Ultraconserved non-coding DNA within insect phyla

Thomas Brody[1*] Amar Yavatkar[2], Alexander Kuzin[1] and Ward F. Odenwald[1*]

[1]Neural Cell-Fate Determinants Section,

[2]Division of Intramural Research, Information Technology Program,
NINDS, NIH, Bethesda, Maryland, USA

Running title: Ultraconserved sequences in insect enhancers

[1]Corresponding authors:

Tel: 301-496-1657

FAX: 301-496-1339

E-mail: brodyt@ninds.nih.gov, odenwaldw@ninds.nih.gov

Abstract

Presence of ultra-conserved sequence elements in vertebrate enhancers suggest that transcription factor regulatory interactions are shared across phylogenetically diverse species. To date evidence for similarly conserved elements among evolutionarily distant insects such as flies, mosquitos, ants and bees, has been elusive. This study has taken advantage of the availability of the assembled genomic sequence of these insects to explore the presence of ultraconserved sequence elements in these phylogenetic groups. To investigate the integrity of fly regulatory sequences across ~100 million years of evolutionary divergence from the fruitfly *Drosophila melanogaster*, we compared *Drosophila* non-coding sequences to those of *Ceratitis capitata*, the Mediterranean fruit fly and *Musca domestica*, the domestic housefly. Using various alignment techniques, Blastn, Clustal, Blat, EvoPrinter and Needle, we show that many of the conserved sequence blocks (CSBs) that constitute *Drosophila cis*-regulatory DNA, recognized by EvoPrinter alignment protocols, are also conserved in *Ceratitis* and *Musca*. We term the sequence elements shared among these species ultraconserved CSBs (uCSBs). The position of the uCSBs with respect to flanking genes is also conserved. The results suggest that CSBs represent the point of interaction of multiple trans-regulators whose functions and interactions are conserved across divergent genera. Blastn alignments also detect putative *cis*-regulatory sequences shared among evolutionarily distant mosquitos *Anopheles gambiae* and *Culex pipiens* and *Aedes aegypti.* We have also identified conserved sequences shared among bee species. Side by side comparison of bee and ant EvoPrints identify uCSBs shared between the two taxa, as well as more poorly conserved CSBs in either one or the other taxon but not in both. Analysis of uCSBs in dipterans, mosquitos and bees will lead to a greater understanding of their evolutionary origin and the function of their conserved sequences.

Introduction

Phylogenetic footprinting of *Drosophila* genomic DNA has revealed that *cis*-regulatory enhancers can be distinguished from other essential gene regions based on their characteristic pattern of conserved sequences (Kuzin et al. 2009; Kuzin et al. 2012) (Odenwald et al. 2005;Pennacchio et al. 2006; Brody et al. 2007; Loots and Ovcharenko, 2007; Hardison, 2000, Bergman et al. 2002). These studies have shown that most enhancers are made up of clusters of conserved sequences that often are comprised of 5 to 30 or more conserved sequence blocks (CSBs). On average, *Drosophila* enhancers span ~1 kb and are flanked by non-conserved DNA of variable length.

Cross-species alignments have also identified conserved non-coding sequence elements associated with vertebrate developmental genes (Thomas et al. 2003; Bejerano et al. 2004), and sequences that are conserved among ancient and modern vertebrates (e. g., the sea lamprey and mammals). These elements conserved between disparate phyla are considered to be 'ultraconserved elements' (McEwen, et al. 2009; Irvine, et al. 2002). Many of these sequences act as *cis*-regulators of transcription (Pennacchio et al. 2006; Visel et al. 2009; McEwen et al. 2009; Visel et al. 2013; Dickel, 2018). Evidence from truncation studies indicates that, in the case of a mammalian Sonic Hedgehog enhancer, the ultraconserved element is not simply a clustering of transcription factor (TF) binding sites but has a structural component that is key to its activity (Lettice et al. 2014), suggesting that such highly conserved sequence blocks fit an enhanceosome model in which multiple adjacent and overlapping transcription factor docking sites act cooperatively to regulate gene expression (Panne, 2008). Previous studies have identified ultra-conserved elements in dipterans [*Drosophila* species and sepsids and mosquitos (Glazov et al. 2005; Hare et al. 2009, Sieglaff et al. 2009, Suryamohan et al, 2016)]. Comparison of consensus transcription factor binding sites, in the spider *Cupiennius salei* and the beetle *Tribolium castaneum,* have been shown to be functional in transgenic *Drosophila* (Ayyar et al. 2010).

Adjacent CSBs within *Drosophila* enhancers exhibit evolutionary conserved spacing. For example, characterization of 19 consecutive *Drosophila* enhancers spanning ~30 Kb between the *vvl* and *Prat2* genes revealed, in many instances, an evolutionarily constrained substructure between sets of enhancer CSBs (Kundu et al. 2013). Linked associations of adjacent CSBs could also be due to fixed spatial requirements for interactions of different transcriptional regulators (see for example Gao, et al. 2008, Panne, 2008).

In this study, we describe sequence conservation between the medfly *Ceratitis capitata,* the house fly *Musca domestica* genomic sequences and *Drosophila* genomic sequences. The house fly and Medfly have each diverged from *Drosophila* for ~100 and ~120 My respectively (Beverley and Wilson, 1984). Our analysis reveals that, in many cases, CSBs that are highly conserved in *Drosophila* are also conserved in *Ceratitis* and *Musca*. Similar to ultraconserved sequences in vertebrates, we consider these cross-phyla conserved sequences to be uCSBs. Additionally, the linear order of these uCSBs with respect to flanking structural genes is also maintained. However, subset of the uCSBs exhibits inverted orientation relative to the *Drosophila* sequence, suggesting that while enhancer location is conserved, their orientation relative to flanking genes is not.

For detection of conserved sequences in mosquitos, we have adapted EvoPrinter algorithms, to include 22 species of *Anopheles* plus *Culex pipens* and *Aedes aegypti*. Use of *Anopheles* species allows for the resolution of CSB clusters that resemble those of *Drosophila*. Comparison of *Anopheles* with *Culex* and *Aedes,* separated by ~150 million years of evolutionary divergence (Krzywinski et al. 2006), reveals uCSBs shared among these taxa. Although mosquitoes are considered to be Dipterans, uCSBs were conserved between mosquito species but not with flies.

In addition, we have developed EvoPrinter tools for sequence analysis of seven bee and thirteen ant species. Both ants and bees belong to the Hymenoptera order and have been separated by ~170 million years (Peters et al. 2017). Within the bees, *Megachile* and *Dufourea* are sufficiently removed from *Apis* and *Bombus* (~100 My; Peters et al. 2017, Elsik et al. 2016) that only portions of CSBs are shared between species; these can be considered to be ultraconserved sequences. uCSBs are found that are shared between ant and bee species, and these are positionally conserved with respect to their associated structural genes. Finally, we discovered ant specific and bee specific CSB clusters that are not shared between the two taxa but are interspersed between shared uCSBs.

Methods

Sequence curation and alignment: *Drosophila melanogaster (Dm), Apis mellifera (Am)* and *Anopheles gambiae (Ag), the fly,* bee and mosquito genomic sequences, were curated from the UCSC genome browser. BLASTn (Altschul et al. 1990) was used to identify non-coding sequences within other species not represented in the UCSC genome browser. Where possible, BLAT (Kent, 2002) and BLASTn were used in comparing the order and orientation of ultra-conserved sequences in reference species with dipteran, bee and mosquito test species. BLAT was not available for the *Culex* comparison to *Aedes*, but we found that the 'align two sequences' algorithm of BLAST, using the 'Somewhat similar sequences (BLASTn)' setting, was comparable to BLAT in sensitivity to sequence homology and was useful in this comparison. Similarly, the pairwise sequence alignment program Needle, which uses the Needleman-Wunsch algorithm (Needleman et al. 1970), aligned shorter regions of near identity that could not be seen by other methods.

Mosquito EvoPrinter: An EvoPrint provides a single uninterrupted view, with near base-pair resolution of conserved sequences as they appear in a species of interest. A prior paper describes protocols for genome indexing, enhanced BLAT alignments and scoring of EvoPrint alignments. Readouts are comparable to those already described (Yavatkar et al. 2008).

To compare 24 *Anopheles, Aedes* and *Culex* genomes, sequences were obtained from VectorBase (https://www.vectorbase.org/genomes). The mosquito EvoPrinter consists of 20 species, including 7 species of the Gambiae subgroup and related species *A. christyi and A. epiroticus*, 5 species of the Neocellia and Myzomyia series (including *A. stephensi, A. maculates, A. calcifacies, A. funestus* and *A. minimus*), 2 species of the Neomyzomyia series (*Anopheles darius* and *Anopheles farauti*), 2 species of subgenus *Anopheles* (*A. sinensus* and *A. atroparvus*), Nyssoryhynchus and other American species, (*A. albimanus* and *A. darling*), and two species of the subfamily Culicinae (*Aedes aegypti* and *Culex quinquefaciatus*). Mosquito genomes are documented by Holt et al. 2002; Nene, et al. 2007; Reddy, et al. 2012, and Neafsey et al. 2014.

Hymenoptera EvoPrinter: We have also formatted seven bee species, including 6 members of the family Apidae and one member of each of the Megachilidae and Halictidae families (Table 1). In addition, we have formatted 13 ant (Formicidae) species, a diverse family of social insects, for EvoPrinter analysis (Table 1). Among these are eight representative of the subfamily Myrmicinae, three representatives of the Formicinae, two of the Ponerinae, and one Dolichoderinae. For consistency, we selected a member of the Myrmicinae as input/reference sequence, and species selection was dependent on the integrity and completeness of the sequence. The ant and bee EvoPrinter consist of the following species, grouped according to their phylogenetic relationships:

Table 1: Ant and Bee species formatted for EvoPrint analysis

Bee species (superfamily **Apoidea**)	Ant species (superfamily **Formicoidea**)
Apidae – social bees	**Myrmicinae –**
	Atta cephalotes – Leafcutter ant
Apis mellifera - Western honey bee	*Acromyrmex echinatior* – Panama leaf-cutting ant
Apis dorsata - Giant honey bee	*Wasmannia auropunctata* - Little fire ant
Apis florea - Dwarf honey bee	*Pogonomyrmex barbatus* - Red harvester ant
	Monomorium pharaonis - Pharaoh ant
Bombus terrestrialis - Buff-tailed bumblebee	*Solenopsis invicta* – Red fire ant
Bombus impatiens - Common eastern bumblebee	*Vollenhovia emeryi* – Japanese ant
Habropoda laboriosa - Southeastern blueberry bee	*Cardiocondyla obscurior* – a tramp ant
Megachilidae	**Formicinae**
Megachile rotundata - alfalfa leafcutter bee – solitary	*Lasius niger* - Black garden ant
	Ooceraea biroi - Army ant – colonial raider ant
	Camponotus floridanus - Florida carpenter ants
	Ponerinae
	Harpegnathos saltator - Jerdon's jumping ant
	Dolichoderinae
	Linepithema humile - Argentine Ant

Results and Discussion

Comparative analysis of dipteran non-coding DNA

Our previous study of 19 consecutive *in vivo* tested *Drosophila* enhancers contained within a 28.9 kb intragenic region located between the *vvl* and *Prat2* genes, revealed that each CSB cluster functioned independently as spatial/temporal *cis*-regulatory enhancer (Kundu et al. 2013). The enhancers possessed a diversity of regulatory functions, including dynamic activation of expression in defined patterns within subsets of cells in discrete regions of the embryo, larvae and/or adult.

Submission of the 29 Kb enhancer field to the RefSeq Genome Database of *Ceratitis capitata* via BLASTn revealed 17 uCSBs; all 17 regions were colinear and located between the *Ceratitis* orthologs of *Drosophila vvl* and *Prat2* genes. In each case the matches between *Ceratitis* and *Drosophila* corresponded to a complete or a portion of a

CSB identified as being highly conserved among *Drosophila* species (Kundu et al. 2013). Submission of the same *Drosophila* region to *Musca domestica* RefSeq Genome Database revealed 13 uCSBs that are colinearly arrayed within the *Musca* genome. Since the structural gene and these conserved uCSBs are currently on different contigs, the absolute orientation of the *Musca* sequences with respect to the *Musca vll* structural gene could not be determined. Nine of these *Ceratitis* and *Musca* CSBs were present in both species and corresponded to CSBs contained in several of the enhancers identified in our previous study of the *Drosophila* enhancer field (Kundu et al., 2013). The conservation within one of these embryonic neuroblast enhancers, vvl-41, is depicted in Fig. 1. Panel A of Fig. 1 is an EvoPrint of vvl-41 annotated to show shared CSBs with *Ceratitis* and *Musca*. Green CSBs are shared 3 ways between the three species, red letters represent bases that are shared between *Dm* and *Ceratitis* and blue letters represent bases that are shared exclusively between *Dm* and *Musca*. Fig. 1B shows two and three-way alignments in vvl-41 between the conserved CSBs in the three species. In many cases the uCSBs contained known DNA motifs for TFs. Each of the CSB elements in vvl-41 that are shared between *Dm* and *Ceratitis* are in the same orientation with respect to the *vvl* structural gene. However, in *Musca*, the orientation of elements with respect to the structural gene is unknown since the structural gene and the CSBs are on different contigs. Supplemental fig. 1 presents three-way alignments of each of the other eight uCSBs within the *vvl* enhancer field that are shared between *Dm, Ceratitis* and *Musca*. The uCSB of vvl-49 in *Ceratitis* is in reverse orientation with respect to the vvl structural gene. Many of the uCSBs in Musca are in a different orientation on the contig than in *Dm*, indicating microinversions. We conclude that, except for microinversions, the order and orientation of highly conserved non-coding sequences in *Drosophila, Ceratitis* and *Musca* with respect to flanking genes is the same.

