Evolution of salivary glue genes in Drosophila species

Jean-Luc Da Lage[1*], Gregg W. C. Thomas[2], Magalie Bonneau[1] and Virginie Courtier-Orgogozo[3]

1 : UMR 9191 Évolution, Génomes, Comportement, Écologie. CNRS, IRD, Université Paris-Sud. Université Paris-Saclay. F-91198 Gif-sur-Yvette, France

2: Department of Biology and Department of Computer Science, Indiana University, Bloomington, IN 47405, USA

3: Institut Jacques Monod - CNRS UMR7592 - Université Paris Diderot, 15 rue Hélène Brion, 75013 Paris, France

*: corresponding author

jean-luc.da-lage@egce.cnrs-gif.fr

grthomas@indiana.edu

magalie.bonneau@egce.cnrs-gif.fr

virginie.courtier@ijm.fr

Abstract

Background: At the very end of the larval stage Drosophila regurgitate a glue secreted by their salivary glands to attach themselves to a substrate while pupariating. The glue is a mixture of apparently unrelated proteins, some of which are highly glycosylated and possess internal repeats. Because species adhere to distinct substrates (i.e. leaves, wood, rotten fruits), glue genes are expected to evolve rapidly. **Results:** We used available genome sequences and PCR-sequencing of regions of interest to investigate the glue genes in 20 *Drosophila* species. We discovered a new gene in addition to the seven glue genes annotated in *D. melanogaster.* We also identified a phase 1 intron at a conserved position present in five of the eight glue genes of *D. melanogaster*, suggesting a common origin for those glue genes. A slightly significant rate of gene turnover was inferred. Both the number of repeats and the repeat sequence were found to diverge rapidly, even between closely related species. We also detected high repeat number variation at the intrapopulation level in *D. melanogaster*. **Conclusion:** Most conspicuous signs of accelerated evolution are found in the repeat regions of several glue genes.

Keywords : Drosophila, glue, internal repeats, Sgs, pupa, adaptation, disordered protein, Eig71Ee, mucin, gene family, gene copy number, salivary gland

Background

Animals interact with their environment (viruses, bacteria, food, chemicals, conspecifics, etc.) in many different ways, particularly through their immune and sensory systems. As animals adapt to new places, the way they interact with their environment is expected to change. Accordingly, the gene families that have been shown to exhibit accelerated rates of gene gain and loss in several animal groups are mostly genes that mediate the interactions with the environment: immune defense, stress response, metabolism, cell signaling, reproduction and chemoreception [1]. Rapid changes in gene copy number can lead to fast phenotypic changes via gene deletion and can provide raw material for genes with new functions via gene duplication [2]; [3]. Rapid turnover of genes within a gene family has also been shown to correlate with fast evolution at the sequence level [4, 5].

We focus here on a gene functional group which mediates the physical interaction of the flies in the genus *Drosophila* with an external substrate during metamorphosis, the Salivary gland secretion (*Sgs*) genes. The *Sgs* genes encode proteins that make up the glue produced by Drosophila larvae that serves to attach the animal to a surface where it can undergo metamorphosis. In *D. melanogaster*, the glue is composed of several salivary gland secretion proteins which accumulate in the salivary glands of late third instar larvae [6]. As the puparium forms, the bloated salivary glands release their contents through the mouth. This secretion then hardens within seconds of contact with the air and becomes a glue which firmly attaches the pupa to the substrate. Metamorphosis is a critical stage of Drosophila development [7] during which the animal is vulnerable and motionless. In Drosophilids pupae are generally attached to a substrate, until the imago leaves the puparium. It is critical for the pupa to be firmly attached in order not to be moved away by some external event (i.e. rain or wind). Furthermore, for the emerging adult to be able to hold on the external substrate and thus get out of the pupal case, it is necessary for the pupa to adhere to a substrate, whether dry or wet. When the pupal case freely moves and is not attached, adults are unable to hatch and eventually die (J. R. David, personal communication).

Pupation sites of Drosophila species in nature have not been extensively characterized but a large diversity of pupation sites has been found. In the wild, *D. melanogaster* pupae have been observed on the dry parts of various rotten fruits, on leaves and on wood (J. R. David, personal communication, [8-10]). *D. mauritiana* pupae may be found on the surface of decaying *Pandanus* fruit, which is hard and lignous (D. Legrand, personal communication). Many Hawaiian Drosophila species pupate several inches deep in the soil [11]. Some other Drosophila species appear to pupate directly within the wet rotten part of fruits, such as *D.*

sechellia, *D. simulans* and the invasive species *D. suzukii* (J. David, personal communication, [12]).

Given the diversity of pupation sites, we hypothesized that *Sgs* genes might evolve rapidly among the Drosophila genus. The glue genes have long been an important model for the regulation of gene expression, through the study in the 1970-80s of puffs on polytene chromosomes, known to be the place of active transcription, some of which being induced by ecdysone (e.g. [13, 14], see also [15]). The protein content of salivary secretion was analyzed in the 70-80s and some major protein coding genes were located and found to correlate with the chromosomal location of major puffs ([16]). On an acid-urea electrophoresis gel the glue was resolved into five major bands, numbered from 1 to 5 in order of increasing electrophoretic mobility [16, 17]. Band 2, which was variable and detected in many other tissues, was considered to be a tissue contamination rather than a true glue protein [16]. Seven glue genes were eventually identified, and their nucleotide sequences are now well characterized: *Sgs1* (band 1, *CG3047*, 2L), *Sgs3* (band 3, *CG11720*, 3L), *Sgs4* (band 4, *CG12181*, X), *Sgs5* (band 5, *CG7596*, 3R), *Sgs7* (*CG18087*, 3L), and *Sgs8* (*CG6132*, 3L) and *Eig71Ee* (also named *geneVII I71-7* or *gp150, CG7604*, 3L) [14, 18-27]. *Eig71Ee*, located at position 71E, is not only expressed in salivary glands but also in hemocytes and in the gut, where it appears to be involved in immunity and clotting [28-30]. Its sequence is longer than *Sgs3*.

