

2012

Evolution of a Large, Conserved, and Syntenic Gene Family in Insects

Neethu Shah

University of Nebraska-Lincoln

Douglas R. Dorer

Martin Methodist College

Etsuko N. Moriyama

University of Nebraska - Lincoln, emoriyama2@unl.edu

Alan C. Christensen

University of Nebraska-Lincoln, achristensen2@unl.edu

Follow this and additional works at: <https://digitalcommons.unl.edu/bioscifacpub>

 Part of the [Biology Commons](#)

Shah, Neethu; Dorer, Douglas R.; Moriyama, Etsuko N.; and Christensen, Alan C., "Evolution of a Large, Conserved, and Syntenic Gene Family in Insects" (2012). *Faculty Publications in the Biological Sciences*. 495.

<https://digitalcommons.unl.edu/bioscifacpub/495>

This Article is brought to you for free and open access by the Papers in the Biological Sciences at DigitalCommons@University of Nebraska - Lincoln. It has been accepted for inclusion in Faculty Publications in the Biological Sciences by an authorized administrator of DigitalCommons@University of Nebraska - Lincoln.

Evolution of a Large, Conserved, and Syntenic Gene Family in Insects

Neethu Shah,* Douglas R. Dorer,[†] Etsuko N. Moriyama,^{*,§} and Alan C. Christensen^{§,1}

*Department of Computer Science and Engineering, University of Nebraska–Lincoln, Lincoln, Nebraska 68588-0115,

[†]Division of Mathematics & Sciences, Martin Methodist College, Pulaski, Tennessee 38478-2716, [‡]Center for Plant Science Innovation, and [§]School of Biological Sciences, University of Nebraska–Lincoln, Lincoln, Nebraska 68588-0666

ABSTRACT The *Osiris* gene family, first described in *Drosophila melanogaster*, is clustered in the genomes of all *Drosophila* species sequenced to date. In *D. melanogaster*, it explains the enigmatic phenomenon of the triplo-lethal and haploinsufficient locus *Tpl*. The synteny of *Osiris* genes in flies is well conserved, and it is one of the largest syntenic blocks in the *Drosophila* group. By examining the genome sequences of other insects in a wide range of taxonomic orders, we show here that the gene family is well-conserved and syntenic not only in the diptera but across the holometabolous and hemimetabolous insects. *Osiris* gene homologs have also been found in the expressed sequence tag sequences of various other insects but are absent from all groups that are not insects, including crustacea and arachnids. It is clear that the gene family evolved by gene duplication and neofunctionalization very soon after the divergence of the insects from other arthropods but before the divergence of the insects from one another and that the sequences and synteny have been maintained by selection ever since.

KEYWORDS

gene duplication
insect
gene family
synteny
Osiris
triplo-lethal

Gene families are commonly found in genomes and are thought to evolve by gene duplication and neofunctionalization. The *Osiris* gene family is a large conserved family first described in *Drosophila melanogaster* (Dorer *et al.* 2003). Although the genes are still of unknown function, they are the molecular basis of the unique *Triplo-lethal* locus in *D. melanogaster*, first described in 1972 (Lindsley *et al.* 1972). The proteins have a secretion signal peptide and four domains that identify them as *Osiris* family members, one of those being a putative transmembrane domain. Twenty-three *Osiris* genes were originally found in the *D. melanogaster* genome, with 20 of them located on chromosome 3R (83E) in a cluster within a 168-kb region, which is both triplo-lethal and haplo-lethal. The *Osiris* gene family was also found in the mosquito *Anopheles gambiae*, maintaining the synteny except for a chromosomal rearrangement that split the cluster (Dorer *et al.* 2003). Subsequent work revealed that the synteny was strongly con-

served among 12 diverse *Drosophila* species (Bhutkar *et al.* 2008). In this work, we report the existence of the *Osiris* gene family, which now includes 24 orthologous gene groups, in a diverse group of insects and report on the evolution of the genes and the conservation of the synteny during a very long evolutionary time frame. The interrupted synteny seen in *Anopheles gambiae* is the exception rather than the rule, and we show that the *Osiris* gene cluster is a well-conserved, insect-specific, and remarkably syntenic gene family.

MATERIALS AND METHODS

Osiris protein sequences used

The 24 *D. melanogaster* *Osiris* protein sequences were downloaded from Flybase (<http://flybase.org/>). Their annotation symbols (CG numbers) are listed in supporting information, Table S1. These protein sequences were used as the queries for searching *Osiris* genes in various organisms.

Insect and other arthropod genomes used

Insect and other arthropod genomes were downloaded from various sources as listed in Table S2. *Daphnia pulex* (a water flea, Subphylum Crustacea) and *Ixodes scapularis* (the deer tick, Class Arachnida) are the two noninsect Arthropoda in which the sequences of complete genomes are available. For insects, we examined in total 23 complete genomes, including 12 species of *Drosophila*, three species of mosquito (*Anopheles gambiae*, *Culex quinquefasciatus*, and *Aedes aegypti*), two hymenoptera (*Apis mellifera* and *Camponotus floridanus*), one

Copyright © 2012 Shah *et al.*

doi: 10.1534/g3.111.001412

Manuscript received October 13, 2011; accepted for publication December 16, 2011
This is an open-access article distributed under the terms of the Creative Commons Attribution Unported License (<http://creativecommons.org/licenses/by/3.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Supporting information is available online at <http://www.g3journal.org/lookup/suppl/doi:10.1534/g3.111.001412/-/DC1>

¹Corresponding author: School of Biological Sciences, E249 Beadle Center, University of Nebraska–Lincoln, Lincoln, NE 68588-0666. E-mail: achristensen2@unl.edu

coleopteran (*Tribolium castaneum*), one lepidopteran (*Bombyx mori*), one phthirapteran (*Pediculus humanus*), and one hemipteran (*Acyrtosiphon pisum*).

BLAST similarity search against NCBI databases

Each of the 24 *D. melanogaster* Osiris protein sequences was used as the query. The blastp protein similarity search (Altschul *et al.* 1997; Camacho *et al.* 2009) was performed against the nonredundant (NR) protein database at the National Center for Biotechnology and Information (NCBI) Web server (<http://www.ncbi.nlm.nih.gov/BLAST/>). The tblastn translated protein similarity search was also performed against NCBI's Expressed Sequence Tag (EST) database. For both searches, the BLOSUM45 scoring matrix was used, and the E-value threshold was set at 0.01. All other options were set to the default.

Osiris gene mining from complete genomes using BLAST similarity search

Using each of the 24 *D. melanogaster* Osiris protein sequences as the query, we performed blastp similarity searches against the 23 complete genomes from insects as well as *Daphnia* and *Ixodes*. The BLAST+ package (version 2.2.24+) installed on our local Linux server was used to prepare the complete set of proteins from each genome and run the blastp program. An E-value threshold of 1 was used, and the effective length of the database was set as 7,500,000 residues (on the basis of the average cumulative number of amino acid residues from all the genomes) for all genomic searches. All other options were set as the default. When gene structures given in the genome project were suspected to be incorrect or incomplete (*e.g.*, missing 5' or 3' exon), we used tblastn, GeneWise version 2-2-0 (Birney *et al.* 2004), and Augustus version 2.5.5 (<http://bioinf.uni-greifswald.de/augustus/>) (Stanke *et al.* 2008) for gene structure prediction. When gene structures were not available in the genome project, we also performed our own prediction using the same strategy. Gene structures different from the ones given in the genome projects are clarified in Table S3.

Profile hidden Markov models (HMMs)

To perform a thorough search of *Osiris* genes, we built a profile HMM based on the alignment of all 24 *Osiris* proteins obtained from the aforementioned blast searches. HMMER version 3.0 (<http://hmm.janelia.org/>) was used to build profile HMMs for all the 24 *Osiris* proteins and to perform searches using these profile HMMs against the entire protein set from each genome. The options used were hmmbuild and hmmsearch, with an E-value threshold of 0.01 and a database size of 20,000 (average of all the sequences from all the genomes). In addition to the 23 genomes, two Annelida genomes, *Capitella teleta* and *Helobdella robusta* (obtained from the Joint Genome Institute <http://genome.jgi-psf.org/>), were searched.

Multiple alignments of Osiris protein sequences

Multiple alignments of *Osiris* protein sequences were generated using MAFFT v6.847b (Katoh and Toh 2008) with the L-INS-i algorithm. Alignments were generated individually for each *Osiris* group, including all *Osiris* sequences at once, and using the profile alignment option. The alignment of the 24 *D. melanogaster* *Osiris* proteins is included in Figure S1.

Phylogenetic tree reconstruction

Phylogenetic trees were reconstructed by FastTree 2 (version 2.1.3), which can infer approximately maximum likelihood phylogenies for

large data sets (Price *et al.* 2010). The default options were used except for “-gamma.” This uses the JTT+CAT (20 fixed-rate categories) model for amino acid substitutions for tree optimization and the discrete gamma model with 20 rate-categories to tree rescaling. Bootstrap analysis with 1000 pseudoreplicates was done using seqboot (Phylip version 3.68) (Felsenstein 2010) and CompareToBootstrap.pl (<http://www.microbesonline.org/fasttree/treecmp.html>).

Motif/domain search in Osiris proteins

Using the 23 *Osiris* protein sequences (Osiris 1 to Osiris 23), we performed the motif search using the Multiple Em for Motif Elicitation (MEME; version 4.6.1 <http://meme.nbcn.net>) (Bailey *et al.* 2006). To find short motifs, the parameters were set to discover up to 10 motifs ranging from six (minimum width) to 30 (maximum width) amino acids and with any number of repetitions. The 10 motifs discovered covered the two-Cys region, the duf1676 domain, and the AQXLY domain. The MEME result is available from our website (<http://bioinfolab.unl.edu/emlab/Osiris>). We used the Motif Alignment and Search Tool (Bailey *et al.* 2009) to search these 10 motifs from all *Osiris* protein candidates to confirm whether these proteins have the *Osiris* signature motifs. We also used the Pfam protein family search at <http://pfam.sanger.ac.uk> (Finn *et al.* 2010) to identify the presence of the duf1676 domain in each candidate. Signal peptide and transmembrane predictions were done by Phobius (version 1.01) (Kall *et al.* 2004). Transmembrane prediction was also confirmed by HMMTOP (version 2.1) (Tusnady and Simon 2001).

Identification of Osiris homologs

Osiris protein candidates found by blast and profile HMM search using the 24 *D. melanogaster* *Osiris* protein sequences were identified as *Osiris* homologs on the basis of the results of 10-motif search using Motif Alignment and Search Tool, duf1676 profile HMM search, as well as reciprocal blast search. For the reciprocal blast, each of the candidate proteins was used as the query and blastp similarity search was performed against the *D. melanogaster* protein set. When none of the known *Osiris* proteins was found to be significantly similar, this candidate was considered not an *Osiris* homolog and excluded from the candidate list.

Classification of Osiris orthologous gene groups

Orthologous groups of *Osiris* genes were identified on the basis of phylogenetic clustering as well as the chromosomal location and order of the *Osiris* gene candidates wherever the gene coordinates on contiguous genomic sequences (*e.g.*, supercontigs) were available. We first aligned all *Osiris* protein sequences and reconstructed a draft phylogeny. On the basis of this phylogeny, preliminary assignment of *Osiris* groups was performed. Alignments and phylogenies were repeatedly refined for each orthologous group individually. Note that as mentioned previously, alignments were generated group by group as well as using all sequences all at once. We confirmed that the phylogenies reconstructed from two versions of alignments were topologically equivalent, and ortholog-grouping was not biased arbitrarily because of the alignment strategy. When the assignment of a gene to a particular paralog group was unclear as the result of weak similarity (unsupported phylogenetic clustering), a nonconserved location, or both, these were called *Osiris*-like genes, and are separately listed as such in Table S3. The final results for all *Osiris* genes we identified are listed in Table S1 and Table S3. The sequences and alignments of all *Osiris* proteins are available from our website (<http://bioinfolab.unl.edu/emlab/Osiris>). The final maximum likelihood phylogeny is shown in Figure S2.

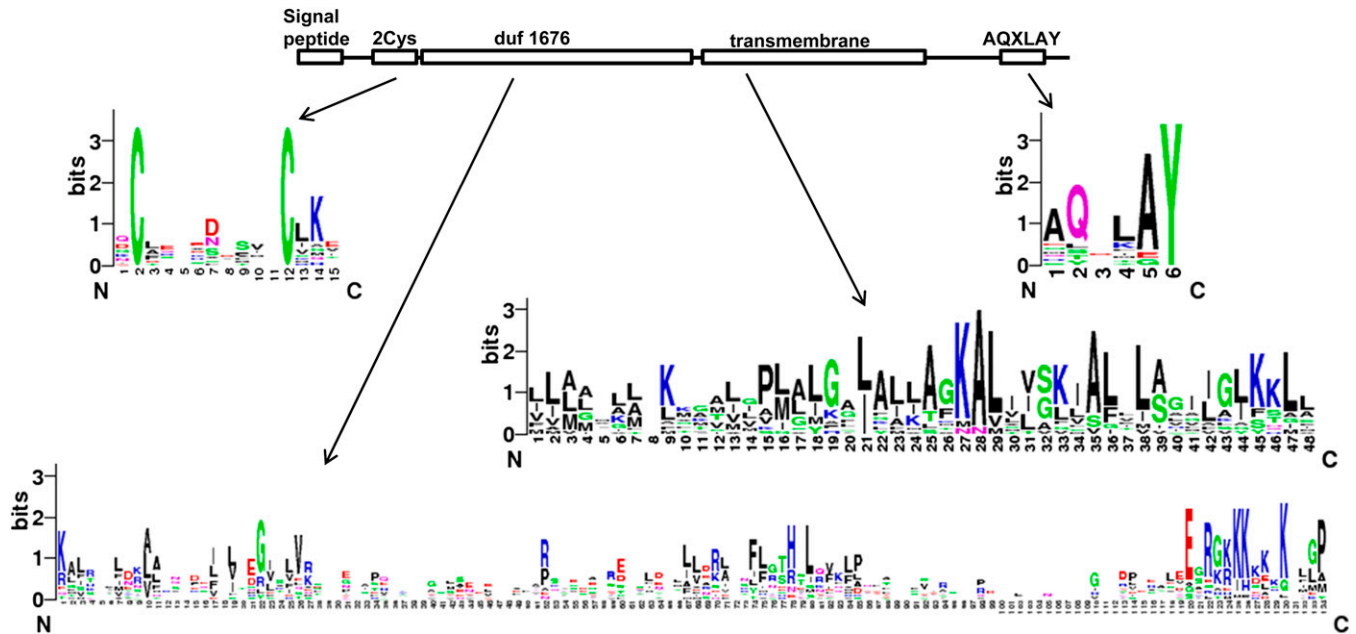


Figure 1 Features of the *Osiris* proteins. The conserved domains within all *Osiris* proteins are indicated as boxes. The regions between these domains are conserved only within orthologous groups and serve to identify the different *Osiris* family members. Some examples of conserved patterns for the ortholog-specific domains are shown below with sequence logos.

Sequence logo of *Osiris* protein sequences

Using a multiple alignment of the *D. melanogaster* *Osiris* 1, 2, 5, 6, 7, 8, 9, 11, 12, 14, 16, 18, 19, 20, and 21 protein sequences, the sequence logo was generated using Weblogo 3 (<http://weblogo.berkeley.edu/>) (Crooks *et al.* 2004). The other paralogs were omitted because they were either missing a domain or have large insertions that make alignments unreliable.

