Znev_10250	SIRT4	Sirt4
Znev_14842	SIRT5	--
Znev_09433	SIRT6	Sirt6
Znev_03848	SIRT7	Sirt7

Supplementary Table 31. Best matching Swissprot proteins for histone acetyltransferases

Z. nevadensis gene ID	Human gene name	*Drosophila* gene name
Znev_12488	HAT1	RE20268
Znev_04968	KAT2	pcaf
Znev_04899	KAT6	MIP13243
Znev_14581	KAT7	Chameau
Znev_10779	NCOAT	O-GlcNAcase
Znev_00128	KAT5	Tip60
Znev_09388	KAT8	MOF
Znev_08401	CBP	nejire
Znev_05083	NCOA1	--
Znev_03596	ELP3	ELP3
Znev_04119	NAA60	NAA60
Znev_04098	CDY	GH11143

Supplementary Table 32. Best matching Swissprot proteins for histone demethylases and methyltransferases

Z. nevadensis gene ID	Human gene name	*Drosophila* gene name/ID
Znev_07618	KDM4C	Histone demethylase 4
Znev_17776	KDM5A	ref|NP_523486.1|
Znev_00365	KDM6A	ref|NP_001188773.1|
Znev_15990	KDM7	--
Znev_02995	DOT1L	grappa
Znev_08209	DPY 30	--
Znev_04615	SMAD 5	mad
Znev_05631	EHMT	--
Znev_13066	CtBP	CtBP
Znev_08621	JARDI1	jumonji
Znev_12225	RBBP5	RBBP5
Znev_09841	RBL1	--
Znev_08510	ASH2	ASH2
Znev_03531	GFI1B	sens 2

Supplementary Table 33. P450s sorted by clans.

P450 Clan	Gene Name	Fragment	Putative subfamily according to BLAST	OrthoDB Classification
CYP2	Znev_00012		15A1	
	Znev_00957		306A1	
	Znev_00958		18A1	18A1
	Znev_04417		307A1	
	Znev_06057		303A1	
	Znev_14286		15A1	
	Znev_14287		15A1	
	Znev_14299	Fragment	15A1	
	Znev_14300		15A1	
	Znev_14301/Znev_14302	Joined fragments	15A1	
	Znev_15869		304	
	Znev_15870	Fragment	304A1	
CYP3	Znev_01139	Fragment	6K1	
	Znev_01838			
	Znev_01867		6B	
	Znev_01868			
	Znev_09132		49A1	
	Znev_04985		9AG4	
	Znev_05339			
	Znev_05340		6A13	
	Znev_06541		3A	
	Znev_06543			
	Znev_08570		6B	
	Znev_13255			
	Znev_14063		6B	
	Znev_14677			
	Znev_14802		9E2	
	Znev_14833			
	Znev_15638			

	Znev_16120	Fragment		
	Znev_16125	Fragment	6K1	
	Znev_16153	Fragment	6BB1V2	
	Znev_16218	Fragment		
	Znev_16398	Fragment		
	Znev_16438			
	Znev_16771		6K1	
	Znev_18486		6K1	
	Znev_18620		6K1	
	Znev_18647		6BQ	
CYP4	Znev_02456			4AA1
	Znev_03004	Fragment		
	Znev_03222			4
	Znev_05390			4
	Znev_05391		4C3	4
	Znev_05398		4G	4
	Znev_06128		4C1	
	Znev_08930			4
	Znev_09012			
	Znev_09478			4
	Znev_09480			4
	Znev_09481			4
	Znev_11665		4G15	
	Znev_13889			4
	Znev_13890/Znev_13891	Joined fragments		
	Znev_13892/Znev_13893	Identical fragments, EST support for first		
	Znev_14143	Fragment	4C1	
	Znev_14590			4
	Znev_14632	Fragment		
	Znev_16223			4
	Znev_17183			
Mitochon	Znev_02808		314A1	

drial	Znev_04827			314
	Znev_07037		353A1	
	Znev_08701		302A1	
	Znev_04232		49	
	Znev_09277		301A1	
	Znev_12901			
	Znev_14659		315A1	
	Znev_16439			
Not classified	Znev_01870	Fragment		
	Znev_06629	Fragment		
	Znev_12912	Fragment		
	Znev_14631	Fragment		

Supplementary Table 34. Sample sizes, means, and confidence intervals of CpG o/e by genomic element

Level	Number	Mean	Lower 95% CI	Upper 95% CI
Gene body	17680	0.6163	0.6120	0.6207
Exon	90183	0.5923	0.5902	0.5944
Intron	73572	0.5016	0.4992	0.5041
Promoter	17183	0.7150	0.7108	0.7193
Genome (1kb window)	481445	0.7943	0.7934	0.7952

Supplementary Table 35. Low CpG o/e (putatively methylated) gene ontology enrichment

Term	Category[a]	Accession	Number of genes	Fold enrichment[b]	FDR	P-Value
Adenyl ribonucleotide binding	F	GO:0032559	369	1.98	4.80E-11	4.67E-13
Meiosis I	P	GO:0007127	34	-	1.12E-06	1.97E-08
ATPase activity, coupled	F	GO:0042623	105	3.12	1.53E-06	2.91E-08
Organelle inner membrane	C	GO:0019866	91	3.27	5.45E-06	1.12E-07
Thiolester hydrolase activity	F	GO:0016790	39	13.32	9.97E-06	2.10E-07
Mitosis	P	GO:0007067	59	5.04	1.14E-05	2.48E-07
Golgi vesicle transport	P	GO:0048193	48	6.56	2.14E-05	4.86E-07
Histone modification	P	GO:0016570	78	3.33	3.06E-05	7.17E-07
Protein binding	F	GO:0005515	608	1.37	3.82E-05	9.12E-07
Helicase activity	F	GO:0004386	46	5.24	1.62E-04	4.18E-06
RNA splicing, via transesterification reactions with bulged adenosine as nucleophile	P	GO:0000377	107	2.36	1.73E-04	4.62E-06
Transcription factor complex	C	GO:0005667	36	8.20	1.77E-04	4.81E-06
Regulation of mitotic cell cycle	P	GO:0007346	53	4.02	2.71E-04	7.74E-06
Spliceosomal complex	C	GO:0005681	85	2.64	2.77E-04	7.96E-06
Chromosome, centromeric region	C	GO:0000775	38	6.49	2.96E-04	8.72E-06
Exonuclease activity	F	GO:0004527	27	18.44	3.09E-04	9.17E-06
Translation factor activity, nucleic acid binding	F	GO:0008135	65	2.96	7.90E-04	2.44E-05
Protein targeting	P	GO:0006605	62	3.02	9.28E-04	2.95E-05
Small conjugating protein ligase activity	F	GO:0019787	61	2.98	0.0012	4.09E-05
DNA-dependent DNA replication	P	GO:0006261	27	9.22	0.0016	5.71E-05
Protein amino acid methylation	P	GO:0006479	33	5.63	0.0019	7.20E-05
Nucleotidyltransferase activity	F	GO:0016779	58	2.64	0.0056	2.20E-04
DNA damage checkpoint	P	GO:0000077	33	4.51	0.0058	2.34E-04
Motor activity	F	GO:0003774	41	3.50	0.0058	2.35E-04
Regulation of synaptic growth at neuromuscular junction	P	GO:0008582	16	-	0.0058	2.39E-04
Centrosome	C	GO:0005813	20	13.66	0.0064	2.71E-04
Coenzyme metabolic process	P	GO:0006732	43	3.26	0.0066	2.81E-04

Structure-specific DNA binding	F	GO:0043566	32	4.37	0.0079	3.45E-04
Chromatin remodeling complex	C	GO:0016585	29	4.95	0.0084	3.70E-04
Phospholipid metabolic process	P	GO:0006644	37	3.61	0.0087	3.91E-04
Oxidoreductase activity, acting on the CH-CH group of donors	F	GO:0016627	37	3.61	0.0087	3.91E-04
Endoplasmic reticulum part	C	GO:0044432	37	3.61	0.0087	3.91E-04
Telomere maintenance	P	GO:0000723	15	-	0.0088	4.03E-04
Nucleotide-excision repair	P	GO:0006289	15	-	0.0088	4.03E-04
Protein localization in organelle	P	GO:0033365	49	2.79	0.0090	4.19E-04
Protein-lysine N-methyltransferase activity	F	GO:0016279	19	12.98	0.0093	4.36E-04
Mitochondrial membrane	C	GO:0031966	53	2.59	0.0104	5.03E-04
ATP biosynthetic process	P	GO:0006754	14	-	0.0131	6.79E-04
Microtubule associated complex	C	GO:0005875	30	4.10	0.0142	7.43E-04
Nuclear envelope	C	GO:0005635	35	3.42	0.0150	7.97E-04
mRNA processing	P	GO:0006397	58	2.33	0.0151	8.04E-04
Phosphate metabolic process	P	GO:0006796	292	1.37	0.0151	8.10E-04
Purine ribonucleoside triphosphate catabolic process	P	GO:0009207	136	1.63	0.0154	8.47E-04
Regulation of DNA metabolic process	P	GO:0051052	24	5.46	0.0157	8.64E-04
Protein ubiquitination	P	GO:0016567	62	2.23	0.0157	8.68E-04

[a] B: biological process, C: cellular component, F: molecular function

[b] '-' indicates inability to calculate fold enrichment due to the presence of 0 genes with the given annotation in the reference set

Supplementary Table 36. High CpG o/e (putatively unmethylated) gene ontology enrichment

Term	Category[a]	Accession	Number of genes	Fold enrichment[b]	FDR	P-Value
DNA metabolic process	P	GO:0006259	317	2.14	1.15E-16	1.34E-19
G-protein coupled amine receptor activity	F	GO:0008227	15	-	2.83E-04	1.31E-06
DNA binding	F	GO:0003677	144	1.69	0.0013	8.99E-06
Structural constituent of cuticle	F	GO:0042302	12	-	0.0026	1.97E-05
Central nervous system development	P	GO:0007417	81	2.01	0.0030	2.40E-05
Polysaccharide binding	F	GO:0030247	27	4.39	0.0033	2.83E-05
Iron ion binding	F	GO:0005506	47	2.55	0.0056	5.36E-05
Regulation of cell morphogenesis involved in differentiation	P	GO:0010769	25	4.07	0.0090	9.83E-05
Tissue homeostasis	P	GO:0001894	26	3.81	0.0100	1.17E-04
Neuropeptide receptor activity	F	GO:0008188	14	10.25	0.0114	1.51E-04
Gonad development	P	GO:0008406	24	3.90	0.0134	1.81E-04
Camera-type eye development	P	GO:0043010	9	-	0.0200	2.97E-04
Proximal/distal pattern formation, imaginal disc	P	GO:0007449	9	-	0.0200	2.97E-04
Integral to plasma membrane	C	GO:0005887	29	3.03	0.0211	3.28E-04
Cullin-RING ubiquitin Ligase complex	C	GO:0031461	30	2.93	0.0221	3.53E-04
Regulation of neuron projection development	P	GO:0010975	12	8.78	0.0401	7.06E-04
Voltage-gated potassium channel activity	F	GO:0005249	8	-	0.0403	7.32E-04
Hormone metabolic process	P	GO:0042445	10	14.64	0.0434	8.36E-04
Cell fate specification	P	GO:0001708	27	2.82	0.0471	9.27E-04

[a] B: biological process, C: cellular component, F: molecular function

[b] '-' indicates inability to calculate fold enrichment due to the presence of 0 genes with the given annotation in the reference set

Supplementary Table 37. Total number of different alternative splicing events

Sample	A3SS	A5SS	AFE	ALE	MXE	IR	ES	sum
E	6415	6192	35	30	1	10714	3373	26760
W2	4846	5177	36	17	1	4046	1974	16097
Ny2	3464	3563	37	24	1	6397	1460	14946
MS2	4664	4912	50	22	1	4570	2171	16390
FS3	4462	4624	42	23	1	4062	2097	15311
MA	3015	3168	45	29	1	5321	1207	12786
MPI	7162	7966	43	21	1	6815	3255	25263
MNR2	5686	6696	45	15	1	4010	2306	18759
MNR3	4460	5124	43	26	2	5957	1957	17569
FA	5954	6216	28	23	7	4177	2290	18695
FPI	6984	7728	32	21	2	3053	2605	20425
FNI2	5094	5867	33	16	1	2859	1742	15612
FNR2	5656	6075	33	27	1	3174	2050	17016

Supplementary Table 38. Number of genes involved in alternative splicing events

Sample	A3SS	A5SS	AFE	ALE	MXE	IR	ES	sum
E	3915	3805	35	30	1	4719	2358	14863
W2	3247	3296	36	17	1	2248	1498	10343
Ny2	2508	2503	37	24	1	3181	1177	9431
MS2	3135	3203	50	22	1	2420	1635	10466
FS3	3086	2999	42	23	1	2144	1572	9867
MA	2249	2198	45	29	1	2685	976	8183
MPI	4246	4407	43	21	1	3420	2257	14395
MNR2	3566	3911	45	15	1	2225	1732	11495
MNR3	3089	3255	43	26	2	3035	1529	10979
FA	3769	3787	28	23	2	2328	1722	11659
FPI	4209	4237	32	21	2	1791	1895	12187
FNI2	3363	3564	33	16	1	1658	1371	10006
FNR2	3664	3662	33	27	1	1773	1556	10716

Supplementary Notes

1 Strain selection

Colony 133 of *Z. nevadensis nuttingi* was collected within a wood log in Pebble Beach near Monterey, California. After transfer of the log into the laboratory, the colony was extracted and, due to its large size, distributed between two nests consisting of several partly pre-cavitated 5mm thick layers of presoaked sheets of spruce pine (*Pinus glabra*) that were bolted together. This artificial nest structure allowed ready access to the termites by disassembling and subsequent reassembling of the wood layers. Nests were kept moist by periodic spraying with distilled water, and were maintained in transparent plastic boxes under a 12L:12D light cycle at 20.5°C. GC-MS was used to confirm species identity by comparison of the cuticular hydrocarbon profile with published profiles of this population [1,2].

Ten *Zootermopsis nevadensis nuttingi* samples from each part of the divided colony were screened at five microsatellite loci; three (Zoot-117, Zoot-73, Zoot-254) previously described by Aldrich and Kambhampati [3] and two (Za-18, Za-123) described by Booth, et al. [4]. PCR conditions followed those described by Aldrich and Kambhampati [3] and Booth, et al. [4]. Following PCR, 5 µl of stop solution (95% Formamide, 20 mM EDTA, Bromophenol blue) was added to each 12 µl reaction. Reactions were subsequently denatured at 90°C for 4 min. A 1 µl aliquot was loaded onto 25 cm 6% 1X TBE polyacrylamide gels, run on a Li-Cor 4300 automated DNA sequencer. Loci were sized using a 50–350 bp standard (Li-Cor Biosciences, Inc.). Five control samples (i.e., samples of known genotype collected from the Pebble Beach natural population) were included in each run to ensure accuracy and consistency of genotyping with previous studies. The GeneProfiler (v4.05) software (Scanalytics, Rockville, MD) was used to collect genotypic data from the LI-COR system.

All loci yielded unambiguous PCR products. Samples from both parts of colony 133 exhibited complete homozygosity at the screened loci (Supplementary Table 1). In comparison, a previous study of the number of alleles observed from the Pebble Beach *Z. nevadensis nuttingi* population by [4] found colonies exhibited between one and three alleles per colony at the loci studied here (Supplementary Table 1). Within the natural population, the expected heterozygosity ranged from 0 to 0.551, with observed heterozygosity ranging from 0 to 0.400. At locus Zoot-117, colony samples exhibited allele 204, which is characteristic of the Pebble Beach population of *Z. nevadensis*. At the same locus in *Z. angusticollis* alleles 173 and 177 are observed (unpubl. data). This confirmed that our study colony was *Z. nevadensis nuttingi* as previously suggested from location and cuticular hydrocarbon profiles (see above). Furthermore, homozygosity indicated that the two samples of the study colony were inbred and represented a single family.

2 Transcriptomics analysis

2.1 Samples

Samples for the transcriptomes were collected from various colonies. Species identity of each colony has been confirmed by cuticular hydrocarbon profile analysis using GC/MS. The number of individuals used per sample depended on availability and previous RNA yield. We tried to get male and female samples of a specific life stage from the same colony, but this was not always possible due to occasional RNA degradation during shipment.

RNA samples were extracted from specific castes and developmental stages as indicated in Supplementary Table 4 and Figure 1 of the main manuscript. Samples of eggs, workers and nymphs are mixed sexes, and all other life stages are separated into male and female. Note that in termites, reproductives (primary or neotenic) can be reproductively active (ovaries produce eggs, testis produce sperm) or inactive (no signs of oocytes in the ovaries, small testes). Since primary reproductives are relatively rare in collected colonies, both the primary reproductive female and primary reproductive male sequenced were inactive. We initially used twelve samples for protein-coding prediction (see Methods in the main manuscript). This initial set contained only one sample for soldiers, with mixed sex (as for workers and nymphs) and a different sample for the alate male: the mixed soldier sample was replaced by a male and a female soldier (MS2 and FS3) while the alate male sample was resequenced due to its low coverage (mapped reads). As a result, 13 transcriptomes were used for the analysis of alternative splicing associated with DNA methylation (see Supplementary Note 11.3). Finally, for the analysis of differential expression, we extended our data to 25 samples by creating replicates for the most interesting samples regarding eusociality: soldiers and reproductives. Additional RNA sequencing was realized on antenna samples for the annotation of chemoreceptors (see Supplementary Notes 7).

2.2 Expression levels

After sequencing, we used TopHat v1.3.3 [5] (based on Bowtie v0.12.7 [6]) to align raw reads against the genome (default parameters except "-r 20 --mate-std-dev 10 -I 10000" to adapt to an insect genome). The statistics of raw data and alignment is listed in Supplementary Table 5.
Raw reads were normalized among all samples (library-size normalization) according to the trimmed mean of M-values method [7] using the edgeR package [8]. Expression values for each sample were calculated using RPKM calculations [9]. We filtered out genes with very low expression levels (RPKM values < 5 in all samples). The resultant 12,972 expressed genes were processed by (i) clustering of gene expression patterns (Supplementary Notes 2.3), (ii) identification of samples with, similar gene expression (Supplementary Notes 2.4) (iii) learning significantly over-/under-expressed genes (Supplementary Notes 2.5), and (iv) characterization of differentially expressed gene families (Supplementary Notes 2.6).

2.3 Clustering

We clustered genes having similar expression patterns using the K-means algorithm as implemented in Cluster3.0 [10]. Two of the most commonly used similarity measures for gene expression clustering are the Euclidean distance and the Pearson correlation [11]. Although the Euclidean distance is known to be

sensitive to scaling and differences in average expression levels, it is possible to circumvent this limitation by normalizing the expression levels of each gene to percentages (each gene expression summing over all 13 samples to 100%). Percentages also allow the conservation of fold-change information. Moreover, we observed more stable local optima with the Euclidean distance than with the Pearson correlation on multiple runs. Ten thousand K-means runs allowed us to find the local optima with the highest likelihood ("-r" option) using the Euclidean distance ("-g 7" option) on the normalized data. We generated a clustering into 50 classes ("-k 50" option) of the expressed genes, and refined the visualisation of these clusters by reordering using a hierarchical clustering of the averaged expression levels of the 50 clusters (mean vectors). This hierarchical clustering was performed with Cluster3.0 using Pearson ranked correlation and pairwise average-linkage ("-e 2 -m a" parameters). It offers a clear visualisation of the samples with similar gene expression (Figure 2 of the main manuscript heatmaps generated using R), further assessed in the next section.

2.4 Types of samples with similar expression patterns

Observation of the 50-class clustering (Figure 2 of the main manuscript) reveals that not only samples with replicates but also different samples share similar expression patterns:

- *Workers* and *Nymphs*, hereafter referred to as *Juveniles* (for example over-expression in cluster 10 or under-expression in clusters 32 to 35);
- *Soldier males and females* (for example over-expression in clusters 14 to 16);
- *Male reproductives* (primary and neotenic, active and inactive, but not the alate) with over-expression in clusters 30 to 35;
- *Female reproductives* (primary and neotenic, active and inactive, but not the alate) with over-expression in clusters 38 to 45.

These observations were confirmed through hierarchical clustering of all the expressed genes across samples with Cluster3.0 using Pearson ranked correlation and pairwise average-linkage ("-e 2 -m a" parameters). The resulting tree topology, represented at the top of Figure 2 of the main manuscript, clearly support the aforementioned groups.

This is especially interesting since these groupings correspond to biological and developmental classes. Another interesting observation is that while reproductives primarily have sex-specific patterns of gene expression, soldiers exhibit a caste-specific expression pattern but no sex-specific pattern.

As a consequence, instead of analysing separately these samples having similar expression pattern, we grouped them into a *type of samples* and consider all members as *replicates* in the following. This notable allowed us to benefit from a larger number of replicates and consider samples without replicates, like the primary reproductives.

Finally, the clustering results indicate that eggs and the two alates have unique expression patterns. For example, while the eggs show specific over-expressed genes, the female alate seem to share over-expressed genes either with the male alate or with the female reproductives. As a consequence, they cannot be classified with other samples, especially in the absence of replicates. These samples have been considered as *unique samples* in the following and required a specific processing given the absence of replicates. We refrained as much as possible from drawing any conclusions about their differentially expressed genes but rather used this information to further confirm the differentially expressed genes in aforementioned *type of samples*.

2.5 Genes showing significant over-/under-expression

The identification of differentially expressed genes was realized with the following protocol:

1. Using the edgeR package [8], we ran pairwise comparisons for the four *types of samples* with replicates: *juveniles* (worker and nymphs*)*, *soldiers* (male and females), *male reproductives males* and *female reproductives* (primary, neotenic). We required a FDR < 0.05 (Benjamini-Hochberg correction for p-values in multiple testing) for statistical significance. We obtained a preliminary list of significantly over-/under-expressed genes in the four *types of samples*;
2. Since we cannot reasonably claim a gene is differentially expressed in a *type of samples* if it is even more differentially expressed in one of the three *unique samples* without replicates (*egg*, *male alate* and *female alate*), we identified differentially expressed genes in these *unique samples*. However, given the absence of replicates we applied a more stringent procedure to identify differentially expressed genes and we refrained to draw major conclusions from expression levels in the *unique samples*. We thus ran pairwise comparisons of each *unique sample* against all four *types of samples*. We also reinforced the detection approach of over-/under-expressed genes to ensure a maximal reliability in such prediction. We applied a more conservative approach to detect differential expression using the DESeq package [12], and then required a more stringent threshold for significance (FDR<0.01).
3. We finally deduced the lists of significant differential expression in the four types of samples by excluding from the preliminary lists (see 1. above), the differentially expressed genes in the three *unique samples* (see 2. above).