A sixth electrophoretic band migrating slightly slower than Sgs3 protein was detected in a few *D. melanogaster* lines [17, 31, 32]. The nucleotide sequence of the corresponding gene *Sgs6* remains unknown but cytogenetic and genetic mapping indicates that *Sgs6* is located in region 71C3-4 and differs from *Eig71Ee* [14, 28, 32]. The three genes *Sgs3*, *Sgs7* and *Sgs8* form a tightly linked cluster on the 3L chromosomal arm at position 68C [33, 34]. All glue genes were found to start with a signal peptide. The largest glue genes, *Sgs1*, *Sgs3* and *Sgs4* and *Eig71Ee* were shown to harbor numerous internal repeats of amino acid motifs, rich in proline, threonine and serine [19, 25, 29, 35]. Molecular studies showed that the number of internal repeats was variable between strains in Sgs3 [36], and Sgs4 [35]. In addition, consistent with missing protein bands, a few laboratory strains were inferred to carry loss-of-function mutations in *Sgs4* [6 , 16, 35, 37], *Sgs5* [27] and *Sgs6* [17, 31, 32].

In the present study, we characterize the diversity and evolution of *Sgs* genes within the Drosophila genus. We inferred loss and gain of glue genes and we investigated repeat number variation and sequence repeat diversity across species and across paralogs.

Results

We used the six *Sgs* genes and *Eig71Ee* annotated in *D. melanogaster* as BLAST queries to identify their putative homologues in 19 other Drosophila species (Table 1). The results are summarized in Figure 1 and Table 2. Note that for most species there is no transcript evidence, which may be due to the narrow time window of expression of the glue genes (late third larval instar) [6]. The organization of the *Sgs* genes was found to be generally conserved across the Drosophila species we investigated (Fig. 1). Proper identification of each ortholog was based on sequence similarity and, when possible, synteny. We describe below our findings for each category of *Sgs* genes.

Gains and losses of Sgs5 *genes*

We found that *Sgs5* had a tandem paralog in *D. melanogaster*, located ca. 300 bp uptream of *Sgs5* (*CG7587*, hereafter named *Sgs5bis*). It is co-expressed with *Sgs5* during late third larval instar in dissected salivary glands, as shown by expression profiles on Gbrowse at flybase.org and FlyAtlas at flymine.org. To our knowledge, this paralog has not been mentioned earlier. Both paralogs harbored two introns in all species. The *Sgs5/5bis* pair was widely distributed and therefore probably ancestral to most of the species studied. The occasional loss of either *Sgs5* or *Sgs5bis* occurred at least four times (loss of *Sgs5bis* in *D. mauritiana*, where a relictual sequence may still be recognized, *D. elegans* and *D. rhopaloa*, loss of *Sgs5* in *D. erecta*, Fig. 1), suggesting that *Sgs5* and *Sgs5bis* can replace each other functionally. In *D. ananassae,* the ortholog of *Sgs5bis* has been withdrawn from the genome annotation (formerly Dana\GF19880) although it is in conserved synteny relative to *D. melanogaster.* In *D. virilis* and *D. pseudoobscura*, a single *Sgs5*/*5bis* gene was identified. A tree of all Sgs5 and Sgs5bis amino acid sequences (Fig. 2) revealed a clear separation in two groups, *Sgs5* and *Sgs5bis*. The *D. virilis* gene annotated as *Sgs5* (Dvir\GJ24445) and the *D. pseudoobscura Sgs5/5bis* gene were clustered with the *Sgs5bis* genes and shared with most other *Sgs5bis* sequences a motif Gln-Ala-Thr in the signal peptide. This suggests that *D. virilis* and *D. pseudoobscura* possess an ortholog of *Sgs5bis*. If the *Sgs5-Sgs5bis* gene duplication arose after the separation of the *D. virilis* and *D. pseudoobscura* lineages, then the ancestral gene before the duplication was probably *Sgs5bis.*