Many of the non-coding regions in dipteran genomes contain uCSBs, especially in and around developmental determinants, and many of these are likely to be *cis*-regulatory elements such as those found in the *vvl* enhancer field. Another example is the prevalence of uCSBs found in the non-coding sequences associated the *Dm hth* gene locus. A previous study identified an ultraconserved regions in *hth* shared between *Drosophila* and *Anopheles* (Glazov et al. 2005). We have identified additional *hth* uCSBs shared among *Dm*, *Ceratitis* and *Musca*. We identified a total of 16 CSBs shared between the three species, 8 CSBs shared between *Dm* and *Ceratitis* but not *Musca*, and 7 CSBs shared between *Dm* and *Musca*, but not *Ceratitis* (fig. 2 and data not shown). Both *Ceratitis* and *Musca* contain uCSBs that were in reversed orientation with respect to the *Drosophila* orthologous regions.

EvoPrint analysis of *Drosophila hth* sequences immediately upstream and including the first exon, revealed a conserved sequence cluster (Fig.2) associated with the transcriptional start site. Fig.2A illustrates correspondence of the *Dm* conserved region in *Ceratitis* and *Musca*. Two of the longer CSBs were conserved in both *Ceratitis* and *Musca*, one shorter CSB was conserved only in *Musca*, and a second shorter CSB was conserved only in *Ceratitis*. Two and three-way alignments as revealed by BLASTn in a comparison of *Dm*, *Ceratitis* and *Musca* are shown in Fig.2B. Each of the uCSBs is in the same orientation with respect to the *hth* structural gene.

Discovery of non-coding conserved sequence elements in mosquitoes

EvoPrinting combinations of species using A. *gambiae* as a reference species and multiple species from the Neocellia and Myzomyia series and the Neomyzomyia provides a sufficient distance from *A. gambiae* to resolve CSBs. The CSB clusters resolved within the *Anopheles* species (data not shown) are similar to those detected using *Dm* as a reference sequence (Brody et al, 2008). Phylogenic analysis has revealed the *Anopheles* species have diverged from ~48 My to ~30 My (Kamali et al, 2014) while *Aedes* and *Culex* diversified from the *Anopheles* lineage in the Jurassic era (~145–200 Mya; Krzywinski et al, 2006) or even earlier.

We sought to identify uCSBs in mosquitos by comparing *Anopheles* species with *Aedes* and *Culex*. We used non-coding sequences associated with the mosquito homolog of the morphogen *wingless* (reviewed by Nusse and Varmus, 1992) to discover associated conserved non-coding sequences. Fig. 3 illustrates a CSB cluster slightly more than 27,000 bp upstream of the *A. gambiae wingless* coding exons. CSB orientation in *A. gambiae* was reversed with respect to the ORF when compared to the orentations of both *Culex* and *Aedes* CSBs. We identified uCSBs, conserved in *Culex* and *Aedes*, coincide with CSBs revealed by EvoPrint analysis of *Anopheles* non-coding sequences. Supplemental fig. 2 illustrates a EvoPrinter scorecard for the non-coding *wingless*-associated CSB cluster described in Fig. 3. Scores for the first four species, all members of the gambiae complex, are similar to that of *A. gambiae* against itself, with subsequent scores reflecting increased divergence from *A. gambiae*. *Culex* and *Aedes* are distinguished from the other species by their belonging to a distinctive branch of the mosquito evolutionary tree, the Culicinae subfamily and their low scores against the *A.*

gambiae input sequence. The mosquito EvoPrinter consists of 20 species, including 7 species of the Gambiae subgroup and related species *A. christyi and A. epiroticus*, 5 species of the Neocellia and Myzomyia series (including *A. stephensi, A. maculates, A. calcifacies, A. funestus* and *A. minimus*), 2 species of the Neomyzomyia series (Anopheles darius and Anopheles farauti), 2 species of subgenus Anopheles (A. sinensus and A. atroparvus), Nyssoryhynchus and other American species, (A. albimanus and A. darling), and two species of the subfamily Culicinae (Aedes aegypti and Culex quinquefaciatus). Mosquito genomes are described by Holt et al., 2002; Nene et al., 2007; Reddy et al., 2012, and Neafsey et al, 2014.

Conserved sequence elements in bees and ants

Bees and ants are members of the Hymenoptera Order, representing the Apoidea (bee) and Vespoidea (ant) super-families. Current estimates suggest that the two families have evolved separately for over 100 million years (Elsik et al. 2015: Hymenoptera Genome Database: integrating genome annotations in HymenopteraMine). To identify conserved sequences shared by bees and ants or unique to each family, we developed EvoPrinter alignment tools for seven bee and 13 ant species (Table 1). Three approaches were employed to identify/confirm conserved elements (both in coding and non-coding sequences) and their positioning within bee and ant orthologous DNAs. First, Evoprinter analysis of bee and ant genes identified conserved sequences in either bees or ants and ultra-conserved sequence elements shared by both families (figs. 4,5). Second, BLASTn alignments of the orthologous DNAs identified/confirmed CSBs that were either bee or ant specific or shared by both (data not shown). Third, side-by-side comparisons of ant and bee EvoPrints and BLASTn comparisons revealed similar positioning of orthologous CSBs relative to conserved exons (figs. 6, S2 and data not shown).

To identify conserved sequences within bee species we initially generated EvoPrints of the honey bee (Apis mellifera) genes using other *Apis* and *Bombus* species. Using EvoPrints of the *Dscam2* locus resolved clusters of conserved sequences (fig. 4). *Dscam2* is implicated in axon guidance in *Drosophila* (Millard et al. 2007) and in regulation of social immunity behavior in honeybees (reviewed by Cremer et al. 2007; Harpur et al. 2019). The EvoPrint scorecard (fig. 4A) reveals a high score (close relationship) with the homologous region in the other two *Apis* species. The more distant *Bombus* species score lower by greater than 50%, and *Habropoda* represents a step down from the more closely related *Bombus* species. *Megachile* shows a significantly lower score reflecting its more distant relationship to *Apis mellifera*. The relaxed EvoPrint readout reveals two CSB clusters (fig. 4b). Only one sequence cluster, the lower 3' cluster, is conserved in all six test species examined, while the 5' cluster is absent present in all species except *Megachile*. BLAST searches confirmed that the 3' cluster was absent from *Megachile*, a more distant species *Dufourea novaeangliae,* and all ant species in the RefSeq genome database (data not shown). BLASTn alignments also revealed conservation of the 3' cluster in the bee species *Dufourea novaeangliae*, the wasp species *Polistes canadensis* and two ant species, *Vollenhavia emeryi* and *Dinoponera quadriceps*.

EvoPrinter analysis of bee and ant genes that are orthologs of the *Drosophila* neural development genes *goosecoid* (*gsc*) and *castor* (*cas*) revealed conserved non-coding DNA that is unique to either bees or ants or conserved in both (fig. 5). The *Drosophila* Gsc homeodomain transcription factor is required for proper axon wiring during embryonic CNS development and has recently been linked to social immunity behavior in honeybees (reviewed by Cremer et al. 2007; Harpur et al. 2019). The *Drosophila* Cas Zn-finger transcription factor has been shown to be essential for neuroblast temporal identity decisions during neural lineage development (Baumgardt et al. 2014; reviewed by Brody and Odenwald 2007). EvoPrints of the Hymenoptera orthologs identify non-coding conserved sequence clusters that contained core uCSBs shared by both ant and bee superfamilies, and these uCSBs are frequently flanked by family-specific conserved clusters (figs. 4, 5, 6 and data not shown). For example, analysis of the non-coding sequence upstream of the *Wasmannia auropunctata* (ant) *cas* first exon identifies both a conserved sequence cluster that contains ant and bee uCSBs and an ant specific conserved cluster that has no counterpart found in bees (fig. 5B and data not shown). It is likely that the ant specific cluster was deleted in bees, since BLASTn searchs of *Wasmannia* against the European paper wasp *Polistes dominula* reveals conservation of a core sequence corresponding to this cluster (data not shown).

The combined evolutionary divergence in the *gsc* and *cas* EvoPrints, accomplished by the using multiple test species, reveals that many of the amino acid codon specificity positions are conserved while wobble positions in their ORFs are not. The lack of wobble conservation indicates that the combined divergence of the test species used to generate the prints afford near base pair resolution of essential DNA.

Cross-group/side-by-side bee and ant comparison of their conserved DNA was performed using bee specific and ant specific EvoPrints and by BLASTn alignments (figs. 6, S2 and data not shown). Fig. 6 highlights the conservation observed among bee and ant exons and flanking sequence of the *glass bottom boat* (*gbb*, *60A*) locus of *Apis melliflera* EvoPrinted with four bee test species (panel A) and the *Wasmannia auropunctata gbb* locus EvoPrinted with three ant species (panel B). Coding sequences are underlined red, non-coding homologous regions are underlined blue, and novel CSBs present in either ants or bees but not both are indicated by the vertical lines to the side of each EvoPrint. Similarly, EvoPrinting a single exon and flanking regions of the *Apis mellifera homothorax* locus with four bee species and generating an ant specific EvoPrint of the orthologous ant sequence of the *Ooceraea biroi homothorax* locus with ten other ant species, reveals CSBs that are conserved in both *Apis* and *Ooceraea*, as well as sequences that are restricted to one of the two Hymenopteran families (supplemental fig. 2).

Summary

Our cross-species comparisons document shared ultraconserved sequences within three separate groups of insects, e.g., flies, mosquitos and Hymenoptera. In each case, CSB clusters were shown to consist of a core of highly conserved CSBs flanked by less well conserved regions. Our previous work in *Drosophila* has shown that most CSB clusters function autonomously as enhancers that control flanking gene expression patterns. This pattern of conservation has been documented for mammalian enhancers and suggests a common structure for cis-regulatory sequences across evolution. In many cases, the uCSBs were flanked by CSBs that were not shared across phyla. We suggest that core uCSBs perform essential cis-regulatory function(s), while flanking conserved sequences, shared only by more closely related species, serve to provide the species specificity to enhancer function. Often these enhancers control a sub-pattern of gene expression. (Perry et al., 2010, Kuzin et al., 2012, Ross et al., 2015)

In the three species groups examined in this study, flies, mosquitos, and ants and bees each have similar clusters of conserved sequences. For example, the alignment of *Apis mellifera* sequences with other *Apis* and *Bombus species,* or of Anopheles gambiae with other Anopheles species resolved clusters of conserved sequences resembling in many aspects BLAT alignment of *Drosophila* Sophophora subgroup (including *D. melanogaster, D. yakuba* and *D. persimilis*) with the *Drosophila* subgroup (including *D. virilis, D grimshawi* and *D mojavensis*). These alignments revealed regions that can be considered to be, in analogy to *Drosophila*, CSB clusters flanked by regions of non-conservation (termed inter-clustal regions) (Kuzin et al. 2009; Ross et al, 2015). Adding more distantly related species, *Ceratitis* and *Musca* for flies, *Aedes* and *Culex* for mosquitos, and *Megachile* and ants for Hymenoptera revealed ultraconserved CSBs, nested within the CSB clusters. Therefore, the general pattern of conservation is the same for all three taxa examined.

In most cases both nBLAST and the EvoPrinter algorithm, based on the eBLAT algorithm had similar sensitivities and gave comparable results, but we recommend that the two techniques should be used in conjunction with one another. The advantage of EvoPrinter is the presentation of an interspecies comparison as a single alignment, while the advantage of nBLAST is that it provides a sensitive detection of sequence homology in a one-on-one alignment. EMBOSSED Needle alignment gives an even more sensitive detection of shorter sequences and is of use once BLAT or EvoPrinter has been used to discover shared CSBs and/or CSB clusters.

Consecutive CSB clusters in distantly related species are often co-linear, in that the order of is maintained with respect to flanking genes. We have documented exceptions to this in both flies and mosquitos in which mini-inversions (rearrangements) occur. The fact that the orientation of CSB clusters with respect to the ORF suggests that such inversions can be tolerated, and that the orientation is irrelevant to their putative enhancer function. However, the co-linear ordering of non-coding CSB clusters suggests that the order of CSB clusters may be important for gene regulation.

The pattern of conservation of CSB clusters in the Hymenoptera suggests that new CSB clusters have their origin not by recombination with other cis-regulatory DNA but random mutational changes. The same is true for mosquitos, in which shared sequences between *Culex* and *Aedes* are often not found in *Anopheles*. We sought to identify ultraconserved CSBs shared among bees and mosquitos that were related to those shared by *Drosophila, Ceratitis* and *Musca*, but failed to find such sequences using conventional alignment protocols. This work provides a basis for future studies to understand unique commonalities and functional differences between taxonomic groups.

Acknowledgments

We would like to acknowledge the editorial expertise and assistance of Judy Brody, Mihaela Serpe, Rosario Vicidomini, and Saumitra Choudhury.

References

Altschul SF, Gish W, Miller, W, Myers, EW. Lipman, DJ. 1990. Basic local alignment search tool. *J Mol Biol*. 215:403-410.

Ayyar S, Negre B, Simpson P, Stollewerk A. 2010. An arthropod cis-regulatory element functioning in sensory organ precursor development dates back to the Cambrian. *BMC Biol,* 8:127.

Bejerano G, Siepel AC, Kent WJ, Haussler D. 2005. Computational screening of conserved genomic DNA in search of functional noncoding elements. *Nat Methods* 2: 535-545.

Bergman CM, Pfeiffer BD, Rincon-Limas DE, Hoskins, RA, Gnirke, A, Mungall, CJ. Wang AM, Kronmiller B, Pacleb J, Park S, et al. 2002. Assessing the impact of comparative genomic sequence data on the functional annotation of the Drosophila genome. *Genome Bio.* 3: RESEARCH0086.

Baumgardt M, Karlsson D, Salmani BY, Bivik C, MacDonald RB, Gunnar E, Thor S. 2014. Global programmed switch in neural daughter cell proliferation mode triggered by a temporal gene cascade. Dev Cell 30(2):192-208.