Final counts of differentially expressed genes are similar to what was observed in the clustering, such as the largest number of differential expression for reproductive females (Supplementary Table 6).

2.6 Over-represented gene families in differentially expressed genes

We identified gene families for which the majority of members are over- or under-expressed in a *type of samples*, or in a *unique sample,* or globally (i.e. cumulating all). These genes would be indicative of termite development or of caste differentiation. For the alternative/null hypothesis, we compared the proportion of differentially expressed genes in the family of interest to this proportion among all expressed proteins (at least one sample with RPKM>5; see Supplementary Notes 2.2). For each *type of samples* or *unique sample S* and for each protein family *F* (defined as all proteins with an identical domain architecture; see Supplementary Notes 3.2), a 2*2 contingency table was built. All proteins *p*, with a measurable expression level, were split into this table according to two criteria:

1. Does *p* belong to family *F*?
2. Does *p* belong to the list of differentially expressed in *S*?

We performed one-tailed Fisher's exact tests on all the contingency tables *(F,S)* for both over- and under- expression. We report in Supplementary Table 7 all associations (*F,S)* that achieve p-value < 1e-04, group them by predicted functional similarity, and we provide additional details about differential expressed genes in the other samples than the significant ones.

2.7 Differential expression of orphan genes

Among the 12,972 expressed genes we identified a set of 2,184 proteins without any significant similarity to any protein in the NCBI non-redundant database (E-value $<10^{-3}$). This set was used to test for the over-representation of *orphan genes* among differentially expressed genes, as recently

suggested [13]. Fisher's exact tests reveal that 202 out of the 637 genes over-expressed in male reproductives are orphan genes (p-value<10^{-20}), as well as 74 out of the 296 genes over-expressed in egg (p-value<2^{-4}). Moreover, despite the absence of significance, we notably observed that 327 (and 441) orphan genes are over- (and under-) expressed in female reproductive, and a total of 1051 orphan genes classified as differentially expressed in at least one of the three *unique sample*s or the four *types of samples*, accounting for half of the orphan genes in *Z. nevadensis*. Hence, many gene emergences in *Z. nevadensis* appear to be driven by lineage-specific adaptations, involving caste- and sex-specific differentiation.

3 Data for evolutionary analysis

3.1 Selected reference genomes

For comparative analyses we used the following arthropod genomes: *D. melanogaster* [14], *T. castaneum* [15], *N. vitripennis* [16], *A. mellifera* [17], *C. floridanus* [18], *H. saltator* [18], *A. pisum* [19], *P. humanus* [20], *D. pulex* [21]. *C. elegans* [22] was used as outgroup (see Supplementary Table 13). Domain [23,24], GO [25] and KEGG [26] annotation of proteins for reference species was conducted using the same approach outlined for *Z. nevadensis* (see Methods in the main manuscript).

3.2 Definition of gene families

Orthologs of all ORFs were used for clustering genes into families. Since protein domain are the evolutionary, structurally and functionally subunits of proteins [27], protein function is principally driven by the domain composition [28], a protein family can be defined as the whole set of proteins with identical domain composition. Accordingly, proteins were considered as sets of domains, without order or repeats taken into account. Furthermore, in order to investigate at the sub-family level in protein families with a specific biological interest, we used a more fine-grained clustering through the OrthoMCL procedure (standard parameters applied on *Z. nevadensis* and the ten reference protein sets mentioned above). Termite proteins with orthology/paralogy relationships to other reference-arthropod proteins are illustrated with Venn diagrams in Supplementary Figure 3.

3.3 Phylogeny analysis for gene families of interest

Several gene families were scrutinized due to their biological interest or specific history in the termite evolution, and relative to the selected reference species described in Supplementary Table 13. The gene family definition was either based on domain arrangement (as described above) or corresponded to a manually curated set of closer orthologs based on orthoMCL clustering. The protocol used to reconstruct the protein family evolution, unless alternative methodology is described, involved the following steps:

1. Protein sequences were aligned with MAFFT [29]
2. Gblocks [30] was used to constrain alignments to only the most conserved positions, if enough were available.
3. Tree topologies for the genes were computed using a maximum likelihood approach. The morePhyML script [31] based on PhyML [32] was applied with default parameters (LG+Γ4+I model [33])
4. Tree visualisation was realized thanks to the TreeDyn software [34].

3.4 Genome quality assessment

Sequencing of a deep-rooted species from a dense clade could limit detectable homologies. However, the *Z. nevadensis* genome proved to have high gene content coverage and gene model quality as revealed by several indicators (see Supplementary Table 14 for summary):

- **CEGMA:** The Core Eukaryotic Genes Mapping Approach (CEGMA) [35] is a classical approach to assess genome completeness and gene structure prediction. From 458 highly conserved genes in

all eukaryotes, it defines a subset of 248 genes for which a low number of in-paralogs is found in a wide range of species (see http://korflab.ucdavis.edu/datasets/genome_completeness/index.html#SCT1). Of these 258 core proteins, 243 (98%) are predicted with a complete gene model (CEGMA alignment length > 70%) in the termite *Z. nevadensis*, which is more than found in *N. vitripennis*, *A. mellifera* or *C. floridanus*. Incorporating partial matches increases the prediction rate to 99.6%.

- **KEGG annotation:** We further used the KO, PATHWAYS and BRITE annotations from the KEGG database produced by the SBH procedure (see Methods in the main manuscript). The termite protein set contains more annotations than other recently sequenced insect genomes such as the jewel wasp and the ants (see Supplementary Table 13). A similar trend is observed for the BBH annotation, which is more restrictive since it focuses only on true orthologs. There, *Z. nevadensis* has the second highest number of unique KEGG ortholog families, second only to *D. pulex* which has many paralogs.
- **OrthoMCL clustering:** Despite the long phylogenetic distance between *Z. nevadensis* and the other insects, the percentage of *Z. nevadensis* proteins clustered by OrthoMCL is relatively high (73%). For example, the Hymenoptera have comparable fraction of clustered genes although their taxon sampling is much denser in the dataset, with divergence times of less than 100 million years. Such dense-taxon sampling usually leads to higher percentage of clustering, especially of the most recent genes, while the termite, given its position, could exhibit a larger proportion of orphan genes. This finding supports the relatively slow evolution anticipated of termites.
- **Domain annotation:** We finally assessed genome quality by investigating protein domains. Models of protein domains (Hidden Markov Models or HMMs) are more sensitive and selective than simple string searches such as BLAST. Furthermore, domains are the units of evolution and domain arrangements show evolutionary dynamics which are largely independent from the proteins (and thus the genes) which carry them. Accordingly, we compared several domain-based indicators of genome completeness between *Z. nevadensis* and reference genomes. *Z. nevadensis* has the largest number of domain families of any insects measured to date, essentially equal to *D. pulex.* However, more domains occur in multi-domain proteins than in *D. pulex* while the ratio of multi- vs. mono-domain proteins is more like all other insects except *D. pulex*. This suggests that there are fewer false positive gene predictions in the *Z nevadensis* genome and genes are less fragmented than in *D. pulex*. These results further corroborate the claim that quality and completeness of the *Z nevadensis* genome ranks among the best of all insect genomes. Finally, to assess gene model quality, we investigated the degree of fragmentation of all domain occurrences. The HMMER3 software is designed to detect local similarity and hence allows a precise identification of domains, even if only fragments are present. Since domains are the basic units of protein function and evolution (see above), unusually short fragments should be rare and would suggest incompleteness of gene models. We first fused fragments of a model that belonged to the same domain family and were found adjacent within a protein. This was done to limit signals due to possible artefacts/imprecision in HMMER3 annotation. Next we computed the average length of occurrences for each domain family in the reference species and in *Z. nevadensis*. We then identified, in each species, fragmented domain occurrences, defined as domains shorter than half the average size of its family. For each species we list domain families that are the most frequently found as short fragments in each species. Again, the *Z. nevadensis* genome seems to be of high quality, only falling behind genomes with a history of many years of annotation and re-annotation.

4 New perspectives on insect-specific gene families

4.1 Osiris genes

Osiris genes form a large family of conserved, syntenic and insect-specific proteins. The Osiris family was initially described for *D. melanogaster* [36], in which 24 members are recorded among which 21 are clustered on a single strand (except one on the opposite strand) within the unique Triplo-lethal region (Tpl). This clustering suggests an evolution by tandem duplication that likely occurred in the ancestor of all insects. Indeed, no Osiris gene has been observed either in more deeply rooted arthropods or in any non-insect species so far.

The Osiris cluster has shown a strong microsynteny in Dipteran species [36,37]. The microsynteny initially observed in Diptera has been confirmed across the holometabolous and hemimetabolous insects with available genomes [38]. Shah and co-workers described 24 Osiris ortholog sub-families. The syntenic region contains the Osiris sub-families ordered as follows: 1-NFRP1-24-2-3-4-5-6-7-8-9-10-11-12-13-14-15-16-17-18-19-20. Note that the NFRP1 gene is not an Osiris gene but encodes a neuropeptide F receptor (containing Pfam domain 7tm_1 - PF00001) but seems embedded in the microsyntenic block. Shah and collaborators deduced from BLASTP searches against the NCBI-NR database that Osiris 5 is diptera-specific while Osiris 1, 5, 13, and 15 are only found in holometabolous insects. In the syntenic region, only Osiris 4 and 13 are poorly conserved (Pfam domain - cf. below - diverges beyond recognition). Such weak conservation is more pronounced for the unclustered members (Osiris 21-22-23), for which orthologs are then more difficult to identify.

Finally, the amino-acid sequences of Osiris proteins are also highly conserved. Osiris proteins are characterized by the presence of a specific Pfam domain of unknown function (DUF1676 - PF07898) and further shorter motifs have been described, notably a secretion signal peptide [38]. However, the Osiris gene function is still largely unknown. In their publication, Shah and collaborators mentioned a possible link between the function and the remarkable microsynteny, which would support a developmental role, based on two observations:

- Maintenance of tandem duplicated genes has been observed in developmental and regulatory proteins to promote their co-expression [39].
- The expression of Osiris genes in Flybase: 1) tissue expression data show that these proteins are expressed in various tissues of the ectoderm except for the nervous system (suggesting a specificity of the insect non-neuronal ectoderm); 2) Osiris genes are all differentially expressed in one or more of the following three stages of development: 12–18 hour-old embryos, second instar larvae, or pupae at 2–3 days post-white-prepupal stage.

In our reference species, we used orthoMCL clustering results and manual inspection (BLAST/orthoDB) to retrieve the members of the distinct 24 sub-families. All species have a fairly well-conserved Osiris family with 17 to 24, except in *P. humanus* with only 13 members. As expected, the crustacean *D. pulex* has no Osiris gene.

In the assembled genome of *Z. nevadensis*, we identified 18 Osiris genes from 17 of the orthologous sub-families. Only one termite ortholog is involved in an unclustered sub-family: Znev_03410 is a putative Osiris 22 and has a well-conserved domain. Concerning the synteny in *Z. nevadensis*, Osiris

genes have been found on four scaffolds:

- Scaffold 335 contains a putative ortholog of Osiris 1 that is located 55kb from the closest border of this scaffold. Interestingly this plays against the previous hypothesis concerning the holometabola-specific aspect of this sub-family;
- Scaffold 2576 is a short scaffold only carrying of Znev_18482, the Osiris 24 ortholog;
- Scaffold 237 contains eleven Osiris genes in sub-families 2 to 17. On the one hand, it includes the duplication of sub-family 11 (interesting *Z. nevadensis*-unique duplication since this is the only insect-wide sub-family mainly found on the opposite strand). On the other hand, it reveals the absence of five sub-families: Osiris 4 (described as Diptera-specific), Osiris 6 (also lost in the beetle and the body louse), Osiris 13 and 15 (previously described as holometabola-specific), and 16 (also lost in the body louse);
- Scaffold 554 contains the last three Osiris genes, 18 to 20 (Znev_12940, Znev_12940, and Znev_18391). The split of these three is frequently observed (e.g. *P. humanus* and *A. pisum*) and they are also transposed in *T. castaneum*;
- Scaffold 10 contains Znev_08485, a putative duplicate of the Osiris 19 sub-family.

Given such fragmentation, it is difficult to come to any conclusion regarding the full microsynteny and orientation of Osiris genes in *Z. nevadensis* but a conserved organization of the clustered genes without orientation change is still possible. We illustrate the Osiris cluster in five reference species (already represented by Shah and collaborators) and *Z. nevadensis* based on orthoMCL subfamily classification (Supplementary Figure 6). Nevertheless, the presence of subfamily members and the large syntenic block of eleven Osiris genes in *Z. nevadensis* reveal interesting features. First, the presence of syntenic termite orthologs to the Osiris 1 and 5 weaken their presumed holometabola-specificity. Another interesting observation is the position of the NFRP1 ortholog (Znev_09703 – differentially expressed in female reproductives unlike the Osiris genes), between Osiris 17 and the border of the scaffold 237. Usually this non-Osiris gene looks captured in the syntenic bloc between Osiris 1 and 24. This is not compatible with its current localization since Osiris 1, 24 and 17 are on distinct scaffolds, but it is interesting to observe that sub-families 24 and 17 have the highest similarity within the Osiris genes. OrthoDB even provides a unique group for these two sub-families while all others are not merged and even sometimes split. We also identified a unique copy of Osiris 7 in *Z. nevadensis* while most reference species (all but *P. humanus* and *C. floridanus*) have two copies. Nevertheless, the termite gene Znev_09689 contains two domain occurrences, as does the *C. floridanus'* gene, suggesting fused gene models that need to be split. Finally, we observe a clear differential expression of Osiris genes in termite eggs. Among the 15 proteins with a detected domain, as an indicator of functional conservation, we obtained expression values for twelve proteins, all being largely over-expressed in the egg sample. Hence, analysis of *Z. nevadensis* genome brings more insights into the evolutionary history of this family thanks to its novel phylogenetic position and supports the potential role of Osiris genes during development in a species other than a drosophilid.

4.2 Yellow genes

The yellow genes are characterized by the presence of the MRJP domain (PF03022), but their function is globally unknown. The name "yellow" originate from the yellow phenotype of adult cuticle and larval mouth parts of *D. melanogaster* [40]. However, despite clear alteration of phenotypes, with observable changes to behaviour and melanic pigmentation of the wings, abdomen and thorax, the mechanisms induced by yellow genes are not clear [41].

The name of the characteristic domain stands for Major Royal Jelly Protein, since the yellow-MRJPs homologs are found in high concentration in royal jelly, the substance fed to those larval bees that develop into queens [42]. In Hymenoptera intensive independent duplications seem to have occurred in the ancestral *yellow-e3* gene and gave rise to the major royal jelly like proteins [43]. MRJPs are involved in regulating reproductive division of labour in the honeybee but this might be a derived trait [44].

The evolutionary history of yellow genes also is an open question. Yellow homologs have been observed so far in bacteria, fungi, insects, and a few other eukaryotes. As a consequence of this varied occurrence, several hypotheses are currently debated, from massive loss in eukaryotic lineages to various scenarios of horizontal gene transfers.

In insect species, twelve sub-families can be distinguished, among which ten were probably present prior to the holometabolan specialization. Based on current sub-family classification, we conducted a phylogenetic analysis (Supplementary Figure 7) coupled with orthoMCL classification in our reference insects and the termite *Z. nevadensis* (no yellow genes in the crustacean *D. pulex*), the results of which are presented in Supplementary Table 15.

Z. nevadensis did not possess orthologs of the clade-specific yellow-x (*T. castaneum* and *D. melanogaster* only) and yellow-MRJP (independent duplication in the Hymenoptera *N. vitripennis*, *A. mellifera*, and the ant *Linepithema humile*) genes. As we did not find any MRJP related genes in *Z. nevadensis*, they cannot play a similar role in *Z. nevadensis* as in Hymenoptera. Moreover, *Z. nevadensis* is missing two of the ten sub-families present in the Endopterygota clade (no trace using TBLASTN): subfamily -y and -g. This leads to speculation about the emergence of these two sub-families subsequent to termite differentiation; alternatively, these orthologs might have been lost in *Z. nevadensis*.

Additionally, we detected an ortholog to the yellow-b subfamily in *Z. nevadensis*. The absence of this subfamily in both *P. humanus* and *A. pisum*, shows the benefits of the termite genome to bring a finer view of the gene content in the common ancestor of these species.

Finally, in previous studies, a microsynteny block containing from five subfamilies – yellow-d, -e, -g, -g2 and -h – has been described [41]. Our analysis extends this block to six subfamilies by adding one putative subfamily of conserved and syntenic orthologs in the genome of *Z. nevadensis, T. castaneum* and *H. saltator*. This suggests that such a large syntenic block was already present in the ancestor of the termite and reference species, despite the absence of orthologs in the majority of species (synteny that could not be confirmed earlier due to the fragmentation of the *A. pisum* genome sequence and multiple losses in the *P. humanus)*. These orthologous proteins have been annotated as member of the -k subfamily due to similarity levels and branching in the phylogeny with the -k gene of *D. melanogaster*. A possible explanation would be that the -k subfamily had relaxed evolutionary constraints that led to multiple losses (in several lineages) and strong divergence in *D. melanogaster*. This offers an alternative hypothesis to -k being a divergent duplication of the –x as advanced by Fergusson and co-workers based on their phylogeny [41].

5 Evolutionary analysis

5.1 Gene families under adaptive and purifying selection

For each orthoMCL clusters, protein sequences were aligned with MAFFT [29] ("L-ins-i" option). The multiple sequence alignments were translated into a distance matrix using the program Protdist within the PHYLIP package [45]. The Jones-Taylor-Thornton (JTT) [46] substitution matrix plus gamma distribution was used to calculate each of the protein sequence distance matrices. The distance matrices were then used to build neighbour joining [47] trees, using the PHYLIP program "Neighbor". NJ trees were only created for clusters which had more than two sequences, thus producing 14,628 trees. To determine which proteins in each tree were the result of duplication or speciation events, the program SoftParsMap4 [48] was used to reconcile each protein tree with the species tree for the eleven species used in the analysis. The species tree was built using the Interactive Tree of Life (iTOL) website (http://itol.embl.de/) [49]. To perform the reconciliation, short branch lengths were assigned a bootstrap value of 1 (length < 0.005) while all branches greater than the threshold value received a bootstrap value of 100. This process helped to minimize the number of duplication events across each lineage. The number of duplications that occurred along each branch of each reconciled tree were counted and mapped onto the species tree. Further, to investigate which lineages may have evolved rapidly, dN/dS ratios were calculated along each branch of each tree using the program PAML4.4 [50]. To allow for dN/dS ratio testing, nucleotide sequences were used and separated into codons using the program PAL2NAL [51]. For each of the trees, two models were tested, a free-ratios model which calculated ω for each branch within the tree, and a single-ratios model where ω was constant for each branch within the tree. A chi-squared analysis was then performed to determine if there was sufficient evidence to support a free-ratios model. Only families exhibiting support for the free-ratios model were included in the results for accelerated evolution. Branches with a dN/dS>1, indicating high rate variation, were then tabulated for each protein family and mapped onto the species tree. 470 proteins showed an accelerated evolution event after the termite speciation (Supplementary Figure 8). It should be noted that this test is aimed at generating the most accurate statistical value of dN/dS on each lineage, but is not a statistical test of significance for dN/dS>1 when that is observed. Most of the 470 proteins were in gene trees, where at least one lineage was saturated in dS. The termite lineage accelerated evolution that was detected generally fell into two categories. One category included lineage-specific duplicates on the termite lineage and these were studies in more detail (see below). The other category included gene trees with different topologies from the species tree. In these cases, there might be interesting convergent evolution at the molecular level, but the specific families had saturation on other lineages, which can lead to mis-specification of ancestral states by PAML and correspondingly to incorrect dN/dS ratios. Without drawing conclusions from this data, there may be interesting gene families to examine in more detail in this group.

Following the gene/species tree reconciliation, the number of duplications that occurred along each branch of each tree were tallied and mapped onto the corresponding branch in the species tree. The same process was performed for each potential accelerated evolution event from the PAML analysis. The tree in Supplementary Figure 8 shows the final number of duplications and selective events that occurred on each branch of the species tree. The branch that experienced the highest number of duplication events was the branch leading to Neoptera (12,806 duplications), while the branch showing the fewest number of duplications was the branch leading to Paraneoptera (9 duplications). The average number of duplications across each branch of the tree is of 3,643. For the number of selective events, the branch that experienced the greatest number of events was again the branch leading to Neoptera

(9,542 events), while the branch with the least number of events was the branch leading to *P. humanus* (82 events). The average number of selective events for each branch in the tree was 1,597. It should be noted that there is large disparity between the results on the branches leading to *P. humanus* and *A. pisum*, respectively. The large number of duplications and selective events for *A. pisum* stand in stark contrast to the relatively few duplications and selective events leading to *P. humanus*. This contrast was investigated and determined to be related to the formation of the protein families. Many of the families for which *A. pisum* proteins were assigned were composed predominately of *A. pisum* proteins, thus increasing the number of duplications through the gene/species tree reconciliation process for the branch leading to this species. It should be noted that the reconciliation was performed using neighbour joining gene trees without bootstrap support (due to the large number of trees calculated) and some of the signal in the analysis may be due to phylogenetic error.

The results from the phylogenetic analysis also reveal that the branch with the largest number of duplications and selection events was that leading to Neoptera. This could explain much of the diversity of *Z. nevadensis* as the first speciation event occurring after the branch leading to Neoptera corresponds to *Z. nevadensis* lineage. This might also explain many of the unique phenotypes that are seen between species within the subclass Neoptera compared to species outside of the subclass. Additionally this might demonstrate a signal for accelerated evolution that caused species within the subclass to develop novel protein functions.