Gains and losses of Sgs3, Sgs7, and Sgs8 *genes*

The genes *Sgs3*, *Sgs7* and *Sgs8* form a tight cluster, 4.5 kb long, on the 3L arm in *D. melanogaster* [33] and they share sequence similarities [19] in their N-terminal and C-terminal parts. *Sgs3* contains internal repeats whereas *Sgs7* and *Sgs8* have no internal repeats. When the internal repeats of *Sgs3* are excluded, the amino acid identity in *D. melanogaster* is 51.3 % between Sgs3 and Sgs7, 48.7 % between Sgs3 and Sgs8, and 46.7 % between Sgs7 and Sgs8. Additionally *Sgs3*, *Sgs7* and *Sgs8* share a phase 1 intron position, interrupting the signal peptide sequence [19]. In the clade *D. yakuba* / *santomea* / *erecta*, *Sgs7* and *Sgs8* are inverted with respect to the *D. melanogaster* arrangement (Fig. 1). In addition, *Sgs7* is duplicated in *D. yakuba* (Dyak\GE20214 and Dyak\GE21218) and *D. santomea* (Fig. 1). The two copies, inverted relative to each other, have only one, nonsynonymous, nucleotide difference. *Sgs8* lies between the two *Sgs7* copies, and has the same orientation as *Sgs3*. In species outside the *D. melanogaster* subgroup, all the *Sgs3*, *Sgs7* and *Sgs8* sequences also have the same intron, with slightly different positions depending on codon indels before the intron. Notably, *D. suzukii* is the only species that has lost *Sgs3*. *D. suzukii* retained *Sgs8* and underwent an amplification of *Sgs7*, three copies of which are identical. In a number of species, *Sgs7* and *Sgs8* could not be identified (Sgs7 and Sgs8 are small proteins, about 75 amino acids in length) : when BLAST search was performed using the *Sgs7* or *Sgs8* sequences of *D. melanogaster*, we retrieved the same target hits as with *Sgs3* (Table 2). In those species, several *Sgs3*-like genes were found instead, i.e. long proteins with internal repeats showing N-terminal and C-terminal parts similar to *Sgs3*. It is tempting to infer that in species where there are several *Sgs3*-like genes, but no *Sgs7* and *Sgs8*, especially in species like *D. pseudoobscura*, where *Sgs3*-like genes occupy the physical location of *Sgs7* and *Sgs8* (Fig. 1), this is because the ancestral *Sgs7* and *Sgs8* would have gained internal repeats. According to such a hypothesis, at least in some cases, the non-repeated parts of those Sgs3-like protein sequences are expected to cluster with Sgs7/8. To disentangle the relationships among *Sgs3* paralogs, and with *Sgs7* and *Sgs8*, we constructed a phylogeny made using an alignment of the non-repeated parts of the protein sequences. The tree (Fig. 3), which does not fit well to the species phylogeny, shows a clear separation between *Sgs3/Sgs3*-like and *Sgs7/Sgs8*, except for *D. bipectinata* and *D. willistoni*, whose *Sgs7/Sgs8* sequences are linked to the *Sgs3* branch, with low support. This would rather suggest that those *Sgs7/Sgs8* sequences are old *Sgs3*-like sequences which have lost their internal repeats. However, the sequence length is far too short to get a reliable tree and we cannot confirm this hypothesis. Likewise, whereas it is more parsimonious to infer that there were two ancestral *Sgs3* and that subsequent losses occurred, the tree topology is not accurate enough to confirm it.

Gains and losses of Sgs1 *genes*

Sgs1 was found only in the *melanogaster* subgroup and in the Oriental subgroups, which suggests that it originated in the ancestor of this clade. No *Sgs1* gene was detected in *D. erecta*, providing evidence for a loss of *Sgs1*. In *D. suzukii*, the putative *Sgs1* is very long, translating into a 2245 amino acid protein. The online sequence at the SpottedWingFly base showed a stop codon in the middle of the repeat region. However, based upon the surrounding repeat sequences, inserting a C at position 1829 (from start) would restore the reading frame. Our analysis of another genome sequence of *D. suzukii* [38] (contig CAKG01017146) showed that in this second strain a C is indeed present at position 1829 and that Sgs1 is 2245 amino acid long. In all the *Sgs1* genes identified, except in *D. elegans*, an intron was found at the same position and phase as in *Sgs3*, *Sgs7* and *Sgs8*. There is also a loose similarity in the N-terminal and C-terminal parts of Sgs1 and Sgs3 (in *D. melanogaster* about 14% identity between Sgs3 and Sgs1 excluding the repeats). This suggests that *Sgs1* belongs to the same family as *Sgs3/Sgs7/Sgs8* genes.

Origin of Sg4 *and* Eig71Ee *genes*

Sgs4, which is intronless, was absent outside the *D. melanogaster* subgroup (Fig. 1, Table 2). The origin of *Sgs4* is unknown. We found no similarity with any other sequence in any genome. Some sequence similarity between *Eig71Ee* and *Sgs4* had been reported [29], but is not convincing since it was in the repeat parts, which are of low complexity. *Eig71Ee* was found in all the *D. melanogaster* subgroup species and in some of the so-called Oriental species, where it has been annotated as *mucin2*, or *extensin* in *D. takahashii*, or even, erroneously, *Sgs3* in *D. suzukii*. We also detected the N-terminal parts of it in the *D. ananassae* group; thus making unclear the phylogenetic distribution of the gene (Table 2). More interestingly, we noticed that *Eig71Ee* harbors an intron at the same position as the one of *Sgs3, Sgs7, Sgs8* and *Sgs1*. This result argues for a certain relatedness among those genes. However, using Eig71Ee as a TBLASTN query did not retrieve any hits from any *Sgs* genes and the Eig71Ee amino acid sequence does not align with the Sgs sequences.

Rate of gene gains and losses in the glue gene families

Our analysis reveals that the seven annotated genes that code for glue proteins can be grouped into three gene families. *Sgs1, Sgs3, Sgs7, Sgs8,* and *Eig71Ee* comprise one of the three families since all of them share a phase 1 intron at the same position, interrupting the signal peptide sequence. *Sgs4* then forms its own family and the *Sgs5* and *5bis* comprise the third family. We used CAFE [39] to reconstruct ancestral copy numbers throughout the Drosophila phylogeny and to test whether these three gene families evolve at an accelerated rate along any Drosophila lineage. For the CAFE analysis *Eig71Ee* was not included due to uncertainties about its presence in some species. We find that the *Sgs4* and *Sgs5-5bis* families do not evolve faster compared to other gene families present in the Drosophila genomes (p=0.58 and p=0.107, respectively; Table S1), however the *Sgs1-3-7-8* family was found to evolve rapidly (p=0.005; Table S1). Overall, this family seems to be prone to duplication and loss (Fig. S1) and we find that this signal for rapid evolution is driven mostly by small changes on many lineages (i.e. a gain or loss of 1 gene) rather than large changes on one or a few particular lineage.