RESULTS

Distribution of *Osiris* genes

The accumulation of genomic and EST sequences from a diversity of insects has allowed us to further characterize the *Osiris* gene

family. *Osiris* family members are characterized by five features: (1) a hydrophobic region at the N-terminus that is likely a secretion signal peptide; (2) a two-Cys region; (3) a domain of unknown function, duf1676 (Pfam family: PF07898) (Finn *et al.* 2010); (4) a hydrophobic putative transmembrane domain, and (5) a region including an AQXLAY motif and often additional nearby tyrosine residues. These domains are illustrated in Figure 1 (see also the alignment in Figure S1). Although these domains are found in most *Osiris* family members, the regions between these domains are what distinguish family members from each other. The regions between the conserved domains are highly variable, in sequence and in length, but are well-conserved within a group of orthologs from different species.

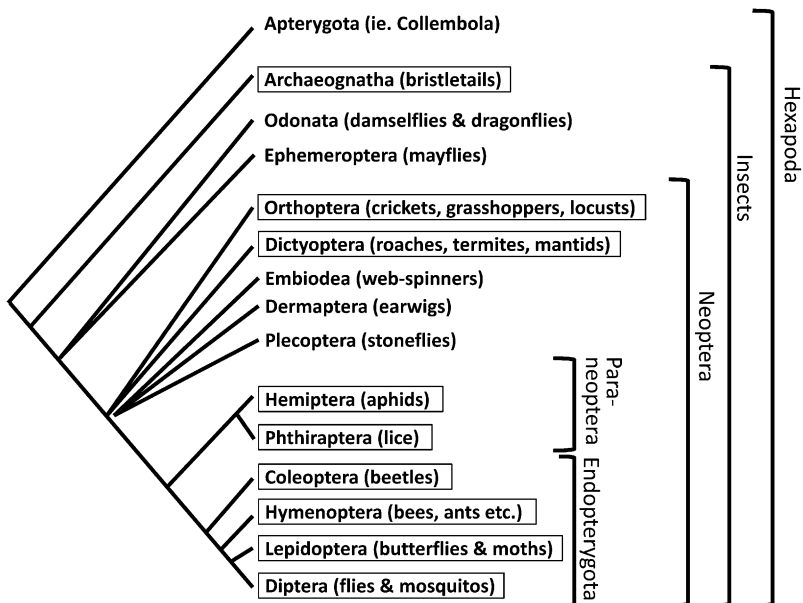


Figure 2 A simplified phylogenetic tree of the insects, indicating the groups shown to have *Osiris* gene homologs. The tree is based on Grimaldi and Engel (2005), Cranston and Gullan (2009), and Whitfield and Kjer (2008). Boxed taxa indicate that at least one *Osiris* family member has definitively been found.

We identified one family member that had not been previously described in *Drosophila melanogaster*, CG15589, which is located between *Osiris 1* and *Osiris 2*. Although the signature domains are not all conserved in the *D. melanogaster* protein (see Figure S1), the orthologs were found to be conserved in other insects as described below. We here rename this gene as *Osiris 24*. With the search strategy and criteria we developed, we did not find any additional *Osiris* genes within the *D. melanogaster* genome.

We searched the NR protein and EST nucleotide sequence databases at NCBI for *Osiris* homologs. The results are summarized in Table S4. *Osiris* homologs were found almost exclusively from insects. Searches using a profile HMM (Durbin *et al.* 1998) for duf1676 against the NR protein database gave us consistent results. Short and weak similarities against two *Osiris* proteins (*Osiris 17* and *24*) were found from EST sequences of Collembola (a springtail) and some crustacea (see Table S4 for details). Similarities found in the EST sequences must be viewed with some caution because they can result from contamination of the material used to make the cDNA library. For example, sequences highly similar to insect *Osiris* sequences were found in several plant EST sequences but not in any complete plant genomic sequence, suggesting that they are attributable to insect contamination of the plant material (a footnote in Table S4 provides some specific examples). To determine whether the *Osiris* family is arthropod-specific or insect-specific, we examined the complete genome sequences of two noninsect arthropod species, a crustacean, *Daphnia pulex* (Colbourne *et al.* 2011) and an arachnid, the deer tick *Ixodes scapularis* (Hill and Wikel 2005; Pagel Van Zee *et al.* 2007). Extensive similarity searching using BLAST (Altschul *et al.* 1997; Camacho *et al.* 2009) and profile HMMs showed only weak similarities (lower than the threshold we used) in these noninsect arthropod genomes and none of them had *Osiris* signature motifs.

In addition to the genomic sequences from the 12 *Drosophila* species (Clark *et al.* 2007), we searched for *Osiris* gene candidates from the following nine insect complete genomes: three species of mosquito, *Anopheles gambiae*, *Culex quinquefasciatus*, and *Aedes aegypti* (Arensburger *et al.* 2010; Nene *et al.* 2007); the honeybee *Apis mellifera* (Honeybee Genome Sequencing Consortium 2006; Munoz-Torres *et al.* 2011); the Florida carpenter ant *Camponotus floridanus* (Bonasio *et al.* 2010); the red flour beetle *Tribolium castaneum* (Tribolium Genome Sequencing Consortium 2008); the silkworm *Bombyx mori* (Xia *et al.* 2004); the pea aphid *Acyrtosiphon pisum* (International Aphid Genomics Consortium 2010); and the body louse *Pediculus humanus* (Kirkness *et al.* 2010). Figure 2 shows the phylogenetic summary of all insects where we found *Osiris* homolog candidates. As shown in Figure 3 and in Table S1 and Table S3, the majority of the members of the *Osiris* family were identified in these species, indicating that the *Osiris* family was present in the common ancestor of the hemimetabolous and holometabolous insects (paraneoptera and endopterygota, respectively).

EST sequences similar to *Osiris* genes have been found in a number of additional distantly related insects, including the German cockroach *Blattella germanica* (order Blattodea), the cricket *Gryllus bimaculatus* (order Orthoptera), and the termite *Reticulitermes flavipes* (order Isoptera). Twenty-three partial sequences have been found in ESTs from the primitive archaeognathan *Lepismachilis y-signata* (jumping bristletails). These *Lepismachilis* EST sequences were compared using BLAST, and those with significant identities were assembled, resulting in 10 unique sequences. This primitive wingless insect clearly has at least 10 different *Osiris* genes. As Figure S2 shows, all these *Lepismachilis* *Osiris* sequences form one cluster along with several other unclassified *Osiris*-like proteins. These sequences may represent ancestral forms of *Osiris* proteins, and none of the currently

	Osi1	NPFR1	Osi24	Osi2	Osi3	Osi4	Osi5	Osi6	Osi7	Osi8	Osi9	Osi10	Osi11	Osi12	Osi13	Osi14	Osi15	Osi16	Osi17	Osi18	Osi19	Osi20	Osi22	Osi23	Osi21
<i>D. melanogaster</i>						*							*												*
<i>D. pseudoobscura</i>						*							*												*
<i>D. virilis</i>						*							*										?	?	
<i>D. grimshawi</i>						*							*										?	?	
<i>An. gambiae</i>						*	?											2	*						2
<i>Ae. aegypti</i>						*	2	?												*		2	?	?	2
<i>Cu. quinquefasciatus</i>						*														*			?		3
<i>B. mori</i>		←	←							←	←6		*					2	*				?		
<i>Ap. mellifera</i>													*												
<i>Ca. floridanus</i>													*												?
<i>T. castaneum</i>							2						*					2*				2	?		
<i>Ac. pisum</i>		?			?									*				6*					?	?	4
<i>P. humanus</i>													*							*					

Figure 3 Presence or absence of *Osiris* family members in various species. Gray indicates presence. Diptera and holometabolous insect-specific presence is shown in red and orange colors, respectively. An asterisk indicates that the gene is inverted relative to *Osiris 2* in each species, and a left arrow indicates that the region including more than one gene is inverted. A question mark indicates when the relative gene direction cannot be inferred because of separated contigs. A number indicates the number of duplicated copies. One non-*Osiris* gene, *NPFR1*, is also included in the figure because it is tightly linked within the *Osiris* gene cluster. In the *D. melanogaster* genome, genes from *Osiris 1* to *Osiris 20* are located within a tightly linked 167.5-kb region on the chromosomal arm 3R (83D-E). Although *Osiris 22* and *Osiris 23* are also on the 3R but located distantly (87E and 99F, respectively), *Osiris 21* is located on the chromosomal arm 2L (32E). In some insects, *Osiris 23* is located on a different chromosome and they are indicated by double-line boxes. See Table S3 for detailed information of all genes we identified.

available *Lepismachilis Osiris* sequences seems to be closely related to presently known *Osiris* subfamilies.

Synteny of *Osiris* genes

The synteny of *Osiris* genes (from *Osi1* to *Osi20*) between *D. melanogaster* and *A. gambiae* that was discovered in 2003 (Dorer *et al.* 2003) is largely maintained within the 12 sequenced *Drosophila* species (see Table S1), and with the two additional mosquito species sequenced. Because the *Osiris* gene cluster is so conserved within the genus *Drosophila*, we have chosen four divergent species (*D. melanogaster*, *D. pseudoobscura*, *D. virilis*, and *D. grimshawi*) as representatives for further analysis.

The synteny is even more striking when sequences from more distantly related species are examined (Figure 4). Both hymenoptera species, the honeybee *A. mellifera* and the ant *C. floridanus*, maintain almost all of the genes in the same order, and they are each transcribed in the same direction in these two species as they are in fruit flies. There are a few interesting exceptions. A neighboring gene *Neuropeptide F Receptor 1 (NPF1)* is missing in these hymenoptera (see Figure 3). *NPF1*, although it is unrelated to *Osiris* genes, is conserved in synteny with the *Osiris* genes (located between *Osi1* and *Osi24*) in all the insects we examined except for the hymenoptera. *Osiris 4* was also not found in *A. mellifera*, nor any other non-dipteran insects examined, although the neighboring genes *Osiris 3* and *Osiris 5* are conserved. Interestingly, *Osiris 4* and *Osiris 11* are transcribed in the opposite direction of all the other genes in *Drosophila* and most other species. *Osiris 1*, *5*, *13*, and *15* are unique to the holometabolous insects we examined. The region between *Osiris 12* and *Osiris 14* is

interesting. There are often *Osiris*-like genes, or *Osiris* fragments in this region, but the similarity is so weak that it is impossible to determine their identity. This region is also apparently a region with lower selection on the synteny – this is the location of the inversion breakpoint in *A. gambiae*, and is also the location where other genes have interposed into the cluster, such as *CG15594* and *CG15597* in *D. melanogaster*. Note that some of the signature domains are not well-conserved in *Osiris 4* and *Osiris 13* (see Figure S1), indicating that their functions may be modified in these more derived paralogs.

Although the complete genome sequences from distantly related insects are less complete and have occasional gaps in the scaffolding, the synteny of the *Osiris* genes is conserved over a remarkably wide range of insect orders. Chromosomal rearrangements such as an inversion in *A. gambiae*, and movements of parts of the cluster in other species (such as the apparent separation of *Osiris 2-12* from *Osiris 14-16* in *T. castaneum*), occasionally disrupt the synteny. Gene duplications have also occurred in some lineages, for example, there are multiple copies of *Osiris 9* in *B. mori*.

DISCUSSION

The *Osiris* gene cluster is a family of genes that is present in all insects but only in insects. Alignment and phylogenetic analysis indicates that the various paralogous members of the gene family each have very distinctive features (see Figure S2). Indeed, the paralogs within *D. melanogaster* are diverged enough to be easily distinguished from one another by the protein sequences between the conserved *Osiris* domains. The *Osiris* paralogs have also diverged enough at the DNA sequence level that there is no evidence for gene conversion in the

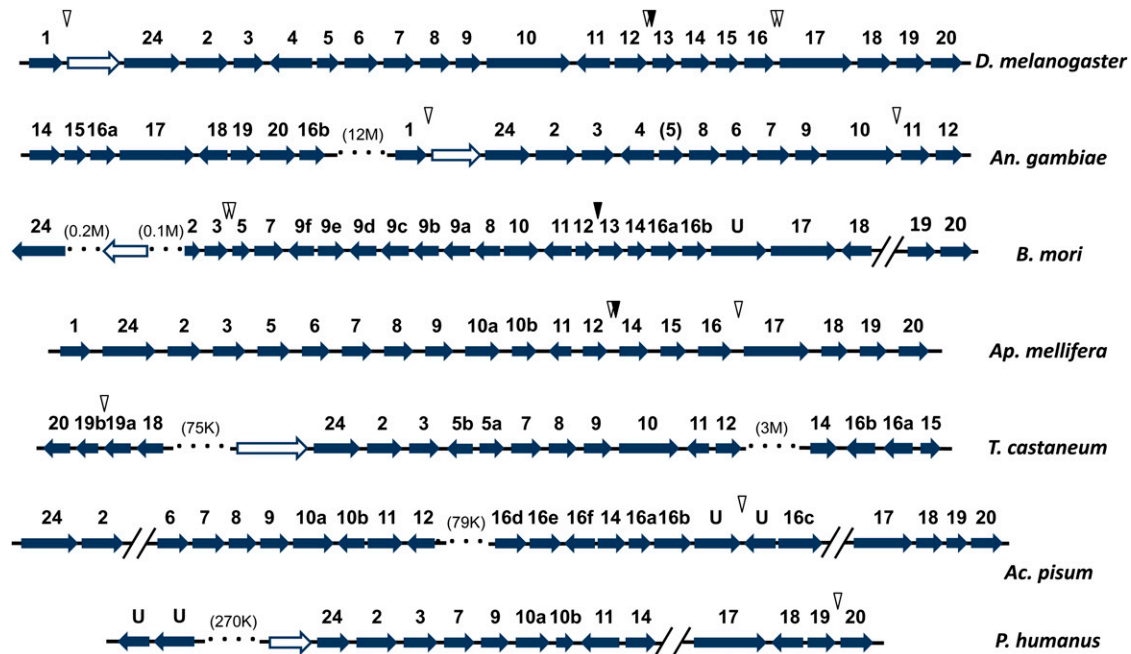


Figure 4 Synteny of the *Osiris* gene cluster. *Osiris* family genes are indicated in their syntenic chromosomal blocks. Paralogs are labeled with numbers, and in the case of duplications with letters, e.g., 9a and 9b. The white arrows indicate the unrelated gene *NPF1*, which is often in the syntenic block. Slashes indicate unknown linkage because of contig gaps, whereas dotted lines indicate large chromosome distances whose sizes are indicated. “U” means that the identity of that *Osiris* paralog is still undetermined. Inverted triangles indicate inserted non-*Osiris* genes, although some have weak similarities. The filled inverted triangles are similar to the *D. melanogaster* gene *CG15594*. Note that *Osiris 10*s in *Drosophila*, mosquitoes, and *T. castaneum* are twice as long as their homologs in other genomes. In the two hymenoptera, *A. pisum*, and *P. humanus* genomes, two *Osiris 10* genes (*10a* and *10b*) are annotated as individual genes, and in *A. pisum*, *10b* is located on the reverse strand. Therefore, it is likely that having a long single *Osiris 10* is a result of incorrect gene prediction. For our analysis, we divided the longer form of *Osiris 10* into two parts and performed alignments and phylogenetic analyses using them as individual proteins.

recent evolutionary history of the species. The differences between the multiple *Osiris* 9 copies in the *Bombyx* lineage (see Figure S2) and the presence of distinct orthologs of the various *Osiris* 9 copies in EST collections of other lepidoptera (D.R. Dorer, unpublished data) indicate that these duplication events occurred in the ancestral lepidopteran lineage since its divergence from the hymenoptera and coleoptera and also show no evidence of recent gene conversion. Any given *Osiris* gene is more similar to the orthologs in other species than to any of the paralogs in the same species. Therefore, the common ancestor of all the Endopterygota must have had an *Osiris* gene family cluster very similar to what is seen today in *D. melanogaster*. The presence of at least seven different members of the family in the Archaeognatha suggests that the evolution of this gene family extends back to the time just after the divergence of insects from other arthropods, but before the divergence of the major insect orders from one another, followed by very strong selection ever since to maintain the diverse members of the gene family as distinct entities. Therefore, the *Osiris* gene family must have evolved by gene duplication and divergence events very early in the radiation of insects, perhaps as many as 400 MYA, and has been subject to strong selection in all insects ever since. Minimal available sequence data on non-insect hexapods, such as the Collembola (springtails), and primitive insects such as the Odonata and Ephemeroptera prevents us from further refining the phylogenetic distribution.