It was observed that several selective events occurred along the *Z. nevadensis* lineage following a duplication event. To further explore if positive selection was acting on termite genes following gene duplications, an analysis was performed on gene families with all termite genes extracted that showed a signal of dN/dS>1 following a duplication event in the earlier analysis. Termite genes following the duplication were placed within their own families and a new multiple sequence alignment was performed using MAFFT. An associated codon alignment was generated using PAL2NAL. Phylogenetic trees were built for each of the new termite specific families using PhyML 3.0 [32]. Model testing was performed on each tree with Prottest 3 [52] to determine the optimal substitution matrix. Ratios for dN/dS were then calculated along each branch using the free-ratios model and the single-ratio model in PAML 4.4. A likelihood ratio test was then performed to determine if there was sufficient support for the free-ratios model over the single-ratio model as indicated previously. Measurements of dS were assessed for saturation and only dS values below 3.0 were maintained in the analysis for positive selection. Furthermore, all branches within the trees had estimated values of dS under 3.0 to reduce the possibility of anomalous dN/dS ratios due to saturation. For families with only two sequences pairwise dN/dS ratios were calculated using PAML 4.4. The pairwise analysis was also assessed for dS saturation with dS values being below 3.0. The results of the evolutionary analysis of the termite genes following a duplication event revealed 87 genes from 18 different orthoMCL clusters. The increase in the number of dN/dS>1 following a gene duplication event is possible due to relaxed selective constraints on the genes accompanied by a period of accelerated evolution (Supplementary Table 16).

Finally, we used the total number of proteins in *Z. nevadensis* (17,737) and the number of proteins that probably experienced an accelerated evolution in the *Z. nevadensis* lineage only (87) to perform Fisher's exact tests on each protein family defined by its domain content (see Supplementary Notes 3.2). We observe two gene families with a significant number of members with accelerated evolution in the termite lineage (Supplementary Table 17). Interestingly, these two families are also significantly differentially expressed in the termite male reproductives (see Supplementary Notes 6) and underwent

frequent duplication during *Z. nevadensis* evolution. BTB-BACK-Kelch proteins are found significantly expanded (see Supplementary Notes 5.5) while Pkinases (with monodomain architecture) are not. Nevertheless, they are still present in larger numbers in *Z. nevadensis* (141) compared to other insects (95-113), and particularly a sub-family of *Z. nevadensis*-specific paralogs (orthoMCL cluster OG2_00770) are exclusively showing differential expression and accelerated evolution. This family might be also involved in the co-expansion of genes linked to male-specific biology of reproduction in *Z. nevadensis*.

5.2 Gene emergence, conservation or HGT in *Z. nevadensis*

Domain-based analysis

At the domain-level, we observed 20 Pfam families unique to the termite *Z. nevadensis*, (i.e. that do not occur in *any* of the nine reference arthropod genomes described in Supplementary Notes 3.1 after using recommended thresholds) and 18 domains that are shared with at least one of the nine reference arthropods. These Pfam domains are listed in Supplementary Table 18.

Among these 38 domains, 23 have a significant hit against EST data from other arthropods. Most cases involve the more deeply rooted species, like the tick *Ixodes scapularis*, and hence suggest ancestral domains that were likely lost in Endopterygota. An intriguing example involves the Formiminotransferase-related domains (PF02971, PF04961 and PF07837), which are all part of the same protein in *Z. nevadensis* and have only one ortholog in *D. pulex*. This protein is linked to the folate pool, while a folate receptor domain (PF03024) is found as specific to the termite, the body louse and *I. Scapularis*. The taxonomic distribution also suggests the more recent appearance of some domains. For example, we confirm the presence in *Z. nevadensis* of a Dictyoptera-specific domain. The Periviscerokin domain (PF08259) has only been found in mantids and coackroaches so far and characterizes neuropeptides found in abdominal perisympathetic organs [53]. Another interesting case is the fungal protease inhibitor domain (PF12190). This domain is currently specific to five lepidopteran species (in four distinct genera) and might play an important role as a natural defense system against invading micro-organisms [54].

The remaining 15 domains without a hit to arthropod ESTs were further investigated. We recorded the taxonomic distribution of these domains according to the Pfam website, and reconstructed phylogenetic trees with the closest representatives from each major lineage. Our analysis suggests that these domains are mainly relics of an ancestral chordate or metazoan protein inventory.

BLAST-based analysis

Proteins or domains lacking detectable orthologs among other arthropods might also come from horizontal gene transfer (HGT). Such cases have already proved to confer important biological functions in the *Nasonia* genome [16]. We searched specifically for such cases as follows:

1. We ran BLASTP search of the termite proteins against NCBI non-redundant database requiring at least 40% coverage and an E-value < 10-5.
2. We retained all termite proteins for which none of the top three species hits belongs to a metazoan species.
3. For the resulting eleven candidate proteins, we searched for remote insect orthologs (with low e-value) to build phylogenies and try to reject the HGT hypothesis. In nine cases, insect

orthologs were found: six led to HGT rejection while three did not lead to any conclusion for or against. These three proteins, two SINA proteins and one cyclin dependent kinase, seem more similar to plants/fungi respectively than to insects and metazoans, but phylogenies showed long-branch attraction and no clear grouping of the candidate gene with putative homologs from the candidate donor taxon.

After these steps, only two HGT candidates remained. For each of these proteins, Znev_12272 and Znev_13267, BLASTP only returned a unique significant hit to a unique virus protein.

Znev_12272 has a high sequence similarity to a capsid protein of the *Nilaparvata lugens commensal X* virus (NLCXV). This virus is a satellite and RNA virus, known for infecting the brown plant hoppers, for which a single capsid protein has been reported [55]. If the integration of RNA viruses into a host genome was initially thought impossible, due to the absence of the intermediary DNA form, now many examples have been reported [56,57]. Such satellite virus is likely integrated by a helper virus, which is not known is the case of NLCXV. The initial BLASTP hit was not highly significant (E-value close to 10e-9) because of a large difference in length (170 amino-acids for Znev_12272 against 460 for the virus protein). However, the gene model could be reviewed to include the adjacent gene on the scaffold (Znev_12273, separated by 15 nucleotides) with a frame shift of +1. This would produce a more complete hit/coverage of the virus protein (roughly 360 amnio-acids, with associated E-value of 10e-15). Znev_12272 and Znev_12273 both have expression data support. Given that the alignment of the termite and NLCXV proteins exhibits an intermediate identity percentage and that several surrounding genes predicted by similarity to insect gene models, a contamination of the sequenced DNA is unlikely. We also discovered a duplication of more than 2,000 nucleotides (including the two aforementioned genes) on scaffold 1125, which suggests the presence of a second copy in the genome of *Z. nevadensis*.

A BLASTP search with Znev_13267 yielded a highly significant hit (10e−30) to a protein of the *Cotesia congregata bracovirus*. Both proteins are about 250 amino-acids long. The virus was described as a symbiont of the parasitic wasp *Cotesia congregata* [58]. It cannot replicate independently from the wasp. The wasp produces virion particles in its ovaries that are co-injected into the wasp's host when it lays its eggs. The virus then manipulates the hosts development and immune system [59,60]. There are few homologies found between proteins from *Cotesia congregata bracovirus* and other viruses [58]. It is unlikely that this case results from contamination given the active expression of Znev_13267 and the surrounding gene neighbourhood on the scaffold. Using TBLASTN to search with Znev_13267 against the *Z. nevadensis* genome itself returns another two complete loci, both without stop codons and one with expression support. Moreover, we identified at least five genomic fragments with more than 90% similarity to Znev_13267 (with their lengths ranging from 90 to 130 translated amino-acids). None of these altogether eight loci has introns and all occur in scaffolds where proteins have been predicted from insect gene models. Such observation could be explained by the nature of region matching in the virus genome, which contain a transposable element and a reverse transcriptase. Bracoviruses have only been found in Braconidae wasps so far [61]. The discovery of Braconidae wasps in a termite nest [62] might suggest a way for horizontal gene transfer, although it is unclear whether Braconidae use termites as a host.

5.3 Possible gene loss

Specific biological features of termites might also be related to the absence of essential proteins that are present in all other arthropods or insects. However, to distinguish a real protein loss from annotation

artefacts is a hard task:

1. Due to the large evolutionary distance to available arthropod genomes, the termite proteins might have diverged beyond the point allowing the recognition of their orthology relationships;
2. An apparent loss could be the consequence of the non-prediction of the gene or its incompleteness in the current version of the genome project;
3. In the worst case, the gene might be located in the genomic region currently not assembled (gap between scaffolds).

To identify *Z. nevadensis*-specific protein loss, we listed all Pfam domains that are recorded in all (or at least eight of) our nine reference arthropods. We initially obtained 23 Pfam families specifically lost in *Z. nevadensis* (present in all reference species) and 21 commonly lost in *Z. nevadensis* and one of the reference species (see footnote p49).

1. To detect divergent domains/proteins, we conducted a domain search by lowering detection thresholds (E-value of 10-1). It resulted in 13 domain families with divergent occurrences that currently escape the Pfam threshold, likely due to the novelty of the termite genome. Such cases should be resolved with Pfam updates integrating more divergent occurrences in the model learning process.
2. To cope with incomplete or unpredicted gene models, we searched in the genome assembly for traces of the remaining 30 "lost" families. We ran TBLASTN searches using the domain sequences of the reference arthropods and manually analysed the regions with the most significant similarity (E-value<10-5). A total of 17 additional proteins, previously thought as lost, were retrieved. However, investigations are still on-going to determine if the initial absence/incompleteness of some gene models correspond to a pseudogenization process (shift, deletion or loss of the ORF).
3. Finally, our analyses left 14 domain families (six present in all references and eight lost jointly with one reference species) for which the loss is likely real (details in Supplementary Table 19). Possible solutions to confirm the loss would be to use BLAST to align to the raw reads or wait for a more complete assembly. Hence, these cases have to be scrutinized in the future for functional implications and confirmation in other termite genomes.

Given the difficulties to confidently predict losses (2/3 of candidate lost domains were easily rejected), analysis of protein losses is still on-going at the sub-family level. We identified clusters for which only the termite *Z. nevadensis* shows an apparent loss (90 clusters) or the termite commonly lost its protein along with one of the reference species (135 clusters, only 43 jointly with *D. pulex*[1]).

5.4 Noteworthy copy loss in the opsin family

The opsin family in *Z. nevadensis* has been reduced to just two genes, the lowest number known for an insect. There is a single long wavelength opsin (Znev_03828) and a UV opsin (Znev_15028). Most insects have one or more representatives of three other opsin lineages; a blue wavelength opsin most similar to the UV opsin, an opsin of unknown function related to the *D. melanogaster* rh7 protein, and the non-visual pteropsin [63]. That these are all older lineages in insects and indeed related arthropods is

[1] Note that, according to the phylogeny, when the domain is commonly lost with *D. pulex*, the loss might actually correspond to a domain emergence that occurred in the Endopterygota ancestor (i.e. after the radiation of termites from the common ancestor of remaining species). However, such cases involve only three domain families for which the taxonomic distribution in Pfam has rejected an inset-specific origin.

shown by the presence of all four opsin lineages in the water flea *D. pulex* [21], therefore these three lineages must have been lost from *Z. nevadensis*, and probably other termites. The next smallest repertoire of opsins in an insect is that of the human body louse, *P. humanus*, which has three opsins, having retained the ortholog of Dmrh7, as well as a long wavelength opsin and a UV opsin [19].

5.5 Expansion/contraction of gene families

To trace the evolution of the *Z. nevadensis* protein repertoire, we investigated the specific and major expansions or contractions of protein families. For each protein family, as defined in Supplementary Notes 3.2, we applied Fisher's exact tests to detect statistically significant differences between the counts in *Z. nevadensis* and each of the reference species (pairwise comparisons). A gene family is then considered as over- or under-represented in *Z. nevadensis* if the signal was:

- Significant: having a p-value below a fixed threshold (typically 1%) when compared to all reference species independently;
- Unambiguous: having a ratio (counts over protein total) pointing in the same direction (inferior or superior) for all the reference species.

We investigated the significant signals of expansion/contraction in *Z. nevadensis* against the nine reference species. Because the genomes of *A. pisum* and *D. pulex* show frequently distorted counts, our criteria were relaxed for the comparison. Note that, we only report in this section the protein families that are present in at least one of the reference species and in *Z. nevadensis*, otherwise, please refer to the previous section on gains and losses (Supplementary Notes 5.2 and 5.3). Significantly expanded architectures are summarized in Supplementary Table 20.

***Z. nevadensis* against all nine reference arthropods**

According to our stringent criteria, four families have been found as expanded in *Z. nevadensis* compared to all nine reference arthropods. One of these domains, the seven-in-absentia (SINA - PF03145) family, corresponds to 33 proteins in *Z. nevadensis* while most reference species ranged from one to four, except *T. castaneum* with 16 proteins. Another family is the zinc finger C2H2-type (PF00096), with 215 proteins in *Z. nevadensis*. Most reference species had 31 to 108, except *A. pisum* which had 210. Even though absolute counts are close in *Z. nevadensis* and in the pea aphid, regarding the protein repertoire size, the expansion is significant. Finally, two protein families correspond to ionotropic receptors (IRs – see Supplementary Notes 7.4) sharing a common domain (Ligand-gated ion channel - PF00060). The mono-domain architecture is found in 134 proteins in *Z. nevadensis* compared to the seven to 29 in the reference insects and 114 in the crustacean *D. pulex*, while the bidomain architecture (PF00060-PF10613) is found in 24 proteins of *Z. nevadensis*, while only in one to three proteins in the reference insects and 14 in *D. pulex*.

***Z. nevadensis* against eight of the reference arthropods**

When comparing to all but one of the reference species, we identified another three expanded protein families. Two families contain Kelch domains: the mono-domain Kelch architecture (PF01344) and the tri-domain BTB-BACK-Kelch (PF00651-PF07707-PF01344) with respectively 20 and 37 proteins in *Z. nevadensis* while only zero to ten and four to ten, respectively, in the reference species. Interestingly both families are even more expanded in the pea aphid (65 and 78 copies, respectively). Further analysis of these proteins suggests that the monodomain architecture likely corresponds to incomplete gene models, frequently shortened by the end of the scaffold. Further, if not considering T. *castaneum*, *Z. nevadensis* shows an expansion of gustatory receptors (monodomain architecture 7tm chemosensory

receptors - PF08395), with 87 proteins compared to the eight to 50 in the eight remaining reference species (165 for the flour beetle). However, this expansion is likely due to the absence of numerous GR genes in the official gene set of the reference species and more generally in GenBank (Hugh Robertson's personal communication). We will therefore not consider GRs as expanded but will discuss their case further in the chemoperception section (see Supplementary Notes 7.3).

***Z. nevadensis* against seven of reference arthropods**

When relaxing our conditions to analyse signals supported by at least seven of the reference species (excluding all possible pairs), we found two more putatively expanded protein families and still no contractions. First, one involves the mono-domain Polycystic kidney disease (PKD) (PF08016) family. The ten proteins found in *Z. nevadensis* create a significant enrichment compared to the unique or double copies of most organisms but is not significant against the three and six copies of *P. humanus* and *D. melanogaster,* respectively. The second case concerns the bi-domain architecture PF00096-PF07776 that are both zinc finger domains (-C2H2 and -associated domain) found in 32 proteins for *Z. nevadensis*. Reference species had one to 14, except in *D. melanogaster* and *T. castaneum*, which had 38 and 40, respectively. These results suggest an independent and more recent expansion in the Diptera-Coleoptera common ancestor.

To cope with domains that might not be currently represented in the Pfam database, we reproduce this analysis with several Interpro databases: SUPERFAMILY, GENE3D, TIGRFAMs, SMART, PROSITE and PRINTS. First, the expansions mentioned above based on Pfam annotation were confirmed by almost all other databases, except PRINTS which uses the shortest motifs. PRINTS is the only database that proposes distinct models for alpha, beta, gamma, delta, epsilon, zeta tubulins, while other databases have a unique model for the tubulins. This explained why the alpha-tubulin family (PR01161–PR01162) was found to be expanded only by the PRINTS database. *Z. nevadensis* exhibits 14 alpha-tubulin proteins in contrast to the four to eight observed in the reference arthropods.

Interestingly, among the nine expanded gene families, five have differential expression patterns in distinct termite samples (see Supplementary Table 7). Two of these five families also show evidence of accelerated evolution subsequent to termite speciation (see Supplementary Table 17). These cumulative signals of a specific evolution are summarized in Table 1 of the main manuscript.

6 Gene expansion and male mating

6.1 Intracolonial mating biology

The range of *Zootermopsis nevadensis* extends into the northwestern USA [64], areas that are subject to periods of freezing. The seasonality of the environment strongly suggests cyclical egg production, since ovarian and testis activity and egg-laying at low temperatures seems unlikely. In the laboratory, eggs were only irregularly encountered when opening the wood blocks in which the colonies were housed (Brent and Liebig, personal observation). Eggs were either absent, or observed in large quantities with or without first instar larvae. Dissections of neotenic reproductives throughout the year revealed highly varied states of gonadal activity. These laboratory observations also strongly support the idea of cyclical egg production in *Z. nevadensis*.

6.2 Phylogenetic analysis of genes jointly differentially expressed in males and expanded in the termite lineage

Among the expanded gene families in the *Z. nevadensis* lineage (Supplementary Table 20), we observed differentially expressed genes occurred predominantly in males (Supplementary Table 7). Since several of these families are known for their involvement in spermatogenesis, we conducted more detailed analyses and phylogenies. The phylogenies found in Supplementary Figure 9 to Supplementary Figure 13 were obtained as described in Supplementary Notes 3.3. They allow us to visualize the evolutionary history of these families, and to distinguish the emergence of those genes differentially expressed in males (in blue) compared to novel termite genes lacking expression bias (in red). PKD channel and monodomain Kelch families are both almost entirely represented in the gene expression data. They are exclusively more highly expressed in males (except one monodomain Kelch differentially expressed in nymphs). After collecting all family members in *Z. nevadensis* and reference species, we built phylogenies displayed for the PKD (Supplementary Figure 9) and Kelch (Supplementary Figure 10) protein families. These analyses show a common origin for eight of the nine PKD proteins, and for 15 of the 20 Kelch monodomain proteins. These two paralogous groups only contain a single non-differentially expressed protein each.

For the tridomain BTB-BACK-Kelch, the analysis revealed 26 proteins differentially expressed in males while the remaining eleven members do not show differential expression. In-depth analysis of the sub-family revealed that these 26 proteins belong to a single orthoMCL cluster while the reference species either had no proteins (*H. saltator*, *D. pulex* and *C. elegans*) or only one. Further, this cluster contains one more termite protein with degenerate domain architecture, BTB-Kelch. Interestingly, we observed that all four termite proteins with BACK-Kelch architecture and four of the five with BTB-Kelch architecture are differentially expressed in males (see Supplementary Table 7). As a consequence, for phylogenetic analysis, we consider all proteins with at least two of the three BTB, BACK or Kelch domains in *Z. nevadensis* (64 proteins) and in the reference species. The tree topology (Supplementary Figure 11) reveals a common origin for the 39 termite proteins differentially expressed in males: 27 proteins have the tridomain architecture (blue and red) and twelve proteins exhibit the degenerate BTB-Kelch, BACK-Kelch and BTB-BACK architectures (light blue and orange).

For the SINA proteins, the expansion involves proteins with distinct differential expression. The largest group involves eleven proteins over-expressed in female reproductives, but the second group

corresponds to six genes over-expressed in males. We present the phylogeny of SINA genes which reveals a common origin of four of these six SINA proteins (Supplementary Figure 12).

For the Alpha-Tubulin family, no significant differential expression is observed but four genes are differentially expressed in males. For phylogenetic analysis we first had to exclude three termite proteins that were short (likely fragments) and exhibited atypical domain architecture (fewer repeats and, after resolving overlaps, only PR01161-Tubulin was retained). Preliminary phylogenies revealed long-branch attraction in nine proteins; two of each from *A. pisum*, *P. humanus*, *D. melanogaster,* and one from *T. castaneum*, *C. floridanus* and *H. saltator.* All but two of these proteins, are characterized by a distinct architecture compared to the usual alpha-tubulins. Indeed, alpha-tubulins are usually composed of C- and N-terminal stretches of five to six alpha tubulins motifs, and a central section of one or two tubulin motifs. Rather than the typical C-terminal alpha tubulin domains, these problematic sequences have a series of standard tubulin motifs. To increase branch-length resolution, these sequences were removed and a new phylogeny was computed to represent the differentially expressed alpha-tubulins in termite males (Supplementary Figure 13).

6.3 Putative interactions and functions of the expanded and male-specifically expressed genes

Two of these gene families, KLHL10 and SINA, have been associated with E3 ubiquitin ligase complexes [65,66] that are responsible for protein degradation. Particular members of such complexes are proteins with a BTB (Bric-a-Brac-Tramtrack-Broad) domain that function as substrate targeting proteins such as KLHL10. Other proteins of these complexes contain a RING (Really Interesting New Gene) domain that binds ubiquitin conjugating enzymes (E2) and a substrate-binding domain that interacts with various proteins sometimes leading to their degradation. SINA genes possess these features and their proteins function as E3 ubiquitin ligases [65,67]. Although the third expanded family, alpha tubulins, is not a component of an E3 ubiquitin ligase complex, they interact with SIAH genes during cell division [68]. Although, no direct link between PKD genes and the other three families exists to our knowledge, eight of the ten PKD-channel genes are co-expressed with KLHL10 in the male reproductive morphs exhibiting high expression levels (Supplementary Table 7).

KLHL10 is involved in formation of a Cullin-3-dependent E3 ubiquitin ligase complex, which plays an important role in spermatogenesis [69]. The Cullin-3–dependent ubiquitin ligase complex is required for caspase activation in spermatids [70], and haploinsufficiency of KLHL10 can lead to infertility in male mice [71]. Similarly, mutations in SINA are associated with male infertility in *Drosophila* [72]. In male mice SIAH2 plays a role in spermatogenesis. It is expressed in spermatids in the testes during meiosis I, and SIAH1a mutants have stalled meiosis at metaphase I through telophase I in spermatocytes [73,74]. Members of the third expanded family, the alpha-tubulins, interact with SIAH-1 and SIAH-2 in a C-terminal region that is highly conserved among SIAH proteins during mitosis suggesting an important general function in cell division [68]. PKD genes encode polycystic proteins that either function as G protein-coupled receptors or Ca^{2+} ion channels with wide-spread expression in testes and a potential role in spermatogenesis [75,76].