Characterization of the repeats in glue proteins

Table 3 summarizes the characteristics of the repeated sequences present within *Sgs* genes. Sgs1, Sgs3 and to a lesser extent, Sgs4 and Eig71Ee, are characterized, besides a signal peptide and a conserved C-terminal part, by long repeats often rich in threonine and prone to O-glycosylations. Although *D. melanogaster* Sgs5 protein is devoid of internal repeats (we checked that it is the case in all populations of the PopFly database), in most other species, even in close relatives, repeats are present, mostly pairs Pro-(Glu/Asp). Sgs5 protein length is highly variable across species. In *D. kikkawai*, there is a long additional stretch (127 amino acids) containing 60 % of acidic residues. The paralog Sgs5bis never has repeats. Sgs7 and Sgs8 are much smaller proteins, without any repeats, and are rich in cysteine (12-14 %). The conserved C-terminal parts are about 120 amino acids long in Sgs1, 50 amino acids in Sgs3, 120 amino acids in Sgs4, 115 amino acids in Sgs5/5bis and 135 amino acids in Eig71Ee. The repeats are quite variable in motif, length and number, even between closely related species, so that, most often, glue proteins may be retrieved only based on their conserved C-terminal part. The longest Sgs protein is Sgs1 of *D. suzukii* (2245 aa), which harbors ca. 63 repeats of a 29 amino acid, threonine-rich motif so that the total content of threonine is 40% ; in *D. melanogaster*, Sgs1 is also very long (1286 aa) due to 86 repeats of a motif of 10 amino acids, also threonine-rich (46%). The shortest Sgs1 protein is the one of *D. sechellia* (492 aa). In all the species where it exists, Sgs1 is also rich in proline (12-18%). Sgs3 has the same kind of amino acid composition. Repeats can also be quite different between paralogs. For example, in *D. eugracilis*, while the two genes are physically neighbors, Sgs3 has about seven repeats of CAP(T)$_9$, whereas Sgs3bis has ca. 65 KPT repeats. In *D. elegans*, the three Sgs3-like proteins also have quite different repeats (Table 3). Sgs4 is richer in proline than in threonine (18% vs. 16% in *D. melanogaster*) and contains 10% cysteine residues.

Interspecific variation in number and sequence of repeats

Between closely related species the number of repeats varied enormously and the repeated sequence diverged sometimes rapidly (Table 3). *D. simulans*, *D. sechellia and D. mauritiana* form a triad of sibling species, which split less than 300,000 years ago [40]. Their *Sgs1* genes harbor the same repeated sequence but the number of repeats ranges from 40 in *D. simulans* to 13 and 22 in *D. mauritiana* and *D. sechellia*, respectively. Sgs3 is very similar in the three species, except in the repeats. There are no repeats in *D. simulans,* but threonine-rich stretches; in the published sequence of *D. mauritiana*, there are three tandem occurrences of CAPPTRPPCTSP(T)$_n$; in *D. sechellia*, several CKP(T)$_6$ repeats. Sgs4 shows shared repeats C(D/N)TEPPT among these species, with many more repeats in *D. mauritiana*. In contrast, in the sibling species *D. yakuba* and *D. santomea*, which diverged 0.5 million years ago [41, 42], *Sgs3*, *Sgs4* and *Sgs5* harbor the same repeat sequences and the same number of repeats (Table 3). *Sgs4* genes show 91% identity at the protein level with the same 23 repeats; Sgs5 97% identity. Another pair of species worth of interest is *D. suzukii*/*D. biarmipes*, considered to have diverged ca. 7.3 mya [43]. As mentioned above, only Sgs1 and Sgs5 can be compared because *D. suzukii* has lost *Sgs3*, and *Sgs4* is limited to the *melanogaster* subgroup. Despite a longer divergence time than for the previous comparisons, the Sgs1 29 amino acid repeats are similar in the two species but *D. suzukii* has many more repeat units. In the non repeat parts, identity is 69.3 %; Sgs5 is well conserved even in the repeat region, with an overall identity of 76.4 % in amino acids, and 84.8 % in the non-repeat parts. A last pair of related species (despite their belonging to different subgroups) is *D. elegans*/*D. rhopalo*a. Their divergent time is unknown. We found that their Sgs proteins are very similar overall, including the repeat parts. This is less striking for the repeats in Sgs3, which exists as four gene copies in *D. rhopalo*a. Their Sgs5 shared a high overall identity (75%), with repeats (Glu-Pro)$_n$. In the non-repeat parts, identity rose to 82%. Indeed we often found more divergence among paralogs within a genome than across orthologous proteins.

Structure prediction programs (IUPred [44], PrDOS [45], disEMBL [46], PONDR [47]) indicate that the repeat regions of Sgs1, Sgs3, Sgs4, Sgs5 and Eig71Ee are intrinsically disordered (Fig. 4). Only IUPred and PrDOS indicate Sgs4 repeats to be ordered, in disagreement with the other predictors.

Intraspecific variation in number of repeats

Owing to the difficulty of short-read sequencing methods to deal with the repeated sequences found in glue genes, we could not get a species-wide insight of repeat number variation (RNV) in *D. melanogaster*. Therefore, we resequenced *Sgs3* and *Sgs4* in strains from various geographic locations using classical Sanger sequencing (Table 4). We found striking inter- and intrapopulation variation in the number of repeats : for *Sgs3* (Fig. S2 and S3, Table 4), there was at least 9 repeat difference between the shortest and the longest allele (22 to 31); for *Sgs4*, 18 to more than 26 repeats (Fig. S4 and S5, Table 4). Regarding the *Sgs4* data from the Drosophila Genome Nexus study (Cairo population), we observed that the repeat region was erroneously reconstituted, often underestimating the repeat number, compared to our Sanger sequencing. We also sequenced the *Sgs3* and *Sgs4* genes in wild-caught *D. mauritiana* individuals. For *Sgs3* we found variation in the number of stretched threonines (10 or 12) and in the number of repeats (Fig S6A and Table 4). For *Sgs4*, we found that the actual sequences were much longer than the sequence available online, and variable in length, even at the intra-population level, ranging from 25 to 35 repeats of the 7 amino acid motif (Fig. S6B and Table 4).