There also appears to be very strong selection on the synteny of the gene family, raising the question of whether the synteny has a function. Recent work has suggested that synteny is more common with developmental regulatory genes and maintained due to selection on co-expression (Quijano *et al.* 2008). However, publicly available expression data on Flybase (<http://www.flybase.org>) indicates the *Osiris* genes are expressed in a variety of tissues, including epidermis, hindgut, foregut, and trachea, suggesting that coexpression is not the selective force. One interesting observation about the tissue expression data is that all of the *Osiris* genes appear to be expressed in tissues derived from ectoderm except for the nervous system. This could be a clue to a specific feature of insect non-neuronal ectoderm compared with other arthropods. The temporal regulation of expression in *D. melanogaster*, although interesting, is probably not sufficiently consistent or unique to explain the conserved synteny either. All of the *Osiris* genes have peaks of expression in one or more of three specific stages of development: 12–18 hr old embryos, second instar larvae, or pupae at 2–3 days post-white-prepupal stage (Gelbart and Emmert 2010). None is expressed to any great extent outside of these three times. However, there is variation within the *Osiris* family, with some being expressed well at all three times, some predominantly in embryos, some predominantly in pupae and at least one predominantly in second instar larvae. This finding indicates that the *Osiris* paralogs maintain their individual functions with some degree of differentiation among them. That the synteny has been conserved so well through such long periods of time, in spite of the high rate of chromosome rearrangement in the genus *Drosophila* (Bhutkar *et al.* 2008; Ranz *et al.* 2001) suggests strong selection on co-localization. In addition, high rearrangement rates were recently shown for coexpressed *Drosophila* genes with short intergene distances (Weber and Hurst 2011), making the synteny of the *Osiris* gene cluster even more remarkable.

As mentioned previously, the *Osiris* signature domains are highly conserved among the majority of the *Osiris* family members. The exceptions are the aforementioned *Osiris* 4 and *Osiris* 13. *Osiris* genes that are not located in the syntenic cluster (*Osiris* 21, *Osiris* 22, and *Osiris* 23) also have more weakly conserved signature domains (Fig-

ure S1). This implies again the possible association between the co-localization of *Osiris* genes and their functions.

The *Osiris* gene family was an early evolutionary innovation in the divergence and spread of insects. Its conserved domains are unique. The extreme dosage effects of the cluster in *Drosophila melanogaster* may be part of the explanation for the selection on synteny. Clearly the *Osiris* gene family is in need of further study.

LITERATURE CITED

- Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang *et al.*, 1997 Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25: 3389–3402.
- Arensburger, P., K. Megy, R. M. Waterhouse, J. Abrudan, P. Amedeo *et al.*, 2010 Sequencing of *Culex quinquefasciatus* establishes a platform for mosquito comparative genomics. *Science* 330: 86–88.
- Bailey, T. L., N. Williams, C. Misleh, and W. W. Li, 2006 MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res.* 34: W369–373.
- Bailey, T. L., M. Boden, F. A. Buske, M. Frith, C. E. Grant *et al.*, 2009 MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* 37: W202–208.
- Bhutkar, A., S. W. Schaeffer, S. M. Russo, M. Xu, T. F. Smith *et al.*, 2008 Chromosomal rearrangement inferred from comparisons of 12 *Drosophila* genomes. *Genetics* 179: 1657–1680.
- Birney, E., M. Clamp, and R. Durbin, 2004 Genewise and enomewise. *Genome Res.* 14: 988–995.
- Bonasio, R., G. Zhang, C. Ye, N. S. Mutti, X. Fang *et al.*, 2010 Genomic Comparison of the Ants *Camponotus floridanus* and *Harpegnathos saltator*. *Science* 329: 1068–1071.
- Camacho, C., G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos *et al.*, 2009 BLAST+: architecture and applications. *BMC Bioinformatics* 10: 421.
- Clark, A. G., M. B. Eisen, D. R. Smith, C. M. Bergman, B. Oliver *et al.*, 2007 Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature* 450: 203–218.
- Colbourne, J. K., M. E. Pfrender, D. Gilbert, W. K. Thomas, A. Tucker *et al.*, 2011 The ecoresponsive genome of *Daphnia pulex*. *Science* 331: 555–561.
- Cranston, P. S., and P. J. Gullan, 2009 Phylogeny of insects, pp. 780–793 in *Encyclopedia of Insects*, Ed. 2, edited by V. Resh, and R. Carde. Academic Press, San Diego.
- Crooks, G. E., G. Hon, J. M. Chandonia, and S. E. Brenner, 2004 WebLogo: a sequence logo generator. *Genome Res.* 14: 1188–1190.
- Dorer, D. R., J. A. Rudnick, E. N. Moriyama, and A. C. Christensen, 2003 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.
- Durbin, R., S. R. Eddy, and A. Krogh, 1998 *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*, Cambridge University Press, Cambridge.
- Felsenstein, J., 2010 PHYLIP (*Phylogeny Inference Package*) version 3.69, Distributed by the author, Department of Genome Sciences, University of Washington, Seattle.
- Finn, R. D., J. Mistry, J. Tate, P. Coggill, A. Heger *et al.*, 2010 The Pfam protein families database. *Nucleic Acids Res.* 38: D211–D222.
- Gelbart, W. M., and D. B. Emmert, 2010 FlyBase High Throughput Expression Pattern Data, Beta Version. Available at: <http://www.flybase.org>.
- Grimaldi, D. A., and M. S. Engel, 2005 *Evolution of the Insects*, Cambridge University Press, Cambridge, U.K.
- Hill, C. A., and S. K. Wikel, 2005 The *Ixodes scapularis* Genome Project: an opportunity for advancing tick research. *Trends Parasitol.* 21: 151–153.
- Honeybee Genome Sequencing Consortium, 2006 Insights into social insects from the genome of the honeybee *Apis mellifera*. *Nature* 443: 931–949.
- International Aphid Genomics Consortium, 2010 Genome sequence of the pea aphid *Acyrtosiphon pisum*. *PLoS Biol.* 8: e1000313.

- Kall, L., A. Krogh, and E. L. Sonnhammer, 2004 A combined transmembrane topology and signal peptide prediction method. *J. Mol. Biol.* 338: 1027–1036.
- Katoh, K., and H. Toh, 2008 Recent developments in the MAFFT multiple sequence alignment program. *Brief. Bioinform.* 9: 286–298.
- Kirkness, E. F., B. J. Haas, W. Sun, H. R. Braig, M. A. Perotti *et al.*, 2010 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.
- Lindsley, D. L., L. Sandler, B. S. Baker, A. T. Carpenter, R. E. Denell *et al.*, 1972 Segmental aneuploidy and the genetic gross structure of the *Drosophila* genome. *Genetics* 71: 157–184.
- Munoz-Torres, M. C., J. T. Reese, C. P. Childers, A. K. Bennett, J. P. Sundaram *et al.*, 2011 Hymenoptera Genome Database: integrated community resources for insect species of the order Hymenoptera. *Nucleic Acids Res.* 39: D658–D662.
- Nene, V., J. R. Wortman, D. Lawson, B. Haas, C. Kodira *et al.*, 2007 Genome sequence of *Aedes aegypti*, a major arbovirus vector. *Science* 316: 1718–1723.
- Pagel Van Zee, J., N. S. Geraci, F. D. Guerrero, S. K. Wikel, J. J. Stuart *et al.*, 2007 Tick genomics: the Ixodes genome project and beyond. *Int. J. Parasitol.* 37: 1297–1305.
- Price, M. N., P. S. Dehal, and A. P. Arkin, 2010 FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS ONE* 5: e9490.
- Quijano, C., P. Tomancak, J. Lopez-Marti, M. Suyama, P. Bork *et al.*, 2008 Selective maintenance of *Drosophila* tandemly arranged duplicated genes during evolution. *Genome Biol.* 9: R176.
- Ranz, J. M., F. Casals, and A. Ruiz, 2001 How malleable is the eukaryotic genome? Extreme rate of chromosomal rearrangement in the genus *Drosophila*. *Genome Res.* 11: 230–239.
- Stanke, M., M. Diekhans, R. Baertsch, and D. Haussler, 2008 Using native and syntetically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* 24: 637–644.
- Tribolium Genome Sequencing Consortium, 2008 The genome of the model beetle and pest *Tribolium castaneum*. *Nature* 452: 949–955.
- Tusnady, G. E., and I. Simon, 2001 The HMMTOP transmembrane topology prediction server. *Bioinformatics* 17: 849–850.
- Weber, C. C., and L. D. Hurst, 2011 Support for multiple classes of local expression clusters in *Drosophila melanogaster*, but no evidence for gene order conservation. *Genome Biol.* 12: R23.
- Whitfield, J. B., and K. M. Kjer, 2008 Ancient rapid radiations of insects: challenges for phylogenetic analysis. *Annu. Rev. Entomol.* 53: 449–472.
- Xia, Q., Z. Zhou, C. Lu, D. Cheng, F. Dai *et al.*, 2004 A draft sequence for the genome of the domesticated silkworm (*Bombyx mori*). *Science* 306: 1937–1940.

Communicating editor: D.-J. De Koning

1_CG15585 .. CR---VVLAL---LLL---VLLHY---CQAKDEGETTQRGAVLVSLT .. -----ENMAKSAGSQL-V--QADPFISRQKQC-----FET-RSLVSCIYKYT
2_CG1148 .. --MAMRALIFL---ALA-----TLVAGEGLRLPDQSSNNIQ .. -----GNDNDPFLARTNSNCL-----G-GDLSECFKTOA
3_CG1150 .. QTFKVCALLAF---CFV-LVSAR .. -----GSKRRDGTVTISESERK-N--IEDFLLAKLKQNCR-----QED---DRACKMVKM
4_CG10303 .. HLVASCILLAL---G-L--NMSLA---AIHKRSGANS--LGVD--- .. -----GKPA-ANPAVS-V--ENTDLLDKLSWK-----ANNASCLYGVA
5_CG15590 .. -----MFRTF-PLLCL---LFLT-----AVRSENC-----DQDAGATLYCRGERA
6_CG1151 .. FVATACILLLLA-AGISADPVKAA .. -----EEQGAFAQCL-----ESD---SISCLQLTL
7_CG1153 .. MASHKVTFGVL---CLV--ALSAA .. -----LPAEETRGRHARNAI-G--GENDIMDSIYSACL-----RKD---SVSCVKYKL
8_CG15591 .. -----MIKYV---WHV--AALMIVFOWLSSARSASYQH .. -----SNPGMGSTGL-W--KDMSMVYRIYQCS-----GDN---MSVCLKVKL
9_CG15592 .. ---MFKFVCL---FAL-IASTA .. -----AATSEA-D--SLLTSALKMVKDCG-----ERS---MVLCKMERA
10a_CG15593 .. -----MSPL-DIVLL--LVSF---HQVIGSDFSAMSLE .. -----FKQCVRG-----SOK-PKIGECLGRSA
10b_CG15593 .. -----CLGSES
11_CG15596 .. --MESRLGLLI---ALL-LAAFQA .. -----WAMEA--PSNYNQNSTE-S--GLLRTVRIHYGQC-----AYS-EDVFWCCKIQG
12_CG1154 .. KQWPWQ--ALI---SLLSVLFLA .. CGAATEQL---GSPPGPSPSQ-S--AGARTLLRVYDECT-----RAE-AGFVPCPKKA
13_CG15595 .. FKSSISIVVLH--LLL--LVSTG--GAYTL .. -----FAAPSVQDNQVEGDNTLGRAARYLGAC-----LES-DDMATCLAVKG
14_CG1155 .. ---MKVFAIA-CVTLL-AASCV .. -----MCACFVC---VVL--LASLV---C-GSMALPSQDNT-----ERDL
15_CG1157 .. -----MLLTKTV-KYLFY-LALFA .. FMCKYATAASVQPTVEP-AVAPETIRIPQRA-ESLLSGC-----EAS-SFSWMCKLKF
16_CG31561 .. ---FLL--LICLIN--SAA---KAD--GTARN--GRLGRHLSTTP .. -----PT---RKPM-D-K--MAPSDSLLLRLARR---FASGNELWDGLVRDCYL-KPDVSCFQKNV
17_CG15598 .. ---MKSTA-ACLIV--ALAAL---STAHS-- .. -----PTEGVV--AP-Q--SATQLALDMYHGC-----L-K-DLSVSCVVRKA
18_CG1169 .. ---MAKLLLIVGVAAL .. VAAGQAAGGSTE--KMQRLIAEEQKNC-----ASG-QDSMACIKERA
19_CG15189 .. AFR---STSLL---AFG--CALLL---VASTSVGAAIENA .. -----V--TPRI-H--SSDELISTIYDKCF-----H---ANAMHCKLKEV
20_CG15188 .. ---KRLEWLL---LLA--LVASV---STAVTPRRRRHSAVESA .. -----PGDWGTAWGL-G--PEMALVRRVYDDC-----QDK-NDFIGCLKQKA
21_CG14925 .. ---MSDLVR--FLL--LSVLC---SSLALAQASSDG .. -----NQETST--VS-Q--EAARGLA-SSYEPEDKQALRKNSHIFMGIYKNIYS--TYLGNKTTSEY
22_CG8644 .. R--VPSVLVT---FLL--GVILV---DRYAAGEDLADK .. -----SW-I--SQMKLRSDLRDCYQSGIHQS-----LWSCFRSRS
23_CG15538 .. NSHRKRRCSPL---LYG--IILLC---KVAMIPAAVPE-SELDAGAL .. S-----PRKQESSESA-A-A-FK-----N-----QTITSI-DAELKGLLEDLL