SIAH-2 function is not restricted to spermatogenesis. In fact, SIAH-2 is expressed in growing oocytes in mice [73]. In *Drosophila*, SINA mutations are also associated with female fertility [72]. This may explain why we also see an expansion of seven SINA genes in two clusters in our termite species that are differentially expressed in female reproductives (Supplementary Figure 12).

7 Chemoperception and sociality

7.1 Odorant binding proteins

The insect odorant binding proteins (OBPs) are a family of small secreted globular proteins generally considered to function in binding and transporting hydrophobic compounds [77]. Originally discovered as highly expressed in insect antennae, the gene family in some insects also contains members expressed in other body parts. Their binding of odorants is usually not highly specific, but they are thought to play an important role in olfaction by transporting hydrophobic ligands from the exterior through the sensillar lymph to the dendrites of olfactory sensory neurons. In some cases OBPs interact directly with olfactory receptors. They are expressed, often at high levels, in the support cells at the base of each sensillum, and secreted into the sensillar lymph. Most insects with complete genome sequences have been shown to encode tens of these proteins. In endopterygotans the family consists of several subfamilies. The classic OBPs usually have six highly conserved cysteines forming disulfide bonds to maintain the tertiary structure; however some have lost two of these cysteines, i.e. one disulfide bond.

OBP family members in *Z. nevadensis* are extremely divergent from each other, and are encoded by long genes with short exons. As a result, TBLASTN searches were largely ineffective at finding these genes. However, because they are generally well expressed, even RNA-seq from whole animals commonly provided full-length EST-contigs, allowing for gene building by the automated annotation effort. Hence most of the termite OBPs described here were discovered by BLASTP searches with aphid OBPs or OBPs previously described for other termite species. Additional genes were discovered by scanning the flanking DNA in the genome browser for EST contigs with similar exon arrangements when aligned to the genome. As a result, any un-modelled singleton gene will have been missed, but there are only three singletons in this set, so it is unlikely that many were overlooked.

Twenty-nine OBPs were modelled (Supplementary Table 21). In stark contrast to the OR and GR genes, 20 of these were already perfectly modelled, presumably as a result of the deep coverage by EST contigs. Another five genes were partially modelled, and only required minor fixes to the model or the assembly, while one gene remains incomplete in the assembly. Only three new gene models are proposed and no pseudogenes were identified. As is commonly the case in other insects, most of these genes are in tandem arrays of 2-7 genes. Their gene structures are fairly complicated, with 5-8 introns. The encoded proteins are generally of typical length for classic OBPs, although OBP24-28 are rather longer and OBP29 has a long insertion between the signal peptide and the corresponding mature protein. OBP3, 9, 23, and 24 have lost two of the conserved cysteines, and presumably one of the three disulfide bonds.

In some insects not all OBPs are expressed in obvious chemosensory organs like the antennae [78]. We nevertheless obtained ESTs from RNAseq data from four different antenna samples (4-8 antennae/sample): male and female worker, male and female pre-alate that were used to identify genes associated with olfaction and gustation. RNAseq was performed as explained in Supplementary Notes 2.2. It appears that almost all OBPs are expressed in the antennae of *Z. nevadensis*, because the vast majority had full-length EST contigs from the RNA-seq for each of the four antennal samples (male/female workers and pre-alates). Only OBP13 was not observed in the antennae, although it was found in whole body samples and might be expressed in other chemosensory organs such as the mouth or cerci. The most interesting observation from this expression data is that OBP1 is expressed only in alates (whole bodies) and pre-alates (antennae), suggesting that it might have a specific role during this

life stage. No sex-specific expression patterns were observed, at least at the level of the available EST contigs.

Only the mature OBP peptides of about 120 amino acids can be confidently aligned, and then only the four regions surrounding the single or paired conserved cysteines can be utilized for phylogenetic analysis, which, given the extremely divergent sequences of these termite OBPs, provides almost no phylogenetic resolution [79]. For this reason a tree is not provided.

7.2 Odorant receptors

The odorant receptor (OR) family of seven-transmembrane proteins found in insects mediates most of olfactory processing [80,81]. Additional contributions to olfaction are made by a subset of the distantly related gustatory receptor (GR) family (e.g., the carbon dioxide receptors in flies [82-84]), and a subset of the more recently described and unrelated ionotropic receptors (IRs) [85-87]. The OR family ranges in size: a low of ten genes in the human body louse [19]; 50-100 in Drosophila flies [88,89], mosquitoes [90-92], the silk moth *Bombyx mori* [93,94], and the pea aphid *A. pisum* [95]; 100-300 in the beetle *T. castaneum* [96], the honey bee *A. mellifera* [97], and Nasonia wasps [98]; and over 300 in several recently sequenced ants [43,99-101]. Although most genes in the Drosophila flies are scattered around the genome, with only a few in small tandem arrays, tandem arrays are more typical of the other species, especially those with large repertoires. It has been inferred that larger repertoires partly result from retention of gene duplicates generated in tandem arrays by unequal crossing over [97].

The OR family was manually annotated using methods employed before for the Drosophila, mosquito, moth, beetle, bee, wasp, aphid, louse, and ant genomes. Briefly, TBLASTN searches were performed using aphid and louse ORs as queries, and gene models were manually assembled in the text editor of PAUP*v4.0b10 [102] or in TextWrangler. Iterative searches were also conducted with each new termite protein as query until no new genes were identified in each major subfamily or lineage. Gene models were confirmed or refined when possible using contigs of ESTs from RNA-seq experiments on whole animals of each sex, castes and stages, as well as antennal samples from both sexes of workers and pre-alates. All of the ZnOr genes and encoded proteins are detailed in Supplementary Table 22. All ZnOr proteins are provided as FASTA format text files (http://termitegenome.org/?q=consortium_datasets). The gene models for these have subsequently been updated in the genome browser, although the fixes to the assembly have not been executed, so these models remain as partials.

Several difficulties typical of draft genome assemblies were encountered in this gene family. These included short gaps in the assembly that could commonly be repaired with raw reads, although some gaps involve long repetitive regions that could not be manually repaired. Occasionally gene models were designed that span scaffolds, with no support other than the agreement of the available exons on both scaffolds, and their appropriate relatedness to similar genes, especially in tandem arrays. These problems are noted in Supplementary Table 22.

Pseudogenes were translated as best as possible to provide an encoded protein that could be aligned with the intact proteins for phylogenetic analysis, and attention was paid to the number of pseudogenizing mutations in each pseudogene. A 200 amino acid minimum was enforced for including pseudogenes in the analysis (roughly half the length of a typical insect OR), and there are several shorter fragments of genes that were not included in Supplementary Table 22 or the analysis, most of which are extremely similar to existing genes, so might be recent duplications or assembly artefacts.

All termite, aphid, and louse ORs were aligned in CLUSTALX v2.0 [103] using default settings and problematic gene models and pseudogenes were refined in light of these alignments.

For phylogenetic analysis, the poorly aligned and variable length N-terminal and C-terminal regions were excluded (specifically ten amino acids before the conserved GhWP motif in the N-terminus and ten after the conserved SYFT motif in the C-terminus), as was the major internal region of long length difference found between the longer DmOr83b orthologs, now known as OrCo proteins [104] (ZnOrCo, ApOr1, and PhOr1) and most of the other Ors. ZnOr1-37 have a similar length of amino acids in this region making them comparable in overall length to OrCo proteins and about 50 amino acids longer than the average insect OR of 400 amino acids, but it is so highly divergent as to be useless for phylogenetics. Other regions of potentially uncertain alignment between these highly divergent proteins were retained, because while potentially misleading for relationships of the subfamilies (which are poorly supported anyway), they provide important information for relationships within subfamilies.

Phylogenetic analysis of this set of 138 proteins was carried out in the same fashion as for previous OR analyses [97,98]. This involved a combination of model-based correction of distances between each pair of proteins, and distance-based phylogenetic tree building. Pairwise distances were corrected for multiple changes in the past using the BLOSUM62 amino acid exchange matrix in the maximum likelihood phylogenetic program TREEPUZZLE v5.2 [105]. These corrected distances were fed into PAUP*v4.0b10 [102] where a full heuristic distance search was conducted with tree-bisection-and-reconnection branch swapping to search for the shortest tree. The resultant tree is shown in Supplementary Figure 14. Bootstrap analysis with 10,000 replications of neighbour-joining using uncorrected distances was performed to assess the confidence of major branches in the tree, and is shown above major branches in the tree.

The ZnOr gene set consists of 69 models. Six models were apparent pseudogenes, ten required repair of assembly gaps, and five were joined across scaffolds. The result is 63 apparently intact OR proteins, with only two of these still missing N-terminal, C-terminal, or internal regions, so their functionality remains uncertain. Less obvious pseudogenes (for example with small in-frame deletions or insertions, crucial amino acids changes, or promoter defects) would not be recognized, so this total might be high. Approximately five gene fragments remain so short and incomplete they were not included, but some might represent intact genes.

The automated gene modelling had access to all available insect ORs in GenBank, for comparative information, as well as EST contigs from the RNA-seq on twelve sexes and life stages, but not the four antennal samples. It succeeded in building at least partial gene models for 26 of these 69 genes. However, as has been true for most other insect genome projects, just one of these was precisely correct, and that was for the highly-conserved and well-expressed OrCo protein. All others required at least one gene-model change, while 29 new gene models were generated (not including pseudogenes or those requiring repair of assembly gaps or joins across scaffolds) (Supplementary Table 22). Because these genes are typically expressed at low levels in only a few cells, they are not always well represented by EST contigs in the whole body RNA-seq. Regardless, most are well covered by full-length EST contigs in the antennal RNA-seq, and so these manually built gene models are highly reliable.

As expected there is a single conserved ortholog of the DmOr83b protein, now called OrCo, sharing 54% amino acid identity with ApOr1, 58% with PhOr1, and 56% with DmOr83b. There are no other

simple orthologous relationships between termite, aphid, and louse Ors. Instead, as is common for these rapidly evolving proteins in such divergent taxa, there are differential gene lineage or subfamily expansions. As described earlier, the aphid ORs consist of three ancient lineages (ApOr2-4), and two relatively recent expansions (ApOr5-13 and 14-79) , while the louse ORs are reduced to a set of eleven proteins, with two divergent lineages represented by single genes (ApOr2/3) and the rest a small cluster of relatively old proteins. The termite ORs are roughly equally split between an old expansion (Or1-37) and a more recent expansion (Or38-69) (Supplementary Figure 14). There was no bootstrap support for the monophyly of the older expansion; however it is largely supported by a Bayesian analysis using MrBayes v3.1 with the JTT substitution model, four chains, one million generations, and two runs, with trees sampled every 100 generations, discarding a burn-in of 250,000 generations. The younger expansion contains most of the gene fragments remaining in the genome, which might be remnants of these duplication events, or might indicate difficulties in assembling these often highly similar sequences. This expansion also contains all six of the pseudogenes, each of which is relatively young with mostly single pseudogenizing mutations. Both of these termite OR expansions contain sets of tandemly duplicated genes, although these tend to be larger in the more recent expansion. The largest expansions consist of eleven (Or38-47) and nine genes (Or59-67), each set spanning two scaffolds.

These two termite OR gene subfamily expansions also differ slightly in their gene structures, as summarized in Supplementary Table 22. ZnOr1-26 have a shared structure of a long first exon, followed by a phase 2 intron, and then four short exons separated by phase 0 introns, which appear to correspond to the widely present final three phase 0 introns in other insect Or genes. ZnOr27-36 have an additional phase 1 intron interrupting the end of this long first exon, and Or30-33 have an additional phase 2 intron interrupting the end of this exon. Or11, 19, and 35 each have an additional idiosyncratic intron. In contrast, the more recent expansion of ZnOr38-69 all have early phase 2 and 0 introns, followed by the same set of 2-0-0-0 introns as the older expansion, except ZnOr69 which lost the final phase 0 intron. The shared set of 2-0-0-0 introns is also shared with the louse and aphid OR genes, however the aphid genes also commonly have earlier introns. This accounting is only for introns within the coding regions. The antennal EST contigs sometimes indicate the presence of an extended 5' UTR with an intron in it, so initiation of transcription and the promoter are commonly far upstream of the start codon.

These termite ORs are so divergent from the aphid and louse ORs, as well as all other insect ORs, that it was possible that additional highly divergent subfamilies might exist, undetected by TBLASTN searches of the genome assembly. Although the long first exon should usually reveal matches, sometimes this is broken up by introns, and the most conserved C-terminal part of the protein is broken up into several exons, lowering the power of TBLASTN searches to find them. Two approaches were employed to check this issue. First, because most of these genes are well represented in the antennal RNA-seq assemblies, which provide full-length proteins for more sensitive searching, these were searched using TBLASTN with several representative termite and aphid ORs, but no additional divergent ORs were discovered. Second, several PSI-BLASTP searches were performed, using as a database all of the "nr" (non-redundant) protein database at GenBank, plus additional unpublished insect ORs, plus the automated termite gene model proteins, seeking proteins that are not in the OGS column in Supplementary Table 22, but none were found. Given that only 26 of these 69 OR genes were even partially annotated, this is not a conclusive test and a few highly divergent genes might have been missed, however it seems unlikely that additional large divergent subfamilies of ORs exist in this genome, because at least a few of them would have automated gene models.

The RNA-seq EST contigs from both sexes and several life stages, provided some support for these gene models, however most were short and covered only a few exons, some were aberrantly spliced, and many genes had no experimental support (Supplementary Table 22). In contrast, the EST contigs from the four antennal RNA-seq experiments from sexed workers and prealates usually provided full-length support for each gene. Those with no antennal RNA support were examined for any indication that they might in fact be pseudogenes, but none were found, and indeed several apparent pseudogenes are transcribed.

Finally, because these ORs are so divergent from both the aphid and louse ORs, and indeed all other insect ORs, no inferences about function can be gleaned by comparison with other insect ORs whose ligands have been determined, primarily those of Drosophila flies, Anopheles mosquitoes, and moths. The more recently expanded subfamily might be implicated in aspects of chemical sensory ecology particular to this group of dampwood termites.

7.3 Gustatory receptors

The gustatory receptor (GR) family of seven-transmembrane proteins in insects mediates most of insect gustation [80,81], as well as some aspects of olfaction, including the carbon dioxide receptors in flies [82-84]. The GR family ranges in size: a low of six genes encoding eight proteins in the human body louse [19], and ten genes in *A. mellifera* [97]; 50-100 receptors in *Drosophila* [88,89], mosquitoes [90,106], *B. mori* [107], *A. pisum* [95], Nasonia wasps [98], and several ant species [43,99,101]. The flour beetle *Tribolium castaneum* is an outlier with over 200 GRs [15]. The GR family is more ancient than the OR family, which was clearly derived from within it, and is found in *D. pulex* [108], the tick *I. scapularis* (HMR, unpublished), the centipede *Strigamia maritima* (HMR, unpublished), and many other animals (HMR, unpublished). This evolutionary history is reminiscent of the ionotropic receptors (IRs) [85,86].

The GR family was manually annotated using methods employed before for the Drosophila, mosquito, moth, beetle, bee, wasp, aphid, louse, and ant genomes and as outlined in Supplementary Notes 7.2. Two genes, GR35 and 43, appear to be alternatively spliced, much as several of the Drosophila and other insect GRs are, with alternative long first exons spliced into shared short C-terminal exons. All of the ZnGr genes and encoded proteins are detailed in Supplementary Table 23. All ZnGrs are provided as FASTA format files (http://termitegenome.org/?q=consortium_datasets).

Several difficulties typical of draft genome assemblies were encountered in this gene family. These included short gaps in the assembly that could commonly be repaired with raw reads, although some gaps involve long repetitive regions that could not be manually repaired. One gene model (Gr17) was designed that spans scaffolds, with no support other than the agreement of the available exons on both scaffolds, and their appropriate relatedness to similar genes in a tandem array. Three genes (Gr23, 56, and 84) appeared to have frameshifting insertions in them, however examination of the raw reads revealed that these were unusual assembly problems and the genes are intact and were repaired. These problems are noted in Supplementary Table 23.

Pseudogenes were translated as best as possible to provide an encoded protein that could be aligned with the intact proteins for phylogenetic analysis, and attention was paid to the number of pseudogenizing mutations in each pseudogene. A 200 amino acid minimum was enforced for including pseudogenes in the analysis (roughly half the length of a typical insect GR), and there are several

shorter fragments of genes that were not included in Supplementary Table 23 or the analysis. All termite and louse GRs, most aphid GRs, representative carbon dioxide and sugar receptors from endopterygotan insects, and all ten honey bee GRs, were aligned in CLUSTALX v2.0 using default settings. Problematic gene models and pseudogenes were refined in light of these alignments. Several aphid GRs were left out of the analysis to simplify it, including several highly similar proteins in two recent expansions (GR47-63 and 67-76), as well as the highly divergent GR15 and a few short proteins that did not align well (GR28 and 29). For phylogenetic analysis, the poorly aligned and variable length N-terminal and C-terminal regions were excluded (specifically 15 amino acids before the conserved GxxP motif in the N-terminus and immediately after the TYhhhhhQF motif in the C-terminus), as was a major internal region of length differences, specifically a long length difference region in the internal loop 2. Other regions of potentially uncertain alignment between these highly divergent proteins were retained, because while potentially misleading for relationships between the subfamilies (which are poorly supported anyway), they provide important information for relationships within subfamilies. Phylogenetic analysis and tree assembly (Supplementary Figure 15) of these 161 proteins was carried out as described for OR analysis in S10.2

The ZnGr gene set consists of 87 models, an average family size for insects. Seven gene models are apparent pseudogenes, eight required repair of assembly gaps or errors, and one was joined across scaffolds. The result is 80 apparently intact GR proteins, with only four of these still missing N-terminal, C-terminal, or internal regions, so their functionality remains uncertain. Less obvious pseudogenes (for example with small in-frame deletions or insertions, crucial amino acid changes, or promoter defects) would not be recognized, so this total might be high. Approximately eleven gene fragments remain that were too short and incomplete to be included, but some might represent intact genes.

The automated gene modelling had access to all available insect GRs in GenBank, for comparative information, as well as EST contigs from the RNA-seq on twelve sexes and life stages, but not the four antennal samples. It succeeded in building at least partial gene models for 32 of these 87 genes, although some of these are large gene models that concatenate two or three genes. However, as has been true for most other insect genome projects, just two of these were precisely correct, and one of these is intronless. All others required at least one change, while 42 new gene models were generated (not including pseudogenes or those requiring repair of the assembly). Unfortunately because these genes are typically expressed at low levels in only a few cells, they are not always well represented by EST contigs in the whole body RNA-seq, and only some are well covered by EST contigs in the antennal RNA-seq (Supplementary Table 23). Regardless, these manually built gene models are highly reliable because there are representative full-length EST contigs for most subfamilies, and the largest subfamily does not even have introns in the coding region (Gr44-87).

The GR family consists of several highly divergent gene subfamilies, the functional roles of most of which are at least generally known, even if the particular ligand for each receptor remains unclear. Additionally, many appear to function as multimeric complexes, making individual roles unclear (but see [109]). The candidate sugar receptors are perhaps the oldest known subfamily, with clear homologs in the aphid [95] and possible relatives in the crustacean *Daphnia pulex* [108]. *Z. nevadensis* has six candidate sugar receptors (GR1-6), which is about average for insects (ranging from 0-14).

The carbon dioxide receptors consist of two proteins in Drosophila flies and three in mosquitoes, the silkmoth, and Tribolium beetles [82-84,110,111]. Surprisingly, this lineage of fairly well conserved proteins is

absent from the sequenced Hymenoptera. Robertson et al. [110] speculated that this lineage was lost in Hymenoptera, but were likely to be found in more basal insects, although neither the human body louse nor the pea aphid have them. Remarkably *Z. nevadensis* has a large GR subfamily unequivocally related to these endopterygotan carbon dioxide receptors (GR7-32). The phylogenetic analysis suggests that this is an old gene lineage in termites and presumably other basal insects and perhaps other arthropods (although Daphnia, Ixodes, and Strigamia have no sign of them), with the carbon dioxide receptors of endopterygotan insects being derived from one sublineage. Unfortunately, there is no bootstrap support for this arrangement. One might anticipate that carbon dioxide receptors should be expressed in antennae, and there is some indication of this in the available EST contigs, however the same is true of subsets of the candidate sugar receptors, and the three-intron and intronless subfamilies (Supplementary Table 23).

The remaining termite GRs are split into two subfamilies, both of which have both older and newer branches. One of these subfamilies, GR33-43, has a simple gene structure of a long first exon, followed by three short exons separated by phase 0 introns. These intron locations are inferred to be ancestral to the family. The remaining subfamily of 43 genes (GR44-87) is intronless, suggesting that they are all derived from a single gene that underwent a gene conversion recombinational event with a cDNA that removed all three ancestral introns, or was the result of an integration of a cDNA copy, a retrogene. Unfortunately the divergence of these two GR subfamilies from all insect GRs precludes assignment of likely functions, although it is believed that most insect GRs are involved in sensing bitter tastants like defensive secondary compounds from plants.

Given that the last two termite GR subfamilies (GR1-6 and GR7-32) are so divergent from the aphid and louse GRs, as well as all other insect GRs, several PSI-BLASTP analyses were performed similar to the search for addition ORs in Supplementary Notes 7.2. However, no additional GRs were found. With only 32 of these 87 GR genes even partially annotated, this is not a conclusive test and a few highly divergent genes might be missed. However, it seems unlikely that additional large divergent subfamilies of GRs exist in this genome, because at least a few of them would have automated gene models.

Two other functional lineages of insect GRs have been identified. The DmGr43a protein is a well-conserved lineage in endopterygotan insects, with a single gene in the available fly and hymenopteran genomes available to date, two genes in the silkmoth, and ten in Tribolium beetle. Recently DmGr43a and one of the silkmoth homologs (BmGr9) was identified as a fructose receptor [112,113], however this gene lineage is not obviously identifiable in *Z. nevadensis*. Another lineage of the Drosophila GRs has been implicated in perception of cuticular hydrocarbons [114], although ligand specificity has yet to be demonstrated, and again *Z. nevadensis* has no simple ortholog of this lineage.