Beverley, SM and Wilson, AC. 1984. Molecular evolution in *Drosophila* and the higher Diptera II. A time scale for fly evolution. *J Mol Evol*. 21:1-13.

Brody T, Rasband W, Baler K, Kuzin A, Kundu M, Odenwald WF. 2007. *cis*-Decoder discovers constellations of conserved DNA sequences shared among tissue-specific enhancers. Genome Biol. 8:R75.

Cremer S, Armitage SA, Schmid-Hempel P. 2007. Social immunity. Curr Biol. 17:R693-702.

Dickel DE, Ypsilanti AR, Pla R, Zhu Y, Barozzi, I, Mannion BJ, Khin YS, Fukuda-Yuzawa Y, Plajzer-Frick I, Pickle CS. et al. 2018. Ultraconserved enhancers are required for normal development. *Cell* 172:491-499.

Elsik CG, Tayal A, Diesh CM, Unni DR, Emery ML, Nguyen HN and Hagen DE. 2016. Hymenoptera Genome Database: integrating genome annotations in HymenopteraMine. *Nucleic Acids Res.* 44:D793-800.

Gao S, Takemura SY, Ting CY, Huang S, Lu Z, Luan H, Rister J, Thum AS, Yang M, Hong ST. et al. 2008. The neural substrate of spectral preference in Drosophila. *Neuron,* 60:328-342.

Glazov EA, Pheasant M, McGraw EA, Bejerano G, Mattick JS. 2005. Ultraconserved elements in insect genomes: a highly conserved intronic sequence implicated in the control of homothorax mRNA splicing. *Genome Res.* 15:800-8.

Hardison RC. 2000. Conserved noncoding sequences are reliable guides to regulatory elements. *Trends Genet.* 16:369-372.

Hare, EE, Peterson, BK, Iyer, VN, Meier R, Eisen MB. 2008. Sepsid *even-skipped* enhancers are functionally conserved in Drosophila despite lack of sequence conservation. *PLoS Genet.* 4: e1000106.

Harpur BA, Guarna MM, Huxter E, Higo H, Moon KM, Hoover SE, Ibrahim A, Melathopoulos AP, Desai S, Currie RW, et al. 2019. Integrative Genomics Reveals the Genetics and Evolution of the Honey Bee's Social Immune System.
Genome Biol Evol pii:evz018.

Holt RA, Subramanian GM, Halpern A, Sutton GG, Charlab R, Nusskern DR, Wincker P, Clark AG, Ribeiro JM, Wides R. et al. 2002. The genome sequence of the malaria mosquito *Anopheles gambiae*. *Science* 298:129-49.

Irvine Q, Carr JL, Bailey, WJ, Kawasaki K, Shimizu N, Amemiya, CT, Ruddle FH. 2002. Genomic analysis of Hox clusters in the sea lamprey *Petromyzon marinus*. *J Exp Zool.* 294:47-62.

Kent WJ. 2002. BLAT--the BLAST-like alignment tool. Genome Res. 12(4): 656-64.

Kamali M, Marek PE, Peery A, Antonio-Nkondjio C, Ndo C, Tu Z, Simard F, Sharakhov IV. 2014. Multigene phylogenetics reveals temporal diversification of major African malaria vectors. *PLoS One* 9:e93580.

Krzywinski J, Grushko OG, Besansky NJ. 2006. Analysis of the complete mitochondrial DNA from *Anopheles funestus*: an improved dipteran mitochondrial genome annotation and a temporal dimension of mosquito evolution. *Mol Phylogenet Evol.* 39:417-423.

Kundu M, Kuzin A, Lin TY, Lee CH, Brody T, Odenwald WF. 2013. Cis-regulatory complexity within a large non-coding region in the *Drosophila* genome. *PLoS One,* 8:e60137.

Kuzin A, Kundu M, Ekatomatis A, Brody T, Odenwald WF. 2009. Conserved sequence block clustering and flanking inter-cluster flexibility delineate enhancers that regulate *nerfin-1* expression during *Drosophila* CNS development. *Gene Expr Patterns* 9:65-72.

Kuzin A, Kundu M, Ross J, Koizumi K, Brody T Odenwald WF. 2012. The cis-regulatory dynamics of the *Drosophila* CNS determinant *castor* are controlled by multiple sub-pattern enhancers. *Gene Expr Patterns* 12:261-272.

Lettice LA, Williamson I, Devenney PS, Kilanowski F, Dorin J, Hill RE. 2014. Development of five digits is controlled by a bipartite long-range cis-regulator. *Development* 141:1715-1725

Liu F, Posakony JW. 2012. Role of architecture in the function and specificity of two Notch-regulated transcriptional enhancer modules. *PLoS Genet.* 8:e1002796.

Loots G, Ovcharenko I. 2007. ECRbase: database of evolutionary conserved regions, promoters, and transcription factor binding sites in vertebrate genomes. *Bioinformatics* 23:122-124.

McEwen, GK, Goode, DK, Parker HJ, Woolfe A, Callaway H, Elgar G. 2009. Early evolution of conserved regulatory sequences associated with development in vertebrates. *PLoS Genet.* 5:e1000762.

Millard SS, Flanagan JJ, Pappu KS, Wu W, Zipursky SL. 2007. Dscam2 mediates axonal tiling in the Drosophila visual system. *Nature* 447:720-4.

Neafsey DE, Waterhouse RM, Abai MR, Aganezov SS, Alekseyev MA, Allen JE, Amon J, Arcà B, Arensburger P, Artemov G. et al. 2015. Mosquito genomics. Highly evolvable malaria vectors: the genomes of 16 Anopheles mosquitoes. *Science* 347: 1258522.

Needleman SB, Wunsch, CD. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol.* 48: 443-453.

Nene V, Wortman JR, Lawson D, Haas B, Kodira C, Tu ZJ, Loftus B, Xi Z, Megy K, Grabherr M, Ren Q. et al. 2007. Genome sequence of Aedes aegypti, a major arbovirus vector. *Science* 316: 1718-1723.

Nusse R, Varmus HE. 1992. Wnt genes. *Cell* 69:1073-87.

Odenwald WF, Rasband W, Kuzin A, Brody T. 2005. EVOPRINTER, a multigenomic comparative tool for rapid identification of functionally important DNA. *Proc Natl Acad Sci U S A.* 102:14700-14705.

Panne D. 2008. The enhanceosome. *Curr Opin Struct Bio.* 18:236-242.

Pennacchio LA, Ahituv N, Moses AM, Prabhakar S, Nobrega MA, Shoukry M, Minovitsky S, Dubchak I, Holt A, Lewis KD et al. 2006. In vivo enhancer analysis of human conserved non-coding sequences. *Nature* 444:499-502.

Perry MW, Boettiger AN, Bothma JP, Levine M. (2010). Shadow enhancers foster robustness of *Drosophila* gastrulation. Curr Biol. 20:1562–1567.

Peters, RS, Krogmann L, Mayer C, Donath A, Gunkel S, Meusemann K, Kozlov A, Podsiadlowski L, Petersen M, Lanfear R, et al. 2017. Evolutionary History of the Hymenoptera. *Curr Biol.* 27:1013-1018.

Reddy, BP, Labbe P, Corbel V. 2012. Culex genome is not just another genome for comparative genomics. *Parasit Vectors* 5:63.

Ross J, Kuzin A, Brody T, Odenwald WF. 2015. cis-regulatory analysis of the *Drosophila pdm* locus reveals a diversity of neural enhancers. BMC Genomics 16:700.

Ross J, Kuzin A, Brody T, Odenwald WF. 2018. Mutational analysis of a *Drosophila* neuroblast enhancer governing nubbin expression during CNS development. Genesis 56:e23237.

Sieglaff DH, Dunn WA, Xie XS, Megy K, Marinotti O, James AA. 2009. Comparative dgenomics allows the discovery of cis-regulatory elements in mosquitoes. *Proc Natl Acad Sci U S A.* 106:3053-3058.

Suryamohan K, Hanson C, Andrews E, Sinha S, Scheel MD, Halfon MS. 2016. Redeployment of a conserved gene regulatory network during Aedes aegypti development. Dev Biol. 416:402-13.

Thomas JW, Touchman JW, Blakesley RW, Bouffard GG, Beckstrom-Sternberg SM, Margulies EH, Blanchette M, Siepel AC, Thomas PJ, McDowell JC, et al. 2003. Comparative analyses of multi-species sequences from targeted genomic regions. *Nature* 424:788-793.

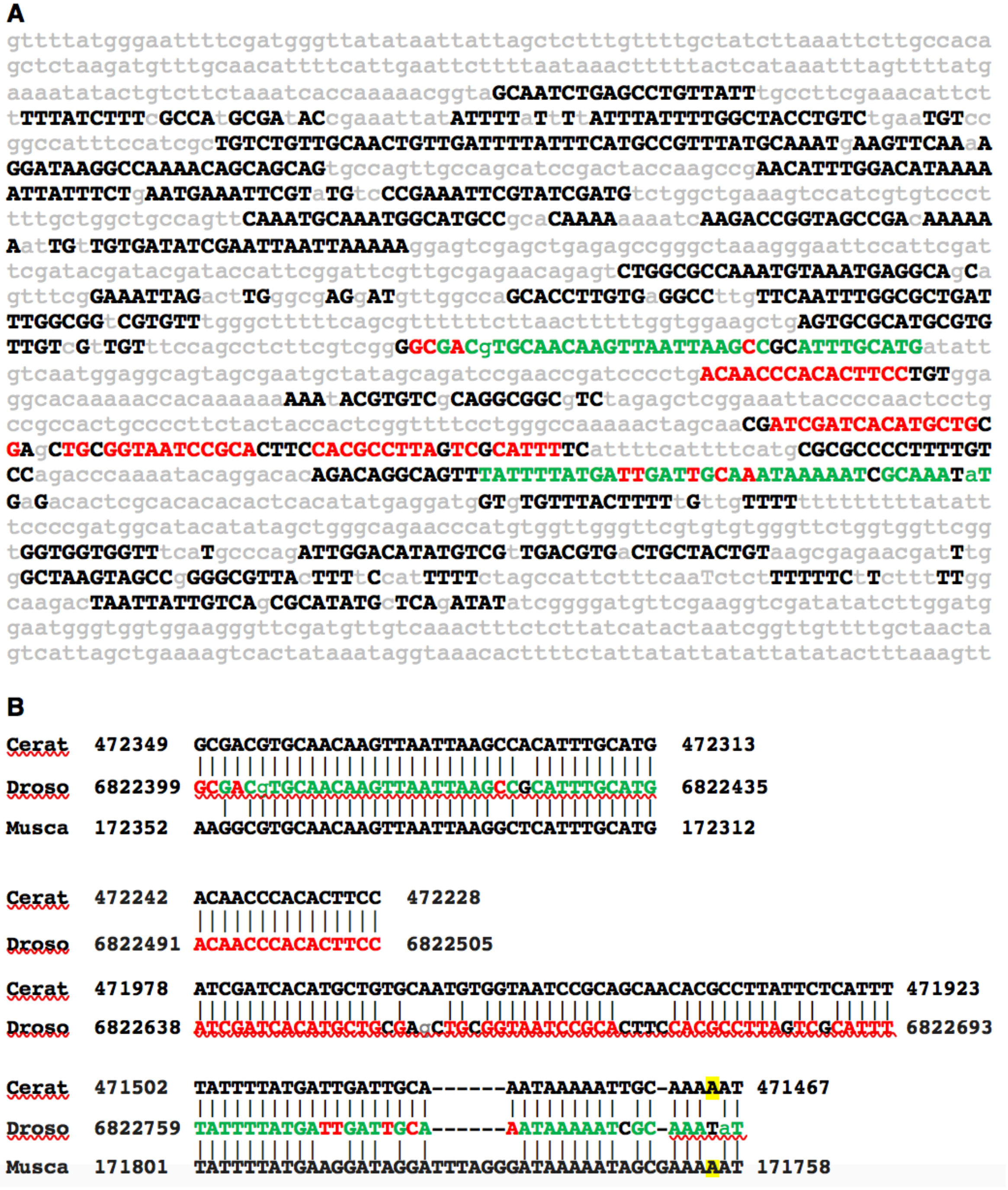
Visel A, Prabhakar S, Akiyama JA, Shoukry M, Lewis KD, Holt A, Plajzer-Frick I, Afzal V, Rubin EM, Pennacchio LA. 2008. Ultraconservation identifies a small subset of extremely constrained developmental enhancers. *Nat Genet.* 40:158-160.

Visel A, Rubin EM, Pennacchio LA 2009. Genomic views of distant-acting enhancers. *Nature* 461:199-205.

Visel A, Taher L, Girgis H, May D, Golonzhka O, Hoch RV McKinsey GL, Pattabiraman K, Silberberg SN, Blow MJ. et al. 2013. A high-resolution enhancer atlas of the developing telencephalon. *Cell* 152:895-908.

Yavatkar AS, Lin Y, Ross J, Fann Y, Brody T, Odenwald WF. 2008. Rapid detection and curation of conserved DNA via enhanced-BLAT and EvoPrinterHD analysis. *BMC Genomics* 9:106.