1_CG15585 S-KLIWKLA---TNSMGFFPSEYGRDLAGD-----RGRWLRVLQV-----GEPADVVVFNDAKSLEGDSSELTM-----
2_CG1148 L-NTFDEIFF-----KDQYK-----LSDFARVRLPE-----TQQR--SLLQEPF-----EYSEPRGDDDEW
3_CG1150 S-IVMNHLY--L-----NTRID-----LGDRFKVTENGN-----ISMV-----PDDPEVNLRSRM-----GSDEET-----
4_CG10303 N-GLMASYRR-----GETLK-----LGLFDLVKLPEDA-----SRKHKW---GTGRG-----
5_CG15590 L-RNVLRLNLN--RS-----DKPLV-----VIRGLEIVPLQN-----NSIS-----D
6_CG1151 F-RKAKSVF--D-----NPQIE-----LFGGVSLVKSNE-----GRQG-----KSLDNSLAVEAAPT-----EARTAE-----
7_CG1153 F-SFVDKVL--GA-----RDQFA-----LTEGVTVVRSPD-----APQQ-----EAARSI-----SGDES-----
8_CG15591 L-TGLEKAF--RS-----AKSLS-----LMEGIQFVSSGG-----ESEE-----TKRAPISEKDIEAVLPRSV-----DAKEQV-----
9_CG15592 L-HYFDAE-----NGDVR-----LTEGIALVKTE-----IPVG-----RSLNEMQLPEEV-----EAREAE-----
10a_CG15593 L-NFIQKLD--E-----SDNVK-----FVEDFVTVKSET-----AAVR-----SLSNVLDTPVDF-----
10b_CG15593 E-QLLDGAT--RD-----NSTWQ-----ITDYLSIEPK-----V--GISKPE-----TRRMD-----
11_CG15596 V-RLLGRAL--K-----VPQLG-----IVDGVSLVRR-----ESFT-----QDTRSGRSSLLESQ-LSNRDLE-----HLSGKS-----
12_CG1154 I-SFIDRLA--P-----IDAIN-----VAEGIKLVRLET-----APRP-----PATSENELESSLPRSG-----SDRDAK-----
13_CG15595 PPVQ-----MGGAIVAAVEQD-----AEQE--AAAEER-----QRVERHWLSMAE
14_CG1155 I-TALNRAA--R-----SNNIE-----LASGVTQORDPA-----SPVS-----RTGKSMSEQDVYALPQNA-----DERTGR-----
15_CG1157 V-NMLNRLD--S-----EESVA-----LFGGLRIDRSES-----GRSF--GASKAV-----
16_CG31561 V-KIMEKLA--E-----QEELN-----VLPGISVVKDEN-----A--TELKTS-----ELMAEVARSYP-----SDPSTR-----
17_CG15598 F-SYLDNVLDV-----QDYN-----VTQRLKFFKNQVDYQVDKEKEH-----SEARAA--SAETPIEE-----
18_CG1169 L-QWFNSAL--R-----QPEVR-----ITERLSIVRTAE-----KVE-----SRS
19_CG15189 M-RFVDNM--S-----KDSFQ-----VSN-LEVRNNGE-----KTPPINEARAS-----
20_CG15188 L-TYLDTVA--NV-----EEEVS-----GRA-----LGDDV-----
21_CG14925 L-HALSRAL--D-----QDSIK-----IVDGLALEKQNG-----SETESILG-----SLTDARQFG--NL
22_CG8644 KKRLRDRVSA-----PQMAENETPEMVEPDQEDQRDSLAEIRQDAQAE-----ALME--TQTESPNYDDSESLAAKRRRKRKRDRNKRDE-----V-----ESETD
23_CG15538 L-HIFEGIM--S-----SPEIS-----IYDGVRLVAA-----P--NS--TD-NATRPDDERKDLKHLT-----
24_CG15589 V-DYVEQFFS-----NGRYE-----PTPGLVLALQON-----HSHP--QSYTG-----KR-----

```

1_CG15585    ---ILKFL---KRAMETFGRNH-GLQLRLN---SEGGARVMEES----- .. E----AR- .. -----LK-RKKKK--WLI---ILPLVILMKIA--HLK--
2_CG1148    --NQLLKYG----LRRARFIKST-ALEVWEP---E-ELTEAGRYEARFIGNDIDGELDLIDDG .. QRAGHFSR- .. -----K-KL-KK---M---IIPLLLVLKIF-KLKL--
3_CG1150    ---FALLM---ANKLWKFIRSR-SLRYKFS--E-NTDFVI---NSDPEGSLNLGVSVRPL .. E----GR- .. -----G-KM-K---N---MGPLIMMAAK---T--
4_CG10303    -----LSGFMDFVTEN-AIRVPVG---P-MVFSVQRAEDSDYIEVALLKKTSSST .. -----GRL .. G--RRRHQH-QDK-KQ-FQ---M---FIPMYLAATTF---G--
5_CG15590    EEPDQEQGL---LDSLSFYLRTH-EINVKLA---D-LLEDES---QVS----- .. E----AR- .. -----KK-DK---G---QGMLLAMALMF---G--
6_CG1151    ---MGNYF---MDNAKSFFAER-SLNFNFA---NAARSVARAI PDDIKADLREL .. E----SR- .. -----TRKK-KL-LK---K---FLPILLGVGAK---I--
7_CG1153    ---FESLA---LNRISFNLNH-TIKVELK---G-ADIVQA--VSSTGR----- .. E----SR- .. -----GKKK-KA-AK---I---LGPILALVALK---A--
8_CG15591    ---LNNMI---LKRVGNFQDH-TLQVKFDN---EANSV----- .. E----GR- .. -----K-KKEKK---G---NGAMIMIPLLL---G--
9_CG15592    ---VDSLL---VERVARFFGTH-TLQFKVP--K-----DSIQDMQRALE----- .. E----SR- .. -----GKKK-EK-KK---Y---LMPLLMLFKLK---M--
10a_CG15593 ---RGI---LENAGAVMQQR-SMEWHMD----- .. T----GR- .. -----V-LT-KQ---Y---LLPFLGLGKFN---L--
10b_CG15593 ---MGL---PGKLELVQGR-ALRLQLP--R-QLTISNAIDDFGSELGLD----- .. Q----GR- .. -----K-KK-DK---D---KNMAMMGMMIM---M--
11_CG15596    ---LDALL---LERFLNFVHSH-QLQVNLN---RLLRFGE--RNVQDWLLHVVGYPMPAS .. E----GR- .. -----KKKDDK---Y---LGPPIAAVLLK-----
12_CG1154    ---LTNML---IERLSYFFNGH-SLQVSFP--K-----LTSDEIGRGLE----- .. E----GR- .. -----G-KM-KK---M---MGMMMGMAMK---M--
13_CG15595    --TQLHSLITDDLSSTEEVNNML----- .. ETWSTEGR- .. --GKH--KKQK-KL-MK---M---VYPLAAVAVA---K--
14_CG1155    ---LVDLA---VSSAADFLSTH-NLEFKLP--A-----ETTQQVARALD----- .. E----GR- .. -----G-KI-KK---M---LGPVALAIGAK---L--
15_CG1157    ---ESF---EDRAERYLETH-ELNLSFS--G-DEQDENSENEYTGRAMD----- .. E----SR- .. -----SK-RM-KK---M---LLPPLLALKLK---K--
16_CG31561    ---LNGYI---VAKLENLLRTR-FLRFRLN---D-DKSLV----- .. E----GR- .. -----KHKFGK-KG-----G---LEALVAAGVMM---K--
17_CG15598    ---VTSAL---YGKSIKAMTHDLEVDLPE-----VM- .. ALLLIKI- .. -----IK-----IKLFWLLPIVIGVGAACKLLLLK
18_CG1169    ---MNPERRL---FDDIDSYLGSN-SLRIQAP--E-YFRTSE-ARSLVPDFLMSNPLTQGGLV .. E----GR- .. -----G-MI-RK---A---VLPFLGLGKLK---T--
19_CG15189    ---SADGF---LDAIENYIRGH-DVSMIDL--L-ADAKVTVSARNLVNQNLSLNLQNGDD .. E----AR- .. GKKGNIFKKGKKH-RL-RK---L---AMPILVLILLK---A--
20_CG15188    ---LDKVI---VDRGLRILNTN-EMRLQLP--Q-TFFAGSVVTVRSRDRG--FDLE-LPKD .. E----KK- .. -----K-DK---L---FLPLLLLMKFK---L--
21_CG14925    ---SPIDRAL---LSKADKLMRTH-TLKIDMDVGGGSDS--VGR----- .. EH- .. -----GHKK-KKHKEGGH---IKYVVAALLTA---M--
22_CG8644    --QPDPAPE---DETIQRYNVGP-GLNVSID--MS-NDIVHVKLDGENLKEIIGARWLTLDNS .. E----GRG .. -----KKYDMITK---VLPFLAIPFLI---Q--
23_CG15538    ---WFDQL---AVSLAKGLTTH-TLQVNLG--K-LTERYLSSDP-----VG----- .. S----AR- .. -----RR-HR-YN---M---IITMMFGVTAL---G--
24_CG15589    ---TTRSI---LE----- .. RLLFFSGL- .. -----KK-----V---MWPIYMLQVLKSVLF--

```



```

1_CG15585    MTLVSMIMG---VLGMNVL .. LVGGVGLIHYLKYKTMCKIHP----- .. WATSKAYNA-HNYLDTISKRIQ----- ..
2_CG1148    LLFLPFIILGIAGLKK--- .. ILGLAAIIVLPGL---FAYFK-----LCRPPGGVGA-FGGGLSGLFGKNT .. -PQEAAYNGYGRNSGKDIVAEQQPQKS ..
3_CG1150    GMVGALLLKGFLFLLAGKAL .. IVSKIALLLAVI---ISLKK-----LL-S--SK----- .. -AQNMAYSG-QQPGKVAQ----- ..
4_CG10303    -GWTMVAAKAVGLLTLKAL .. ILSKIAFVVAAI---VLIKK-----LMDN----- .. HM--PYRS----- ..
5_CG15590    KMMAVMGLGGIAALAMKAL .. GVSIVALMMAGM---LGLKT---AA-Q--HG----- .. -LAYRG-WD----- ..
6_CG1151    AVLGVGSIFGLLFLAKKAL .. VVSVIAFFLALA---AGASS-----GLGRIGSGG-GGGLLGLGGLFGGK .. -AQTIAQQG-YKQARR----- ..
7_CG1153    AALLPPLLGAIALIAGKAL .. LIGKIALVLSAV---IGLKK-----LL-S--QE----- .. -AQDLAYGA-QKPVQA----- ..
8_CG15591    GTIVPLAYGALAMLAGKAL .. IVSKLALVLSI---IGIKK-----LL-S--GGGG-G----- .. -ELDTAYSG-WKPAAKESAGSAKSL--- ..
9_CG15592    AALLPLAIGFLALISFKAL .. VIGKIALLLSGI---IGLKK-----LL-E--SK----- .. -AQQLAYAA-YKQ----- ..
10a_CG15593 VALVPLIFAGICLLLKSL .. FLVKLAIYVSSF---LGLGG-----IVGGL- .. TVFGKQDEFHH-----QYD ..
10b_CG15593 ATLAQMFGLKGVILIAGSAF .. IMAKIALVISLL---GSLKK-----GS-T--GHS----- .. GSHMEYYQA-YQMEPLKRR----- ..
11_CG15596    TAILKMAYHSIAIVAGKAL .. IVGKIALIISAI---IGLKK-----LV-G--HDGG----- .. -MQDKAYQA-WMPHVAASPSVAKGS--- ..
12_CG1154    MGMPIAMGALYILAGKAL .. IISKIALLLAGI---IGLKK-----LM-S--GKSS-GGSSGWSS----- .. -AQELAYRA-HHQEQVAHAQSRPQ--- ..
13_CG15595    VVLLPLILKWLTALESTSSF .. VMGKIALVTSGI---LALKW-----ILSGGAHDRL-E----- .. -IISH--APLVKGLH---ASDLSS ..
14_CG1155    FAVIPLVLGFLALLTFKAV .. IVAKLAFFLAIL---VGGSR-----LL-G--GFGN-KFGG---NSFAGAY .. -AQQLAYAG-QQQQ----- ..
15_CG1157    AVVVKIMFTTIKFIKAL .. AISFLALILAGA---TFFKD-----LL-A--KKK----- .. -AADLAYNH-YGLAQPF----- ..
16_CG31561    GMLMAMGLGALAMGKAL .. MTALMALTSVSGV---LGLKS-----LA-G--GGG----- .. -FGYGG-YARSLKVDQSANKI--- ..
17_CG15598    LLFLFPALSHLFLKCSHYQ .. ILLKIDKIIIEQL-----GVKN--DLCKERIVCSMYKDPATYSP-HSNFISAEISRDT .. -YRLIQAARDQDQDKCQSLYPQCN ..
18_CG1169    TVLVPLALGLIALKTKWAM .. TLGLLSLVLSGA---LVIFK-----IAKP----- .. -AQDLAYAGQK----- ..
19_CG15189    ITVIPMAIGILKIKAFNAL .. ALGFFSFIVSG---LAIFQ-----LCKK----- .. LGQALAYQA-YA----- ..
20_CG15188    KVIMPILLALIGLKATKAL .. ILSKIAIKLVLG---FLIYN-----LI-Q--KLG--GMK--MNMVMPMA .. -SQNLAYSS-YHPSSSSSYSSGSSGSS ..
21_CG14925    GIAGPLGLKALAAIAGKAL .. VISKVALTIAGI---IALKK-----LFSH----- .. -DPYRYYYEYHQ----- ..
22_CG8644    SAIVPFLVTKLKLKLVKSI .. LVGKLAIFLLII---SAIKN----- .. SAAAAAYNG-YRVEGKPTTWIS----- ..
23_CG15538    AILVPMGFQMLSIIVSGKAL .. LLAKMALLLAI---NGLKR-----VA----- .. -NNGLHYGLYHPGE---HLGGY-YDRG ..
24_CG15589    AMFLPTIISVSRLIGKGI . ATLSRY . LLSRLDSVFAQL-----KLNPNENACREKLICLMYANPAKYAP-YSNLVAQLSRELN .. -FKYMRRAKDQGDVDCDESFACK ..

```

Figure S1 Alignment of *D. melanogaster* Osiris protein sequences. Multiple alignment was generated using the entire Osiris sequences and aligned *D. melanogaster* sequences are extracted and presented above. The five Osiris signatures are color-coded: predicted signal peptide in red, 2-Cys region in orange, duf1676 region in blue, predicted transmembrane

region in green, and AQXLAY motif in purple. Long insertion regions are excluded and indicated by "..". Osiris 10 proteins are divided into two parts, and included in the alignment as 10a and 10b.