7.4 Ionotropic receptors

In addition to the OR and GR families in the insect chemoreceptor superfamily [88], there is a second completely different family of olfactory and gustatory receptors in insects, the ionotropic receptors [85], which clearly evolved from the ionotropic glutamate receptors involved in synaptic transmission [86]. These proteins are somewhat larger than the ORs and GRs, and have three transmembrane domains. They function as obligate heterodimers, usually consisting of two and sometimes three different proteins. While some of these IRs are highly conserved, and have been implicated in olfaction, others

are highly divergent and some are implicated in gustation. Like the ORs, and probably many GRs, the divergent IRs function in complexes with some of the conserved proteins, specifically IR8a and/or IR25a [87]. Given the relatively small sizes of the OR and GR repertoires in *Z. nevadensis* genome, we also characterized the IR family and discovered it is comparable in size to the OR and GR families combined.

The IR family and related ionotropic glutamate receptors were manually annotated using methods employed before for the chemoreceptors in Drosophila, mosquito, moth, beetle, bee, wasp, aphid, louse, and ant genomes as described in Supplementary Notes 7.2. All of the ZnIR genes and encoded proteins are detailed in Supplementary Table 24. All *Z. nevadensis* IR and iGluR proteins along with improved models for the *A. pisum* and the human body louse *P. humanus,* are provided as a FASTA format text file (http://termitegenome.org/?q=consortium_datasets).

In order to do an analysis of the phylogeny of this gene family rooted with the more conserved ionotropic glutamate receptors from which they evolved, it was necessary to annotate the NMDA, AMPA and KAINATE receptors (numbers for the KAINATE receptors are independent of those for the Drosophila KAINATE receptors (GluRIIa-c) because they are independent duplications). Gene models for these proteins were also refined as necessary from the available aphid and louse annotations. In addition, two new IRs were recognized in the louse genome and eight more in the aphid genome. These were numbered from IR320-IR327, following the last IR named by Benton et al. [85] in ants, but two of the new aphid IRs are closely related to the existing ApIR75d1/2 proteins, so they were named ApIR75d3/4.

Naming and numbering of the termite IRs is complicated. Following the example of the Benton et al. [85], the conserved orthologs of several IRs in other insects are given those names, specifically 8a, 21a, 25a, 68a, 93a and 76b. There are five paralogs of DmIR41a, so those were named IR41a1-5, and there are 17 paralogs of the DmIR75a-d group, which is also expanded in other insects. These were named IR75a-q. Continuing the Benton et al. [85] approach of naming each IR from each new insect with a consecutive number will eventually become cumbersome, so the remaining termite IRs were numbered starting from 101, which avoids any confusion with the existing DmIRs which are numbered through 100a according to their cytological position in the genome.

Several difficulties typical of draft genome assemblies were encountered in this gene family. These included short gaps in the assembly that could commonly be repaired with raw reads, although some gaps involve long repetitive regions that could not be manually repaired. There were also three insertions in the assembly not present in the raw reads. These problems are noted in Supplementary Table 24.

Pseudogenes were translated as best as possible to provide an encoded protein that could be aligned with the intact proteins for phylogenetic analysis, and attention was paid to the number of pseudogenizing mutations in each pseudogene. All termite, louse, aphid, and *D. melanogaster* IRs were aligned in CLUSTALX v2.0 [103] using default settings. Problematic gene models and pseudogenes were refined in light of these alignments.

For phylogenetic analysis, the poorly aligned and variable length N-terminal and C-terminal regions were excluded, along with several internal regions of highly length-variable sequence. Other regions of potentially uncertain alignment between these highly divergent proteins were retained, because while

potentially misleading for relationships between subfamilies, they provide important information for relationships within subfamilies.

Phylogenetic analysis of this set of 280 proteins was carried out in the same fashion as for previous chemoreceptor analyses [97,98] and as for the OR analysis outlined in Supplementary Notes 7.2. Over 2.4 million trees were examined and the resultant tree is shown in Supplementary Figure 16.

The ZnIR gene set consists of 150 models, the largest repertoire known for any insect to date, and roughly ten-fold more than the pea aphid (19) and body louse (13). Thirteen (9%) of these are apparent pseudogenes, while twelve gene models required repair of assembly gaps or insertions. The result is 137 apparently intact IR proteins, with only nine of these still missing N-terminal, C-terminal, or internal regions, so their functionality remains uncertain. Less obvious pseudogenes (for example with small in-frame deletions or insertions, crucial amino acids changes, or promoter defects) would not be recognized, so this total might be high. Approximately five gene fragments remained so short and incomplete they were not included, but some might represent intact genes.

The automated gene modelling had access to all available insect IRs in GenBank, for comparative information, as well as EST contigs from the RNA-seq on twelve sexes and life stages, but not the four antennal samples. It succeeded in building at least partial gene models for 92 of these 150 genes, although some of these are large gene models that concatenate multiple genes. However, just 13 of these were precisely correct, and six of these are intronless. All others required at least one change. A total of 46 new gene models were generated, not including pseudogenes or those requiring repair of the assembly. Unfortunately because these genes are typically expressed at low levels in only a few cells, they are not always well represented by EST contigs in the whole body RNA-seq, although many are well covered by EST contigs in the antennal RNA-seq (Supplementary Table 24). Nevertheless, these manually built gene models are highly reliable because they are representative full-length EST contigs for most subfamilies, and the largest subfamily is intronless in the coding region (IR156-222).

For the iGluR genes, the expected simple orthologs were found for NMDA receptors 1-3 and the AMPA receptor, however *Z. nevadensis* also has a recently duplicated paralog of NMDAR2, here called NMDAR4, that aphid, louse and Drosophila do not have. Drosophila species have lost NMDAR3. The KAINATE receptors are independently duplicated in flies versus these three exopterygotan insects, each of which has five or six of them. The phylogenetic relationships and the equally highly conserved IR25a and IR8a lineages, which together were declared the outgroup in which the tree was rooted, are not quite right in Supplementary Figure 16. Separate analysis of them shows that the AMPA receptors cluster outside of the DmGLuRIIa-c and KAINATE receptors, in agreement with Croset et al. [86].

The IR family contains several conserved orthologous genes shared across insects. As noted above, the co-receptor IR25a and 8a genes are unusually highly conserved and cluster confidently with the iGluRs from which they clearly evolved [86]. The other orthologous lineages are rather more rapidly evolving, including IR93a, 76b, 68a, and 21a. Like many other insects, the IR41a and 75 lineages are duplicated, with the 17 IR75 paralogs in *Z. nevadensis* an unusually high number [86]. There was no obvious ortholog for the IR40a that has a representative in aphid and louse. Croset et al. [86] named an IR100a ortholog in the aphid, but in this analysis its relationships with DmIR100a is less clear, and the termite relatives are members of a large intronless expansion that is unlikely to be orthologous to the IR100a lineage. Like the louse (IR145, 226, 320 and 321) and aphid (IR100a, 322-327), *Z. nevadensis* has a few highly divergent IRs, specifically IR101-104, which weakly cluster with ApIR324, and IR105-107,

which in turn weakly cluster with PhIR226. These are either ancient IR lineages in these three organisms that have not been duplicated much, or they are recent and rapidly evolving. There is nothing obvious about the expression of these divergent IRs to give clues to their functions, with some having ESTs in both whole body and antennal RNA-seq, some in one or the other, and some in neither.

The remainder of the termite IRs form two distinctive large subfamilies without close relatives in other insects. The first consists of 48 genes, IR108-155, and they form a phylogenetically separate subfamily. These genes all share the same 9-exon gene structure, which, to various extents, is also shared by the divergent IRs above and the related genes in other species. The largest subfamily of 66 genes, IR156-222, however, is largely intronless, and appears to have resulted from a retrogene that lost all of its introns. Six of these genes have subsequently acquired single idiosyncratic introns that do not match those of the 48-gene subfamily in location or phase (two in IR212, but one is shared with IR211). Most of these novel introns interrupting the coding region are at the 5' ends of the genes, and might originally have been 5' UTR introns. These last two large termite IR subfamilies are so divergent from the aphid and louse IRs, as well as all other insect IRs, that it was possible that additional highly divergent subfamilies might exist, undetected by TBLASTN searches. Several PSI-BLASTP searches were performed, using as a database all of the "nr" (non-redundant) protein database at GenBank, additional unpublished arthropod IRs, and the automated termite gene model proteins. We sought proteins that are not in the OGS column in Supplementary Table 24. This led to the discovery of IR105-107, but revealed no additional large divergent subfamilies.

If the termite only had the conserved IR lineages and IR101-107, it would have 13 IRs, a number comparable to aphid and louse. With the expansions of IR41a, IR75, the nine-exon and the intronless sets, the termite IR repertoire is unusually large. To check that comparable expansions of divergent IRs have not gone undetected in aphid and louse, the above PSI-BLASTP output was also searched for aphid and louse proteins from their automated annotations, but no new IRs were discovered. It appears then that the termite IR expansions are indeed unusual. However, the expansions of IRs in fly genomes are comparable. Drosophila IRs mostly cluster in two subfamilies (Supplementary Figure 16), with branches longer than those of most termite subfamilies. This difference might simply reflect the generally higher rates of molecular evolution in these short-generation flies.

Like the ORs and GRs, these IRs evolve by a birth-and-death process, however the preponderance of single genes per scaffold (Supplementary Table 24) indicates that many of them are old in the genome without recent duplications. The largest set of arrayed genes is ten nine-exon genes in scaffold399, although not all in tandem, while seven of the 17 IR75 paralogs are in scaffold468. The overall impression is that these expansions are old in this genome lineage, implying that large IR repertoires will be common in termites.

Unfortunately, the ligand-specificity is known for only a few IRs in Drosophila, and most of those do not have simple termite orthologs. For example, Grosjean *et al*. [115] report that DmIR84a along with IR8a is responsible for perception of phenylacetic acid and phenylacetaldehyde, but IR84a is a fly-specific receptor and has no simple termite ortholog, being related to the IR75 expansions. In Drosophila, IR75a-c along with IR8a are implicated in perception of proprionic acid. And IR75a (which is related to the IR41a expansion), IR76b (a reasonably conserved potential co-receptor), and the co-receptor IR25a form a functional receptor for phenylethyl amine. As noted in the main manuscript, the relatively low number of glomeruli in the antennal lobe (Supplementary Notes 7.5), suggests that most of these IRs function in gustation.

7.5 Olfactory glomeruli in workers

Odours are sensed by olfactory receptor neurons (ORNs) that come in contact with odorants (volatile chemical compounds) in the environment (see Supplementary Notes 7.2). In arthropods, the sensory processes (dendrites) of these ORNs reside within sensilla, often hair-like cuticular structures located on the insects' antennae. Each sensillum houses one or more ORN whose central fibers (axons) project to the brain and together form the antennal nerve.

The dendritic membranes of olfactory sensory cells each present a particular kind of chemoreceptor protein that interacts with a limited number of chemical compounds. Interaction of the chemoreceptor protein with an appropriate odorant results in electrical activity in the receptor cells [116]. Olfactory receptor proteins are each coded by a specific chemoreceptor gene and recent advances in genomic techniques have established the number of these genes for a range of animals [117]. Generally, each ORN expresses only one kind of chemoreceptor gene, which thus determines the respective neuron's range of chemical compounds by which it can be activated.

ORNs on the antennae send their axons into the antennal lobe, the primary olfactory centre in the insect brain. In the antennal lobe, sensory axons from the periphery terminate in globular structures, referred to as olfactory glomeruli. Each specific glomerulus receives input from those particular ORNs on the antenna that express the same chemoreceptor gene. Hence each glomerulus presumably represents the odour specificity of one particular chemoreceptor protein and across phyla the number of chemoreceptor genes implicated in olfaction roughly matches the number of olfactory glomeruli in most insects studied so far – e.g. 47 glomeruli and 62 OR genes in *Drosophila melanogaster* [118,119], 160–165 glomeruli [120] and 163 OR genes in honey bees [97], 60 glomeruli [121] and 79 OR genes in the mosquito *Anopheles gambiae* [90], and 259 glomeruli and 220 OR genes in the parasitic wasp *Nasonia vitripennis* [98].

It is generally assumed that the number of olfactory chemoreceptor genes expressed in an animal (or the number of olfactory glomeruli) gives some indication of the range and precision of different odours that a particular animal species can discriminate. Most natural odours are physiologically represented by the co-activation of different subsets of glomeruli, forming "odortopic" activity maps across the olfactory lobe. Here we wanted to assess the number of olfactory glomeruli in the antennal lobes of *Z. nevadensis* to see if their number matches the number of presumed olfactory chemoreceptor genes, as is the case in many other invertebrates.

Zootermopsis workers were taken from laboratory colonies 124 (n=7), 136 (n=4), and 142 (n=5). Individual termites were sexed on the basis of the shape and size of their terminal abdominal sternites [122]. Additional dissection for gonadal identification was done if necessary. Termites were decapitated and a large hole was cut into the rostral part of the head capsule, removing the mandibles and allowing fixative and embedding medium to penetrate the tissue. Heads were then fixed overnight in 4% formaldehyde in phosphate buffer (pH 6.8), rinsed in buffer and stained in 1% aqueous osmiumtetroxide for 2 hours at 4 °C and for an additional 30 minutes at room temperature. Heads were then rinsed, dehydrated, plastic-embedded (Spurr's low viscosity medium; EMS, Pennsylvania) and polymerized at 65°C. Heads were sectioned on a sliding microtome at 8µm thickness, assuring that most glomeruli were represented in at least three consecutive sections. Sections were mounted, cover-slipped and photographed (SPOTflex digital camera, Zeiss Axioplan microscope). Images of individual sections were then aligned manually (Adobe Photoshop CS3) or automatically (using open source

software Fiji for ImageJ; http://fiji.sc/wiki/index.php/Fiji) and stacked (Adobe Photoshop CS3). For counting glomeruli, each glomerulus' cross-section was marked and compared with the previous and subsequent sections to assure that each glomerulus was only counted once (Supplementary Figure 17c-e).

Glomeruli were spherical or slightly elongate and the size of individual glomeruli varied considerably, from small (long axis / short axis = 12µm / 10µm; approximate volume 600µm^3) to several large ones (40µm / 30µm; volume ca. 19,000µm^3), with one particularly large glomerulus (50µm / 50µm; volume 65,000µm^3) that seemed to be from a fusion of several smaller ones (and which we counted as a single glomerulus). The majority of glomeruli were about 20µm in diameter (approximate volume 4,000µm^3).

Most of the glomeruli were well separated and unambiguously identifiable, but a few glomeruli were very close to each other and not always easy to discriminate (see arrow in Supplementary Figure 17d). While such ambiguities were generally settled by repeatedly comparing adjoining sections, in some cases they could not be resolved and led to variation in the number of glomeruli counted across individuals (or within individuals comparing the left and right antennal lobe). We evaluated 15 brains (7 female, eight male workers), but in some of these only one antennal lobe was analysed. Overall, we counted 72.9 ± 4.15 (mean ± s.e.) glomeruli in male workers (13 antennal lobes counted) and 72.0 ± 3.9 glomeruli in female workers (11 antennal lobes counted). We also compared the number of glomeruli in the right and left antennal lobes, which served as a control under the assumption that the number of glomeruli should be the same on the left and right side of the brain. We counted 70.8 ±2.8 glomeruli for the left antennal lobes (n=12) and 74.1 ± 4.4 glomeruli for the right antennal lobes (n= 12). This difference thus indicates the error margin of our analysis. None of the differences were statistically significant (Student's t-test), suggesting that the left and right antennal lobes and, more importantly, male and female worker brains are comprised of the same number of glomeruli.

The number of glomeruli found in *Z. nevadensis* is in the upper range of glomeruli found in most insect families in general (e.g. 44 in the sawfly *Neodiprion* [123], 48 in *Drosophila melanogaster* [124], 61 in the mosquito *Anopheles gambiae* [121], 63 and 67 in the moths *Manduca sexta* [125] and *Mamestra brassicae* [126], respectively, 62 in the butterfly *Pieris brassicae* [126]). However, *Zootermopsis* workers have fewer glomeruli compared to the related cockroaches, and other insects in the superorder Dictyoptera (109 in *Blaberus craniifer* [127] and 125 in *Periplaneta americana* [128]), probably suggesting that *Z. nevadensis*' narrow dietary specialization on wood and insular lifestyle reduced the necessity for discriminating a wide range of odorants, as is indispensable for the wide ranging and omnivorous cockroaches.

8 Immune genes

For the identification of immune-related genes in the *Z. nevadensis* genome, we utilized protein sequence datasets of known immune-related insect genes described in the literature and in the ImmunoDB database [129] (http://cegg.unige.ch/Insecta/immunodb). We selected genes covering all aspects of immunity, including pattern recognition (i.e. GNBP and BGRP), modulation, signalling pathways (both Toll and IMD), signal transduction, NF-kappaB pathway, melanisation immune response and known effector molecules. Deduced protein sequences of genes known to be part of the above pathways were selected from different neopteran insect species, such as Diptera, Lepidoptera and Coleoptera, and non-neopteran insects, such as the phylogenetically more distantly related *Thermobia domestica* belonging to the order Thysanura. BLAST searches against the predicted *Z. nevadensis'* proteins were conducted on a local server using the National Center for Biotechnology Information (NCBI) BLASTALL program and the Seqtools 8.4 program (http://www.seqtools.dk/). In addition, BLAST searches (TBLASTN) of unannotated genomic scaffolds were performed to obtain immune-related and effector genes not found in the gene prediction models.

All positive hits were manually curated, testing both for correct open reading frames and completeness of the coding sequences (i.e. truncations). Identified *Z. nevadensis'* immune-related candidate genes were assigned putative functions and pathways based on three criteria: BLAST matches (e-value and bitscore) to known immune genes, identification of conserved domains utilizing InterPro, and identification of conserved amino acid residues such as correctly spaced cysteine residues in the antimicrobial peptide defensin. Immune genes are listed in Supplementary Table 25.

For most of the immune gene families, their copy numbers in *Z. nevadensis* lie in the range of other insects. The only exception is the GNBP (gram negative binding protein) family. GNBPs are characterized by the Pfam domain PF00722 (Glyco_hydro_16). We found six GNBPs in *Z. nevadensis*, while none in *P. humanus*, one in *A. pisum*, two in the bee and ants, three in *N. vitripennis*, *T. castaneum* and *D. melanogaster*, and ten in *D. pulex*. All these proteins are composed of a single copy of the Pfam domain, except in *A. pisum* where two additional N-terminal domains WSC (PF01822 carbohydrate binding domain) are observed. Among the six termite proteins, three are almost adjacent (Znev_03257, Znev_03259, Znev_03260) and the remaining two (Znev_00932, Znev_00933) are directly adjacent. Regarding the expression levels, Znev_03257 is over-expressed in workers, while Znev_03259 and Znev_03260 are not differentially expressed. Znev_02878 is over-expressed in male reproductives. Znev_00932 and Znev_00933, despite their adjacency, are differentially expressed in workers and female reproductives, respectively.

We conducted phylogenetic analysis by collecting available GNBP proteins in Isoptera species from Uniprot: three in the grasshopper *Locusta migratoria*, and two in several termite species (*Nasutitermes corniger*, *Reticulitermes flavipes* and *Reticulitermes virginicus*). Further, we included the unique GNBP known in the crustecean *Artemia sinica*. The resulting phylogeny (Supplementary Figure 18) shows a group of termite-specific GNBPs. This group, which includes the two isolated GNBPs in termites so far, involves Znev_03257 and Znev_03259 in *Z. nevadensis*. Accordingly, these two GNBPs seem to have arisen from a termite-specific duplication. There might have been more termite-specific duplication, as indicated by the position of Znev_02878, Znev_00932 and Znev_00933 in the topology. The close position of GNBPs from *T. castaneum* and *D. pulex* is surprising and likely an artefact of long-branch attraction. We observed only one of the three grasshopper GNBPs grouping with the

termite-specific GNBPs. The two that remain likely emerged from a duplication specific in the lineage of *L. migratoria* since GNBP1 Lm and GNBP3 Lm grouped together first, and then with Znev_03260 after the grouping of all insect GNBPs.

9 Reproductive division of labour

9.1 Insulin/insulin-like growth factor signalling pathway (IIS)

In addition to its normal metabolic functions the IIS pathway plays a major role in regulating life history traits. The IIS is highly conserved in both function and structure among organisms ranging from the nematode *C. elegans* to humans. It adjusts body functions (growth, maintenance, reproduction) of an organism in accord with resource availability [130]. In the honeybee it is also involved in worker division of labour: down-regulation of IIS signalling delays the age-related transition from nursing to foraging.

Amino acid sequences of genes from the IIS and TOR pathway from *D. melanogaster* were used with BLASTP to search for orthologous genes in the reference species. Up to five sequences were chosen if there was no difference in e-values between the top hit and these subsequent ones. To ensure the integrity of BLAST results, an additional round of BLAST was performed using the previously obtained hits as queries against the NCBI database. We found all major components of the IIS signalling pathway in *Z. nevadensis* (Supplementary Table 26). Further research needs to show whether the links between its components are differently regulated in termites compared to solitary insects, as seems to be the case in the honeybee.

9.2 Vitellogenins

Biological background

Vitellogenins (Vgs) are glycolipoproteins, first discovered in the silkmoth *Hyalophora cecropia* [131]. Vgs are a precursor to the yolk protein vitellin that forms the egg yolk from which the embryo is nourished. Vgs are able to function as storage proteins in males as well [132]. Recent studies in the honey bee *A. mellifera* showed that Vgs also function in the organisation and coordination of caste-specific behaviour and that interacts with juvenile hormone [133-135] It has also been implicated in being a key factor in the regulation of the queen's longevity [136].