Nonsense mutations in the *Sgs* genes

Despite the rather low quality of sequences in the Drosophila Genome Nexus data set, we searched for putative premature stop codons (PSC) in *Sgs* genes of *D. melanogaster*, which could lead to non-functional proteins. The search was limited to non-repeat regions. We found PSC in *Sgs4* of several lines, that truncated the protein at the beginning of its conserved C-terminal part. We confirmed experimentally the presence of this PSC in 10 lines of the Cairo population EG (K165stop) (Fig. S5 and Table 4). We also found putative PSC for *Sgs5* in a few lines (W161stop, that is sub-terminal, and maybe not detrimental), and experimental verification confirmed it in one Ethiopian line (EF66N); in *Sgs5bis*, we found a putative PSC (C33stop) in six African lines from Rwanda (RG population) and Uganda (UG population). We also found a putative PSC for *Sgs1* in a few lines from USA and Cairo (P49stop), which was confirmed by resequencing the Egyptian line EG36N. This nonsense mutation required two substitutions from CCA to TAA in all cases. Interestingly, EG36N has also a truncated Sgs4. Therefore its glue should be investigated more carefully. In *Sgs3*, no PSC was found. Putative PSC were found for Eig71Ee in two lines, EA90N (S345stop) and RAL894 (W380stop), both in the C-terminal region. One putative PSC was found in *Sgs7* (Q47stop, line USI33), but was not checked experimentally. No PSC was found in *Sgs8* sequences. Stretches of Ns found in non-repeat regions could possibly, at least in some cases, turn out to be true deletions, which deserves further investigation.

Evolutionary rate of *Sgs* protein sequences

Given than glue proteins harbor RNV and given our hypothesis that they could be putative targets for fast selection, we wanted to test whether glue gene coding sequences evolve quickly. To this end, we computed substitution rates of *Sgs* genes between *D. melanogaster* and *D. simulans*, the genomes of which are well annotated. We did not include *Sgs3*, because the internal repeats were very different and not alignable between the two species. This, at any rate, shows that this particular gene evolved rapidly. However, we were able to make an estimate for *Sgs1*, although it had the biggest size and the highest number of repeats, because the repeats were rather similar in *D. melanogaster* and *D. simulans.* We removed the unalignable parts before computation, therefore underestimating the real evolutionary rate. We performed similarly for *Eig71Ee*, *Sgs4* and *Sgs5*. The results are shown on Table 5. We plotted dN and dN/dS for *Sgs* genes on the genome-wide distribution of dN and dN/dS between these species (Fig. 5) using the data of the flyDIVaS database [48]. All dN values were within the highest quartile, and *Sgs1*, *Sgs4* and *Sgs8* were within the highest three centiles. Furthermore, high dN/dS values were found for *Sgs1* (dN/dS=1.393) and *Sgs8* (dN/dS=1.259), indicating accelerated protein evolution. The dN value of *Sgs8* (0.1789) contrasts with the one of its close relative *Sgs7* (0.0475). We wondered if *Sgs8* had also evolved faster than *Sgs7* in other pairs of related species. Table 6 shows the results for other species pairs known to be close relatives : *D. melanogaster/D. sechellia*, *D. simulans/D. sechellia* ; *D. yakuba/D. erecta* ; *D. biarmipes/D. suzukii*. Whereas the latter two pairs showed no evolutionary rate difference between *Sgs7* and *Sgs8*, comparing *D. simulans* and *D. sechellia* showed a ten times higher dN for *Sgs7* relative to *Sgs8*, a situation opposite to *D. simulans* vs. *D. melanogaster*. In fact, *D. sechellia Sgs7* is more divergent than *D. simulans* from *D. melanogaster Sgs7*, whereas *Sgs8* did not diverged further. Obviously, the small number of substitutions points to a high variance, and the difference may be not significant.

To test for adaptive evolution after the “out of Africa” event of *D. melanogaster* [49], we measured the nucleotide diversity π and divergence D_{xy} between one population from Zambia, (ZI) thought to be within the original geographical area of *D. melanogaster*, another African population (EF, Ethiopia) and two derived populations, from France (FR) and USA (Raleigh, RAL). This study was limited to the coding sequences of *Sgs5* and *Sgs5bis*, due to the absence of internal repeats and to the gene size, not too short, (*Sgs7* and *Sgs8* were too short). Due to the numerous residual unidentified nucleotides in the Drosophila Genome Nexus data, the number of sites taken into account could be much smaller than the sequence size, e.g. for *Sgs5bis*, 278 sites left over 489 in RAL. We compared the overall π and D_{xy} between these populations [50]. The results are shown on Table 7. Roughly, for both genes π is higher in ZI than in EF, FR and RAL, as for the whole genome and as expected for the region of origin of this species, but divergences D_{xy} are less than expected from the whole genome, except for the ZI/EF comparison of *Sgs5*. Both genes gave similar results. Therefore, the glue genes *Sgs5* and *Sgs5bis* do not show particular divergence across populations, which could have been related to a change in population environment.

We also searched for episodic diversifying selection (EDS) among species for the three genes entirely devoid of repeats, *Sgs5bis*, *Sgs7* and *Sgs8*. The branch-site REL test of the HyPhy package was used. No accelerated evolution was detected for *Sgs5bis*, whereas one branch (*D. santomea-D. yakuba* clade) underwent EDS for *Sgs7* (corrected p-value 0.012) and one branch (*D. erecta-D. yakuba-D. santomea)* underwent EDS for *Sgs8* (corrected p-value 0.015) (Fig 6). These results must be considered with caution given the small size of the data set, but anyway do not favor a specific selection regime, regarding single nucleotide (or amino acid) polymorphism.