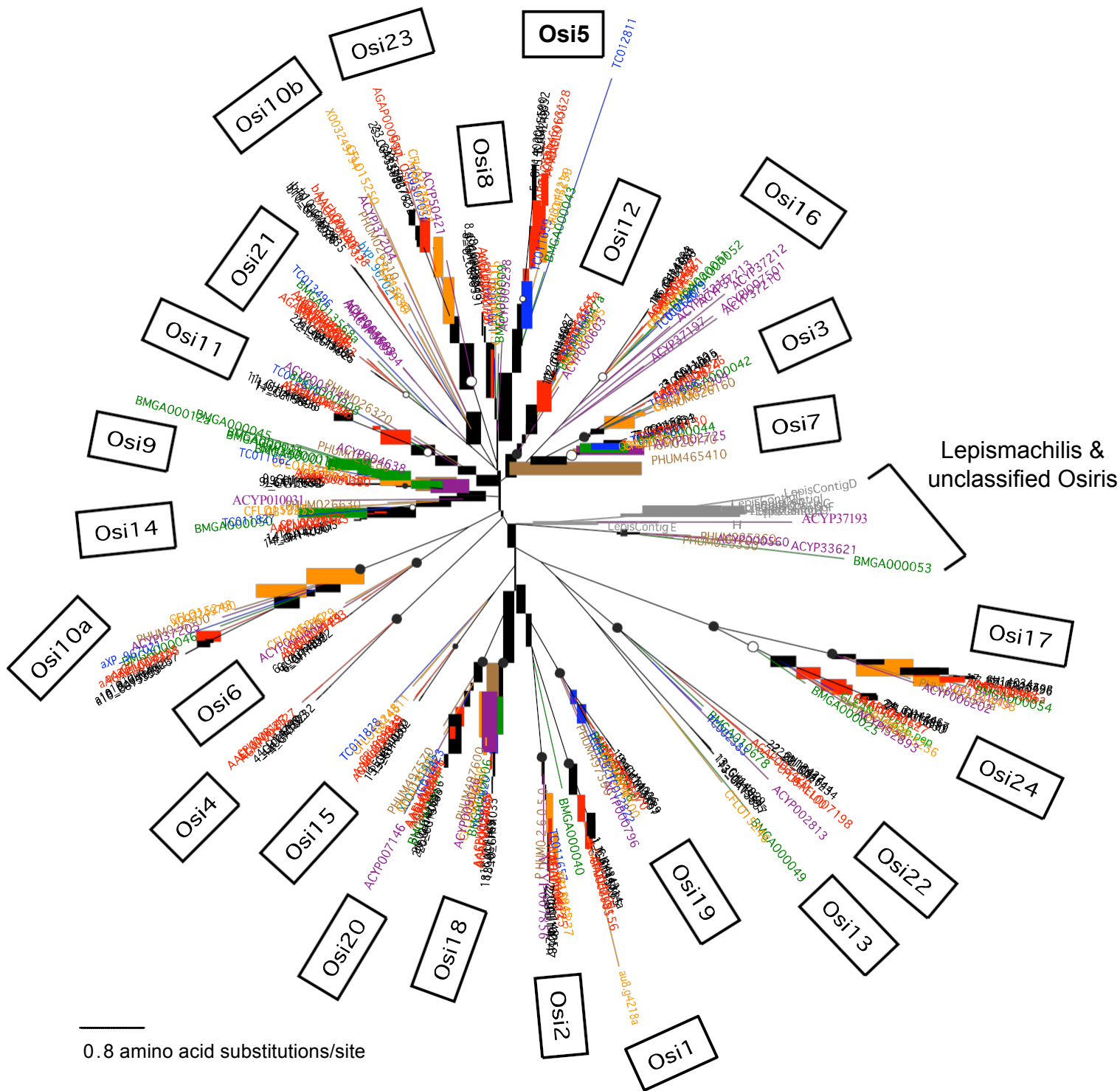


Figure S2 The maximum likelihood phylogeny of Osiris proteins reconstructed by FastTree. See Supplementary Table S2 for the sequences included in this phylogeny. Sequence labels and branches are color-coded based on species: four *Drosophila* species in black, three mosquito species in red, *B. mori* in green, two hymenoptera species in orange, *T. castaneum* in blue, *A. pisum* in purple, and *P. humanus* in brown. Ten sequences assembled from *Lepismachilis y-signata* ESTs are also included and shown in grey. Osiris 10 sequences were divided into two parts and aligned individually. These sequences are shown as "Osi10a" and "Osi10b" above. Major clusters supported by higher than 90% or 70% bootstrap supporting values are shown with solid and open circles, respectively. Different sizes of the circles are used for the supported clusters that include all species in the group (large circles), only holometabolous insects (middle circles), or only dipteran species (small circles).

Table S1 Organisms found to have sequences similar to Osiris proteins.^{a, b}

[Phylum] Subphylum; Class [Arthropoda]	Superorder	Order	Number of species and list of genera
Hexapoda; Insecta			
Dicondylia; Pterygota; Neoptera			
	Endopterygota	Diptera	25 <i>Drosophila, Aedes, Anopheles, Culex, Cochliomyia, Glossina, Haematobia, Lucilla, Phlebotomus, Polypedilum, Simulium, Sitodiplosis, Teleopsis</i>
		Lepidoptera	16 <i>Bicyclus, Danaus, Heliconius, Choristoneura, Tineola, Papilio, Bombyx, Antheraea, Manduca, Spodoptera, Mamestra, Heliopsis, Trichoplusia, Ostrinia</i>
		Hymenoptera	9 <i>Nasonia, Bombus, Apis, Megachile, Acromyrmex, Solenopsis, Camponotus, Harpegnathos</i>
		Coleoptera	5 <i>Callosobruchus, Dendroctonus, Tribolium, Diabrotica, Onthophagus</i>
	Paraneoptera	Hemiptera	11 <i>Acyrtosiphon, Adelphocoris, Aphis, Diaphorina, Homalodisca, Myzus, Maconellicoccus, Nilaparvata, Peregrinus, Rhopalosiphum, Toxoptera</i>
		Phthiraptera	1 <i>Pediculus</i>
	Dictyoptera	Blattodea	1 <i>Blattella</i>
		Isoptera	2 <i>Reticulitermes, Coptotermes</i>
	Orthopterida	Orthoptera	3 <i>Schistocerca, Locusta, Gryllus</i>
	Monocondylia	Archaeognatha	1 <i>Lepismachilis</i>
Hexapoda; Entognatha			
	Collembola	Entomobryomorpha	1 <i>Folsomia</i> ^c
Crustacea			
	Malacostraca	Decapoda	2 <i>Penaeus</i> ^c , <i>Homarus</i> ^c
	Branchiopoda	Cladocera	2 <i>Daphnia</i> ^c
[Platyhelminthes]			1 <i>Clonorchis</i> ^d

^aSearch was done using blastp and tblastn using each of *D. melanogaster* Osiris protein sequences as queries against the non-redundant protein and EST databases at NCBI. The E-value threshold is 0.01.

^bHighly similar sequences were identified from several plant EST and cDNA sequences. However, these sequences are not included in this table. Such similarities are not found from any complete plant genomic sequences suggesting insect contamination of plant materials. Examples of these cases include: CX523685.1 (an EST from *Medicago truncatula*) with higher than 95% protein similarity ($E \sim 10^{-66}$) against *Drosophila* Osi7, and BT086177.1 (a complete cDNA from *Zea mays*) with 64% protein similarity ($E \sim 10^{-69}$) against aphid Osi6.

^cWeakly similar EST sequences were found from Collembola ($E \sim 10^{-4}$) and Crustacea ($E = 10^{-18} \sim 10^{-4}$). Reciprocal blast search indicated possible homologous relationships with *Drosophila* Osi17 and Osi24.


^dOne Platyhelminthes EST sequence (FS162944) had high protein similarities against *Drosophila* Osi7 (84%, $E \sim 10^{-46}$).

Table S2 List of all *Osiris* genes identified from the 13 complete insect genomes.

ID: All IDs are for the first transcript/peptide (-RA/-PA) unless noted otherwise

Direct: 1/-1 show the direction relative to Osi2 (1 as the same direction).

Start/End: Start and end positions for the coding sequences. For *A. pisum*, transcript start/end positions are shown. When genes are not on the same chromosome or supercontig, only the strand information (+/-) is shown.

Color codes:  annotation changed by us in this report

(*Osiris* group assignment)

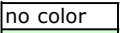

 based on both synteny and similarity
 based on synteny but similarity not conclusive

Table S2 List of all *Osiris* genes identified from the 13 complete insect genomes.

Osiris 1								
Genome	ID	Length (AA)	Chromosome (linkage group)	Direct	Start	End	# exons	Note
<i>D. melanogaster</i>	CG15585	308	3R	1	1999367	2000555	3	
<i>D. pseudoobscura</i>	GA13829	304	2	1	15819977	15818610	3	GA13829-PB (FBpp0298266)
<i>D. virilis</i>	GJ14243	323	scaffold_12822	1	1897146	1895783	3	
<i>D. grimshawi</i>	GD14014a	307	scaffold_14624	1	2876740	2875141	3	based on the last 2 exons of GH14014 GeneWise+Augustus model 2875141..2875544,2875805..2875927,2876344..2876740
<i>An. gambiae</i>	AGAP004121	284	2R	1	50327269	50328282	3	
<i>Ae. aegypti</i>	AAEL015156	282	supercont1.1530	1	22088	5471	3	
<i>Cu. quinquefasciatus</i>	CPIJ006985	293	Supercontig3.144	1	369648	358007	3	
<i>B. mori</i>	None							
<i>Ap. mellifera</i>	au8.g4218a	276	Group15.14	1	546442	548486	4	Based on au8.g4218 ex1-ex4: (546442..546663, 547241..547508,547660..547920, ,548407..548486) Similarity is weak, no signature domains; but cluster on
<i>Ca. floridanus</i>	None							
<i>T. castaneum</i>	None							
<i>Ac. pisum</i>	None							
<i>P. humanus</i>	None							

Table S2. List of

Genome	NPFR1							Note
	ID	Length (AA)	Chromosome (linkage group)	Direct	Start	End	# exons	
<i>D. melanogaster</i>	CG1147	485	3R	1	2014920	2019871	4	4 isoforms, all identical aa seqs
<i>D. pseudoobscura</i>	GA11019	498	2	1	15808917	15807000	3	
<i>D. virilis</i>	GJ14245	516	scaffold_12822	1	1885599	1883380	3	
<i>D. grimshawi</i>	GH14016	511	scaffold_14624	1	2866563	2864216	3	
<i>An. gambiae</i>	AGAP004123	425	2R	1	50350559	50354332	3	
<i>Ae. aegypti</i>	AAEL010626	351	supercont1.491	1	826221	794272	3	
<i>Cu. quinquefasciatus</i>	CPIJ006984	400	supercont3.144	1	326156	314299	3	
<i>B. mori</i>	BGIBMGA00001	383	nscaf1071	-1	456503	450075	4	Located after Osi24
<i>Ap. mellifera</i>	None							
<i>Ca. floridanus</i>	None							
<i>T. castaneum</i>	TC011655	619	ChLG9	1	20609466	20598308	6	TCOGS2:GLEAN_11655 (XP_967689 contains NPFR1+Osi24)
<i>Ac. pisum</i>	ACYPI007664	391	GL349633	+	115595	205777	6	
<i>P. humanus</i>	PHUM025830	364	DS235004.1	1	376037	377170	2	

Table S2. List of

Osiris 24								
Genome	ID	Length (AA)	Chromosome (linkage group)	Direct	Start	End	# exons	Note
<i>D. melanogaster</i>	CG15589	533	3R	1	2030167	2032903	5	
<i>D. pseudoobscura</i>	GA13830	537	2	1	15796968	15795081	5	
<i>D. virilis</i>	GJ14246	472	scaffold_12822	1	1850145	1847962	6	
<i>D. grimshawi</i>	GH14017	497	scaffold_14624	1	2854383	2852256	5	
<i>An. gambiae</i>	AGAP004124	409	2R	1	50384354	50387218	5	
<i>Ae. aegypti</i>	AAEL010627	445	supercont1.491	1	449215	396633	5	
<i>Cu. quinquefasciatus</i>	CPIJ006981	427	supercont3.144	1	144904	128341	5	
<i>B. mori</i>	BGIBMGA000025	479	nscaf1071	-1	273107	262711	7	Located before NPFR1
<i>Ap. mellifera</i>	GB18863	473	Group15.14	1	552773	555471	4	Amel_2.0_OGSv1_gmap (XP_001121409 LG15:4223598..4226296, no longer in NCBI)
<i>Ca. floridanus</i>	CFLO15236	471	scaffold309	1	116909	101296	6	OGSv1.0 (EFN64650/GL441542.1)
<i>T. castaneum</i>	GLEAN_11656	418	ChLG9	1	20593022	20590333	4	Glean_5_19_06 (XP_967689 contains NPFR1+Osi24)
<i>Ac. pisum</i>	ACYPI52893	501	GL349761	1	61451	21695	7	
<i>P. humanus</i>	PHUM026040	295	DS235004.1	1	403565	405857	5	5'-end missing

Table S2. List of

Genome	Osiris 2							Note
	ID	Length (AA)	Chromosome (linkage group)	Direct	Start	End	# exons	
<i>D. melanogaster</i>	CG1148-PA	390	3R	1	2037184	2039100	4	isoform
	CG1148-PB	390	3R	1	2037184	2039100	4	isoform
<i>D. pseudoobscura</i>	GA13830	404	2	1	15789619	15788078	4	
<i>D. virilis</i>	GJ14247	397	scaffold_12822	1	1842804	1841325	4	
<i>D. grimshawi</i>	GH14018	406	scaffold_14624	1	2847740	2845905	4	
<i>An. gambiae</i>	AGAP004125	369	2R	1	50398489	50399671	2	
<i>Ae. aegypti</i>	AAEL010623	365	supercont1.491	1	232553	231401	2	missing the 3' end
<i>Cu. quinquefasciatus</i>	CPIJ008149	362	Supercontig 3.179	1	692137	690987	2	
<i>B. mori</i>	BGIBMGA000040	142	nscaf1071	1	584271	584699	1	missing 3' half, weak similarity
<i>Ap. mellifera</i>	GB15845	325	Group15.14	1	558576	559878	4	Amel_2.0_OGSv1_gmap (XP_001121449.1, LG15:4536711..4539563)
<i>Ca. floridanus</i>	CFLO15237	317	scaffold309	1	97058	94081	4	OGSv1.0 (EFN64649/GL441542.1)
<i>T. castaneum</i>	TC011657	319	ChLG9	1	20588465	20587053	5	TCOGS2, XP_967536.1
<i>Ac. pisum</i>	ACYPI007856	377	GL349761	1	7604	120	3	
<i>P. humanus</i>	PHUM026050	369	DS235004.1	1	413407	415968	3	

Table S2. List of

Osiris 3								
Genome	Acc#	Length (AA)	Chromosome (linkage group)	Direct	Start	End	# exons	Note
<i>D. melanogaster</i>	CG1150	288	3R	1	2041786	2043639	3	
<i>D. pseudoobscura</i>	GA11035	289	2	1	15785557	15784397	3	
<i>D. virilis</i>	GJ14248	282	scaffold_12822	1	1837891	1836711	3	
<i>D. grimshawi</i>	GH14019	284	scaffold_14624	1	2842701	2841453	3	
<i>An. gambiae</i>	AGAP004126	295	2R	1	50423391	50425056	3	
<i>Ae. aegypti</i>	AAEL010624	287	supercont1.491	1	153272	137568	3	
<i>Cu. quinquefasciatus</i>	CPIJ008148	269	supercont3.179	1	677815	663250	3	
<i>B. mori</i>	BGIBMGA000042	212	nscaf1071	1	620089	626738	4	missing regions at both ends
<i>Ap. mellifera</i>	GB19141	294	Group15.14	1	565917	567581	2	AmeL_2.0_OGSv1_gmap (XP_001121482.1 LG15:4544707..4547498)
<i>Ca. floridanus</i>	CFLO15238	278	scaffold309	1	85571	84078	2	OGSv1.0 (EFN64648/GL441542.1)
<i>T. castaneum</i>	TC011658	276	ChLG9	1	20584063	20582557	4	TCOGS2, XP_967452.1
<i>Ac. pisum</i>	ACYPI001904	303	GL349870	+	639943	646593	3	
<i>P. humanus</i>	PHUM026160	300	DS235004.1	1	431224	433412	5	