Domain architecture, copy number and phylogeny

Vgs are characterized by the presence of three Pfam domains: Vitellogenin_N at the N-terminal region (PF01347), a von Willebrand growth factor type D domain (PF00094) and a C-terminal domain of unknown function DUF1943 (PF09172). In our study, the domain-based and orthoMCL definition of the Vg family do not cover all members. The main reason is the divergence of some domains beyond Pfam detection thresholds. We investigated proteins with at least two of the three characteristic domains by searching for a divergent occurrence of the third ones (lowering the recommended thresholds). This way we identified the following number of gene copies in the reference species: one in the fruit fly and Hymenoptera (bee and ants), two in all remaining Neoptera species, and eleven in the crustacean *D. pulex*. *Z. nevadensis'* genome contained four Vgs; Znev_07681 and Znev_07682 were adjacent and had a complete domain architecture, and Znev_06771 and Znev_08605 which both had a divergent C-terminal DUF1943 domain. In previous studies, Vgs have been reported in other Blattodea: two copies in the American cockroach *Periplaneta americana*, two proteins in the Madeira cockroach *Rhyparobia maderae* and one protein in the German cockroach *Blattella germanica* [137]. To construct protein trees, we utilized these sequences from the Uniprot database and added a termite Vg from *Cryptotermes secundus*, and a protein from *Locusta migratoria* which seems to be a fusion of a

Vg and another protein.

In the phylogenetic analysis (Supplementary Figure 19), we observed two groups of sequences: on the one side the highly divergent Vgs of the pea aphid, the fly, *Locusta migratoria* and *Z. nevadensis* protein Znev_06771, on the other side the most conserved Vgs, with *D. pulex* paralogs as outgroups and Znev_08605 preceding all other insect Vgs. Hence, the two adjacent Vgs Znev_07681 and Znev_07682 seem to have emerged from a recent duplication in the Blattodea ancestor. Znev_07681 groups with Vit-2 of *P. americana* and the Vgs of *B. germanica* and *R. maderae*. Znev_07682 groups with the Vit-1 of *P. americana* and the Vg of the termite *C. secundus.* The globally low agreement of the gene tree with the species tree can be explained by the extreme divergence of the proteins but also by a very fast evolution with multiple duplications and losses as suggested by the presence of only species-specific paralogs.

Finally, Znev_07681 and Znev_07682 were clearly over-expressed in neotenic female reproductives, both being in cluster 43 which is composed of 14 hyper-specific genes expressed at this stage (93-94% for FNR). These two are the best candidates for "true" vitellogenins used in eggs. Znev_08605 is also over-expressed in neotenic female reproductives. Znev_06771 does not show differential expression.

Vitellogenin receptors

Vitellogenin receptors (VgRs) are a specialized subfamily of the low-density lipoprotein receptor (LDLR) superfamily that transports external lipids into a recipient cell [138]. VgRs also have been implicated in transovarial transmission of the Babesia parasite in the ixodid tick, *Haemaphysalis longicornis* Neumann [139]. The first vitellogenin receptor in a hemimetabolous species was characterized in the cockroach *Periplaneta Americana* by [140]. The authors further scrutinized VgRs and lipophorin receptors (LpRs), another subfamily of LDLRs, using two additional cockroach species and other model insects [141]. Insect VgRs are characterized by a multi-domain architecture composed of:

1. A first region of ligand binding domains (LBD) that consists of repeats of low-density lipoprotein receptor domain class A (PF00057 in Pfam named Ldl_recept_a);
2. A first calcium-binding EGF domain (PF07685 in Pfam named EGF_CA);
3. Repeats of low-density lipoprotein receptor domain class B (PF00058 in Pfam named Ldl_recept_b);
4. A second region of LBD (repeats of the low-density lipoprotein receptor domain class A);
5. A final calcium-binding EGF domain.

We identified one putative VgR in the *Z. nevadensis* genome, Znev_02120, of which the gene model has been refined. The corrected protein sequence exhibits a similar domain architecture to model insect VgRs. Interestingly, almost all insect VgRs contain a LI internalization signal close to the C-terminal end, while LpRs and a majority of other LDLRs display a NPXY motif. Since only *Solenopsis invicta* VgR displayed both an NPXY and a LI motif, Tufail and Takeda [137] highlighted the interesting case of an NPTF motif in the three cockroaches and suggested it as a candidate for an alternative internalization signal. Interestingly, *Z. nevadensis* VgR has a different motif when aligned to well-conserved cockroach sequences: NPAF. This suggests a relaxed constraint on the third position of the motif (NPXF) similar to the one observed in LpRs and supports this motif as a candidate functional internalization signal.

9.3 JHIII Biosynthetic Pathways and Juvenile Hormone Binding Proteins

Juvenile hormones (JHs) control growth, development, metamorphosis, and reproduction in insects. These sesquiterpenoids are synthesized *de novo* in specialized endocrine glands called the corpora allata (CA) [142]. In social insects, JH additionally regulates other developmental processes such as polyphenisms and caste determination [143,144]. There are several forms of JH that have been fully characterized chemically and physiologically, with JH III found in termites as well as other insects. The biosynthetic pathway of JH III is divided into two parts:

1. The initial steps of JH biosynthesis follow the isoprenoid pathway[2] from acetyl-CoA to farnesyl pyrophosphate (FPP) [145].
2. In the later part, FPP is hydrolyzed by a pyrophosphatase to farnesol, which is then oxidized to farnesal and farnesoic acid (FA) [146-148]. The order of the final two enzymes in JH synthesis differs depending on the insect. In Lepidoptera, epoxidation by the P450 monooxygenase precedes esterification by a juvenile hormone acid methyltransferase (JHAMT) [149]. In Orthoptera, Dictyoptera, Coleoptera and Diptera, epoxidation follows methylation [150-154].

To identify and assign functionality to *Z. nevadensis* genes involved in the JH biosynthetic pathway, we used as a reference each enzyme from the insect JH pathway so far studied [155]. Using the amino-acid sequences of the enzymes in reference species (*D. melanogaster*, *A. aegypti* and *A. gambiae*), we identified orthologs and recent paralogs in the *Z. nevadensis* lineage using reciprocal BLAST searches. These assignments were confirmed using the orthoDB database and orthoMCL ortholog groups (see Supplementary Notes 3.2). Finally, gene models were refined when required and possible by analysing *Z. nevadensis* assembly to match corresponding insect genes using the Apollo Genome Annotation and Curation Tool [156]. The annotation allowed the assignment of putative functions to the complete list of the enzymes of the JH III biosynthetic pathway (see Supplementary Table 27).

1. Every enzyme involved in the isoprenoid pathway from acetyl-CoA thiolase to FPP was characterized, which includes: an acetoacetyl-CoA thiolase, two hydroxymethylglutaryl-CoA synthases (HMG-S), a hydroxymethylglutaryl-CoA reductase (HMG-R), a mevalonate kinase, a phosphotransferase (that phosphorylates mevalonic acid to 5-phosphomevalonic acid), an isopentenyl-diphosphate delta-isomerase (that catalyses the interconversion of isopentenyl diphosphate (IPP) and dimethylallyl diphosphate (DMAPP)), and a farnesyl pyrophosphate synthetase (a prenyltransferase that catalyses the condensation of isopentenyl pyrophosphate with geranylpyrophosphate to FPP, which is the last product of the isoprenoid pathway; [155]).
2. Following the isoprenoid pathway, two metabolic routes are possible, i.e., the sterol branch and the JH branch, of which the sterol branch has been lost in insects. Insects and other arthropods do not synthesize cholesterol *de novo* [157], due to the absence of genes encoding squalene synthase and other subsequent enzymes of the sterol branch [155]. No squalene synthase homologs were found in our search of the *Z. nevadensis* genome suggesting that only the JH branch exists in this termite. On the other hand, the JH branch, which spans from FPP to JH III, appears to be conserved in *Z. nevadensis* since we found matches for all five major enzymes of the JH branch. However, it should be noted that no absolute functional assignments (based on their demonstrated synthesis of the precursors of JH III) are currently available for insect farnesyl diphosphate pyrophosphatase, farnesol oxidase, and farnesal dehydrogenase (marked with an asterisk in Supplementary Table 27).

[2] Also called *mevalonate pathway* by opposition to the alternative *non-mevalonate pathway* in plants and apicomplexan protozoa that produce isoprenoids (terpenoids) in the plastids.

9.4 Neofem genes

Amino acid sequences of Neofem1-4 (ABN05619.1, ABN05620.1, ABN05621.1, ABN05622.1) were used with BLASTP to search for orthologous genes in the reference species. Up to five sequences were utilized if there was no difference in their e-values from that of the top hit. To ensure the integrity of BLAST results the found sequences were re-mapped to the NCBI database with an additional round of BLAST searches. The phylogenies (Supplementary Figure 20) were obtained following the procedure described in Supplementary Notes 3.3.

9.5 Neuropeptides

Understanding the endocrinology of caste differentiation has been challenging because of our poor understanding of termite hormones. In non-social insects, development, growth, and even behaviour is controlled by neuropeptides and interactions between the insect-exclusive sesquiterpenoid, juvenile hormone (JH; see Supplementary Notes 9.3), and the hexa-hydroxylated steroid, 20-hydroxyecdysone [158,159]. In a search of the *Zootermopsis nevadensis* genome, 27 neuropeptides were found (Supplementary Table 28). Also of great interest is the presence of neuropeptides in termite reproductives which previously have only been associated with the regulation of insect moulting. This could be a previously unknown function for this fundamental pathway.

9.6 Biogenic amine receptors

In insects, the biogenic amines, octopamine, tyramine, dopamine and serotonin, modulate a number of processes including metabolism [160], sensory processing [161-163], learning and memory [164-167] and locomotion [168,169]. In addition, the mechanisms underlying signalling via biogenic amines is of particular interest for those working on eusocial insects, as they have been implicated in division of labour [170,171], reproductive behaviour [172-176], responses to pheromonal cues [177] and nestmate recognition [178]. Although the evolution of some receptor subtypes predates the protostomia/deuterostomia split [179,180], the lack of high levels of octopamine in the mammalian nervous system has led to production of insecticides that target the octopamine receptor [181]. Thus, characterization of the octopamine receptors of pest species such as termites is of particular interest.

The genome of *Z. nevadensis* contains putative orthologs of each category of biogenic amine receptors found in *Drosophila melanogaster* (see Supplementary Table 29). *Z. nevadensis* does not appear to have a close ortholog of the type II tyramine receptor found in *Drosophila* (CG16766); however, this receptor subtype appears to be missing from other insect genomes as well [99]. Interestingly, type 1 tyramine receptor (Tyr1) appears to have undergone recent gene duplication in *Z. nevadensis*. Not unexpectedly, the two Tyr1 orthologs are closely related to the cockroach (*Periplaneta americana*) tyramine receptor [182].

9.7 Histone modifying enzymes

We identified histone acetyltransferases, deacetylases, methyltransferases, and demethylases by performing BLASTP searches using orthologous protein sequences from human and Drosophila databases on Uniprot. The orthologs were identified as follows: 1) homology search with BLASTP with $E < 10^{-5}$; 2) in case of multiple matches for each gene, selection was based on most homologous match; 3) regions with less than 30% identity with the query protein were excluded.

Histone acetyltransferases (HATs) and histone deacetylases (HDACs) are involved in the regulation of gene expression through chromatin modification by adding and removing acetyl groups on histone tails [183]. Based on domain organization and sequence identity, HDACs of higher eukaryotes are divided into four classes that consist of HDAC 1 through 11 and the NAD^+ dependent sirtuin family of proteins (sirtuin 1-7) [184]. Fewer have been found in eusocial hymenopterans, ranging from 4-5 HDACs and six sirtuins in honeybees [17] and ants [18]. We identified six HDACs and seven sirtuins in *Z. nevadensis* (Supplementary Table 30). Although sirtuins have been associated with NAD^+ dependent deacetylase activity with functions in lengthening lifespan [185], they presumably have many different functions not limited to histone deacetylation [186]. Sirtuin 7 has a very limited deacetylation activity, but is involved in the regulation of RNA polymerase I activity. We found that sirtuin 6 and 7 are significantly overexpressed across all the female reproductive stages (Supplementary Figure 21 and Supplementary Figure 22). Interestingly, an up-regulation of sirtuin deacetylases is associated with *H. saltator* reproductives [18]. Whether this indicates a common mechanism involved in increasing lifespan of reproductive castes in social insects or these are unrelated coincidental findings will be an interesting question for future studies. We also identified twelve orthologs of HATs in *Z. nevadensis* (Supplementary Table 31). However, we did not find any correlation of their gene expression with caste or life stage.

Histone methylation is involved in the regulation of transcriptional activation and repression [187]. Although most of the methylated histones are associated with repression, some of these histone modifications have also been found in the context of transcription activation. We identified homologs for lysine-specific histone demethylases (KDM4C, KDM5, KDM6A and KDM7; see Supplementary Table 32). KDM4C is significantly overexpressed in female reproductive castes (Supplementary Figure 23). These lysine-specific histone demethylases are involved in processes including stem cell renewal and germ cell development [188].

The role of the upregulation of several histone acetyl- and methyltransferases in reproductive females is unclear (Supplementary Figure 24 and Supplementary Figure 25).

9.8 Elongases and desaturases

Cuticular hydrocarbon profiles are correlated with reproductive status in many eusocial insects including *Z. nevadensis* [189] and have been shown to play an important role as fertility signals in the regulation of reproduction in some of these species (reviewed e.g. in [190-193]). Despite their importance, very little is known about the relationship between reproductive status and biosynthesis. Candidate genes involved in the biosynthesis of hydrocarbons are elongases and desaturases [194]. We identified these genes in *Z. nevadensis* and then analysed their expression patterns across the life stages and castes investigated.

Using *D. melanogaster* proteins (Desat1 and Desat2), a BLASTP search against *Z. nevadensis* proteome results in ten putative desaturase-like proteins. Desaturases are characterized by a specific domain architecture involving a single Pfam domain FA_desaturase (Fatty acid desaturase – PF00487). In two of the ten proteins, the domain has not been confirmed: the domain end seems highly divergent or partially lost in Znev_01236, and the gene model of Znev_18502 might be incomplete due to the cut off at the end of the scaffold. Using TBLASTN searches, we identified a candidate locus for another desaturase protein at the end of scaffold1789. The prediction of this gene revealed an incomplete

protein which likely corresponds to the missing beginning of the previously mentioned Znev_18502. Since no complete gene model could be built from these gapped scaffolds, Znev_18502 has not been included in the phylogenetic reconstructions. However, this protein and Znev_01236 are good candidates for desaturase proteins in *Z. nevadensis*, since they are tandem duplicates of two complete desaturases (belonging to the eight with a complete/detected domain).

Phylogenetic reconstructions were conducted to distinguish the sub-families of desaturases in the reference species and *Z. nevadensis* (Supplementary Figure 26). First, we observed an intermediate number of desaturases in the termite, compared to the larger range (~14-15) found in *T. castaneum*, *N. vitripennis* and *H. saltator*, and to the reduced sets (~6-7) of *D. pulex*, *P. humanus*, *D. melanogaster* and *A. mellifera*. Second, most of the reference species exhibit several lineage-specific duplications while the termite exhibits just one ortholog for each sub-family. This suggests a dynamic evolution of desaturases, while it also confirms the presumed relatively limited divergence of termites from their Neoptera ancestor.

Using the eloF protein from *D. melanogaster*, a BLASTP search resulted in 16 putative elongase-like proteins in *Z. nevadensis*. These proteins are characterized by a single Pfam domain ELO (GNS1/SUR4 family - PF01151), which is not found in any other termite protein. The phylogenetic analysis (Supplementary Figure 27) demonstrated that *Z. nevadensis* displays an ortholog for each of the insect elongase families. Most interesting is the duplication of James Bond proteins in *Z. nevadensis* and the pea aphid only, with three and six copies, respectively, while all other species had just one. In *Z. nevadensis*, two of the three James Bond elongases (Znev_14342 and Znev_14343) are adjacent and at the end of a scaffold, while the third (Znev_04596) is also at the border of a scaffold. Given domain information and BLAST alignments, Znev_14343 and Znev_04596 might actually be fused into a unique protein (by fusing scaffolds) which would lead to only two James Bond elongases in *Z. nevadensis* that arose from tandem duplication.

Additionally, the scaffold that contains Znev_14342 and Znev_14343, also contains three other elongases in a reduced genomic region: Znev_18616, Znev_14337 and Znev_14336 (this later seems incomplete but a TBLASTN search indicated a good candidate for completeness by fusing with contig C13861411). These three proteins emerged from recent duplication according to the phylogenetic reconstruction. Another cluster of three elongases could be found in *Z. nevadensis* (Znev_07404, Znev_07408 and Znev_07409). Such groups of tandemly repeated elongases could be found in all reference species, suggesting a potential evolutionary/regulatory advantage.

The expression analysis revealed reproductive-specific expression of the fused elongase genes Znev_14343/Znev_04596 and the desaturase Znev_14774 (Supplementary Figure 28).

10 Caste differentiation

10.1 P450s

P450 genes are of particular interest in insects because they detoxify many xenobiotics, and some have evolved to confer pesticide resistance. Furthermore, some genes of the cytochrome P450 family 4 (CYP4) are thought to be important for caste development in some termite species. The gene Neofem4 is supposed to be responsible for the regulation of soldier development in colonies of *Reticulitermes flavipes*.

The InterPro entry for P450 (IPR001128) comprises domain signatures from 3 different databases: Pfam, Gene3D and SUPERFAMILY. We considered all proteins of *Z. nevadensis* and the reference species with this InterPro entry as P450 candidates. It results in 76 proteins in *Z. nevadensis*, which is intermediate to the 39 of the body louse and the 136 of *C. floridanus*, but is equivalent to the number observed in most arthropod species (e.g. 76 in *D. pulex* and 85 in *D. melanogaster*).

Subfamily classification of the P450s, based on Nelson's works [195,196], was determined through BLAST searches using all candidate genes as queries against all arthropod sequences in NCBI nr. Only the best five hits were retained and the GenBank IDs were used to extract the description that contains the family using the Entrez package from the Biopython Project. Unambiguous annotations were directly assigned to the proteins while ambiguous ones and non-annotated genes were checked manually.

After noticing some genes with very low similarity to any others, we realized that many candidate genes have less than half of the domain (shorter than 250 amino acids while the average size in arthropods is more than 500). For some observable cases, Znev_14301/Znev_14302 or Znev_13890/Znev_13891 and Znev_13892/Znev_13893, adjacent genes contain complementary parts of the domain and the termite genome browser shows transcripts spanning both genes and the intergenic region, suggesting these gene models need correcting. Such faulty gene models are due to the usage of ant gene models (*H. saltator* and *C. floridanus*) in which large numbers of P450s and fragmented occurrences have been reported (Supplementary Notes 3.4).

Finally, of the 76 termite P450 genes, 55 have at least one complete P450 domain. The largest numbers are found in families 4, 6 and 9. In many genome papers, the numbers of CYP4 genes are reported because of expansion or contraction of this family and its special significance in terms of insecticide resistance. We observe 20 genes annotated as CYP4 in *Z. nevadensis*, of which three only have domain fragments (Supplementary Table 33). This number is similar to that found in Drosophila (21) but distinct from the expansion observed in the ant *C. floridanus* (53), or the contraction in the honey bee (6).

10.2 Hexamerins

Five hexamerin proteins have been identified in *Z. nevadensis*: Znev_05598 and the four adjacent proteins Znev_04925, Znev_18795, Znev_18796 and Znev_18797. These proteins exhibit specific tridomain architecture (PF03722-PF00372-PF03723, called Hemocyanin N, Hemocyanin M and Hemocyanin C) which are specific to hexamerins. We collected hexamerin orthologs in the reference species using orthoMCL clusters: ten proteins in *Tribolium*, seven in *Drosophila*, four to five in ants

and wasp, two to three in body louse, bee and pea aphid and, interestingly, only one in *Daphnia* (none in the worm). Almost all of these proteins have the expected domain architecture, except for two proteins of *T. castaneum* (TC015848 and
TC014908) and one of *C. floridanus* (Cflo_12692) that have been split in the middle of the central domain and only consist of the end of the usual architecture, and one protein in *N. vitripennis* (NV12559) that has an additional N-terminal domain (PF00201 - UDP glycosyltransferase). We then collected additional proteins from Blattodea (*Periplaneta americana*, *Blaptica dubia* and *Cryptotermes secundus*) using the hexamerins previously described in *R. flavipes* [197] as a query for a BLASTP search against the NCBI-nr database.

From our phylogenetic analysis (Supplementary Figure 29), we first observe that Znev_18795 and Znev_04925 form two ortholog groups with the two hexamerins described in the two roaches and the termite *C. secundus*. The absence of any other insect hexamerin before their common ancestor suggests that these two sub-families arose from a tandem duplication in the roach-termite ancestor. Moreover, the topology clearly supports the orthology of Znev_18796 and Znev_18797 with the two *R. flavipes* hexamerins known to be involved in the soldier differentiation process. These two sub-families seem specific to termites so far but share a common ancestry with the two previously mentioned (all four are adjacent) and cluster within one group of insect hexamerins. The hexamerin in *Z. nevadensis*, Znev_05598, is closer to the second group of insect hexamerins.

11 DNA methylation

11.1 The DNA methylation machinery

DNA methyltransferases (DNMTs) are a conserved family of enzymes characterized by the presence of a Pfam domain DNA methylase (PF00145). Their role involves the covalent addition of methyl groups to the 5'-carbon of cytosine, which is an important process in epigenetic regulation of gene expression. In mammals, three active DNMTs have been described: DNMT1 is responsible for maintaining DNA methylation patterns across cell generations. DNMT3A and DNMT3B are the active methyltransferases in *de novo* methylation while DNMT3L is an accessory protein. Previously described DNMT2 is not a DNA methyltransferase, but methylates $tRNA^{Asp}$.

In the termite *Z. nevadensis*, DNMT1 (Znev_18516) and DNMT3 (Znev_11906) are present,as is the case in *D. pulex* and the three Hymenoptera from our reference species. However, copy number and gene presence varies among reference species. DNMT1 is absent from *D. melanogaster* and has been duplicated in *N. vitripennis* (three copies), *A. pisum, P. humanus* and *A. mellifera* (two copies). DNMT3 is absent from *P. humanus*, *T. castaneum*, *D. melanogaster* and *H. saltator,* according to orthoMCL clustering and Nasonia genome sequence paper [16] OrthoMCL did not find evidence of a DNMT3 homolog in *H. saltator*, although DNMT3 is present according to Bonasio et al [18] and our domain analysis. We found the termite DNMT3 clustered with the DNMT3 of *N. vitripennis*, *C. floridanus*, *A. pisum, D. pulex* and *A. mellifera* (the honey bee has two copies). Moreover, we detected an additional fragment of a DNA methylase domain in another termite protein (Znev_06587, 94 amino-acid long, not clustered by OrthoMCL). A phylogenetic analysis revealed that this domain is likely to be a partial and recent duplication of the DNMT3 protein. Further efforts of completing the gene model of the fragmented protein at the scaffold level did not succeed. However, it revealed a duplication and an inversion of ~760 nucleotides that include this fragmented DNMT3. Separated by 935 nucleotides and with expression support, this could indicate a putative third protein (fragmented) of DNMT3 in *Z. nevadensis*.