Discussion and conclusion

We have investigated the presence and characteristics of *Sgs* genes and proteins in several Drosophila species belonging to the two main subgenera *Sophophora* and *Drosophila*, with particular emphasis on species closer to *D. melanogaster*. We have identified the various *Sgs* genes through sequence similarity with *D. melanogaster*. Therefore it is possible that we may have missed glue genes completely different from the ones of *D. melanogaster*. Clearly, getting the full collection requires transcriptional evidence from late larval salivary gland RNA for each species studied. Interestingly, according to our census, the seven genes characterized for years in *D. melanogaster* are far from being always present in the other genomes, although the seven members are generally preserved in the *D. melanogaster* subgroup. Our results are in disagreement with the succinct interspecific study of Farkaš [15]. We also propose here a eighth member, *Sgs5bis*, a tandem paralog of *Sgs5*, based on its close sequence homology and its co-expression with *Sgs5*. We notice that *Sgs5bis* never contains internal repeats whereas *Sgs5* often harbors more or less developped repeat motifs, although not in *D. melanogaster*. Given our data, and notwithstanding the unbalanced taxonomic sampling which may mislead us, we suggest that the ancestor of the species studied here had only *Sgs3* and *Sgs5bis* (Fig 1). It is likely that *Sgs7*, *Sgs8*, and maybe also *Sgs1* and *Eig71Ee*, originated from duplications of *Sgs3*. The important differences in repeat motifs between duplicate *Sgs3* (e.g. in *D. eugracilis*) are striking and suggest a high rate of evolution, or independent acquisition of repeats from a repeatless or repeat-poor parental gene. A part of the sequence we named *Sgs3-like* in *D. willistoni* is reported in FlyBase as GK28127, with transcription on the opposite strand, and without an homologue in *D. melanogaster*. Thus, it is not impossible that some duplicates of *Sgs3* may have been actually recruited for other functions, different from making the glue. In this respect, it is also possible that Eig71Ee, which has been studied mostly for its immune functions, could be an ancient glue protein, which gained new functions.

The repeat-containing glue proteins are typical of secreted mucins. Mucins are highly glycosylated proteins found in animal mucus and they protect epithelia from physical damage and pathogens [51]. In *D. melanogaster*, more than 30 mucin-like proteins have been identified [52] but the precise function of most of them remain unknown. It would be interesting to compare the glue genes with the other mucin-like genes in terms of protein domains and sequence evolution. In *D. melanogaster*, repeats similar to those of Sgs3 (KPTT) are found in the mucin gene *Muc12Ea*. The high level of glycosylation is thought to favor solubility at high concentration while accumulating in salivary glands ([15]). The richness in cysteins suggests that, upon release in the environment through regurgitation, disulfide bridges between glue proteins may be formed by cystein oxidation by air, making a complex fibrous matrix. Intramolecular disulfide bonds can also be predicted ([15]). Examination of the amino acid composition of the glue proteins suggests that the numerous prolines may induce a zigzag-like shape ; serine and threonine, which are very abundant, besides being prone to O-glycosylation, make them very hydrophilic and favor interaction with the solvent and then solubility while preventing folding. The presence of regularly scattered arginines or lysines (or sometimes aspartic and glutamic acids) would add charge repulsion, helping the thread structure to be maintained flat and extended. This is similar to linkers found between mobile domains in some proteins [53]. The shorter Sgs7/Sgs8 would, considering their richness in cystein, bind the threads together through disulfide bonding.

In the frame of an intrinsically disordered structure (Fig. 4), it is not surprising to observe a high level of repeat number variation (RNV) even at the intra-population level. It has been reported ([54, 55]) that in proteins with internal domain or motif repeats, if these repeats form disordered regions and do not interact with the rest of the protein chain (for a cooperative folding for example), they are more prone to indels which are better tolerated, and favored by the genetic instability of repeated sequences. It is likely that, within a certain repeat number range, variations in repeat numbers might have little effect on the chemical and mechanical properties of the glue. In fact it is likely that the differences in repeat motif sequences rather than the number of repeats would change the mechanical and physical properties of the glue. Accordingly, we measured rather fast rates of evolution, but found no clear indication of positive selection. One reason why the evolution of the repeats is fast (across related species or across paralogs) might be that the constraints to maintain disorder and the thread-like shape are rather loose ([54])

We do not know the respective roles of the different Sgs proteins in the final glue. Farkaš (2016) mentioned that Sgs1 could have chitin-binding properties, which is in line with the function of the glue. He also proposed roles of specific components before regurgitation, inside salivary gland granules, related to packaging, solubility... The absence of some glue components may have consequences on its properties and may play a role in adaptation, as suggested by [15]. Gene loss, gene duplication, or repeat sequence change may modify the strength of the glue or its resistance to water or moisture, to acidity (of a fruit) and therefore might be linked to pupation site preference. For instance, *D. suzukii* lacks several glue components. In contrast to its closely related species which prefer rather dry pupation sites, *D. suzukii* animals pupate within ripe and wet fruits such as cherries or raspberries, the pupa half protruding. An efficient and strong glue might not be necessary within the wet medium of a ripe fruit. Shivanna et al. ([56]) have related pupation site preference to the quantity of glue and, counter-intuitively, have reported that species that prefer to pupate on the food medium in the laboratory produce more glue than species that pupate on the glass walls of the vials. However, the chemical glue content was not investigated. [57] compared pupation site preferences between the sibling species *D. mauritiana*, *D. sechellia* and *D. simulans*. While *D. simulans* populations from the native region share pupation preference in fruits with *D. mauritiana* and *D. sechellia*, worldwide populations preferably pupariate off-fruit, i.e. on a drier and harder substrate. Although the QTL associated with pupation site preference in *D. simulans* and *D. sechellia* do not map to glue genes [57], it would be interesting to see whether, secondarily, significant variations in glue composition or quantity occurred and might be contrasted across *D. simulans* populations. Given its worldwide expansion associated with adaptation to multiple local environments including diverse pupation sites, *D.*

melanogaster is an interesting model to study the intraspecific evolution of *Sgs* genes in relation to adaptation. Interestingly, absence of Sgs4 protein was reported in a few strains from Japan and USA [35], most likely due to deletions or mutations in the promoter region. Our resequencing of a few Nexus lines revealed nonsense mutations within the coding sequence at position 165 in *Sgs4*, deleting the well conserved C-terminal part. The consequences for final glue properties remains unknown.