Table S2. List of

Genome	Osiris 4							
	ID	Length (AA)	Chromosome (linkage group)	Direct	Start	End	# exons	Note
<i>D. melanogaster</i>	CG10303	393	3R	-1	2047384	2044255	4	
<i>D. pseudoobscura</i>	GA10232	405	2	-1	15780992	15783643	4	
<i>D. virilis</i>	GJ14490	399	scaffold_12822	-1	1832406	1835997	4	
<i>D. grimshawi</i>	GH14283	403	scaffold_14624	-1	2837162	2840675	4	
<i>An. gambiae</i>	AGAP004127	306	2R	-1	50430646	50429581	3	
<i>Ae. aegypti</i>	AAEL010625	312	supercont1.491	-1	88165	101794	3	
<i>Cu. quinquefasciatus</i>	CPIJ008147	317	supercont3.179	-1	651118	653972	3	
<i>B. mori</i>	None							
<i>Ap. mellifera</i>	None							
<i>Ca. floridanus</i>	None							
<i>T. castaneum</i>								
<i>Ac. pisum</i>	None							
<i>P. humanus</i>	None							

Table S2. List of

Osiris 5								
Genome	ID	Length (AA)	Chromosome (linkage group)	Direct	Start	End	# exons	Note
<i>D. melanogaster</i>	CG15590-PA	202	3R	1	2053428	2054461	2	isoform
	CG15590-PB	201	3R	1	2053428	2054458	2	isoform
<i>D. pseudoobscura</i>	GA13832	212	2	1	15775460	15774575	2	
<i>D. virilis</i>	GJ14249	210	scaffold_12822	1	1823924	1822411	2	
<i>D. grimshawi</i>	GH14020	219	scaffold_14624	1	2830195	2828423	2	
<i>An. gambiae</i>	AGAP012556	228	UNKN	-	15940121	15939378	2	
<i>Ae. aegypti</i>	AAEL010628	257	supercont1.491	-1	36946	50288	2	
	AAEL010631	249	supercont1.491	1	12724	1865	2	
<i>Cu. quinquefasciatus</i>	CPIJ008146	251	supercont3.179	1	634575	623303	2	
<i>B. mori</i>	BGIBMGA00004	163	nscaf1071	1	688056	693867	3	Weak similarity
<i>Ap. mellifera</i>	au8.g4221a	293	Group15.14	1	572124	573215	3	Based on au8.g4221.t1 ex1-ex3, (572124..572628,572715..572822,572947..573215) (XP_001121508 LG15:4551244..4552330, no
<i>Ca. floridanus</i>	CFLO15239	152	scaffold309	1	78191	77108	3	OGSv1.0 (EFN64647/GL441542.1) missing 5' region
<i>T. castaneum</i>	TC011659	229	ChLG9	1	20578022	20576613	2	5a, TCOGS2, XP_976093.1
	TC012811	232	ChLG9	-1	20581262	20582060	5	5b, TCOGS2, Osi5-like?
<i>Ac. pisum</i>	None							
<i>P. humanus</i>	None							

Table S2 (continued)

Osiris 6								
Genome	ID	Length (AA)	Chromosome (linkage group)	Direct	Start	End	# exons	Note
<i>D. melanogaster</i>	CG1151	312	3R	1	2060458	2062024	2	
<i>D. pseudoobscura</i>	GA26494	288	2	1	15768074	15766825	2	
<i>D. virilis</i>	GJ14250	294	scaffold_12822	1	1813629	1812417	2	
<i>D. grimshawi</i>	GH14022	297	scaffold_14624	1	2820796	2819537	2	
<i>An. gambiae</i>	AGAP004129	237	2R	1	50491387	50492715	2	
<i>Ae. aegypti</i>	AAEL014433	240	supercont1.1116	+	181441	189732	2	
<i>Cu. quinquefasciatus</i>	CPIJ008144	240	supercont3.179	1	530566	525882	2	
<i>B. mori</i>	None							
<i>Ap. mellifera</i>	GB19629	257	Group15.14	1	580714	581723	3	Amel_2.0_OGSv1_gmap (XP_001121541.1 LG15:4559825..4560843)
<i>Ca. floridanus</i>	CFLO15240	273	scaffold309	1	70694	69570	3	OGSv1.0 (EFN64646/GL441542.1)
<i>T. castaneum</i>	None							
<i>Ac. pisum</i>	ACYPI000840	287	GL349857	-	605198	602013	4	
<i>P. humanus</i>	None							

Table S2 (continuu

Genome	Osiris 7							Note
	ID	Length (AA)	Chromosome (linkage group)	Direct	Start	End	# exons	
<i>D. melanogaster</i>	CG1153	288	3R	1	2074929	2075795	1	
<i>D. pseudoobscura</i>	GA11054	287	2	1	15754289	15753426	1	
<i>D. virilis</i>	GJ14251	288	scaffold_12822	1	1797298	1796432	1	
<i>D. grimshawi</i>	GH14023	291	scaffold_14624	1	2805530	2804655	1	
<i>An. gambiae</i>	AGAP004130	297	2R	1	50520138	50521031	1	
<i>Ae. aegypti</i>	None							
<i>Cu. quinquefasciatus</i>	CPIJ008141	288	supercont3.179	1	431192	430326	1	
<i>B. mori</i>	BGIBMGA000044	261	nscaf1071	1	709297	710082	1	
<i>Ap. mellifera</i>	GB13419	277	Group15.14	1	587513	589282	3	Amel_2.0_OGSv1_gmap (XP_624937.1 LG15:4566421..4568940)
<i>Ca. floridanus</i>	CFLO15241a	275	scaffold309	1	63969	61248	3	CFLO15241:ex1-ex3 AUGUSTUS model 61248..61630,63407..63582,63701..63969 (EFN64645=Osi7+Osi8)
<i>T. castaneum</i>	TC011660	278	ChLG9	1	20571810	20570974	1	TCOGS2, XP_967197.1
<i>Ac. pisum</i>	ACYPI002725	298	GL349857	-	577533	574085	3	
<i>P. humanus</i>	PHUM026170	272	DS235004.1	1	443236	446107	3	

Table S2 (continuu

Osiris 8								
Genome	ID	Length (AA)	Chromosome (linkage group)	Direct	Start	End	# exons	Note
<i>D. melanogaster</i>	CG15591	274	3R	1	2080997	2082114	2	
<i>D. pseudoobscura</i>	GA13833	287	2	1	15747924	15746989	2	
<i>D. virilis</i>	GJ14253a	276	scaffold_12822	1	1786185	1785242	3	GJ14253+GJ14254 AUGUSTUS: 1785242..1785634,1785675..1785839,1785913..1786185
<i>D. grimshawi</i>	GH14024	281	scaffold_14624	1	2798067	2797160	2	
<i>An. gambiae</i>	AGAP004128	282	2R	1	50448638	50449565	2	
<i>Ae. aegypti</i>	AAEL004275	247	supercont1.113	-	492815	491930	2	extra 3 aa's before M
<i>Cu. quinquefasciatus</i>	CPIJ008140	280	supercont3.179	1	388401	387495	2	
<i>B. mori</i>	BGIBMGA000009	236	nscf1071	-1	828580	820371	3	
<i>Ap. mellifera</i>	GB10057	264	Group15.14	1	592556	593593	2	Amel_2.0_OGSv1_gmap (XP_003249793.1 LG15:4571385..4573071)
<i>Ca. floridanus</i>	CFLO15241b	259	scaffold309	1	56944	55895	2	CFLO15241:ex4-ex5 AUGUSTUS model 55895..56178,56447..56944 (starts with 'atatt') (EFN64645=Osi7+Osi8)
<i>T. castaneum</i>	TC011661	254	ChLG9	1	20567569	20566676	2	TCOGS2, XP_967101.1
<i>Ac. pisum</i>	ACYPI005238	254	GL349857	-	550960	544609	3	
<i>P. humanus</i>	None							

Table S2 (continuation)

Genome	Osiris 9							
	ID	Length (AA)	Chromosome (linkage group)	Direct	Start	End	# exons	Note
<i>D. melanogaster</i>	CG15592	233	3R	1	2086135	2087371	3	
<i>D. pseudoobscura</i>	GA13834	232	2	1	15742961	15741569	3	
<i>D. virilis</i>	GJ14255	231	scaffold_12822	1	1780683	1779314	3	
<i>D. grimshawi</i>	GH14025	231	scaffold_14624	1	2792525	2791171	3	
<i>An. gambiae</i>	AGAP004131	238	2R	1	50568998	50570643	3	
<i>Ae. aegypti</i>	AAEL004280	237	supercont1.113	-	282815	276138	3	
<i>Cu. quinquefasciatus</i>	CPIJ008139	235	supercont3.179	1	291863	289386	3	
<i>B. moriyama</i>	BGIBMGA000013	239	nscaf1071	-1	731188	722003	3	9f
	BGIBMGA000045	241	nscaf1071	1	756455	761654	3	9e
	BGIBMGA000012b	243	nscaf1071	-1	776934	775097	3	9d 000012 contains two Osi9 Augustus model nscaf1071:775097..775440,776272..776524,776800..776934
	BGIBMGA000012a	341	nscaf1071	-1	780388	777716	5	9c 000012 contains two Osi9 Augustus model nscaf1071:777716..778070,779043..779240,779318..779351,779496..779742,780197..780388
	BGIBMGA000011	235	nscaf1071	-1	795081	788764	3	9b
	BGIBMGA000010	240	nscaf1071	-1	813664	807853	3	9a
<i>Ap. mellifera</i>	GB19626	255	Group15.14	1	599571	600600	3	Amel_2.0_OGSv1_gmap (XP_001121625.1 LG15:4578466..4580519)
<i>Ca. floridanus</i>	CFLO15242	250	scaffold309	1	49954	48426	3	OGSv1.0 (EFN64644/GL441542.1)
<i>T. castaneum</i>	TC011662	250	ChLG9	1	20564637	20562939	3	TCOGS2, XP_001808361.1
<i>Ac. pisum</i>	ACYPI004638	264	GL349857	-	502336	498348	2	
<i>P. humanus</i>	PHUM026280	258	DS235004.1	1	457230	458434	3	

Table S2 (continuu

Osiris 10								
Genome	ID	Length (AA)	Chromosome (linkage group)	Direct	Start	End	# exons	Note
<i>D. melanogaster</i>	CG15593-PB	741	3R	1	2089581	2093117	4	possibly 10a+10b
	CG15593-PA	576	3R	1	2089698	2093117	5	possibly 10a+10b
<i>D. pseudoobscura</i>	GA13835-PC	634	2	1	15738413	15735593	5	possibly 10a+10b
	GA13835-PB	767	2	1	15738488	15735593	4	possibly 10a+10b
<i>D. virilis</i>	GJ14257	592	scaffold_12822	1	1775378	1771016	5	possibly 10a+10b
<i>D. grimshawi</i>	GH14026	509	scaffold_14624	1	2787800	2784545	5	possibly 10a+10b missing 5' end
<i>An. gambiae</i>	AGAP004132	610	2R	1	50579082	50585938	5	possibly 10a+10b
<i>Ae. aegypti</i>	AAEL004303	565	supercont1.113	-	205352	160117	5	possibly 10a+10b
<i>Cu. quinquefasciatus</i>	CPIJ008138	564	supercont3.179	1	260063	230396	5	possibly 10a+10b
<i>B. mori</i>	BGIBMGA000046	314	nscaf1071	1	839498	846172	4	10a or 10b only
<i>Ap. mellifera</i>	XP_003249790.1	314	Group15.14	1	602439	603849	2	10a, NCBI_RefSeq
	GB13856	223	Group15.14	1	606624	607462	3	10b, Amel_2.0_OGSv1_gmap
<i>Ca. floridanus</i>	CFL015243	283	scaffold309	1	46630	44997	2	10a, OGSv1.0 (EFN64643/GL441542.1)
	CFL015244	219	scaffold309	1	40662	39172	3	10b, OGSv1.0 (EFN64644/GL441542.1)
<i>T. castaneum</i>	XP_967021.2	533	ChLG9	1	20561128	20557379	8	LOC655388, TC011663+TC011664 possibly 10a+10b
<i>Ac. pisum</i>	ACYPI37205	360	GL349857	-	494658	484212	4	10a
	ACYPI37204	240	GL349857	+	477903	479043	2	10b
<i>P. humanus</i>	PHUM026300	307	DS235004.1	1	470753	471897	4	10a
	PHUM026310	165	DS235004.1	1	476,424	477503	4	10b

Table S2 (continued)

Osiris 11								
Genome	ID	Length (AA)	Chromosome (linkage group)	Direct	Start	End	# exons	Note
<i>D. melanogaster</i>	CG15596	302	3R	-1	2094410	2093502	1	
<i>D. pseudoobscura</i>	GA13838	317	2	-1	15734475	15735428	1	
<i>D. virilis</i>	GJ14489	321	scaffold_12822	-1	1769902	1770867	1	
<i>D. grimshawi</i>	GH14282	326	scaffold_14624	-1	2783391	2784371	1	
<i>An. gambiae</i>	AGAP004134	263	2R	1	50594209	50595000	1	
<i>Ae. aegypti</i>	AAEL004298	263	supercont1.113	+	85610	86401	1	
<i>Cu. quinquefasciatus</i>	CPIJ008136	262	supercont3.179	1	198762	197867	1	
<i>B. mori</i>	BGIBMGA00000	253	nscaf1071	-1	851744	850983	1	
<i>Ap. mellifera</i>	XP_003249794.	202	Group15.14	-1	609315	608097	3	NCBI-RefSeq model corresponding to GB14757, but without extra 12 aa at 5' end
<i>Ca. floridanus</i>	CFLO15250	193	scaffold309	-1	35342	37142	3	OGSv1.0, missing 5' region (EFN64641/GL441542.1)
<i>T. castaneum</i>	TC012810	279	ChLG9	-1	20555016	20556390	2	TCOGS2 (XP_996928.2 missing 5' and 3')
<i>Ac. pisum</i>	ACYPI003143	316	GL349857	-	465923	448880	5	
<i>P. humanus</i>	PHUM026320	346	DS235004.1	-1	482961	481380	3	

Table S2 (contin

Genome	Osiris 12							Note
	ID	Length (AA)	Chromosome (linkage group)	Direct	Start	End	# exons	
<i>D. melanogaster</i>	CG1154	295	3R	1	2104429	2107526	2	
<i>D. pseudoobscura</i>	GA11059	310	2	1	15724932	15722318	2	
<i>D. virilis</i>	GJ14258	296	scaffold_12822	1	1759247	1755541	2	
<i>D. grimshawi</i>	GH14027	306	scaffold_14624	1	2773783	2770028	2	
<i>An. gambiae</i>	AGAP013195	245	2R	1	50611644	50612588	2	
<i>Ae. aegypti</i>	AAEL002957a	232	supercont1.73	+	533897	545601	2	based on AAEL002957 ex4-ex5:
<i>Cu. quinquefasciatus</i>	CPIJ008135	342	supercont3.179	1	128851	120079	5	
<i>B. mori</i>	BGIBMGA000047	234	nscaf1071	1	860527	877819	3	BGIBMGA00004 missing 3' end; Augustus model nscaf1071:860527..860929,866403..866530,877646..877819
<i>Ap. mellifera</i>	XP_001121769.1	263	Group15.14	1	614359	615150	1	NCBI_RefSeq model corresponding to GB18845, but with extra 41 aa at N-term (LG15:4593479..4594270)
<i>Ca. floridanus</i>	CFLO15245	257	scaffold309	1	31139	30366	1	OGSv1.0 (EFN64640/GL441542.1)
<i>T. castaneum</i>	TC011665	237	ChLG9	1	20552342	20551573	2	TCOGS2, XP_966834.1
<i>Ac. pisum</i>	ACYPI000603	255	GL349857	+	445712	449651	2	
<i>P. humanus</i>	None							