11.2 DNA methylation and gene expression

Overall GC ratio of the whole genome assembly is 38.18%, which is similar to other insects. But the range of the GC ratios (window size 5K, Supplementary Figure 30) is narrower than other insects. We used the specific distribution of CpG dinucleotides (5' – 3' cytosine followed by guanine) to determine putative DNA methylation levels. Our computational methods to determine the putative level of genome-wide DNA methylation rely on the fact that DNA methylation predominantly targets CpG dinucleotides in animal genomes. Because methylated cytosines undergo spontaneous deamination to thymine with high frequency [198], depletion of normalized CpG content (CpG o/e) represents a reliable evolutionary signature of DNA methylation in animal genomes.

Normalized CpG dinucleotide content (CpG o/e) was calculated using the equation:

$$CpG\ o/e = length \times \frac{CpG\ count}{C\ count \times G\ count}.$$

For visualization of CpG o/e distributions, all CpG o/e values equal to zero were dropped. GO annotation and enrichment was determined using Blast2GO [199]. The threshold used to differentiate

putatively methylated (low CpG o/e) and putatively unmethylated (high CpG o/e) genes for analysis of functional enrichment was the mean CpG o/e value across all gene bodies (CpG o/e = 0.616).

The expression level measure used for analysis with CpG o/e was calculated as the mean of expression level (RPKM) as measured across all *Z. nevadensis* samples. The morph specificity index used for analysis with CpG o/e was calculated in the same manner as a previously published tissue-specificity index [200], but using whole-organism samples rather than tissues, as:

$$Morph\ specificity\ index = \frac{\sum_{j=1}^{n} 1-(E_j/E_{max})}{n-1},$$

where n is the number of morphs, E_j is the expression level of the gene in the jth tissue and E_{max} the maximum expression level of the gene across the n morphs.

We observed a parabolic relationship between gene expression level and the signature of intragenic DNA methylation in *Z. nevadensis,* where moderately expressed genes appear to be most highly methylated (Supplementary Figure 31A). Such a relationship has been observed in several eukaryotes [201]. This highlights the fact that intragenic DNA methylation is generally not associated with transcriptional repression in eukaryotes.

DNA methylation levels of putative promoter regions (upstream of *Z. nevadensis* genes) exhibited a similar relationship with gene expression as intragenic DNA methylation (Supplementary Figure 31C). This suggests that promoter methylation is not widely associated with gene repression in insects or alternatively, that indications of promoter methylation in *Z. nevadensis* may be driven by incomplete gene annotation.

We calculated the variation in gene expression among the 13 different *Z. nevadensis* morphs leading to a metric of "morph specificity". High values of morph specificity indicate that a particular gene tends to be expressed in relatively few morphs, whereas low levels of morph specificity are indicative of genes expressed relatively equally among morphs. We found that morph specificity of gene expression was positively correlated with CpG o/e (Supplementary Figure 31D), indicating that DNA methylation is preferentially targeted to genes ubiquitously expressed among morphs and stages in *Z. nevadensis*, as in other investigated insects [202,203].

Analysis of phylogenetic conservation was performed with BLAST searches (TBLASTX, e-value threshold: 10^{-50}) [204] between the official *Z. nevadensis* transcript set and official transcript sets from eight species of varying phylogenetic distance (*A. mellifera*, *D. melanogaster, Ciona intestinalis*, *I. scapularis*, *Danio rerio*, *Homo sapiens*, *Neurospora crassa*, and *Arabidopsis thaliana*). We assigned a number between zero and four to convey varying degrees of phylogenetic conservation, as follows: if a gene had no hits above the similarity threshold it was assigned 0, if it had hits in insects only (*A. mellifera* and/or *D. melanogaster*) it was assigned 1, if in invertebrates only (one insect as well as *I. scapulari* and/or *C. intestinalis*) it was assigned 2, if in animals (one insect, one invertebrate and *D. rerio* and/or *H.sapiens*) it was assigned 3, and if present across diverse eukaryotes (one in each previous category as well as *N. crassa* and/or *A. thaliana*) it was assigned 4.

Genes which exhibit ubiquitous expression among tissues generally evolve at a lower rate than genes with tissue-specific expression [205]. As expected given that DNA methylation was targeted to genes with

ubiquitous expression across tissues and morphs, we found that DNA methylation in *Z. nevadensis* was preferentially targeted to genes exhibiting broad phylogenetic conservation (Supplementary Figure 31). The targeting of methylation to highly conserved genes is common to diverse invertebrate taxa [203,206]. We also observed highly significant enrichment of cellular 'housekeeping' functions among putatively methylated genes, including those related to transcription and translation (Supplementary Table 35) [203,206]. Conversely, putatively unmethylated genes were enriched for many terms related to development and response to external stimuli (Supplementary Table 36) [203].

11.3 DNA methylation and alternative splicing

We first mapped all RNA-seq reads to the assembled scaffold using TopHat v1.3.3 [5]. The putative splice variants were determined by comparison of the observed junction sites with the annotated gene models (v2.2 gene set). We identified seven categories of alternative splicing events according to the classification of Wang and Burge [207]: exon skipping (ES), intron retention (IR), mutually exclusive exon (MXE), alternative 5' splice site (A5SS), alternative 3' splice site (A3SS), alternative first exon (AFE), and alternative last exon (ALE). The AS events and the genes involved for 13 termite samples are shown in Supplementary Table 37 and Supplementary Table 38, respectively. To prevent cryptic unannotated exons from being erroneously included in the intron-retention (IR) class, we only counted as *bona fide* IR events those observed in gene models for which the affected intron region had a read coverage >90%.

Analysis of CpG o/e versus the number of alternative splice events per gene was performed by calculating the number of unique alternative splice events detected across twelve samples (including exon-skipping, intron-retention, alternative 5' splice site, alternative 3' splice site, mutually exclusive exon, alternative first exon, and alternative last exon).

Supplementary References

1 Haverty, M. I., Page, M., Nelson, L. J. & Blomquist, G. J. Cuticular hydrocarbons of dampwood termites, *Zootermopsis*: Intra-and intercolony variation and potential as taxonomic characters. *J. Chem. Ecol.* **14**, 1035-1058 (1988).

2 Aldrich, B. & Kambhampati, S. Population structure and colony composition of two *Zootermopsis nevadensis* subspecies. *Heredity* **99**, 443-451 (2007).

3 Aldrich, B. T. & Kambhampati, S. Microsatellite markers for two species of dampwood termites in the genus *Zootermopsis* (Isoptera: Termopsidae). *Mol. Ecol. Notes* **4**, 719-721 (2004).

4 Booth, W. *et al.* Population genetic structure and colony breeding system in dampwood termites (*Zootermopsis angusticollis* and *Z. nevadensis nuttingi*). *Insect. Soc.* **59**, 127-137 (2012).

5 Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105-1111 (2009).

6 Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).

7 Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* **11**, R25 (2010).

8 Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139-140 (2010).

9 Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* **5**, 621-628 (2008).

10 de Hoon, M. J. L., Imoto, S., Nolan, J. & Miyano, S. Open source clustering software. *Bioinformatics* **20**, 1453-1454 (2004).

11 D'haeseleer, P. How does gene expression clustering work? *Nat. Biotechnol.* **23**, 1499-1502 (2005).

12 Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11**, R106 (2010).

13 Tautz, D. & Domazet-Lošo, T. The evolutionary origin of orphan genes. *Nat. Rev. Genet.* **12**, 692-702 (2011).

14 Adams, M. D. *et al.* The genome sequence of *Drosophila melanogaster*. *Science* **287**, 2185-2195 (2000).

15 Richards, S. *et al.* The genome of the model beetle and pest *Tribolium castaneum*. *Nature* **452**,

949-955 (2008).

16 Werren, J. H. *et al.* Functional and evolutionary insights from the genomes of three parasitoid *Nasonia* species. *Science* **327**, 343-348 (2010).

17 The Honeybee Genome Sequencing Consortium. Insights into social insects from the genome of the honeybee *Apis mellifera*. *Nature* **443**, 931-949 (2006).

18 Bonasio, R. *et al.* Genomic comparison of the ants *Camponotus floridanus* and *Harpegnathos saltator*. *Science* **329**, 1068-1071 (2010).

19 Kirkness, E. F. *et al.* Genome sequences of the human body louse and its primary endosymbiont provide insights into the permanent parasitic lifestyle. *Proc. Natl. Acad. Sci. USA* **107**, 12168-12173 (2010).

20 TIAGC, T. I. A. G. C. Genome sequence of the pea aphid *Acyrthosiphon pisum*. *PLoS Biol.* **8**, e1000313 (2010).

21 Colbourne, J. K. *et al.* The ecoresponsive genome of *Daphnia pulex*. *Science* **331**, 555-561 (2011).

22 Coulson, A. The *Caenorhabditis elegans* genome project. *Biochem. Soc. Trans.* **24**, 289 (1996).

23 Finn, R. D. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **36**, D281-D288 (2008).

24 Zdobnov, E. M. & Apweiler, R. InterProScan–an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* **17**, 847-848 (2001).

25 Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25 (2000).

26 Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A. C. & Kanehisa, M. KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res.* **35**, W182-W185 (2007).

27 Moore, A. D. & Bornberg-Bauer, E. The dynamics and evolutionary potential of domain loss and emergence. *Mol. Bio. Evol.* **29**, 787-796 (2012).

28 Hegyi, H. & Gerstein, M. Annotation transfer for genomics: measuring functional divergence in multi-domain proteins. *Genome Res.***11**, 1632-1640 (2001).

29 Katoh, K., Kuma, K., Toh, H. & Miyata, T. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.* **33**, 511-518 (2005).

30 Castresana, J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Bio. Evol.* **17**, 540-552 (2000).

31 Criscuolo, A. morePhyML: Improving the phylogenetic tree space exploration with PhyML 3.

Mol. Phylogenet. Evol. **61**, 944-948 (2011).

32 Guindon, S. *et al.* New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* **59**, 307-321 (2010).

33 Le, S. Q. & Gascuel, O. An improved general amino acid replacement matrix. *Mol. Bio. Evol.* **25**, 1307-1320 (2008).

34 Chevenet, F., Brun, C., Bañuls, A. L., Jacq, B. & Christen, R. TreeDyn: towards dynamic graphics and annotations for analyses of trees. *BMC Bioinformatics* **7**, 439 (2006).

35 Parra, G., Bradnam, K. & Korf, I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23**, 1061-1067 (2007).

36 Dorer, D. R., Rudnick, J. A., Moriyama, E. N. & Christensen, A. C. A family of genes clustered at the Triplo-lethal locus of *Drosophila melanogaster* has an unusual evolutionary history and significant synteny with *Anopheles gambiae*. *Genetics* **165**, 613-621 (2003).

37 Bhutkar, A. *et al.* Chromosomal rearrangement inferred from comparisons of 12 *Drosophila* genomes. *Genetics* **179**, 1657-1680 (2008).

38 Shah, N., Dorer, D. R., Moriyama, E. N. & Christensen, A. C. Evolution of a large, conserved, and syntenic gene family in insects. *G3(Betheseda)* **2**, 313-319 (2012).

39 Quijano, C. *et al.* Selective maintenance of *Drosophila* tandemly arranged duplicated genes during evolution. *Genome Biol.* **9**, R176 (2008).

40 Biessmann, H. Molecular analysis of the yellow gene (y) region of *Drosophila melanogaster*. *Proc. Natl. Acad. Sci. USA* **82**, 7369-7373 (1985).

41 Ferguson, L. C., Green, J., Surridge, A. & Jiggins, C. D. Evolution of the insect yellow gene family. *Mol. Biol. Evol.* **28**, 257-272 (2011).

42 Santos, K. S. *et al.* Profiling the proteome complement of the secretion from hypopharyngeal gland of Africanized nurse-honeybees (*Apis mellifera* L.). *Insect Biochem. Mol. Biol* **35**, 85-91 (2005).

43 Smith, C. D. *et al.* Draft genome of the globally widespread and invasive Argentine ant (*Linepithema humile*). *Proc. Natl. Acad. Sci. USA* **108**, 5673-5678 (2011).

44 Fischman, B. J., Woodard, S. H. & Robinson, G. E. Molecular evolutionary analyses of insect societies. *Proc. Natl. Acad. Sci. USA* **108**, 10847-10854 (2011).

45 Felsenstein, J. PHYLIP–Phylogeny Inference Package (Version 3.2). *Cladistics* **5**, 164–166 (1989).

46 Jones, D. T., Taylor, W. R. & Thornton, J. M. The rapid generation of mutation data matrices from protein sequences. *Comput. Appl.Biosci.* **8**, 275-282 (1992).

47 Saitou, N. & Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Bio. Evol.* **4**, 406-425 (1987).

48 Berglund-Sonnhammer, A. C., Steffansson, P., Betts, M. J. & Liberles, D. A. Optimal gene trees from sequences and species trees using a soft interpretation of parsimony. *J. Mol. Evol.* **63**, 240-250 (2006).

49 Letunic, I. & Bork, P. Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. *Nucleic Acids Res.* **39**, W475-W478 (2011).

50 Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Bio. Evol.* **24**, 1586-1591 (2007).

51 Suyama, M., Torrents, D. & Bork, P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.* **34**, W609-W612 (2006).

52 Darriba, D., Taboada, G. L., Doallo, R. & Posada, D. ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics* **27**, 1164-1165 (2011).

53 Predel, R., Kellner, R., Baggerman, G., Steinmetzer, T. & Schoofs, L. Identification of novel periviscerokinins from single neurohaemal release sites in insects MS/MS fragmentation complemented by Edman degradation. *Eur. J. Biochem.* **267**, 3869 (2000).

54 Shrivastava, B. & Ghosh, A. K. Protein purification, cDNA cloning and characterization of a protease inhibitor from the Indian tasar silkworm, *Antheraea mylitta*. *Insect Biochem. Mol. Biol.* **33**, 1025-1033 (2003).

55 Nakashima, N., Kawahara, N., Omura, T. & Noda, H. Characterization of a novel satellite virus and a strain of Himetobi P virus (*Dicistroviridae*) from the brown planthopper, *Nilaparvata lugens*. *J. Invertebr. Pathol.* **91**, 53-56 (2006).

56 Geuking, M. B. *et al.* Recombination of retrotransposon and exogenous RNA virus results in nonretroviral cDNA integration. *Science* **323**, 393-396 (2009).

57 Katzourakis, A. & Gifford, R. J. Endogenous viral elements in animal genomes. *PLoS Genet.* **6**, e1001191 (2010).

58 Drezen, J. M. *et al.* The few virus-like genes of *Cotesia congregata* bracovirus. *Arch. Insect Biochem. Physiol.* **61**, 110-122 (2006).

59 Friedman, R. & Hughes, A. L. Pattern of gene duplication in the *Cotesia congregata* Bracovirus. *Infect. Genet. Evol.* **6**, 315-322 (2006).

60 Bézier, A. *et al.* Polydnaviruses of braconid wasps derive from an ancestral nudivirus. *Science* **323**, 926-930 (2009).

61 Strand, M. R. & Burke, G. R. Polydnaviruses as symbionts and gene delivery systems. *PLoS*

Pathogens **8**, e1002757 (2012).

62 Belokobylskij, S. A. Two new Oriental genera of Doryctinae (Hymenoptera, Braconidae) from termite nests. *J. Nat. History* **36**, 953-962 (2002).

63 Velarde, R. A., Sauer, C. D., O Walden, K. K., Fahrbach, S. E. & Robertson, H. M. Pteropsin: A vertebrate-like non-visual opsin expressed in the honey bee brain. *Insect Biochem. Mol. Biol.* **35**, 1367-1377 (2005).

64 Thorne, B. L., Haverty, M. I., Page, M. & Nutting, W. L. Distribution and biogeography of the North-American termite genus *Zootermopsis* (Isoptera: Termopsidae). *Ann. Entomol. Soc. Am.* **86**, 532-544 (1993).

65 Lorick, K. L. *et al.* RING fingers mediate ubiquitin-conjugating enzyme (E2)-dependent ubiquitination. *Proc. Natl. Acad. Sci. USA* **96**, 11364-11369 (1999).

66 Kaplan, Y., Gibbs-Bar, L., Kalifa, Y., Feinstein-Rotkopf, Y. & Arama, E. Gradients of a ubiquitin E3 ligase inhibitor and a caspase inhibitor determine differentiation or death in spermatids. *Dev. Cell* **19**, 160-173 (2010).

67 House, C. M. *et al.* A binding motif for Siah ubiquitin ligase. *Proc. Natl. Acad. Sci. USA* **100**, 3101-3106 (2003).

68 Germani, A. *et al.* SIAH-1 interacts with alpha-tubulin and degrades the kinesin Kid by the proteasome pathway during mitosis. *Oncogene* **19**, 5997-6006 (2000).

69 Wang, S. H., Zheng, H. L., Esaki, Y., Kelly, F. & Yan, W. Cullin3 is a KLHL10-interacting protein preferentially expressed during late spermiogenesis. *Biol. Reprod.* **74**, 102-108 (2006).

70 Arama, E., Bader, M., Rieckhof, G. E. & Stellar, H. A ubiquitin ligase complex regulates caspase activation during sperm differentiation in *Drosophila*. *PloS Biol.* **5**, 2270-2285 (2007).

71 Yan, W., Ma, L., Burns, K. H. & Matzuk, M. M. Haploinsufficiency of kelch-like protein homolog 10 causes infertility in male mice. *Proc. Natl. Acad. Sci. USA* **101**, 7793-7798 (2004).

72 Carthew, R. W. & Rubin, G. M. *seven in absentia*, a gene required for specification of R7 cell fate in the *Drosophila* eye. *Cell* **63**, 561-577 (1990).

73 Della, N. G., Bowtell, D. D. L. & Beck, F. Expression of *Siah*-2, a vertebrate homologue of *Drosophila sina*, in germ cells of the mouse ovary and testis. *Cell Tissue Res.* **279**, 411-419 (1995).

74 Dickins, R. A. *et al.* The ubiquitin ligase component Siah1a is required for completion of meiosis I in male mice. *Mol. Cell. Biol.* **22**, 2294-2303 (2002).

75 Chen, Y. *et al.* Expression of Pkd2l2 in testis is implicated in spermatogenesis. *Biol. Pharm. Bull.* **31**, 1496-1500 (2008).

76 Zhou, J. Polycystins and primary cilia: Primers for cell cycle progression. *Annu. Rev. Physiol.*

71, 83-113 (2009).

77 Pelosi, P., Zhou, J. J., Ban, L. P. & Calvello, M. Soluble proteins in insect chemical communication. *Cell. Mol. Life Sci.* **63**, 1658-1676 (2006).

78 Iovinella, I. *et al.* Differential expression of odorant binding proteins in the mandibular glands of the honey bee according to caste and age. *J. Proteome Res.* **10**, 3439-3449 (2011).

79 Gotzek, D., Robertson, H. M., Wurm, Y. & Shoemaker, D. 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** (2011).

80 Su, C. Y., Menuz, K. & Carlson, J. R. Olfactory Perception: Receptors, Cells, and Circuits. *Cell* **139**, 45-59 (2009).

81 Touhara, K. & Vosshall, L. B. Sensing odorants and pheromones with chemosensory receptors. *Annu. Rev. Physiol.* **71**, 307-332 (2009).

82 Jones, W. D., Cayirlioglu, P., Kadow, I. G. & Vosshall, L. B. Two chemosensory receptors together mediate carbon dioxide detection in *Drosophila*. *Nature* **445**, 86-90 (2007).

83 Kwon, J. Y., Dahanukar, A., Weiss, L. A. & Carlson, J. R. The molecular basis of CO_2 reception in *Drosophila*. *Proc. Natl. Acad. Sci. USA* **104**, 3574-3578 (2007).

84 Lu, T. *et al.* Odor coding in the maxillary palp of the malaria vector mosquito *Anopheles gambiae*. *Curr. Biol.* **17**, 1533-1544 (2007).

85 Benton, R., Vannice, K. S., Gomez-Diaz, C. & Vosshall, L. B. Variant ionotropic glutamate receptors as chemosensory receptors in *Drosophila*. *Cell* **136**, 149-162 (2009).

86 Croset, V. *et al.* Ancient protostome origin of chemosensory ionotropic glutamate receptors and the evolution of insect taste and olfaction. *PLoS Genetics* **6** (2010).

87 Abuin, L. *et al.* Functional architecture of olfactory ionotropic glutamate receptors. *Neuron* **69**, 44-60 (2011).

88 Robertson, H. M., Warr, C. G. & Carlson, J. R. Molecular evolution of the insect chemoreceptor gene superfamily in *Drosophila melanogaster*. *Proc. Natl. Acad. Sci. USA* **100**, 14537-14542 (2003).

89 Nozawa, M. & Nei, M. Evolutionary dynamics of olfactory receptor genes in *Drosophila* species. *Proc. Natl. Acad. Sci. USA* **104**, 7122-7127 (2007).

90 Hill, C. A. *et al.* G protein coupled receptors in *Anopheles gambiae*. *Science* **298**, 176-178 (2002).

91 Bohbot, J. *et al.* Molecular characterization of the *Aedes aegypti* odorant receptor gene family. *Insect Mol. Biol.* **16**, 525-537 (2007).

92 Arensburger, P. *et al.* Sequencing of *Culex quinquefasciatus* establishes a platform for mosquito comparative genomics. *Science* **330**, 86-88 (2010).

93 Wanner, K. W. *et al.* Female-biased expression of odourant receptor genes in the adult antennae of the silkworm, *Bombyx mori*. *Insect Mol. Biol.* **16**, 107-119 (2007).