Figure 1



A

```
gttttatgggaattttcgatgggttatataattattagctctttgttttgctatcttaaattcttgccaca
gctctaagatgtttgcaacattttcattgaattcttttaataaactttttactcataaatttagttttatg
aaaatatactgtcttctaaatcaccaaaaacggtaGCAATCTGAGCCTGTTATTtgccttcgaaacattct
tTTTATCTTTcGCCAtGCGAtACcgaaattatATTTTaTtTtATTTATTTTGGCTACCTGTCtgaaTGTcc
ggccatttccatcgcTGTCTGTTGCAACTGTTGATTTTATTTCATGCCGTTTATGCAAATgAAGTTCAAaA
GGATAAGGCCAAAACAGCAGCAGtgccagttgccagcatccgactaccaagccgAACATTTGGACATAAAA
ATTATTTCTgAATGAAATTCGTaTGtcCCGAAATTCGTATCGATGtctggctgaaagtccatcgtgtccct
tttgctggctgccagttCAAATGCAAATGGCATGCCgcaCAAAAaaaatcAAGACCGGTAGCCGAcAAAAA
AatTGtTGTGATATCGAATTAATTAAAAAggagtcgagctgagagccgggctaaagggaattccattcgat
tcgatacgatacgataccattcggattcgttgcgagaacagagtCTGGCGCCAAATGTAAATGAGGCAgCa
gtttcgGAAATTAGactTGggcgAGgATgttggccaGCACCTTGTGaGGCCttgTTCAATTTGGCGCTGAT
TTGGCGGtCGTGTTtgggcttttcagcgtttttcttaacttttggtggaagctgAGTGCGCATGCGTG
TTGTcGtTGTttccagcctcttcgtcggGGCGACgTGCAACAAGTTAATTAAGCCGCATTTGCATGatatt
gtcaatggaggcagtagcgaatgctatagcagatccgaaccgatcccctgACAACCCACACTTCCTGTgga
ggcacaaaaaccacaaaaaaAAAtACGTGTCgCAGGCGGCgTCtagagctcggaattaccccaactcctg
ccgccactgccccttctactaccactcggttttcctgggccaaaaactagcaaCGATCGATCACATGCTGC
GAgCTGCGGTAATCCGCACTTCCACGCCTTAGTCGCATTTTCattttcattttcatgCGCGCCCCTTTTGT
CCagacccaaaatacaggacacAGACAGGCAGTTTATTTTATGATTGATTGCAAATAAAAATCGCAAATaT
GaGacactcgcacacacactcacatatgaggatgGTgTGTTTACTTTTtGttgTTTTtttttttttatatt
tccccgatggcatacatatagctgggcagaacccatgtggttgggttcgtgtgtgggttctggtggttcgg
tGGTGGTGGTTtcaTgcccagATTGGACATATGTCGtTGACGTGaCTGCTACTGTaagcgagaacgatTtg
gGCTAAGTAGCCgGGGCGTTAcTTTtCcatTTTTctagccattctttcaaTctctTTTTTCtTctttTTgg
caagacTAATTATTGTCAgCGCATATGcTCAgATATatcggggatgttcgaaggtcgatatatcttggatg
gaatgggtggtggaagggttcgatgttgtcaaactttctcttatcatactaatcggttgttttgctaacta
gtcattagctgaaaagtcactataaataggtaaacacttttctattatattatattatatactttaaagtt
```

B

```
Cerat   472349   GCGACGTGCAACAAGTTAATTAAGCCACATTTGCATG   472313
                 |||||||||||||||||||||||||| ||||||||||
Droso   6822399  GCGACgTGCAACAAGTTAATTAAGCCGCATTTGCATG   6822435
                   | |||||||||||||||||||| | ||||||||||
Musca   172352   AAGGCGTGCAACAAGTTAATTAAGGCTCATTTGCATG   172312


Cerat   472242   ACAACCCACACTTCC   472228
                 |||||||||||||||
Droso   6822491  ACAACCCACACTTCC   6822505

Cerat   471978   ATCGATCACATGCTGTGCAATGTGGTAATCCGCAGCAACACGCCTTATTCTCATTT 471923
                 ||||||||||||||| |   || |||||||||||    ||||||||| || |||||
Droso   6822638  ATCGATCACATGCTGCGAgCTGCGGTAATCCGCACTTCCACGCCTTAGTCGCATTT 6822693

Cerat   471502   TATTTTATGATTGATTGCA------AATAAAAATTGC-AAAAAT 471467
                 |||||||||||||||||||      ||||||||| || ||| ||
Droso   6822759  TATTTTATGATTGATTGCA------AATAAAAATCGC-AAATaT
                 ||||||||||  ||| | |       |||||||| || ||| ||
Musca   171801   TATTTTATGAAGGATAGGATTTAGGGATAAAAATAGCGAAAAAT 171758
```

Figure 2

A

```
tatttgagggccgcggagggaatttctttcaacgaggggatcccgccgatccccccgcaaaaacaaagaaa
cacgcaaattgccgtcgcaaaaacacaacagaatatattgcattatatattaacaatataatttgttcatc
tcTAGTaacgTAGTATGTAGAAGCAGCTGTGATAAATAGTGCGCATATTATACGAAaaaaGCAAAAAgAGT
ATACGACAaACAAataaaaattaaaAgcAAccacAActactactataAAAataaccaaggaaaaggcaagg
aaactcgcaaaacgagagagagaggaaaagggaaaaaacgccATGGATTTATGACAACAACAAatgcaacc
cagtaacaaatgatgttgcagCAACtactacaatacacgagtataccaagaacatcaACAACaaCAActgC
AACGGCAACcAAcAACAAATTGTATACGATTTTTATAAACCAAACAAAtgcTATGTAAATAaTtaTtatAA
CAaacacacaagtaAAAAAaacAttcAAAAccAGATCTCCCGCTctagaaaacaaacAAcaAACGCGCG
TAACaaaaacgACAACAtCAACAatACAACAAGtAtACaAacgtaatAATAATAcATAAAaTCTaAACaAA
TGTGCCTGTGTTAAAAaAAAAgAAagcaAAAtcAAGaagcaaagcgttcaaaaagcaaaggcttcccaata
catgaaaaaagtgcaaaacatatataaaaatccaagaaaataacatccaaaactgaagaactggaaaagcc
aagagtaaatctaaggactccacacgggcagctacagtagagcttccctaaaccgaagcagttgcagcgg
cggagcgggacgtcggtcagtgtgttgcgcCCCcatttgttgttcctACtgtaAAAAGTGGCACAAATCGG
GTTAGACGTCGCTCCCAGAGCATTAATGGCTCAGCCcAGGGTAAGTAgatccaccaaagttcactttctaa
atatatagaatccatatagaatggtcccccgagagtgttgttatccttaaacaagttccgacggctaaaaa
```

B

```
Cerat 387653   GTTGAAGAAGCTGTGAT-ATTTGTGAGCATAT 387626
               || |||| ||||||||| | | ||| ||||||
Droso 10613790 GTAGAAGCAGCTGTGATAAATAGTGCGCATAT 10613756


Droso 10613564 cATGGATTTATGAC 10613577
               ||||||||||||||
Musca 35636    CATGGATTTATGAC 35649


Cerat 386904    TCAAAACCACATATACAAGCATTGTATACGATTTT----AAAGCAACAAAAT--TAT-TAAATTATT 386865
                  |   | ||  | |  |  |||||||||||||||    ||| |||  ||||  ||| ||||| |||
Droso 10613458  CAACGGCAACcAAcAACAA-ATTGTATACGATTTTT-ATAAACCAAACAAAtgcTATGTAAATAaTt 10613522
                ||||  |||||| |||||| ||||||||||||||||  |||| |||| | ||  ||  |||  ||||
Musca 35742     CAACAACAACCA-CAACAAAATTGTATACGATTTTTTCTAAAACAAAAATATAATAATTAA--AATT 35805


Cerat 386251   AAAAGTGGCACA-GGTCGGGCTGGACATCGCTCCCAGAGCATTAATGGCTCAACCTAGGGTAAGTACAT 386184
               ||||||||||||   ||||||| ||| ||||||||||||||||||||||||| || |||||||||| ||
Droso 10613042 AAAAGTGGCACA-AATCGGGTTAGACGTCGCTCCCAGAGCATTAATGGCTCAGCCcAGGGTAAGTAgat 10612975
               |||||||||||||||||||||| |||  |||||| ||||||||||||||||| || ||||||||||
Musca 383950   AAAAAGTGGCACAAATCGGGTTGGACACCGCTCCTAGAGCATTAATGGCTCAACCTAGGGTAAGTATTG 384012
```

Figure 3

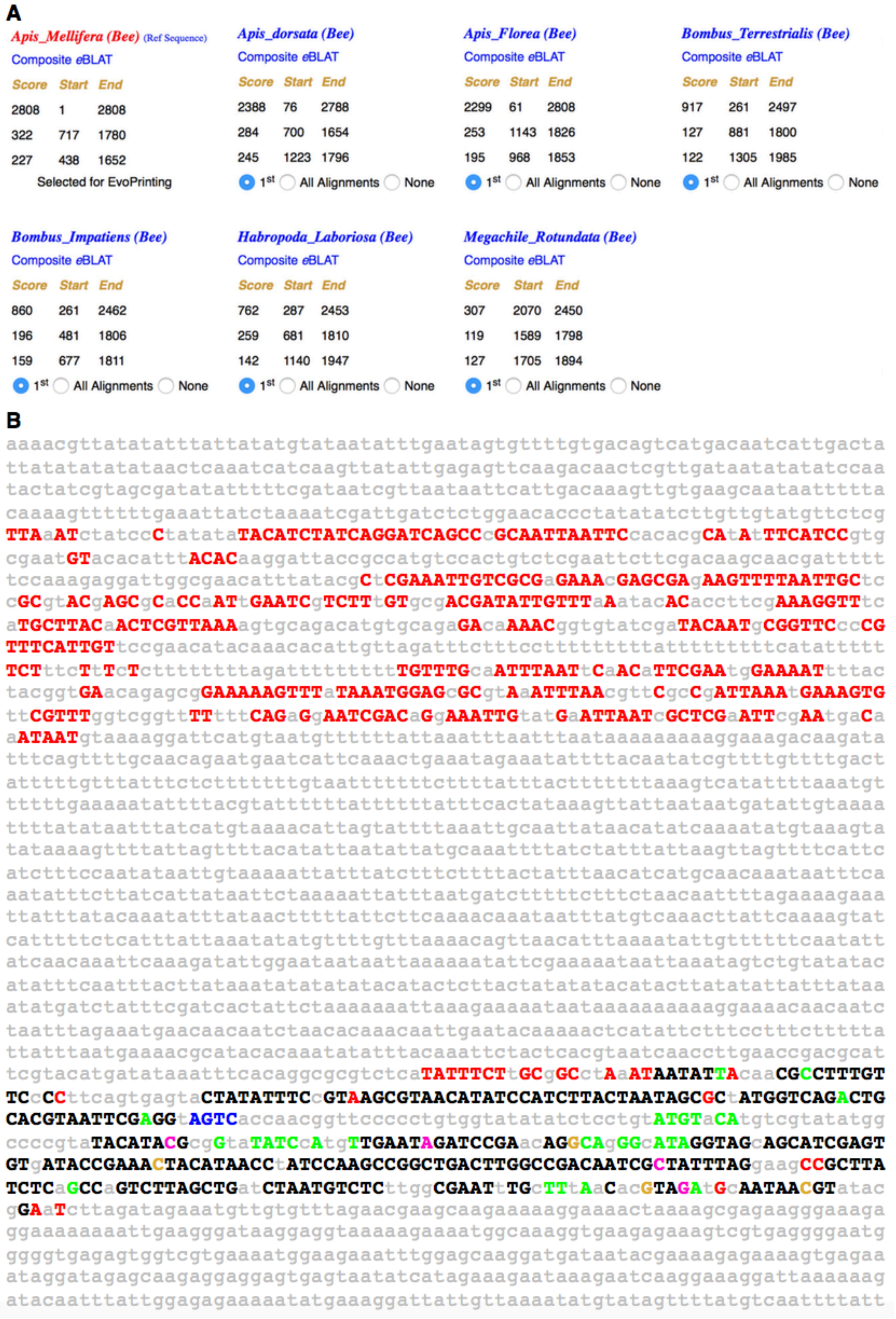
A

```
tattgaataaagaaacgtaggtttaaatcgaaaattgatgatgatatttttatcaaagaaagacattattg
tccctttatcctcccccaaaagaaacacccatcgtagcttattaagttcccttcggccccgcacttgggag
ttGAGTCATTAAATAATTGttcGAgGATAAAtACAcAAtcccgccagcaggcgaccaaaccccttcgccgt
ctgccagaagagccaatacggtgctagttgGAGAAGTGCTGCTAATGTGTTATGACTCAATTTACTGTGAA
AATTTAAATATTGATTCGAagcagcaccgccagtgctgagagagaagaaaGGTggAaATAAAATCATAAAA
ATACCACACaccggcacttgggccgcactcatgagcgcTAGAAAAGTGCATtcaaatgtgcagtgaagagc
atccttcgtctcggtggaatgaagagaggggtgattgtaaagaaagcaaggggagagaTTGTTTGaaGGg
AAAACACTTTAGAAATTAAATATGAgcATTCTATCGgGCTGACTTTTATCTCCATAAATAACGATTTTATG
AACTGACGCGCCTgccCGgggcctgCcgCACCtttggagaaatggaacgcacgaagaggctccgtgggat
ttctcaccgctacagggcgagcggttaGATAAACAtgcCGGcCCCTTTTGTTTTCGtttcgatcgtggaat
tggaaaagcgcagcgaacgaggttggtgatgaagcGTgTGTTTAGAAcaggcgacggaacgagccggcgtg
acGAGTGACATTCTCCGaGCgaaaaatggtcgaaagaagtgcacggggcacacgatcgcgtttgagagtct
ttggagtgacaggtgcatatctgttacaagatcgttgtgatcttctccatgtgctcgatttaaccggtctg
ggatactgatttttttgaagatttcgatcgatccggtttgatcgtaaagtatccgtaatagtgtttaaact
gcgggacgccctattaccgGtCGgcaaaACCGGAtGCtggtaTAATTAGGTTAGCGCCACCGagcgccacc
tGGCGccacccaaatagagttgaaatggcgtgaattggtataattttTGTCTtctatgtttcactccggta
gcagttattTGTCCTATAAGTTAACaATTTTATtgcccggccggcgtcgtcatagagctgtgagccgaga
tattatgagttagatacagacggacgGCGaacatttagctGGaaTTctGTCTGGAGCCagtgtagcaaagt
ctaaatgctgtgtctgatttttcttggtttgatgagatttcagaagaaatgacttgccctcccgccaaaaa
attctccaggaacaaaacacctgtttaatgttttccctgaactcaacgtttagtttaaacggaatggatag
```