Table S2 (contin

Osiris 13								
Genome	ID	Length (AA)	Chromosome (linkage group)	Direct	Start	End	# exons	Note
<i>D. melanogaster</i>	CG15595	210	3R	1	2117624	2118491	2	
<i>D. pseudoobscura</i>	GA13837	219	2	1	15713662	15712936	2	
<i>D. virilis</i>	GJ14260	207	scaffold_12822	1	1745092	1744398	2	
<i>D. grimshawi</i>	GH14029	208	scaffold_14624	1	2759717	2758601	2	
<i>An. gambiae</i>	None							
<i>Ae. aegypti</i>	None							
<i>Cu. quinquefasciatus</i>	None							
<i>B. mori</i>	BGIBMGA00004	222	nscaf1071	1	889090	891513	3	Weak similarity
<i>Ap. mellifera</i>								
<i>Ca. floridanus</i>	CFLO15246	225	scaffold309	1	23657	21613	3	OGSv1.0 (EFN64638/GL441542.1) weak similarity
<i>T. castaneum</i>	None							
<i>Ac. pisum</i>	None							
<i>P. humanus</i>	None							

Table S2 (contin

Osiris 14								
Genome	ID	Length (AA)	Chromosome (linkage group)	Direct	Start	End	# exons	Note
<i>D. melanogaster</i>	CG1155	268	3R	1	2124973	2126299	3	
<i>D. pseudoobscura</i>	GA11061	269	2	1	15707310	15706315	3	
<i>D. virilis</i>	GJ14261	277	scaffold_12822	1	1737239	1736231	3	
<i>D. grimshawi</i>	GH14030	270	scaffold_14624	1	2750767	2749290	3	
<i>An. gambiae</i>	AGAP003465	280	2R	1	38016606	38018316	3	
<i>Ae. aegypti</i>	AAEL002962	290	supercont1.73	+	921829	928099	3	
<i>Cu. quinquefasciatus</i>	CPIJ009837	273	supercont3.253	-	539625	536876	3	
<i>B. mori</i>	BGIBMGA00005	189	nscaf1071	1	899025	905630	4	
<i>Ap. mellifera</i>	GB19255	267	Group15.14	1	625746	627315	3	Amel_2.0_OGSv1_gmap (XP_392084.1 LG15:4604768..4606632)
<i>Ca. floridanus</i>	CFLO15247	262	scaffold309	1	18470	16528	3	OGSv1.0 (EFN64637/GL441542.1)
<i>T. castaneum</i>	TC011827	245	ChLG9	1	17784120	17783271	3	TCOGS2, XP_972144.1
<i>Ac. pisum</i>	ACYPI010031	258	GL349857	-	367021	365776	3	
<i>P. humanus</i>	PHUM026630	271	DS235004.1	1	514933	515929	3	

Table S2 (contin

Osiris 15								
Genome	ID	Length (AA)	Chromosome (linkage group)	Direct	Start	End	# exons	Note
<i>D. melanogaster</i>	CG1157	214	3R	1	2127635	2128897	3	
<i>D. pseudoobscura</i>	GA11070	214	2	1	15704652	15703720	3	
<i>D. virilis</i>	GJ14262	215	scaffold_12822	1	1734711	1733651	3	
<i>D. grimshawi</i>	GH14031	214	scaffold_14624	1	2747671	2746647	3	
<i>An. gambiae</i>	AGAP003466	206	2R	1	38023425	38027635	3	
<i>Ae. aegypti</i>	AAEL002949	206	supercont1.73	+	947347	977880	3	
<i>Cu. quinquefasciatus</i>	CPIJ009835	208	supercont3.253	-	520208	503607	3	
<i>B. mori</i>	None							
<i>Ap. mellifera</i>	GB14511	233	Group15.14	1	629349	631777	2	Amel_2.0_OGSv1_gmap (XP_001121861.1 LG15:4608426..4611097)
<i>Ca. floridanus</i>	CFLO15248	234	scaffold309	1	14885	12348	2	OGSv1.0 (EFN64636/GL441542.1)
<i>T. castaneum</i>	TC011828	188	ChLG9	1	17771168	17770175	2	TCOGS2, XP_971987.1
<i>Ac. pisum</i>	None							
<i>P. humanus</i>	None							

Table S2 (continued)

Genome	Osiris 16							
	ID	Length (AA)	Chromosome (linkage group)	Direct	Start	End	# exons	Note
<i>D. melanogaster</i>	CG31561	278	3R	1	2130992	2131887	2	
<i>D. pseudoobscura</i>	GA16326	277	2	1	15701265	15700368	2	
<i>D. virilis</i>	GJ14264	278	scaffold_12822	1	1731967	1730888	2	
<i>D. grimshawi</i>	GH14033	276	scaffold_14624	1	2745403	2744257	2	
<i>An. gambiae</i>	AGAP003467	238	2R	1	38030429	38031217	2	16a
	AGAP003472	238	2R	1	38077169	38077957	2	16b
<i>Ae. aegypti</i>	AAEL002961	254	supercont1.73	+	1025518	1038012	2	
<i>Cu. quinquefasciatus</i>	CPIJ009834	243	supercont3.253	-	495771	494975	2	
<i>B. mori</i>	BGIBMGA00005	248	nscaf1071	1	916070	920540	2	16a
	BGIBMGA00005	222	nscaf1071	1	924686	927766	2	16b
	See Osiris-like located here							
<i>Ap. mellifera</i>	GB16524	303	Group15.14	1	636935	638007	2	Amel_2.0_OGSv1_gmap (XP_001121887.1 includes extra 3')
<i>Ca. floridanus</i>	CFLO15249	294	scaffold309	1	6971	5865	2	OGSv1.0 (EFN64635/GL441542.1)
<i>T. castaneum</i>	TC012679	257	ChLG9	-1	17774547	17775320	1	16b, TCOGS2, XP_972042.1
	TC012680	260	ChLG9	-1	17778206	17781508	2	16a, TCOGS2, XP_972093.1
<i>Ac. pisum</i>	ACYPI37215	395	GL349857	-	264785	258197	3	16c
	ACYPI37213	335	GL349857	-	335265	325498	3	16b
	ACYPI37212	206	GL349857	-	347582	342664	3	16a
	ACYPI37197	271	GL349857	+	386913	389207	2	16f, located before Osi14
	ACYPI37210	284	GL349857	-	398236	395754	2	16e, located before Osi14
	ACYPI007501	287	GL349857	-	402294	399369	2	16d, located before Osi14
<i>P. humanus</i>	None							

Table S2 (contin

Osiris 17								
Genome	ID	Length (AA)	Chromosome (linkage group)	Direct	Start	End	# exons	Note
<i>D. melanogaster</i>	CG15598	648	3R	1	2141609	2152519	8	
<i>D. pseudoobscura</i>	GA26496-PA	740	2	1	15692057	15683149	9	Isoforms
	GA26496-PC	654	2	1	15692183	15683149	9	Isoforms
<i>D. virilis</i>	GJ14266	705	scaffold_12822	1	1717024	1703545	9	
<i>D. grimshawi</i>	GH14034a	761	scaffold_14624	1	2735733	2723660	9	last exon removed. Revised model: (2723660..2723801,2726507..2726685,2726749..2726859,2728780..2728978,2729765..2730073,2730663..2730850,2733027..2733280,2733378..2733611,2735064..2735
<i>An. gambiae</i>	AGAP003468	661	2R	1	38042427	38056441	8	
<i>Ae. aegypti</i>	AAEL002966a	598	supercont1.73	+	1118311	1246764	8	AAEL002966 + AAEL002952 genewise model: CDS(1118311..1118617, 1155934..1156176,1166316..1166569,1182188..1182342,1183777..1184010, 1193705..1193876,1215056..1215336, 1246614..1246764), start codon not found
<i>Cu. quinquefasciatus</i>	CPIJ009830a	669	supercont3.253	-	449432	366128	8	CPIJ009830+CPIJ009831+CPIJ009832+CPIJ009833; genewise model: CDS(449432..449153,417630..417388, 413069..412816,406802..406639,403596.. 403396,393345..393201,379175..378895, 366278..366128); Need to check
<i>B. mori</i>	BGIBMGA000054	588	nscaf1071	1	968425	985551	10	
<i>Ap. mellifera</i>	XP_001121915.2	586	Group15.14	1	655434	658194	4	NCBI-RefSeq model corresponding to GB16817, but with extra 91 aa at N-term (LG15:4634123..4638274)
<i>Ca. floridanus</i>	CFLO14498	560	scaffold767	-	348549	345974	5	OGSv1.0 (EFN71496/GL436778.1)
<i>T. castaneum</i>	None							
<i>Ac. pisum</i>	ACYPI006202	525	GL349975	+	190119	206662	7	
<i>P. humanus</i>	PHUM497620a	645	DS235830.1	-	141828	133937	8	PHUM497620+PHUM497610; GeneWise model: 141828..141297,138866..138492, 135455..135226,135153..134991, 134907..134726,134649..134541, 134461..134257,134078..133937

Table S2 (contin

Genome	Osiris 18							Note
	ID	Length (AA)	Chromosome (linkage group)	Direct	Start	End	# exons	
<i>D. melanogaster</i>	CG1169	306	3R	1	2156073	2157285	2	
<i>D. pseudoobscura</i>	GA11143	313	2	1	15676247	15675240	2	
<i>D. virilis</i>	GJ14267	309	scaffold_12822	1	1700037	1699044	2	
<i>D. grimshawi</i>	GH14035	318	scaffold_14624	1	2720195	2719171	2	
<i>An. gambiae</i>	AGAP003469	265	2R	-1	38058328	38057430	2	
	AGAP012548	106	UNKN	-1	14659597	14654022	2	Almost identical copy in unknown location; possible assembly mistake; see
<i>Ae. aegypti</i>	AAEL002965	266	supercont1.73	-	1259787	1258929	2	
<i>Cu. quinquefasciatus</i>	CPIJ009829	267	supercont3.253	+	356854	357720	2	
<i>B. mori</i>	BGIBMGA00000	270	nscaf1071	-1	993294	992334	3	
<i>Ap. mellifera</i>	GB16900	249	Group15.14	1	659884	660816	3	Amel_2.0_OGSv1_gmap (XP_001121942.2 LG15:4638990..4640092)
<i>Ca. floridanus</i>	CFLO14499	246	scaffold767	-	342467	341404	3	OGSv1.0 (EFN71495/GL436778.1)
<i>T. castaneum</i>	TC012820	246	ChLG9	-1	20684703	20685544	3	TCOGS2, XP_968992.1
<i>Ac. pisum</i>	ACYPI009026	242	GL349975	+	217232	219193	4	
<i>P. humanus</i>	PHUM497600	285	DS235830.1	+	123337	124403	3	

Table S2 (contin

Osiris 19								
Genome	ID	Length (AA)	Chromosome (linkage group)	Direct	Start	End	# exons	Note
<i>D. melanogaster</i>	CG15189-PA	266	3R	1	2161165	2162662	4	isoform
	CG15189-PB	257	3R	1	2161165	2162659	4	isoform
<i>D. pseudoobscura</i>	GA13557	273	2	1	15672361	15670940	4	
<i>D. virilis</i>	GJ14268	266	scaffold_12822	1	1695141	1693812	4	
<i>D. grimshawi</i>	GH14036	262	scaffold_14624	1	2716195	2714924	4	
<i>An. gambiae</i>	AGAP003470	232	2R	1	38061429	38062320	3	
	AGAP012549	247	UNKN	+	14662691	14663580	3	Almost identical copy in unknown location; possible assembly mistake; see
<i>Ae. aegypti</i>	AAEL002960	250	supercont1.73	+	1285435	1286307	3	
<i>Cu. quinquefasciatus</i>	CPIJ009828	251	supercont3.253	-	344600	343725	3	
<i>B. mori</i>	BGIBMGA01071	258	nscaf3003	-	3992132	3991021	3	
<i>Ap. mellifera</i>	GB16804	247	Group15.14	1	663088	663932	2	Amel_2.0_OGSv1_gmap (XP_001121961.1 LG15:4642085..4643841)
<i>Ca. floridanus</i>	CFLO14500	247	scaffold767	-	339201	338256	2	OGSv1.0 (EFN71494/GL436778.1)
<i>T. castaneum</i>	TC012821	243	ChLG9	-1	20686916	20687749	3	19a, TCOGS2, XP_969068.1
	TC012822	196	ChLG9	-1	20689368	20690009	2	19b, TCOGS2, XP_969215.1
<i>Ac. pisum</i>	ACYPI000796	190	GL349975	+	226304	227943	4	
<i>P. humanus</i>	PHUM497590	255	DS235830.1	-	122029	121084	3	

Table S2 (contin

Genome	Osiris 20							Note
	ID	Length (AA)	Chromosome (linkage group)	Direct	Start	End	# exons	
<i>D. melanogaster</i>	CG15188	280	3R	1	2165910	2166818	2	
<i>D. pseudoobscura</i>	GA13556	276	2	1	15666849	15665952	2	
<i>D. virilis</i>	GJ14269	277	scaffold_12822	1	1689890	1688986	2	
<i>D. grimshawi</i>	GH14037	280	scaffold_14624	1	2711690	2710778	2	
<i>An. gambiae</i>	AGAP003471	320	2R	1	38065791	38066897	2	
<i>Ae. aegypti</i>	AAEL002955	292	supercont1.73	+	1305642	1313330	2	
	AAEL014727	292	supercont1.1240	-	188349	182326	2	20b, Unlinked Osi20 copy; almost identical
<i>Cu. quinquefasciatus</i>	CPIJ009827	297	supercont3.253	-	330657	327540	2	
<i>B. mori</i>	BGIBMGA010717	293	nscaf3003	-	3982740	3981708	3	
<i>Ap. mellifera</i>	XP_001121985.1	270	Group15.14	1	665962	666847	2	NCBI_RefSeq model corresponding to GB15865, but with extra 120 aa at 5' end (LG15:4644922..4646726)
<i>Ca. floridanus</i>	CFLO14501	267	scaffold767	-	335705	334509	2	OGSv1.0 (EFN71493/GL436778.1)
<i>T. castaneum</i>	TC012823	248	ChLG9	-1	20691474	20692314	3	TCOGS2, XP_969284.1
<i>Ac. pisum</i>	ACYPI007146	281	GL349975	+	238670	239799	2	
<i>P. humanus</i>	PHUM497570	290	DS235830.1	-	117238	116133	4	

Table S2 (continued)