94 Tanaka, K. *et al.* Highly selective tuning of a silkworm olfactory receptor to a key mulberry leaf volatile. *Curr. Biol.* **19**, 881-890 (2009).

95 Smadja, C., Shi, P., Butlin, R. K. & Robertson, H. M. Large gene family expansions and adaptive evolution for odorant and gustatory receptors in the pea aphid, *Acyrthosiphon pisum*. *Mol. Bio. Evol.* **26**, 2073-2086 (2009).

96 Engsontia, P. *et al.* The red flour beetle's large nose: An expanded odorant receptor gene family in *Tribolium castaneum*. *Insect Biochem. Mol. Biol.* **38**, 387-397 (2008).

97 Robertson, H. M. & Wanner, K. W. The chemoreceptor superfamily in the honey bee, *Apis mellifera*: Expansion of the odorant, but not gustatory, receptor family. *Genome Res.***16**, 1395-1403 (2006).

98 Robertson, H. M., Gadau, J. & Wanner, K. W. The insect chemoreceptor superfamily of the parasitoid jewel wasp *Nasonia vitripennis*. *Insect Mol. Biol.* **19**, 121-136 (2010).

99 Smith, C. R. *et al.* Draft genome of the red harvester ant *Pogonomyrmex barbatus*. *Proc. Natl. Acad. Sci. USA* **108**, 5667-5672 (2011).

100 Wurm, Y. *et al.* The genome of the fire ant *Solenopsis invicta*. *Proc. Natl. Acad. Sci. USA* **108**, 5679-5684 (2011).

101 Zhou, X. *et al.* Phylogenetic and transcriptomic analysis of chemosensory receptors in a pair of divergent ant species reveals caste-specific signatures of odor coding. *PLoS Genetics* **8**, e1002930 (2012).

102 Swofford, D. L. *PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods)*. Vol. Version 4 (Sinauer Associates, 2003).

103 Larkin, M. A. *et al.* Clustal W and clustal X version 2.0. *Bioinformatics* **23**, 2947-2948 (2007).

104 Vosshall, L. B. & Hansson, B. S. A unified nomenclature system for the insect olfactory coreceptor. *Chem. Senses* **36**, 497-498 (2011).

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

106 Kent, L. B., Walden, K. K. O. & Robertson, H. M. The Gr family of candidate gustatory and olfactory receptors in the yellow-fever mosquito *Aedes aegypti*. *Chem. Senses* **33**, 79-93 (2008).

107 Wanner, K. W. & Robertson, H. M. The gustatory receptor family in the silkworm moth *Bombyx*

mori is characterized by a large expansion of a single lineage of putative bitter receptors. *Insect Mol. Biol.* **17**, 621-629 (2008).

108 Penalva-Arana, D. C., Lynch, M. & Robertson, H. M. The chemoreceptor genes of the waterflea *Daphnia pulex*: many Grs but no Ors. *BMC Evol. Biol.* **9** (2009).

109 Miyamoto, T., Chen, Y., Slone, J. & Amrein, H. Identification of a *Drosophila* Glucose receptor using Ca2+ imaging of single chemosensory neurons. *PloS one* **8**, e56304 (2013).

110 Robertson, H. M. & Kent, L. B. Evolution of the gene lineage encoding the carbon dioxide receptor in insects. *J. Insect Sci.* **9** (2009).

111 Erdelyan, C. N. G., Mahood, T. H., Bader, T. S. Y. & Whyard, S. Functional validation of the carbon dioxide receptor genes in *Aedes aegypti* mosquitoes using RNA interference. *Insect Mol. Biol.* **21**, 119-127 (2012).

112 Sato, K., Tanaka, K. & Touhara, K. Sugar-regulated cation channel formed by an insect gustatory receptor. *Proc. Natl. Acad. Sci. USA* **108**, 11680-11685 (2011).

113 Miyamoto, T., Slone, J., Song, X. & Amrein, H. A. Fructose receptor functions as a nutrient sensor in the *Drosophila* brain. *Cell* **151**, 1113-1125 (2012).

114 Wang, L. M. *et al.* Hierarchical chemosensory regulation of male-male social interactions in *Drosophila*. *Nat. Neurosci.* **14**, 757-U392 (2011).

115 Grosjean, Y. *et al.* An olfactory receptor for food-derived odours promotes male courtship in *Drosophila*. *Nature* **478**, 236-U123 (2011).

116 Hildebrand, J. G. & Shepherd, G. M. Mechanisms of olfactory discrimination: Converging evidence for common principles across phyla. *Annu. Rev. Neurosci.* **20**, 595-631 (1997).

117 Nei, M., Niimura, Y. & Nozawa, M. The evolution of animal chemosensory receptor gene repertoires: roles of chance and necessity. *Nat. Rev. Genet.* **9**, 951-963 (2008).

118 Laissue, P. P. *et al.* Three-dimensional reconstruction of the antennal lobe in *Drosophila melanogaster*. *J. Comp. Neurol.* **405**, 543-552 (1999).

119 Fishilevich, E. & Vosshall, L. B. Genetic and functional subdivision of the *Drosophila* antennal lobe. *Curr. Biol.* **15**, 1548-1553 (2005).

120 Galizia, C. G. & Menzel, R. The role of glomeruli in the neural representation of odours: results from optical recording studies. *J. Insect Physiol.* **47**, 115-130 (2001).

121 Ghaninia, M., Hansson, B. S. & Ignell, R. The antennal lobe of the African malaria mosquito, *Anopheles gambiae* - innervation and three-dimensional reconstruction. *Arthropod Struct. Dev.* **36**, 23-39 (2007).

122 Thompson, C. B. The castes of termopsis. *J. Morphol.* **36**, 495-535 (1922).

123 Dacks, A. M. & Nighorn, A. J. The organization of the antennal lobe correlates not only with phylogenetic relationship, but also life history: a basal hymenopteran as exemplar. *Chem. Senses* **36**, 209-220 (2011).

124 Stocker, R. F. The organization of the chemosensory system in *Drosophila melanogaster*: a rewiew. *Cell Tissue Res.* **275**, 3-26 (1994).

125 Rospars, J. & Hildebrand, J. Anatomical identification of glomeruli in the antennal lobes of the male sphinx moth *Manduca sexta*. *Cell Tissue Res.* **270**, 205-227 (1992).

126 Rospars, J. P. Invariance and sex-specific variations of the glomerular organization in the antennal lobes of a moth, *Mamestra brassicae*, and a butterfly, *Pieris brassicae*. *J. Comp. Neurol.* **220**, 80-96 (1983).

127 Chambille, I., Masson, C. & Rospars, J. P. The deutocerebrum of the cockroach *Blaberus craniifer* Burm. Spatial organization of the sensory glomeruli. *J. Neurobiol.* **11**, 135-157 (1980).

128 Prillinger, L. Postembryonic development of the antennal lobes in *Periplaneta americana* L. *Cell Tissue Res.* **215**, 563-575 (1981).

129 Waterhouse, R. M. *et al.* Evolutionary dynamics of immune-related genes and pathways in disease-vector mosquitoes. *Science* **316**, 1738 (2007).

130 Flatt, T. & Heyland, A. *Mechanisms of life history evolution*. (Oxford University Press, 2011).

131 Telfer, W. H. Immunological studies of insect metamorphosis II. The role of a sex-limited blood protein in egg formation by the cecropia silkworm. *J. Gen. Physiol.* **37**, 539-558 (1954).

132 Weil, T., Rehli, M. & Korb, J. Molecular basis for the reproductive division of labour in a lower termite. *BMC Genomics* **8**, 198 (2007).

133 Nelson, C. M., Ihle, K. E., Fondrk, M. K., Page, R. E. & Amdam, G. V. The gene vitellogenin has multiple coordinating effects on social organization. *PLoS Biol.* **5**, e62 (2007).

134 Hartfelder, K. & Engels, W. Social insect polymorphism: hormonal regulation of plasticity in development and reproduction in the honeybee. *Curr. Top. Dev. Biol.* **40**, 45-77 (1998).

135 Corona, M. *et al.* Vitellogenin, juvenile hormone, insulin signaling, and queen honey bee longevity. *Proc. Natl. Acad. Sci. USA* **104**, 7128-7133 (2007).

136 Seehuus, S.-C., Norberg, K., Gimsa, U., Krekling, T. & Amdam, G. V. Reproductive protein protects functionally sterile honey bee workers from oxidative stress. *Proc. Natl. Acad. Sci. USA* **103**, 962-967 (2006).

137 Tufail, M. & Takeda, M. Molecular characteristics of insect vitellogenins. *J. Insect Physiol.* **54**, 1447-1458 (2008).

138 Rodenburg, K. W., Smolenaars, M. M., Van Hoof, D. & Van der Horst, D. J. Sequence analysis of the non-recurring C-terminal domains shows that insect lipoprotein receptors constitute a

distinct group of LDL receptor family members. *Insect Biochem. Mol. Biol.* **36**, 250-263 (2006).

139 Boldbaatar, D. B. D. *et al.* Tick vitellogenin receptor reveals critical role in oocyte development and transovarial transmission of *Babesia* parasite. *Biochem. Cell Biol.* **86**, 331-344 (2008).

140 Tufail, M. & Takeda, M. Molecular cloning, characterization and regulation of the cockroach vitellogenin receptor during oogenesis. *Insect Mol. Biol.* **14**, 389-401 (2005). *Insect Mol. Biol.*

141 Tufail, M. & Takeda, M. Insect vitellogenin/lipophorin receptors: molecular structures, role in oogenesis, and regulatory mechanisms. *J. Insect Physiol.* **55**, 88-104 (2009). J. Insect Physiol.

142 Schooley, D. A. & Baker, F. C. Juvenile hormone biosynthesis. *Comprehensive Insect Physiology, Biochemistry and Pharmacology* **7**, 363-389 (1985).

143 Hartfelder, K. & Emlen, D. Endocrine control of insect polyphenism. *Insect Endocrinology*, 464-522 (2012).

144 Nijhout, H. F. *Insect hormones*. (Princeton University Press, 1998).

145 Bellés, X., Martín, D. & Piulachs, M.-D. The mevalonate pathway and the synthesis of juvenile hormone in insects. *Annu. Rev. Entomol.* **50**, 181-199 (2005). *Annu. Rev. Entomol.*

146 Cao, L., Zhang, P. & Grant, D. F. An insect farnesyl phosphatase homologous to the N-terminal domain of soluble epoxide hydrolase. *Biochem. Biophys. Res. Commu.* **380**, 188-192 (2009).

147 Mayoral, J. G. *et al.* Molecular and functional characterization of a juvenile hormone acid methyltransferase expressed in the corpora allata of mosquitoes. *Insect Biochem. Mol. Biol.* **39**, 31-37 (2009).

148 Baker, F. C., Mauchamp, B., Tsai, L. W. & Schooley, D. A. Farnesol and farnesal dehydrogenase (s) in corpora allata of the tobacco hornworm moth, *Manduca sexta*. *J. Lipid Res.* **24**, 1586-1594 (1983).

149 Reibstein, D., Law, J., Bowlus, S. & Katzenellenbogen, J. in *The juvenile hormones* (ed L.I. (Ed.) Gilbert) 131-146 (Plenum Press, 1976).

150 Feyereisen, R., Pratt, G. & Hamnett, A. Enzymic synthesis of juvenile hormone in locust corpora allata: evidence for a microsomal cytochrome P-450 linked methyl farnesoate epoxidase. *Eur. J. Biochem.* **118**, 231 (1981).

151 Tobe, S. S. & Pratt, G. E. The influence of substrate concentrations on the rate of insect juvenile hormone biosynthesis by corpora allata of the desert locust *in vitro*. *Biochem. J.* **144**, 107-113 (1974).

152 Hammock, B. D. NADPH dependent epoxidation of methyl farnesoate to juvenile hormone in the cockroach *Blaberus giganteus* L. *Life Sci.* **17**, 323-328 (1975).

153 Weaver, R., Pratt, G., Hamnett, A. & Jennings, R. The influence of incubation conditions on the rates of juvenile hormone biosynthesis by corpora allata isolated from adult females of the

beetle *Tenebrio molitor*. *Insect Biochem.* **10**, 245-254 (1980).

154 Li, Y., Hernandez-Martinez, S., Unnithan, G. C., Feyereisen, R. & Noriega, F. G. Activity of the corpora allata of adult female Aedes aegypti: effects of mating and feeding. *Insect Biochem. Mol. Biol.* **33**, 1307-1315 (2003).

155 Noriega, F. *et al.* Comparative genomics of insect juvenile hormone biosynthesis. *Insect Biochem. Mol. Biol.* **36**, 366-374 (2006).

156 Lewis, S. E. *et al.* Apollo: a sequence annotation editor. *Genome Biol.* **3**, 1-14 (2002).

157 Clark, A. & Bloch, K. The absence of sterol synthesis in insects. *J. Biol. Chem.* **234**, 2578-2582 (1959).

158 Gilbert, L. I., Granger, N. A. & Roe, R. M. The juvenile hormones: historical facts and speculations on future research directions. *Insect Biochem. Mol. Biol.* **30**, 617-644 (2000).

159 Klowden, M. J. *Physiological systems in insects*. 2nd edn, (Elsevier, 2007).

160 Libersat, F. & Pflueger, H. J. Monoamines and the orchestration of behavior. *Bioscience* **54**, 17-25 (2004).

161 Dasari, S. & Cooper, R. L. Modulation of sensory-CNS-motor circuits by serotonin, octopamine, and dopamine in semi-intact *Drosophila* larva. *Neurosci. Res.* **48**, 221-227 (2004).

162 Kloppenburg, P. & Mercer, A. R. Serotonin modulation of moth central olfactory neurons. *Annu. Rev. Entomol.* **53**, 179-190 (2008).

163 Rein, J., Mustard, J. A., Strauch, M., Smith, B. H. & Galizia, C. G. Octopamine modulates activity of neural networks in the honey bee antennal lobe. *J. Comp. Physiol. A* **199**, 947-62 (2013).

164 Hammer, M. The neural basis of associative reward learning in honeybees. *Trends Neurosci.* **20**, 245-252 (1997).

165 Schwaerzel, M. *et al.* Dopamine and octopamine differentiate between aversive and appetitive olfactory memories in *Drosophila*. *J. Neurosci.* **23**, 10495-10502 (2003).

166 Unoki, S., Matsumoto, Y. & Mizunami, M. Roles of octopaminergic and dopaminergic neurons in mediating reward and punishment signals in insect visual learning. *Eur. J. Neurosci.* **24**, 2031-2038 (2006).

167 Wright, G. A. *et al.* Parallel reinforcement pathways for conditioned food aversions in the honeybee. *Curr. Biol.* **20**, 2234-2240 (2010).

168 Fussnecker, B. L., Smith, B. H. & Mustard, J. A. Octopamine and tyramine influence the behavioral profile of locomotor activity in the honey bee (*Apis mellifera*). *J. Insect Physiol.* **52**, 1083-1092 (2006).

169 Mustard, J. A., Pham, P. M. & Smith, B. H. Modulation of motor behavior by dopamine and the D1-like dopamine receptor AmDOP2 in the honey bee. *J. Insect Physiol.* **56**, 422-430 (2010).

170 Schulz, D. J. & Robinson, G. E. Biogenic amines and division of labor in honey bee colonies: behaviorally related changes in the antennal lobes and age-related changes in the mushroom bodies. *J. Comp. Physiol. A* **184**, 481-488 (1999).

171 Taylor, D. J., Robinson, G. E., Logan, B. J., Laverty, R. & Mercer, A. R. Changes in brain amine levels associated with the morphological and behavioral development of the worker honeybee. *J. Comp. Physiol. A* **170**, 715-721 (1992).

172 Boulay, R. & Lenoir, A. Social isolation of mature workers affects nestmate recognition in the ant *Camponotus fellah*. *Behav. Proc.* **55**, 67-73 (2001).

173 Cuvillier-Hot, V. & Lenoir, A. Biogenic amine levels, reproduction and social dominance in the queenless ant *Streblognathus peetersi*. *Naturwissenschaften* **93**, 149-153 (2006).

174 Harano, K., Sasaki, K. & Nagao, T. Depression of brain dopamine and its metabolite after mating in European honeybee (*Apis mellifera*) queens. *Naturwissenschaften* **92**, 310-313 (2005).

175 Harris, J. W. & Woodring, J. Elevated brain dopamine levels associated with ovary development in queenless worker honey bees (*Apis mellifera* L.). *Comp. Biochem. Physiol. C* **111**, 271-279 (1995).

176 Sasaki, K., Yamasaki, K. & Nagao, T. Neuro-endocrine correlates of ovarian development and egg-laying behaviors in the primitively eusocial wasp (*Polistes chinensis*). *J. Insect Physiol.* **53**, 940-949 (2007).

177 Beggs, K. T. *et al.* Queen pheromone modulates brain dopamine function in worker honey bees. *Proc. Natl. Acad. Sci. USA* **104**, 2460-2464 (2007).

178 Vander Meer, R. K., Preston, C. A. & Hefetz, A. Queen regulates biogenic amine level and nestmate recognition in workers of the fire ant, *Solenopsis invicta*. *Naturwissenschaften* **95**, 1155-1158 (2008).

179 Le Crom, S., Kapsimali, M., Barôme, P. O. & Vernier, P. Dopamine receptors for every species: gene duplications and functional diversification in Craniates. *J. Struct. Funct. Genomics* **3**, 161-176 (2003).

180 Peroutka, S. J. & Howell, T. A. The molecular evolution of G protein-coupled receptors: Focus on 5-hydroxytryptamine receptors. *Neuropharmacology* **33**, 319-324 (1994).

181 Casida, J. E. & Durkin, K. A. Neuroactive insecticides: targets, selectivity, resistance and secondary effects. *Annu. Rev. Entomol.* **58**, 99-117 (2013).

182 Rotte, C. *et al.* Molecular characterization and localization of the first tyramine receptor of the American cockroach (*Periplaneta americana*). *Neuroscience* **162**, 1120-1133 (2009).

183 Wang, Z. B. *et al.* Genome-wide mapping of HATs and HDACs reveals distinct functions in active and inactive genes. *Cell* **138**, 1019-1031 (2009).

184 Gregoretti, I. V., Lee, Y. M. & Goodson, H. V. Molecular evolution of the histone deacetylase family: Functional implications of phylogenetic analysis. *J. Mol. Biol.* **338**, 17-31 (2004).

185 Finkel, T., Deng, C. X. & Mostoslavsky, R. Recent progress in the biology and physiology of sirtuins. *Nature* **460**, 587-591 (2009).

186 Tsai, Y. C., Greco, T. M., Boonmee, A., Miteva, Y. & Cristea, I. M. Functional proteomics establishes the interaction of SIRT7 with chromatin remodeling complexes and expands its role in regulation of RNA polymerase I transcription. *Mol. Cell. Proteomics* **11**, 60-76 (2012).

187 Kouzarides, T. Chromatin modifications and their function. *Cell* **128**, 693-705 (2007).

188 Agger, K., Christensen, J., Cloos, P. A. C. & Helin, K. The emerging functions of histone demethylases. *Curr. Opin. Genetics Dev.* **18**, 159-168 (2008).

189 Liebig, J., Eliyahu, D. & Brent, C. Cuticular hydrocarbon profiles indicate reproductive status in the termite *Zootermopsis nevadensis*. *Behav. Ecol. Sociobiol.* **63**, 1799–1807 (2009).

190 Monnin, T. Chemical recognition of reproductive status in social insects. *Ann. Zool. Fenn.* **43**, 515-530 (2006).

191 Le Conte, Y. & Hefetz, A. Primer pheromones in social hymenoptera. *Annu. Rev. Entomol.* **53**, 523-542 (2008).

192 Peeters, C. & Liebig, J. in *Organization of Insect Societies: From Genome to Socio-Complexity* (eds J. Gadau & J. Fewell) 220-242 (Harvard University Press, 2009).

193 Liebig, J. in *Insect Hydrocarbons: Biology, Biochemistry, and Chemical Ecology* Ch. The evolution of hydrocarbon profiles as dominance and fertility signals in social insects (Cambridge University Press, 2010).

194 Wicker-Thomas, C. & Chertemps, T. in *Insect hydrocarbons: Biology, biochemistry, and chemical ecology Molecular biology and genetics of hydrocarbon production* (Cambridge University Press, 2010).

195 Nelson, D. R. Metazoan cytochrome P450 evolution. *Comp. Biochem. Physiol. C* **121**, 15-22 (1998).

196 Nelson, D. R., Goldstone, J. V. & Stegeman, J. J. The cytochrome P450 genesis locus: the origin and evolution of animal cytochrome P450s. *Phil. Trans.R. Soc. B* **368** (2013).

197 Zhou, X., Tarver, M., Bennett, G., Oi, F. & Scharf, M. Two hexamerin genes from the termite *Reticulitermes flavipes*: Sequence, expression, and proposed functions in caste regulation. *Gene* **376**, 47-58 (2006).

198 Bird, A. P. DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Res.* **8**,

1499-1504 (1980).

199 Götz, S. *et al.* High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Res.* **36**, 3420-3435 (2008).

200 Yanai, I. *et al.* Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics* **21**, 650-659 (2005).

201 Zemach, A., McDaniel, I. E., Silva, P. & Zilberman, D. Genome-wide evolutionary analysis of eukaryotic DNA methylation. *Science* **328**, 916-919 (2010).

202 Glastad, K. M., Hunt, B. G., Yi, S. V. & Goodisman, M. A. D. DNA methylation in insects: on the brink of the epigenomic era. *Insect Mol. Biol.* **20**, 553-565 (2011).

203 Hunt, B. G., Brisson, J. A., Yi, S. V. & Goodisman, M. A. D. Functional conservation of DNA methylation in the pea aphid and the honeybee. *Genome Biol. Evol.* **2**, 719-728 (2010).

204 Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389-3402 (1997).

205 Duret, L. & Mouchiroud, D. Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. *Mol. Bio. Evol.* **17**, 68-74 (2000).

206 Sarda, S., Zeng, J., Hunt, B. G. & Yi, S. V. The evolution of invertebrate gene body methylation. *Mol. Bio. Evol.* **29**, 1907-1916 (2012).

207 Wang, Z. & Burge, C. B. Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. *RNA* **14**, 802-813 (2008).