In conclusion, the pupal glue appears as a genetically and phenotypically simple model system for investigating the genetic basis of adaptation. The present work provides a first exploration of the evolution of glue genes across *Drosophila* species and paves the way for future studies on the functional and adaptive consequences of glue composition variation in relation to habitat and geographic and climatic origin.

Methods

Identification of Sgs *genes in Drosophila species*

The seven annotated glue genes of *D. melanogaster (Sgs1* (CG3047) *; Sgs3* (CG11720) *; Sgs4* (CG12181) *; Sgs5* (CG7596) *; Sgs7* (CG18087) *; Sgs8* (CG6132)) and *Eig71Ee* (CG7604) were used as BLAST queries for retrieving their orthologs in 19 other Drosophila species. The genome data used for each species is indicated in Table 1. BLAST searches were performed directly through GenBank, FlyBase [58], the SpottedWingFly base for *D. suzukii* [59] or using local BLAST program (v2.2.25) after downloading the genomes for *D. santomea* [60] and *D. mauritiana* [61]. The BLASTP and TBLASTN programs were used [62], without filtering for low complexity, which otherwise would have missed the repeated regions. Repeats, when present, were often quite different from the repeats present in *D. melanogaster Sgs* sequences. Consequently, BLAST results were often limited to the C-terminal part of the targeted gene, which was the most conserved part of the proteins, and to a lesser extent to the N-terminal end. For each species, a nucleotide sequence containing large regions upstream and downstream of the BLAST hits was downloaded from InsectBase [63] or from species-specific websites when genome data was not present in InsectBase (Table 1). We used Geneious (Biomatters Ltd.) to identify by eye the coding regions, the start of which was identified by the signal peptide sequence. Putative introns were also identified manually, guided by the intron-exon structure of the *D. melanogaster* orthologs. In cases of uncertainties or missing sequence data, we extracted DNA from single flies of the relevant species (Table 4) and the questionable gene regions were amplified with primers chosen in the reliable sequence parts (Table S2), and sequenced by the Sanger method using an ABI 3130 sequencer. For instance, we characterized the exact sequence corresponding to N stretches in the published sequence of *D. mauritiana Sgs4* ; we found that the published premature stop codon (PSC) of *D. biarmipes Sgs3* was an error and that three frameshifts found within 50 bp in *D. sechellia Sgs1* were erroneous.

Evolutionary relationships between genes and estimate of evolutionary rates

Alignments of DNA or protein sequences were done using MUSCLE [64] implemented in Geneious and protein trees were computed using PhyML, as implemented in the online server Phylogeny.fr [65] and drawn using iTOL [66]. The substitution rates dN and dS values for over 10,000 coding sequences computed for *D. melanogaster/D. simulans* comparisons were retrieved from the flyDIVaS database [48] but *Sgs* genes were not included in this dataset. Thus, dN and dS were computed using yn00 in the PAML package ([67]), removing the unalignable parts. We tested for episodic diversifying selection across species using the branch-site random effect likelihood (BS-REL) algorithm implemented in the HyPhy package [68, 69] at the Datamonkey website (classic.datamonkey.org) [70]. We used only genes devoid of repeats to ensure reliable aligments, and we supplied species trees for the analysis.

Test for accelerated gene turnover

To infer ancestral gene counts in the three newly classified *Sgs* gene families and to determine whether the three newly classified *Sgs* gene families are evolving rapidly we first need to determine the average rate of gene gain and loss (λ) throughout *Drosophila*. Previous studies have estimated λ from 12 *Drosophila* genomes and found rates of 0.0012 gain/losses per million years [4] and 0.006 gains/losses per million years after correcting for assembly and annotation errors [39]. However, since those studies numerous additional *Drosophila* genomes have been published. In order to update the gene gain/loss rate (λ) for this genus, we obtained 25 available *Drosophila* peptide gene annotations from NCBI and FlyBase. The latest versions at the time of study for the genomes of the original 12 sequenced species (*ananassae* v1.05*, erecta* v1.05*, grimshawi* v1.3*, melanogaster* v6.10*, mojavensis* v1.04*, persimilis* v1.3*, pseudoobscura* v3.04*, sechellia* v1.3*, simulans* v2.02*, virilis* v1.06*, willistoni* v1.05*i,* and *yakuba* v1.05) were downloaded from FlyBase [71] and 13 other species (*arizonae, biarmipes, bipectinata, busckii, elegans, eugracilis, ficusphila, kikkawai, miranda, navojoa, rhopaloa, suzukii,* and *takahashii*) were downloaded from NCBI [72].

To ensure that each gene from the 25 *Drosophila* species was counted only once in our gene family analysis, we used only the longest isoform of each protein in each species. We then performed an all-vs-all BLAST search [73] on these filtered sequences. The resulting e-values from the search were used as the main clustering criterion for the MCL (Markov cluster algorithm) program to group peptides into gene families [74].This resulted in 17,330 clusters. We then removed all clusters not present in the *Drosophila* ancestor, resulting in 9,379 gene families. An ultrametric phylogeny with branch lengths in millions of years was inferred using MCL orthogroups in a similar fashion, with the addition of the genome of the house fly, *Musca domestica*, as an outgroup and utilizing single-copy orthogroups between all 26 species [75].