B

```
Culex      610788      AATGTGTTATGTAATGTGTTATGATTCAATTACCCAGAAAAATTCAAATATTGAT   610746
                       |          ||||||||||||| ||||||  |    |||||| ||||||||||
Anopheles  42001058    AGAAGTGCTGCTAATGTGTTATGACTCAATTTACTGTGAAAATTTAAATATTGAT   42001004
                       |||||  ||| ||||||||||||| |||||||||    |||||| |||||||||
Aedes      406893776   AGAAGACCTGTTAATGTGTTATGATTCAATTTACCAGAAAAATTCAAATATTGAA   406893721


Culex      610664      GAAAACACTTAAGAAATTAAATATGT--ACGTTCTATCG--CTGACTTTCATCTCCATAAAT
                       |||||||||| ||||||||||||||    | ||||||||  |||||||| ||||||||||||
Anopheles  42000806    gAAAACACTTTAGAAATTAAATATGA--GCATTCTATCGGGCTGACTTTTATCTCCATAAAT
                       |||||||||| |||||||||||||||  |  ||||||||  ||||||||||||| |||||||
Aedes      406893622   GAAAACACTTAAGAAATTAAATATGACTGTGTTCTATCG--CTGACTTTTATCTTCATAAAT


Culex      (cont'd)    AACGCTTTT-ATTAACTTTATGCATCGTCGTCCC   610591
                       |||| |||| || ||||              ||
Anopheles  (cont'd)    AACGATTTT-ATGAACTGAC--------GCGCCT   42000731
                       ||||||||| || ||||           ||||||
Aedes      (cont'd)    AACGATTTTTATTAACTTTACATGCAGTGCGCCT   406893546


Culex      605426     GTCGGACAGACAGGAAGCGTGGGATAATTAGGTTAGCGCCACCAAG   605381
                      |||||  | || ||| || ||| |||||||||||||||||||| ||
Anopheles  42000289   GtCGgcaaaACCGGAtGC-tggtaTAATTAGGTTAGCGCCACCGaG   42000245


Culex      605259     TGTCCCATAAGTTAGCGA-TTTAT   605234
                      ||||| |||||||| | | |||||
Anopheles  2000157    TGTCCTATAAGTTAACaATTTTAT   2000134
```

Figure 4



A

Apis_Mellifera (Bee) (Ref Sequence)

Composite *e*BLAT

Score	Start	End
2808	1	2808
322	717	1780
227	438	1652

Selected for EvoPrinting

Apis_dorsata (Bee)

Composite *e*BLAT

Score	Start	End
2388	76	2788
284	700	1654
245	1223	1796

● 1st ○ All Alignments ○ None

Apis_Florea (Bee)

Composite *e*BLAT

Score	Start	End
2299	61	2808
253	1143	1826
195	968	1853

● 1st ○ All Alignments ○ None

Bombus_Terrestrialis (Bee)

Composite *e*BLAT

Score	Start	End
917	261	2497
127	881	1800
122	1305	1985

● 1st ○ All Alignments ○ None

Bombus_Impatiens (Bee)

Composite *e*BLAT

Score	Start	End
860	261	2462
196	481	1806
159	677	1811

● 1st ○ All Alignments ○ None

Habropoda_Laboriosa (Bee)

Composite *e*BLAT

Score	Start	End
762	287	2453
259	681	1810
142	1140	1947

● 1st ○ All Alignments ○ None

Megachile_Rotundata (Bee)

Composite *e*BLAT

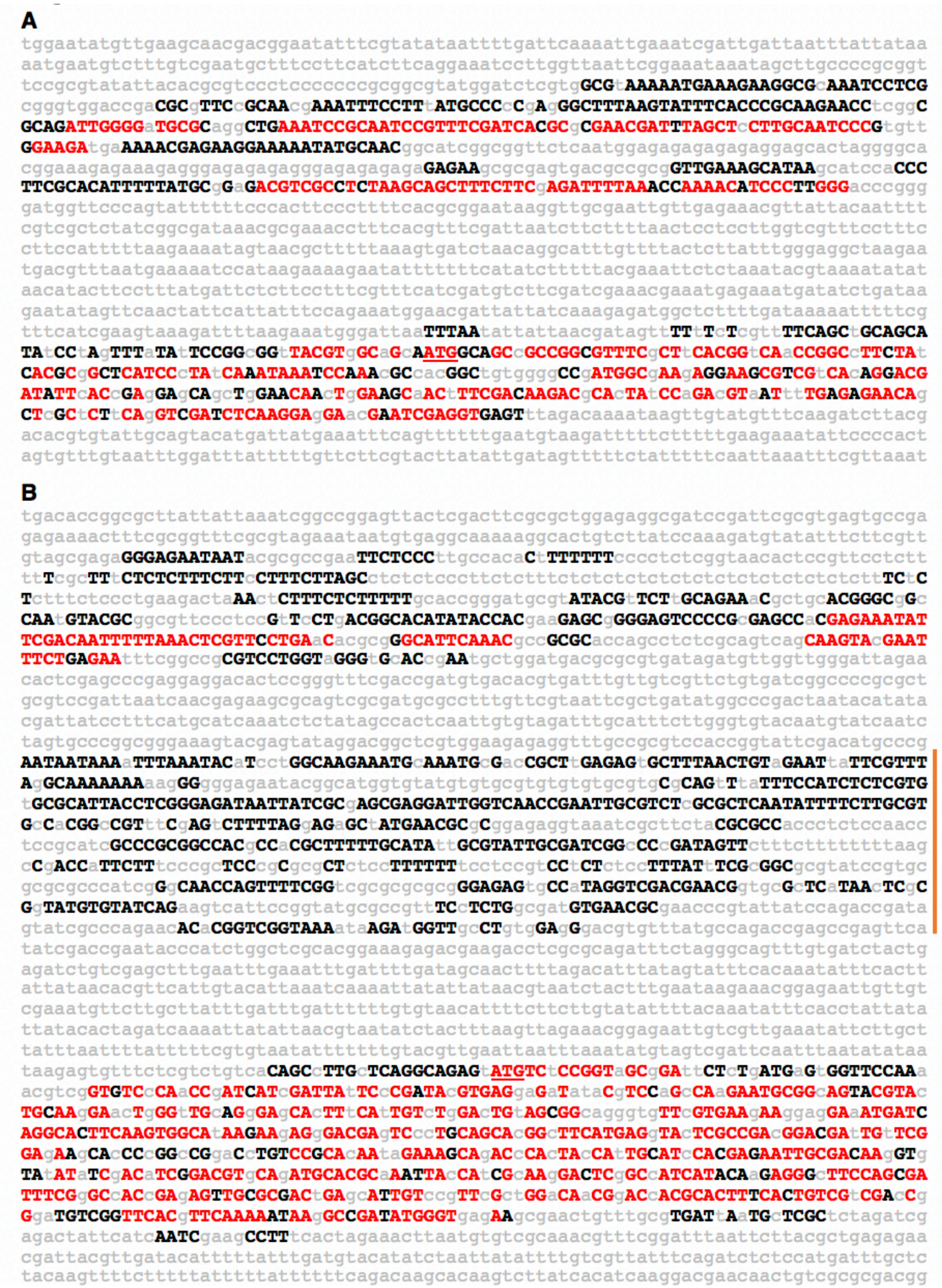
Score	Start	End
307	2070	2450
119	1589	1798
127	1705	1894

● 1st ○ All Alignments ○ None

B

```
aaaacgttatatatttattatatgtataatatttgaatagtgttttgtgacagtcatgacaatcattgacta
ttatatatatataactcaaatcatcaagttatattgagagttcaagacaactcgttgataatatatatccaa
tactatcgtagcgatatatttttcgataatcgttaataattcattgacaaagttgtgaagcaataattttta
caaaagtttttgaaattatctaaaatcgattgatctctggaacaccctatatatcttgttgtatgttctcg
TTAaATctatccCtatataTACATCTATCAGGATCAGCCcGCAATTAATTCcacacgCAtAtTTCATCCgtg
cgaatGTacacatttACACaaggattaccgccatgtccactgtctcgaattcttcgacaagcaacgatttt
tccaaagaggattggcgaacatttatacgCtCGAAATTGTCGCGaGAAAcGAGCGAgAAGTTTAATTGCtc
cGCgtACgAGCgCaCCaATtGAATCgTCTTtGTgcgACGATATTGTTTaAatacACaccttcgAAAGGTTtc
aTGCTTACaACTCGTTAAAagtgcagacatgtgcagaGAcaAAACggtgtatcgaTACAATgCGGTTCccCG
TTTCATTGTtccgaacatacaaacacattgttagatttctttccttttttttttattttttttttcatatttt
TCTttcTtTcTctttttttttagatttttttttTGTTTGcaATTTAATtCaACaTTCGAAtgGAAAATtttac
tacggtGAacagagcgGAAAAAGTTTaTAAATGGAGcGCgtAaATTTAAcgttCgcCgATTAAAtGAAAGTG
ttCGTTTggtcggttTTtttCAGaGgAATCGACaGgAAATTGtatGaATTAATcGCTCGaATTcgAAtgaCa
aATAATgtaaaaggattcatgtaatgtttttttattaaatttaatttaataaaaaaaaaggaaagacaagata
tttcagttttgcaacagaatgaatcattcaaactgaaatagaaatattttacaatatcgttttgttttgact
atttttgtttatttctcttttttttgtaatttttttcttttatttactttttttaaagtcatattttaaatgt
tttttgaaaaatatttttacgtatttttttattttttatttcactataaagttattaataatgatattgtaaaa
ttttatataatttatcatgtaaaacattagtattttaaattgcaattataacatatcaaaatatgtaaagta
tataaaagttttattagttttacatattaatattatgcaaattttatctatttattaagttagttttcattc
atctttccaatataattgtaaaaattatttatctttcttttactatttaacatcatgcaacaaataatttca
aatatttcttatcattataattctaaaaattatttaatgatctttttctttctaacaattttagaaaagaaa
ttatttatacaaatatttataacttttttattcttcaaaacaaataatttatgtcaaacttattcaaaagtat
catttttctcatttattaaatatatgttttgtttaaaacagttaacatttaaaatattgttttttcaatatt
atcaacaaattcaaagatattggaataataattaaaaaatattcgaaaaataattaaatagtctgtatatac
atatttcaatttacttataaatatatatatacatactcttactatatatacatacttatatattatttataa
atatgatctatttcgatcactattctaaaaaaattaaagaaaaataataaaaaaaaaaggaaaacaacaatc
taatttagaaatgaacaacaatctaacacaaacaattgaatacaaaaactcatattctttcctttcttttta
ttatttaatgaaaacgcatacacaaatatatttacaaattctactcacgtaatcaaccttgaaccgacgcat
tcgtacatgatataaatttcacaggcgcgtctcaTATTTCTtGCgGCctAaATAATATTAcaaCGCCTTTGT
TCcCCttcagtgagtaCTATATTTCcGTAAGCGTAACATATCCATCTTACTAATAGCGCtATGGTCAGACTG
CACGTAATTCGAGGtAGTCaccaacggttcccctgtggtatatatctcctgtATGTaCAtgtcgtatatgg
ccccgtaTACATACGcgGtaTATCcAtgTTGAATAGATCCGAacAGGCAgGGcATAGGTAGcAGCATCGAGT
GTgATACCGAAACTACATAACCtATCCAAGCCGGCTGACTTGGCCGACAATCGCTATTTAGgaagCCGCTTA
TCTCaGCCaGTCTTAGCTGatCTAATGTCTCttggCGAATtTGcTTtAaCacGTAGAtGcAATAACGTatac
gGAaTcttagatagaaatgttgtgtttagaacgaagcaagaaaaaggaaaactaaaagcgagaagggaaaga
ggaaaaaaaattgaagggataaggaggtaaaaagaaaatggcaaaggtgaagagaaagtcgtgaggggaatg
ggggtgagagtggtcgtgaaaatggaagaaatttggagcaaggatgataatacgaaaagagaaaagtgagaa
ataggatagagcaagaggaggagtgagtaatatcatagaaagaataaagaatcaaggaaaggattaaaaaag
atacaatttattggagagaaaaatatgaaaggattattgttaaaatatgtatagttttatgtcaattttatt
```