Genome	Osiris 22							Note
	ID	Length (AA)	Chromosome (linkage group)	Direct	Start	End	# exons	
<i>D. melanogaster</i>	CG8644	332	3R	1	9116438	9117497	2	
<i>D. pseudoobscura</i>	GA21234	342	2	1	9987757	9986666	2	
<i>D. virilis</i>	GJ10414	363	scaffold_12855	-	3036310	3035156	2	
<i>D. grimshawi</i>	GH18427	339	scaffold_14906	-	5747016	5745936	2	
<i>An. gambiae</i>	AGAP003420	366	2R	1	37510453	37511640	2	
<i>Ae. aegypti</i>	AAEL007198	359	supercont1.243	+	914044	915223	2	
<i>Cu. quinquefasciatus</i>	CPIJ011100	345	supercont3.316	+	372903	374003	2	
<i>B. mori</i>	BGIBMGA010678	163	nscaf2998	+	1317487	1322074	2	Middle region missing
<i>Ap. mellifera</i>	None							
<i>Ca. floridanus</i>	None							
<i>T. castaneum</i>	TC002385	151	unknown	+	15336400	15336898	2	Middle region missing
<i>Ac. pisum</i>	ACYPI002813	196	GL350227	-	167081	164772	3	Middle region
<i>P. humanus</i>	None							

Table S2 (continue)

Osiris 23								
Genome	ID	Length (AA)	Chromosome (linkage group)	Direct	Start	End	# exons	Note
<i>D. melanogaster</i>	CG15538	250	3R	-1	26270756	26269871	3	
<i>D. pseudoobscura</i>	GA13796	270	2	-1	1254748	1255676	3	
<i>D. virilis</i>	GJ22762	265	scaffold_13047	-	15824251	15823319	3	
<i>D. grimshawi</i>	GH17432	264	scaffold_14830	+	4185984	4186905	3	
<i>An. gambiae</i>	AGAP000957	293	X	-	18368810	18367147	4	
<i>Ae. aegypti</i>	None							Short similarity against AGAP000957 and CG15538 found in supercont1.202:1394474-1394301
<i>Cu. quinquefasciatus</i>	Cqui_Osi23	189	supercont3.42	+	474599	480673	3	GeneWise+Augustus model: 474599..474855,474955..475082,480489..480673; still missing 3'
<i>B. mori</i>	None							
<i>Ap. mellifera</i>	XP_001120227.2	254	Group8.6	-	87657	86288	3	NCBI_RefSeq model corresponding to GB14285, but with extra 14 aa at N-term (LG8:2309823..2307814)
<i>Ca. floridanus</i>	CFLO22710	249	scaffold799		210020	211808	3	OGSv1.0 (EFN70025/GL437711.1)
<i>T. castaneum</i>	TC030703a	226	ChLG2	+	11496092	11497260	3	TC030703:TCOGS2 with extra N/C-term. XP_970499.1 (non-Osi protein) includes this as part of its exons. AUGUSTUS model: 11496092..11496258,11496701..11496798,11496845..11497260
<i>Ac. pisum</i>	ACYPI50421	217	GL350440	-	80703	67062	3	
<i>P. humanus</i>	None							

Table S2 (continue)

Genome	Osiris 21							Note
	ID	Length (AA)	Chromosome (linkage group)	Strand	Start	End	# exons	
<i>D. melanogaster</i>	CG14925	282	2L	-	11285603	11284586	2	
<i>D. pseudoobscura</i>	GA13356	280	4_group3	-	6710686	6709780	2	
<i>D. virilis</i>	GJ14913	277	scaffold_12963	+	2228486	2229378	2	
<i>D. grimshawi</i>	GH13529	291	scaffold_15126	+	1313369	1314317	2	
<i>An. gambiae</i>	AGAP005899	256	2L	-	23383765	23382399	3	21a
	AGAP010354	242	3L	-	2060900	2060107	2	21b
<i>Ae. aegypti</i>	AAEL004788	264	supercont1.130	+	1236881	1240903	3	21a
	AAEL014873	242	supercont1.1318	-	61163	25952	2	21b
<i>Cu. quinquefasciatus</i>	CPIJ006501	255	supercont3.125	-	457478	456268	2	21c
	CPIJ016349	255	supercont3.798	+	131376	132525	2	21b
	CPIJ004911	270	supercont3.80	+	580797	581778	3	21a
<i>B. mori</i>	BGIBMGA013568a	216	nscaf3078	+	404693	405837	4	BGIBMGA013568+BGIBMGA013569 GeneWise model: 404693..404875,404959..405151, 405257..405437,405754..405837; need to check
<i>Ap. mellifera</i>	None							
<i>Ca. floridanus</i>	None							
<i>T. castaneum</i>	TC013496	250	ChLG5	-	10767500	10766748	1	TCOGS2, XP_976416.1
<i>Ac. pisum</i>	ACYPI008994	252	GL350028	-	167549	164571	2	21c
	ACYPI43145	245	GL350028	-	154956	151423	2	21b
	ACYPI005783	258	GL350028	-	135395	126447	2	21a
	ACYPI064603	141	GL356106	+	0	3876	2	21a2, missing 5' and 3'-ends
<i>P. humanus</i>	None							

Table S2 (continued)

Genome	Osiris like							Note
	ID	Length (AA)	Chromosome (linkage group)	Direct	Start	End	# exons	
<i>D. melanogaster</i>								
<i>D. pseudoobscura</i>								
<i>D. virilis</i>								
<i>D. grimshawi</i>								
<i>An. gambiae</i>								
<i>Ae. aegypti</i>								
<i>Cu. quinquefasciatus</i>								
<i>B. mori</i>	BGIBMGA000053	493	nscaf1071	1	931016	939783	7	located after Osi16b
<i>Ap. mellifera</i>								
<i>Ca. floridanus</i>								
<i>T. castaneum</i>								
<i>Ac. pisum</i>	ACYPI000560	419	GL349857	+	296605	308820	5	Located between 16b and c
	ACYPI37193	286	GL349857	+	266124	275397	3	Located between 16b and c
	ACYPI33621	262	GL350199	+	284013	299232	2	
<i>P. humanus</i>	PHUM025350	285	DS235004.1	-1	94973	93537	3	
	PHUM025360	360	DS235004.1	-1	110230	108271	4	
	PHUM465410	327	DS235812.1	-1	294251	290973	3	

Table S2 (continued)

Genome	CG15594 like							
	ID	Length (AA)	Chromosome (linkage group)	Direct	Start	End	# exons	Note
<i>D. melanogaster</i>								
<i>D. pseudoobscura</i>								
<i>D. virilis</i>								
<i>D. grimshawi</i>								
<i>An. gambiae</i>								
<i>Ae. aegypti</i>								
<i>Cu. quinquefasciatus</i>								
<i>B. mori</i>	BGIBMGA000048272		nscf1071	1	883002	886205	2	
<i>Ap. mellifera</i>	XP_003249795.1	169	Group15.14	-1	617796	616879	2	LG15:4597217..459575
<i>Ca. floridanus</i>	CFLO10930	225	scaffold309	-1	27850	28890	2	EFN64639/GL441542
<i>T. castaneum</i>								
<i>Ac. pisum</i>	ACYPI007103	364	GL349679	-1	542850	539522	6	
<i>P. humanus</i>								

Table S3 The list of 24 *Osiris* genes in the *D. melanogaster* genome and those identified from the 11 other *Drosophila* genomes.

	<i>D. melanogaster</i>	<i>D. simulans</i>	<i>D. sechellia</i>	<i>D. yakuba</i>	<i>D. erecta</i>
<i>Osi1</i>	CG15585	GD19841	GM10858	GE10183	GG13122
<i>Osi2</i>	CG1148	GD19845	GM10863	GE10186	GG13136
<i>Osi3</i>	CG1150	GD19846	GM10864	GE10187	GG13139
<i>Osi4</i>	CG10303	GD19554	GM10564	GE24111	GG10555
<i>Osi5</i>	CG15590	GD19847	GM10866	GE10188	GG13143
<i>Osi6</i>	CG1151	GD19848	GM10867	GE10189	GG13147
<i>Osi7</i>	CG1153	Dsim 7 [‡]	GM10868	GE10190	GG13151
<i>Osi8</i>	CG15591	GD19849	GM10869	GE10191	GG13155
<i>Osi9</i>	CG15592	GD19850	GM10870	GE10192	GG13161
<i>Osi10</i>	CG15593*	GD19851	GM10871	GE10194	GG13170
<i>Osi11</i>	CG15596	GD19553	GM10562	GE24110	GG10544
<i>Osi12</i>	CG1154	GD19852	GM10872	GE10195	GG13181
<i>Osi13</i>	CG15595	GD19856	GM10874	GE10197	GG13203
<i>Osi14</i>	CG1155	GD19857	GM10875	GE10198	GG13215
<i>Osi15</i>	CG1157	GD19858	GM10877	GE10199	GG13226
<i>Osi16</i>	CG31561	GD19859	GM10879	GE10201	GG13248
<i>Osi17</i>	CG15598	GD19854	GM10881	GE10205	GG13280
<i>Osi18</i>	CG1169	GD19861	GM10882	GE10206	GG13291
<i>Osi19</i>	CG15189	GD19862	GM10883	GE10207	GG13302
<i>Osi20</i>	CG15188	GD19863	GM10884	GE10208	GG13313
<i>Osi21</i>	CG14925	GD22180	GM11091	GE12926	GG10314
<i>Osi22</i>	CG8644	GD18894	GM24095	GE26258	GG19544
<i>Osi23</i>	CG15538	GD17079	GM12155	GE23387	GG11936
<i>Osi24</i>	CG15589	GD19844	GM10862	GE10185	GG13130

	<i>D. persimilis</i>	<i>D. pseudoobscura</i>	<i>D. willistoni</i>	<i>D. virilis</i>	<i>D. mojavensis</i>	<i>D. grimshawi</i>
<i>Osi1</i>	GL24048	GA13829	GK13029	GJ14243	GI24386	GH14014** [†]
<i>Osi2</i>	GL24055	GA11025	GK13033	GJ14247	GI24391	GH14018
<i>Osi3</i>	GL24056	GA11035	GK13034	GJ14248	GI24392	GH14019
<i>Osi4</i>	GL23480	GA10232	GK14207	GJ14490	GI23175	GH14283
<i>Osi5</i>	GL24057	GA13832	GK13035	GJ14249	GI24393	GH14020
<i>Osi6</i>	GL24058	GA26494	GK13036	GJ14250	GI24394	GH14022
<i>Osi7</i>	GL24059	GA11054	GK13037	GJ14251	GI24395	GH14023
<i>Osi8</i>	GL24060	GA13833	GK13038	GJ14253**	GI24396	GH14024
<i>Osi9</i>	GL24061	GA13834	Dwil ₉ [#]	GJ14255	GI24397	GH14025
<i>Osi10</i>	GL24062	GA13835	GK13040	GJ14257	GI24398	GH14026
<i>Osi11</i>	GL23479	GA13838	GK14206	GJ14489	GI23173	GH14282
<i>Osi12</i>	GL24063	GA11059	GK13041	GJ14258	GI24400	GH14027
<i>Osi13</i>	GL24066	GA13837	GK13043	GJ14260	GI24402	GH14029
<i>Osi14</i>	GL24067	GA11061	GK13044	GJ14261	GI24403	GH14030
<i>Osi15</i>	GL24068	GA11070	GK13045	GJ14262	GI24404	GH14031
<i>Osi16</i>	GL24070	GA16326	GK13048	GJ14264	GI24405	GH14033
<i>Osi17</i>	GL24071	GA26496	GK13049	GJ14266	GI24407	GH14034**
<i>Osi18</i>	GL24072	GA11143	GK13051	GJ14267	GI24408	GH14035
<i>Osi19</i>	GL24073	GA13557	GK13052	GJ14268	GI24409	GH14036
<i>Osi20</i>	GL24074	GA13556	GK13053	GJ14269	GI24411	GH14037
<i>Osi21</i>	GL18575	GA13356	GK21104	GJ14913	GI18143	GH13529
<i>Osi22</i>	GL24463	GA21234	GK13933	GJ10414	GI24328	GH18427
<i>Osi23</i>	GL14048	GA13796	GK13129	GJ22762	GI22761	GH17432
<i>Osi24</i>	GL24053	GA13830	GK13032	GJ14246	GI24390	GH14017

**Osi10* has two alternative transcripts. For our alignment and phylogenetic analysis, *Osi10*-PA (CG15593-PA) protein sequence was used.

**The annotations for these genes are corrected in this study. See Supplementary Table S2 for details.

[†] These genes are not annotated as *Osi* orthologs in Flybase.

[‡] *D. simulans Osi7* is not annotated in FlyBase, but the gene is identified in the genome (3R:2090001..2094500).

[#] *D. willistoni Osi9* is not annotated in FlyBase, but the gene is identified in the initial region of GK13040.

Table S4 The list of data sources for the arthropod genomes used in this study.

Species	Source	Release
<i>Drosophila melanogaster</i>	FlyBase (http://flybase.org/)	R5.40
<i>Drosophila simulans</i>	FlyBase (http://flybase.org/)	R1.3
<i>Drosophila sechellia</i>	FlyBase (http://flybase.org/)	R1.3
<i>Drosophila yakuba</i>	FlyBase (http://flybase.org/)	R1.3
<i>Drosophila erecta</i>	FlyBase (http://flybase.org/)	R1.3
<i>Drosophila ananassae</i>	FlyBase (http://flybase.org/)	R1.3
<i>Drosophila pseudoobscura</i>	FlyBase (http://flybase.org/)	R2.23
<i>Drosophila persimilis</i>	FlyBase (http://flybase.org/)	R1.3
<i>Drosophila willistoni</i>	FlyBase (http://flybase.org/)	R1.3
<i>Drosophila mojavensis</i>	FlyBase (http://flybase.org/)	R1.3
<i>Drosophila virilis</i>	FlyBase (http://flybase.org/)	R1.3
<i>Drosophila grimshawi</i>	FlyBase (http://flybase.org/)	R1.3
<i>Anopheles gambiae</i>	VectorBase (http://www.vectorbase.org/SequenceData/)	AgamP3.6
<i>Aedes aegypti</i>	VectorBase (http://www.vectorbase.org/SequenceData/)	AaegL1.2
<i>Culex quinquefasciatus</i>	VectorBase (http://www.vectorbase.org/SequenceData/)	CpipJ1.2
<i>Bombyx mori</i>	Silkworm Genome Database (http://silkworm.genomics.org.cn)	

		v2.0
<i>Apis mellifera</i>	BeeBase (http://hymenopteragenome.org/beebase/)	Amel_4.5
<i>Camponotus floridanus</i>	Ant Genomes Portal (http://hymenopteragenome.org/ant_genomes/)	V3.3
<i>Tribolium castaneum</i>	BeetleBase (http://beetlebase.org/)	Tcas_3.0
<i>Acyrtosiphon pisum</i>	AphidBase (http://www.aphidbase.com/aphidbase)	Acyr_2.0
<i>Pediculus humanus</i>	VectorBase (http://www.vectorbase.org/SequenceData/)	PhumU1.2
<i>Daphnia pulex</i>	Joint Genome Institute (http://www.jgi.doe.gov)	v1.0
<u><i>Ixodes scapularis</i></u>	<u>VectorBase (http://www.vectorbase.org/SequenceData/)</u>	<u>IscaW1.1</u>