Figure 5



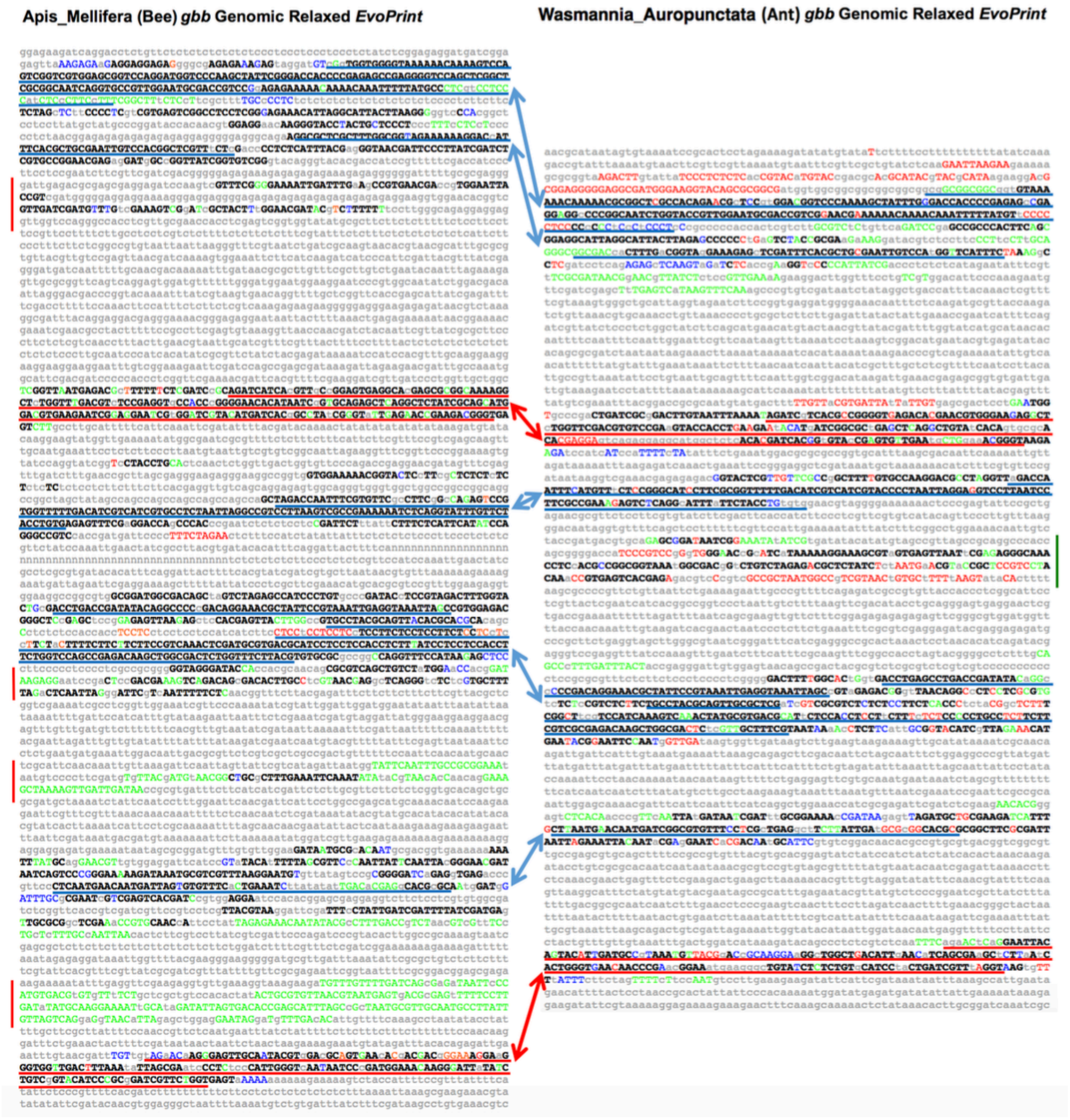
A

```
tggaatatgttgaagcaacgacggaatatttcgtatataattttgattcaaaattgaaatcgattgattaatttattataa
aatgaatgtctttgtcgaatgctttccttcatcttcaggaaatccttggttaattcggaaataaatagcttgccccgcggt
tccgcgtatattacacgcgtccctccccccgcggcgtatggatctcgtgGCGtAAAAATGAAAGAAGGCGcAAATCCTCG
cgggtggaccgaCGCgTTCcGCAAcgAAATTTCCTTtATGCCCcCgAgGGCTTTAAGTATTTCACCCGCAAGAACCtcggC
GCAGATTGGGGaTGCGCaggCTGAAATCCGCAATCCGTTTCGATCACGCgCGAACGATTTAGCTcCTTGCAATCCCGtgtt
GGAAGAtgaAAAACGAGAAGGAAAAATATGCAACggcatcggcggttctcaatggagagagagagaggagcactaggggca
cggaaagagaaagagggagagagagggagagagagaGAGAAgcgcgagtgacgccgcgGTTGAAAGCATAAgcatccaCCC
TTCGCACATTTTTATGCgGaGACGTCGCCTCTAAGCAGCTTTCTTCgAGATTTTAAACCAAAACATCCCTTGGGacccggg
gatggttcccagtatttttcccacttcccttttcacgcggaataaggttgcgaattgttgagaaacgttattacaatttt
cgtcgctctatcggcgataaacgcgaaacctttcacgtttcgattaatcttcttttaactcctccttggtcgtttcctttc
cttccattttaagaaaatagtaacgctttttaaagtgatctaacaggcatttgttttactcttatttgggaggctaagaa
tgacgtttaatgaaaaatccataagaaaagaatatttttttcatatctttttacgaaattctctaaatacgtaaaatatat
aacatacttcctttatgattctcttcctttcgtttcatcgatgtcttcgatcgaaacgaaatgagaaatgatatctgataa
gaatatagttcaactattcattatttccagaaatggaacgattattatcaaagagatggctcttttgataaaaatttttcg
tttcatcgaagtaaagattttaagaaatgggattaaTTTAAtattattaacgatagttTTtTcTcgttTTCAGCtGCAGCA
TAtCCtAgTTTaTAtTCCGGcGGtTACGTgGCaGcAATGGCAGCcGCCGGCGTTTCgCTtCACGGtCAaCCGGCcTTCTAt
CACGCgGCTCATCCcTAtCAAATAAATCCAAAcGCcacGGCtgtggggCCgATGGCgAAgAGGAAGCGTCGtCAcAGGACG
ATATTcACcGAgGAgCAgcTgGAACAAcTgGAAGCaACtTTCGACAAGACgCAcTAtCCaGAcGTaATttTGAGAGAACAg
CTcGCtCTtCAgGTCGATCTCAAGGAgGAacGAATCGAGGTGAGTttagacaaaataagttgtatgtttcaagatcttacg
acacgtgtattgcagtacatgattatgaaatttcagtttttgaatgtaagatttttctttttgaagaaatattccccact
agtgtttgtaatttggatttatttttgttcttcgtacttatattgatagtttttctatttttcaattaaatttcgttaaat
```

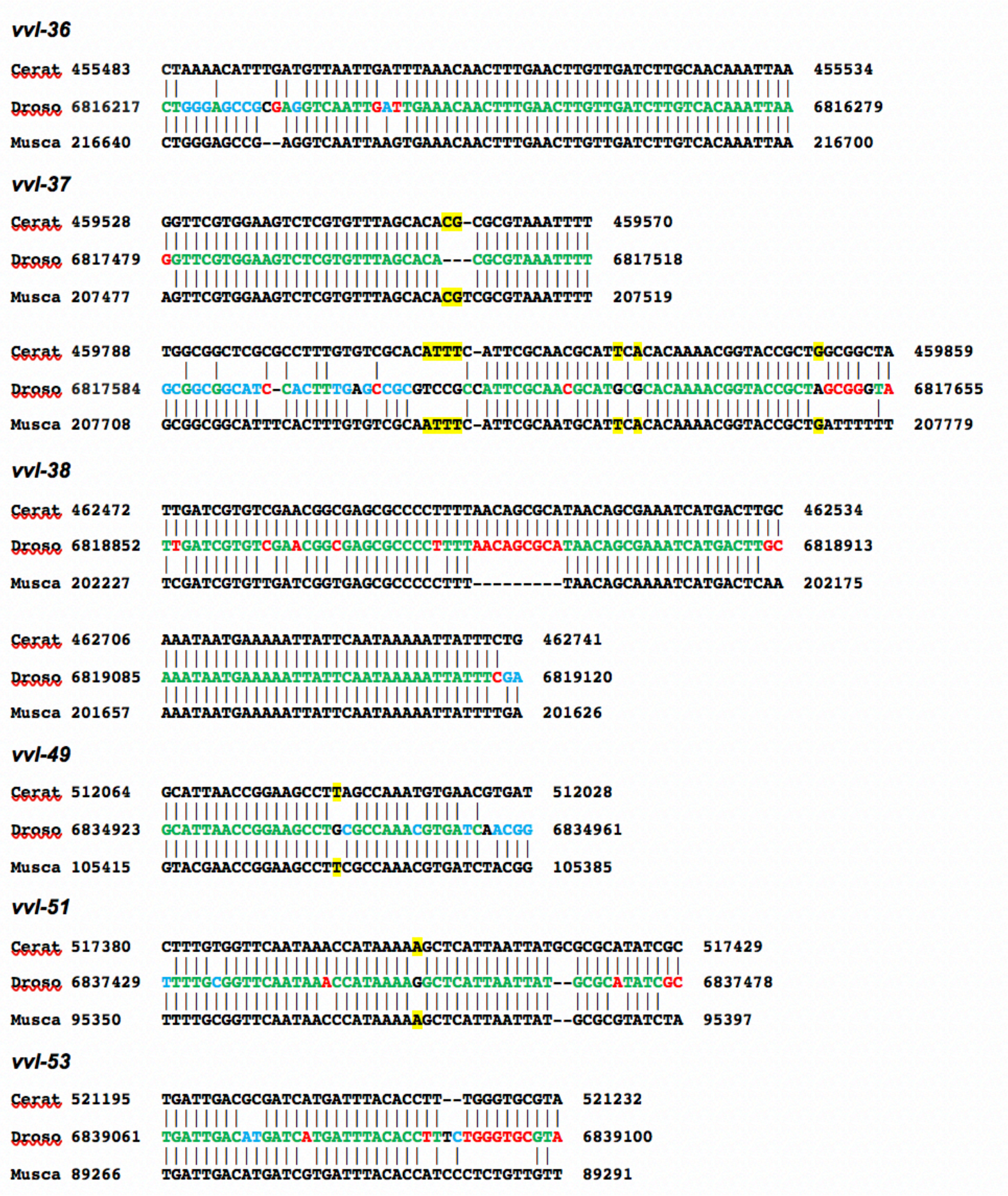
B

```
tgacaccggcgcttattattaaatcggccggagttactcgacttcgcgctggagaggcgatccgattcgcgtgagtgccga
gagaaaactttcgcggtttcgcgtagaaataatgtgaggcaaaaaggcactgtcttatccaaagatgtatatttcttcgtt
gtagcgagaGGGAGAATAATacgcgccgaaTTCTCCCttgccacaCtTTTTTTcccctctcggtaacactccgttcctctt
ttTcgcTTtCTCTCTTTCTTcCTTTCTTAGCctctctcccttctctttctctctctctctctctctctctctcttTCtC
TctttctccctgaagactaAActCTTTCTCTTTTTgcaccgggatgcgtATACGtTCTtGCAGAAaCgctgcACGGGCgGc
CAAtGTACGCggcgttccctccGtTcCTgACGGCACATATACCACgaaGAGCgGGGAGTCCCCGcGAGCCaCGAGAAATAT
TCGACAATTTTTAAACTCGTTCCTGAaCacgcgGGCATTCAAACgccGCGCaccagcctctcgcagtcagCAAGTAcGAAT
TTCTGAGAAtttcggccgCGTCCTGGTaGGGtGcACcgAAtgctggatgacgcgcgtgatagatgttggttgggattagaa
cactcgagcccgaggaggacactccgggtttcgaccgatgtgacacgtgatttgttgtcgttctgtgatcggccccgcgct
gcgtccgattaatcaacgagaagcgcagtcgcgatgcgcctttgttcgtaattcgctgatatggcccgactaatacatata
cgattatcctttcatgcatcaaatctctatagccactcaattgtgtagatttgcatttcttgggtgtacaatgtatcaatc
tagtgcccggcgggaaagtacgagtataggacggctcgtggaagagaggtttgccgcgtccaccggtattcgacatgcccg
AATAATAAAaTTTAAATACaTcctGGCAAGAAATGcAAATGcGacCGCTtGAGAGtGCTTTAACTGTaGAATtaTTCGTTT
AgGCAAAAAAAaagGGgggagaatacggcatggtgtatgtgtgcgtgtgtgtgcgtgCgCAGtTtaTTTCCATCTCTCGTG
tGCGCATTACCTCGGGAGATAATTATCGCgAGCGAGGATTGGTCAACCGAATTGCGTCTcGCGCTCAATATTTTCTTGCGT
GcCaCGGcCGTttCgAGtCTTTTAGgAGaGCtATGAACGCgCggagaggtaaatcgcttctaCGCGCCaccctctccaacc
tccgcatcGCCCGCGGCCACgCCaCGCTTTTTGCATAttGCGTATTGCGATCGGcCCcGATAGTTctttctttttttaag
cCgACCaTTCTTtcccgcTCCcgCgcgCTctccTTTTTTtcctccgtCCtCTctccTTTATtTCGcGGCgcgtatccgtgc
gcgcgcccatcgGgCAACCAGTTTTCGGtcgcgcgcgcgGGAGAGtgCCaTAGGTCGACGAACGgtgcGcTCaTAAcTCgC
GgTATGTGTATCAGaagtcattccggtatgcgccgttTCcTCTGgcgatGTGAACGCgaacccgtattatccagaccgata
gtatcgcccagaacACaCGGTCGGTAAAataAGAtGGTTgcCTgtgGAgGgacgtgtttatgccagaccgagccgagttca
tatcgaccgaatacccatctggctcgcacggaaagagacgaagacctcgcgcagatttctagggcagtttgtgatctactg
agatctgtcgagctttgaatttgaaatttgattttgatagcaactttagacatttatagtatttcacaaatatttcactt
attataacacgttcattgtacattaaatcaaaattatattataacgtaatctactttgaataagaaacggagaattgttgt
cgaaatgttcttgcttatttgatttgatttttgtgtaacatttcttcttgtatatttacaaatatttcacctattata
ttatacactagatcaaaattatattaacgtaatatctactttaagttagaaacggagaattgtcgttgaaatattcttgct
tatttaattttatttttcgtgtaatattttttttgtacgttgaattaatttaaatatgtagtcgattcaatttaatatataa
taagagtgtttctcgtctgtcaCAGCcTTGcTCAGGCAGAGtATGTCtCCGGTaGCgGAttCTcTgATGaGtGGTTCCAAa
acgtcgGTGTCcCAaCCgATCATcGATTAtTCcCGATAcGTGAGgaGaTataCgTCCaGcCAaGAATGCGGcAGTACGTAc
TGCAAgGAacTgGGtTGcAGgGAgCAcTTtCAtTGTcTgGAcTGtAGCGGcagggtgTTcGTGAAgAAggagGAaATGATC
AGGCACTTCAAGTGGCAtAAGAAgAGgGACGAgTCccTGCAGCAcGGcTTCATGAGgTAcTCGCCGAcGGACGAtTGtTCG
GAgAAgCAcCCcGGcCGgaCcTGTCCGCAcAAtaGAAAGCAgACcCAcTAcCAtTGCATcCACGAGAATTGCGACAAgGTg
TAtATaTCgACaTCGGACGTgCAgATGCACGCaAATTAcCAtCGcAAgGACTCgGCcATCATACAaGAGGGcTTCCAGCGA
TTTCGgGCcACcGAgAGTTGCGCGACtGAgcATTGTccgTTcGctGGaCAaCGgACcACGCACTTTCACTGTCGtCGAcCg
GgaTGTCGGTTCACgTTCAAAAATAAgGCCGATATGGGTgagAAgcgaactgtttgcgTGATtAaTGcTCGCtctagatcg
agactattcatcAATCgaagCCTTtcactagaaacttaatgtgtcgcaaacgtttcggatttaattcttacgctgagagaa
cgattacgttgatacatttttatttgatgtacatatctaattatatttgtcgttatttcagatctctccatgatttgctc
tacaagttttctttttatttttatttttcagacaagcacaagtcttatcacatcaaggacgaacaactgtcgcgcgacgg
```

Figure 6


Apis_Mellifera (Bee) gbb Genomic Relaxed EvoPrint
Wasmannia_Auropunctata (Ant) gbb Genomic Relaxed EvoPrint

Supplemental Figure 1



vvl-36

```
Cerat 455483     CTAAAACATTTGATGTTAATTGATTTAAACAACTTTGAACTTGTTGATCTTGCAACAAATTAA   455534
                 ||   |     || |||||||| ||||||||||||||||||||||||||||||||||||||||
Droso 6816217    CTGGGAGCCGCGAGGTCAATTGATTGAAACAACTTTGAACTTGTTGATCTTGTCACAAATTAA   6816279
                 ||||||||||  ||||||||| | |||||||||||||||||||||||||||||||||||||||
Musca 216640     CTGGGAGCCG--AGGTCAATTAAGTGAAACAACTTTGAACTTGTTGATCTTGTCACAAATTAA   216700
```

vvl-37

```
Cerat 459528     GGTTCGTGGAAGTCTCGTGTTTAGCACACG-CGCGTAAATTTT   459570
                 ||||||||||||||||||||||||||||   ||||||||||||
Droso 6817479    GGTTCGTGGAAGTCTCGTGTTTAGCACA---CGCGTAAATTTT   6817518
                  |||||||||||||||||||||||||||   ||||||||||||
Musca 207477     AGTTCGTGGAAGTCTCGTGTTTAGCACACGTCGCGTAAATTTT   207519

Cerat 459788     TGGCGGCTCGCGCCTTTGTGTCGCACATTTC-ATTCGCAACGCATTCACACAAAACGGTACCGCTGGCGGCTA   459859
                   |  |    | |  ||    |        | ||||||||||||| | ||||||||||||||||| |||| ||
Droso 6817584    GCGGCGGCATC-CACTTTGAGCCGCGTCCGCCATTCGCAACGCATGCGCACAAAACGGTACCGCTAGCGGGTA   6817655
                 ||||||||||  ||||||| | |||     | |||||||| |||| | |||||||||||||||||      |
Musca 207708     GCGGCGGCATTTCACTTTGTGTCGCAATTTC-ATTCGCAATGCATTCACACAAAACGGTACCGCTGATTTTTT   207779
```

vvl-38

```
Cerat 462472     TTGATCGTGTCGAACGGCGAGCGCCCCTTTTAACAGCGCATAACAGCGAAATCATGACTTGC   462534
                 ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Droso 6818852    TTGATCGTGTCGAACGGCGAGCGCCCCTTTTAACAGCGCATAACAGCGAAATCATGACTTGC   6818913
                 | |||||||| || ||| ||||||||| |||         ||||||||||||||||||||
Musca 202227     TCGATCGTGTTGATCGGTGAGCGCCCCCTTT---------TAACAGCAAAATCATGACTCAA   202175

Cerat 462706     AAATAATGAAAAATTATTCAATAAAAATTATTTCTG   462741
                 ||||||||||||||||||||||||||||||||||
Droso 6819085    AAATAATGAAAAATTATTCAATAAAAATTATTTCGA   6819120
                 ||||||||||||||||||||||||||||||||| ||
Musca 201657     AAATAATGAAAAATTATTCAATAAAAATTATTTTGA   201626
```

vvl-49

```
Cerat 512064     GCATTAACCGGAAGCCTTAGCCAAATGTGAACGTGAT   512028
                 |||||||||||||||||  |||||| |||| |
Droso 6834923    GCATTAACCGGAAGCCTGCGCCAAACGTGATCAACGG   6834961
                 ||||||||||||||||| |||||||||||||| ||||
Musca 105415     GTACGAACCGGAAGCCTTCGCCAAACGTGATCTACGG   105385
```

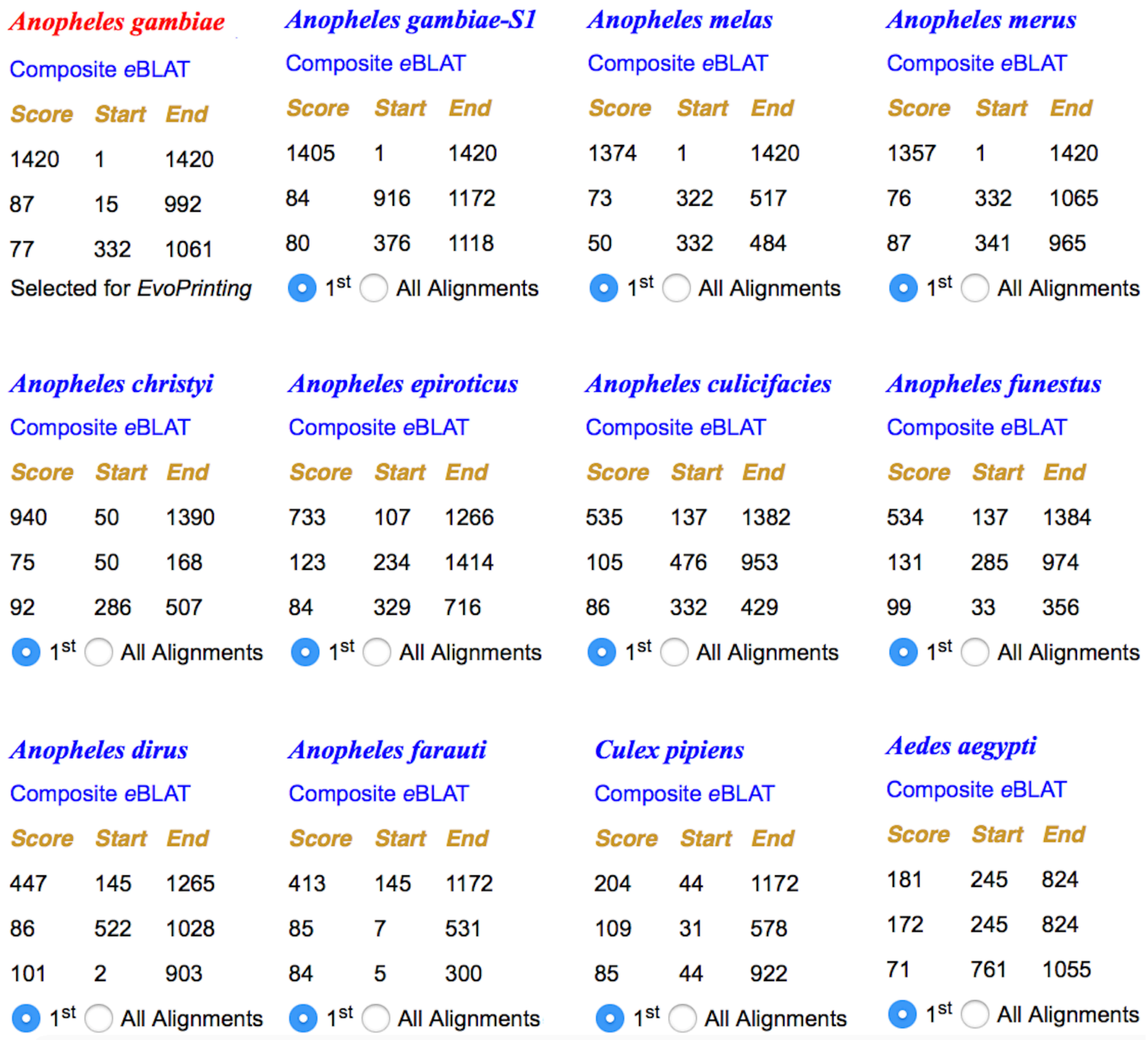
vvl-51

```
Cerat 517380     CTTTGTGGTTCAATAAACCATAAAAAGCTCATTAATTATGCGCGCATATCGC   517429
                  |||| ||||||||||||||||||| |||||||||||||  |||||||||||
Droso 6837429    TTTTGCGGTTCAATAAACCATAAAAGGCTCATTAATTAT--GCGCATATCGC   6837478
                 |||||||||||||||| |||||||| |||||||||||||  |||| ||||
Musca 95350      TTTTGCGGTTCAATAACCCATAAAAAGCTCATTAATTAT--GCGCGTATCTA   95397
```

vvl-53

```
Cerat 521195     TGATTGACGCGATCATGATTTACACCTT--TGGGTGCGTA   521232
                 ||||||||  ||||||||||||||||||  ||||||||||
Droso 6839061    TGATTGACATGATCATGATTTACACCTTTCTGGGTGCGTA   6839100
                 |||||||||||||| ||||||||||| | |       ||
Musca 89266      TGATTGACATGATCGTGATTTACACCATCCCTCTGTTGTT   89291
```

Supplemental Figure 2



Anopheles gambiae

Composite *e*BLAT

Score	Start	End
1420	1	1420
87	15	992
77	332	1061

Selected for *EvoPrinting*

Anopheles gambiae-S1

Composite *e*BLAT

Score	Start	End
1405	1	1420
84	916	1172
80	376	1118

● 1st ○ All Alignments

Anopheles melas

Composite *e*BLAT

Score	Start	End
1374	1	1420
73	322	517
50	332	484

● 1st ○ All Alignments

Anopheles merus

Composite *e*BLAT

Score	Start	End
1357	1	1420
76	332	1065
87	341	965

● 1st ○ All Alignments

Anopheles christyi

Composite *e*BLAT

Score	Start	End
940	50	1390
75	50	168
92	286	507

● 1st ○ All Alignments

Anopheles epiroticus

Composite *e*BLAT

Score	Start	End
733	107	1266
123	234	1414
84	329	716

● 1st ○ All Alignments

Anopheles culicifacies

Composite *e*BLAT

Score	Start	End
535	137	1382
105	476	953
86	332	429

● 1st ○ All Alignments

Anopheles funestus

Composite *e*BLAT

Score	Start	End
534	137	1384
131	285	974
99	33	356

● 1st ○ All Alignments

Anopheles dirus

Composite *e*BLAT

Score	Start	End
447	145	1265
86	522	1028
101	2	903

● 1st ○ All Alignments

Anopheles farauti

Composite *e*BLAT

Score	Start	End
413	145	1172
85	7	531
84	5	300

● 1st ○ All Alignments

Culex pipiens

Composite *e*BLAT

Score	Start	End
204	44	1172
109	31	578
85	44	922

● 1st ○ All Alignments

Aedes aegypti

Composite *e*BLAT

Score	Start	End
181	245	824
172	245	824
71	761	1055

● 1st ○ All Alignments

Supplemental Figure 3

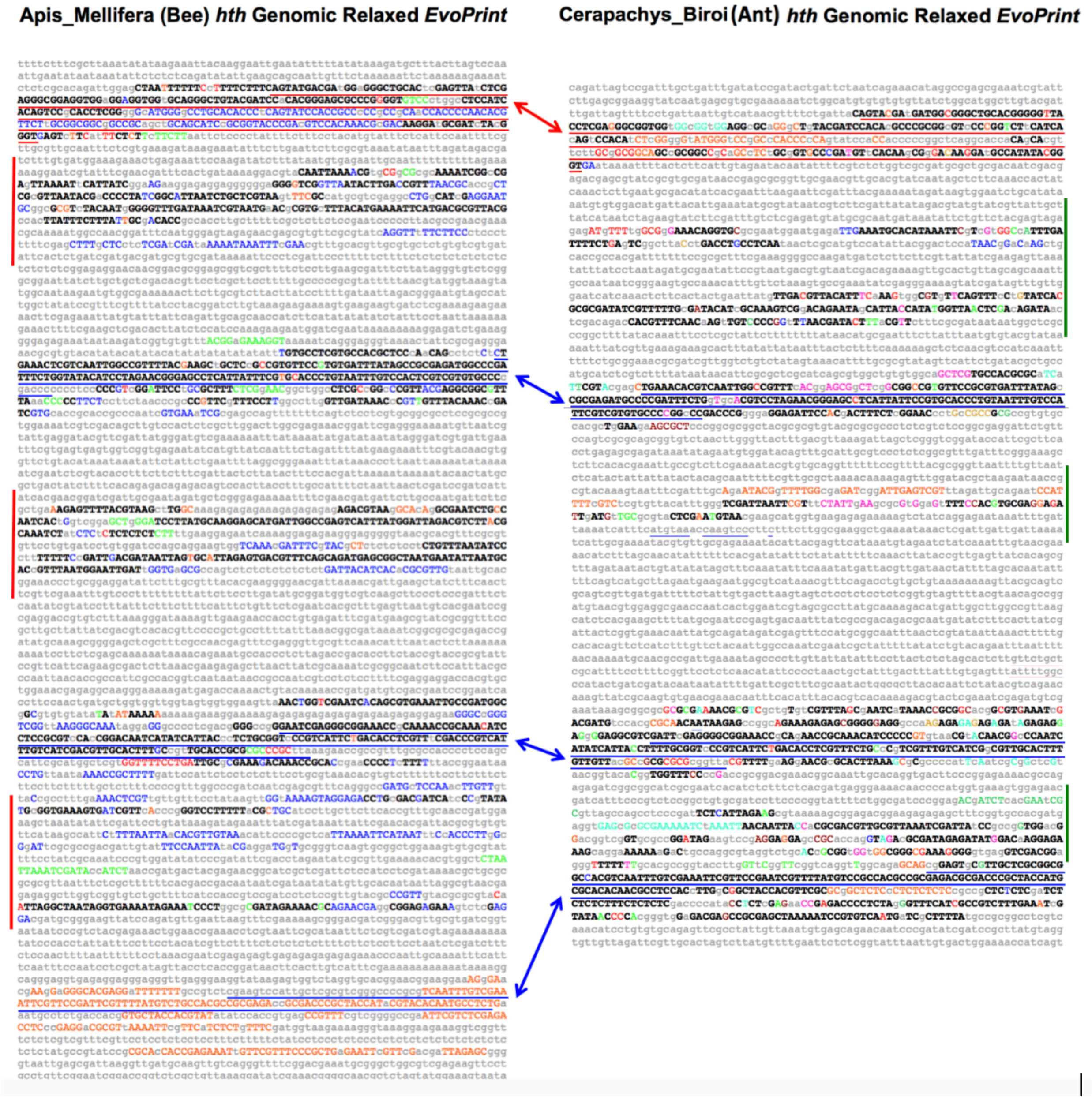

Apis_Mellifera (Bee) *hth* Genomic Relaxed *EvoPrint*
Cerapachys_Biroi (Ant) *hth* Genomic Relaxed *EvoPrint*

Figure Legends

Figure 1. Ultra-conserved sequences shared among a *Drosophila ventral veins lacking* enhancer and orthologous DNA within the *Ceratitis capitata* and *Musca domestica* genomes.

A) An *EvoPrint* of the *D. melanogaster vvl-41* neuroblast enhancer showing 1,775 bp, located 26.6 kb 3' of the *vvl* transcribed sequence. Capital letters represent bases in the *D. melanogaster* reference sequence that are conserved in *D. simulans, D. sechellia, D. yakuba, D. erecta, D. ananassae, D. persimilis, D. grimshawi, D. mojavensis and D. virilis* orthologous DNAs. Lower case grey bases that are not conserved in one or more of these species. Conserved sequence blocks (CSBs) shared with *Ceratitis* and *Musca*, as detected using BLASTn, DNA Block Aligner and the *EvoPrinter* CSB aligner are shown in Green text while red bases are shared between *D. melanogaster* and *Ceratitis* but not with *Musca*. **B**) Two and three-way alignments between of the ultra-conserved CSBs using BLASTn alignments. Green and red font annotations in the *Drosophila* CSBs are as describe above. Yellow highlighted bases in *Ceratitis* and *Musca* are not shared in *Drosophila*. Flanking BLASTn designator numbers indicate genomic sequence positions.

Figure 2. Ultra-conserved *Drosophila*, *Ceratitis capitata* and *Musca domestica* sequences within the *homothorax locus*.

A) A 1,065bp *EvoPrint* of the *D. melanogaster homothorax locus* that includes 5' non-transcribed sequence, its 5' UTR, the first five codons of its encoded protein and 102bp of the first intron. Capital letters represent bases in the *D. melanogaster* reference sequence that are conserved in *D. simulans, D. sechellia, D. yakuba, D. erecta, D. ananassae, D. persimilis, D. grimshawi, D. mojavensis and D. virilis* orthologous DNAs. Lower case grey letters represent bases that are not conserved in one or more than one of the test species. *Drosophila* CSBs that are shared with *Ceratitis* and *Musca*, as detected in figure 1, are shown in green. Red bases are shared only between *Drosophila* and *Ceratitis* and blue text represent bases shared exclusively between *Drosophlia* and *Musca*. The translation start codon is marked by an underlined ATG. **B**) BLASTn two and three-way alignments of the ultra-conserved CSBs. Font color annotations are as in panel A. Yellow highlighted bases in *Ceratitis* and *Musca* are not shared in *Drosophila*. Flanking BLASTn designator numbers indicate genome base positions.

Figure 3. *EvoPrint* analysis of the intragenic region adjacent to the *Anopheles Wnt-4* and *wingless* genes identifies ultra-conserved sequences shared with the evolutionary distant *Culex pipiens* and *Aedes aegypti* genomes.

A) *Anopheles gambiae* genomic *EvoPrint* that spans 1,420 bp, located 10.2 kb upstream of the *Wnt-4* gene and 27.5 kb upstream of the wingless gene which is transcribed in the opposite orientation of *Wnt-4* transcription. Capital letters (all font colors) represent bases conserved in all or all but one of the following *Anopheles* test species: *A. gambiae-S1, A. merus, A. melas, A. epiroticus, A. christyi, A. funestus, A. culicifacies, A. dirus or A. farauti.* Lower case grey letters represent bases that are not conserved in two or more of the Anopheles species included in the relaxed *EvoPrint*. Green uppercase bases indicate sequences are conserved in the *Anopheles* species, *Culex pipiens* and *Aedes aegypti,* blue font indicates *Anopheles* sequences that are shared only between *Culex pipiens* but not with *Aedes aegypti* and red font sequences are present only in *Anopheles* and *Culex*. **B**) To confirm the shared ultra-conserved CSBs, two and three-way BLASTn alignments of the shared sequences are shown. Color coding is as in panel A and yellow highlighted bases in the three-way alignments indicate identity between *Culex* and *Aedes* that is not present in *Anopheles*. Flanking BLASTn designator numbers indicate genome base positions.

Figure 4. Conserved sequence clusters within the honeybee *dscam2* gene second intron.

EvoPrinter analysis reveals *Apis mellifera* non-coding sequence elements that are conserved in other bee species or only in a subset of species. **A**) Alignment data generated from one-on-one comparisons of a 2.8 kb sequence from the honeybee 16 kb *dscam2* second intron. For each species, the top three independent *e*BLAT alignment scores are listed. Scores indicate the total number of bases within the reference sequence, the *Apis mellifera dscam2* intron, that align with the test species genome. The test species; *Apis dorsata*, *Apis florea*, *Bombus terrestrialis*, *Bombus impatiens*, *Habropoda laboriosa* and *Megachile rotundata* are listed (L -> R) based on their highest alignment score in descending order. Website links to individual *e*BLAT alignments and superimposed composite *e*BLATs are indicated in either red or blue font colors. As indicated in the alignment scorecard by the blue selection buttons, the top (highest scoring alignment) for each test species has been selected for *EvoPrinting*.

B) A color-coded relaxed *EvoPrint* of the 2.8 kb honeybee *dscam2* second intron generated from the alignment data shown in panel A. Black uppercase letters indicate bases conserved in all test species. Font colors represent sequences conserved in all species except for *Apis dorsata*, *Apis florea*, *Bombus terrestrialis*, *Bombus impatiens*, *Habropoda laboriosa* *or Megachile rotundata*. Gray lowercase nucleotides are not conserved in at least two of the test species.

Figure 5. Combined Ant and Bee *EvoPrints* identify ultra-conserved *Hymenoptera* DNA

A) An *Apis mellifera goosecoid (gsc) EvoPrint* generated with four evolutionary divergent bee genomes and then overlaid with a print that includes the four bee genomes plus four divergent ant species. The *Apis* honeybee *gsc* DNA (1,701 bp) includes 5' non-coding, the first exon and intron sequences. All uppercase bases (both black and red font) are conserved in bees and sequences that are conserved in both bees and ants are denoted with red-font uppercase bases. Lowercase gray-colored bases are not conserved in one or more of the bee test genomes. Bee test genomes: *Bombus terrestrialis, Bombus impatiens, Habropoda laboriosa* and *Megachile rotundata*. Ant test genomes: *Linepithema humile, Monomorium pharaonis, Wasmannia auropunctata and Atta cephalotes*. **B**) *EvoPrints* of the ant *Wasmannia auropunctata castor* (*cas*) gene locus. The 3,078 bp *Wasmannia* genomic DNA includes *cas* 5' non-coding, the first exon and flanking intron genomic sequences. The initial *Evoprint* was generated with four evolutionary divergent ants and then super-imposed with a print that included these four ants plus four bee genomes. All uppercase bases (both black and red font) are conserved in the ants *Cerapachys biroi, Linepithema humile, Atta cephalotes and Vollenhovia emeryi*. Sequences that is conserved in both ants and bees (*Apis florea, Bombus impatiens, Habropoda laboriosa and Megachile rotundata*) are shown as red colored uppercase bases. Lowercase gray-colored bases are not conserved in one or more of the ant test species. The translation initiation codon is underlined. The left flanking vertical brown bar indicates an ant-specific conserved DNA cluster that is not found in bees. Note, in the exon ORF most, but not all, of the conserved codons do not have conserved wobble positions indicating that the cumulative evolutionary divergence of the test species used to generate the *EvoPrint* afford near base pair resolution of essential DNA.

Figure 6. Side-by-Side comparison of conserved sequences within in the bee and ant *glass bottom boat loci* identify clusters of conserved and species-specific sequences.

A) Relaxed *EvoPrint* of *Apis mellifera* genomic DNA that includes the *glass bottom boat* (*gbb*) second and third exons (red underlined sequences) plus flanking intronic sequences (6.6 kb). Black uppercase bases are conserved in all test bee species and colored uppercase bases are conserved in all but one of the color-coded test species: *Bombus terrestrialis*, *Habropoda laboriosa, Megachile rotundata* and *Bombus impatiens*. First and second exons sequences underlined red. Blue underlined sequences are homologous to underlined sequences in panel B. Vertical red bars flanking the *EvoPrint* indicate conserved bee-specific sequences that are not found in ants. **B**) Relaxed *EvoPrint* of *Wasmannia auropunctata* DNA that spans the second and third exons of the *gbb* gene including their flanking intronic sequences (5.1 kb). As in panel A, black uppercase bases are conserved in all test ant species and colored uppercase bases are conserved in all but one of the color-coded species: *Cardiocondyla obscurior, Cerapachys biroi* and *Linepithema humile*. Red and blue underlined sequences are respectively homologous coding and non-coding sequences in panel A and the green vertical bar flanking the *EvoPrint* indicates ant-specific conserved sequences that are not found in bees.

Supplemental Figure 1. Ultra-conserved DNA in *Drosophila vvl* enhancers identified in *Ceratitis capitata* and *Musca domestica* orthologous DNAs.

Three-way *Ceratitis-Drosophila-Musca* BLASTn alignments of CSBs within six different *in vivo* tested *Drosophila vvl* enhancers. *Drosophila* sequences that are shared with *Ceratitis* and *Musca* are shown in green. Red bases are shared only between *Drosophila* and *Ceratitis* and blue text represent bases shared exclusively between *Drosophila* and *Musca*. Yellow highlighted *Ceratitis* and *Musca* bases are not shared in *Drosophila*. Flanking BLASTn designator numbers indicate genomic base positions.

Supplemental Figure 2. Conservation within the mosquito *wingless* gene second intron.

EvoPrinter analysis reveals *Anopheles gambiae* non-coding sequence elements located between the mosquito homologs of Drosophila *wg* and *wnt4* that are conserved in other mosquito species. Alignment data generated from one-on-one comparisons of a 1420 base sequence from the *A. gambiae* genome. For each species, the top three independent *e*BLAT alignment scores are listed. Scores indicate the total number of bases within the reference sequence that align with the test species genome. In this analysis, 11 of the 19 mosquito test species present in the database are illustrated. The test species are listed (L -> R) based on their highest alignment score in descending order. Website links to individual *e*BLAT alignments and superimposed composite *e*BLATs are indicated in either red or blue font colors. As indicated in the alignment scorecard by the blue selection buttons, the top (highest scoring alignment) for each test species has been selected for *EvoPrinting*.

Supplemental Figure 3. Side-by-side comparison of conserved sequences within ant and bee *homothorax* loci identifies shared exon/intron architecture and species-specific conserved sequences.

EvoPrints of bee and ant genomic DNA that includes *homothorax* (*hth)* encoding an exon isologous to the 2nd exon of *Drosophila hth* plus flanking intronic sequences. Blue and red underlined regions are coding and non-coding sequences, respectively, and align with homologous regions in the two panels. Black uppercase bases are conserved in all test species and colored uppercase bases are conserved in all but one of four bee tests species in panel A and all but one of three ant test species in panel B. **A)** Relaxed *EvoPrint* of *Apis mellifera* genomic sequences (6.3kb; Group5:7,111,526-7,117,900). Vertical red bars flanking the *EvoPrint* indicate conserved bee-specific sequences that are not found in ants. Colored uppercase bases are conserved in all but one of the color-coded test species: *Apis florea*, *Habropoda laboriosa*, *Bombus terrestrialis* and *Bombus impatiens*. **B**) Relaxed *EvoPrint* of *Cerapachys biroi* genomic DNA (5.1kb; 6532628-6527517, *Ooceraea biroi* isolate clonal line C1 chromosome 14, Obir_v5.4). The green vertical bar flanking the *EvoPrint* indicates ant-specific conserved sequence that in absent in bees. Black uppercase bases are conserved in all test ant species and colored uppercase bases are conserved in all but one of the color-coded test species: *Monomorium pharaonis, Atta cephalotes, Vollenhovia emeryi*, *Acromyrmex echinatior*, *Lasius niger*, *Pogonomyrmex barbatus*, *Wasmannia auropunctata*, *Cardiocondyla obscurior or Linepithema humile*.