With the gene family data and ultrametric phylogeny as input, we estimated gene gain and loss rates (λ) with CAFE v3.0 [4]. This version of CAFE is able to estimate the amount of assembly and annotation error (ε) present in the input data using a distribution across the observed gene family counts and a pseudo-likelihood search. CAFE is then able to correct for this error and obtain a more accurate estimate of λ. We find an ε of about 0.04, which implies that 4% of gene families have observed counts that are not equal to their true counts. After correcting for this error rate, we find $\lambda = 0.0034$. This value for ε is on par with those previously reported for *Drosophila* (Table S3; [39]). However, this λ estimate is much higher than the previous reported from 12 *Drosophila* species (Table S3; [4, 39]), indicating a much higher rate of error distributed in such a way that CAFE was unable to correct for it, or a much higher rate of gene family evolution across Drosophila than previously estimated. The 25 species *Drosophila* phylogeny was then manually pruned and modified to represent the 20 *Drosophila* species in which *Sgs* gene families have been annotated. Some *Sgs* gene families are not present in the ancestor of all 20 species, so additional pruning was done to the phylogeny for each family as necessary (see Table S1). The phylogeny, *Sgs* gene copy numbers, and the updated rate of gene gain/loss ($\lambda = 0.0034$) were then used by CAFE to infer p-values in each lineage of each family (Table S4). Low p-values (< 0.01) may indicate a greater extent of gene family change along a lineage than is expected with the given λ value, and therefore may represent rapid evolution.

Search for polymorphism and repeat number variation in D. melanogaster *and* D. mauritiana

Polymorphism in *D. melanogaster* was investigated in the coding regions, especially the repeat number variation (RNV). We intended to use the data from the Drosophila Genome Nexus study ([50, 76], available at the Popfly web site [77]) to assess RNV. This database contains resequenced and aligned genomes of hundreds of *D. melanogaster* lines from about 30 populations from all over the world. Those data, like most *D. melanogaster* populations' and other species' genomes were obtained using NGS technologies, which yielded short reads. The data were often not accurate in repeat regions, likely because short reads may be not properly assembled when there are numerous short tandem repeats, and thus could not be used for counting RNV. Thus, experimentally, using single-fly DNAs, we amplified and sequenced the repeat-containing *Sgs3* and *Sgs4* from one or a few individual flies from several strains or natural populations available at the laboratory (French Guyana, Ethiopia, France, Benin, Ivory Coast, India, Comores, and the laboratory strain Canton S), and from a number of lines used in the Drosophila Genome Nexus study (Table 4). In addition, we investigated the occurrence of possible premature stop codons in gene alignments from the Drosophila Nexus database [50, 76], available at the Popfly web site [77] and checked the results by PCR in *Sgs4* and *Sgs5* (Table 4). We also used data from the Drosophila Nexus database to study polymorphism and divergence in *Sgs5* and *Sgs5bis*, which are devoid of repeats, and are not too short. Four populations represented by numerous lines were retained for analysis : ZI (Siavonga, Zambia), for the ancestral geographical range, EF (Fiche, Ethiopia), which shows overall rather large differentiation (Fst) with most other populations [50], and FR (France) and RAL (Raleigh, USA) for the worldwide populations. Diversity and divergence indices were computed with DnaSP [78]. Experimental sequences were deposited to GenBank with accessions MH019984-MH020055.

Ethics approval and consent to participate :

Not applicable

Consent for publication :

Not applicable

Availability of data and material :

Available upon request to the authors

Competing interests :

The authors declare that they have no competing interest

Funding :

The research leading to this paper has received funding from the CNRS to JLDL, MB and VCO and from the European Research Council under the European Community's Seventh

Framework Program (FP7/2007-2013 Grant Agreement no. 337579) to VCO. GWCT is supported by NSF DBI-1564611.

Authors' contributions :

VCO and JLDL designed the study and analyzed data; JLDL and MB performed experimental work; GWCT performed CAFE analysis; JLDL, VCO and GWCT wrote the manuscript.

Acknowledgments :

The authors thank Dr Georges Feller for comments on the disordered protein regions, and Dr Amir Yassin for critical reading of the manuscript.

References :

1. Demuth JP, Hahn MW: **The life and death of gene families**. *Bioessays* 2009, **31**(1):29-39.
2. Sánchez-Gracia A, Vieira FG, Rozas J: **Molecular evolution of the major chemosensory gene families in insects**. *Heredity* 2009, **103**(3):208-216.
3. Innan H, Kondrashov F: **The evolution of gene duplications: classifying and distinguishing between models**. *Nature Rev Genet*, **11**(2):97-108.
4. Hahn MW, Han MV, Han S-G: **Gene family evolution across 12 Drosophila genomes**. *PLoS Genetics* 2007, **3**(11):e197.doi:110.1371/journal.pgen.0030197.
5. Chen FC, Chen CJ, Li WH, Chuang TJ: **Gene family size conservation is a good indicator of evolutionary rates**. *Mol Biol Evol* 2010, **27**(8):1750-1758.
6. Beckendorf SK, Kafatos F: **Differentiation in the salivary glands of Drosophila melanogaster: characterization of the glue proteins and their developmental appearance**. *Cell* 1976, **9**:365-373.
7. Sameoto DD, Miller RS: **Selection of pupation site by *Drosophila melanogaster* and *D. simulans***. *Ecology* 1968, **49**:177-180.
8. Sokolowski MB: **Genetics and ecology of *Drosophila melanogaster* larval foraging and pupation behavior**. *J Insect Physiol* 1985, **31**:857-864.
9. Beltrami M, Medina-Munoz MC, Arce D, Godoy-Herrera R: **Drosophila pupation behavior in wild**. *Evolutionary ecology* 2010, **24**:347-358.
10. Del Pino F, Jara C, Godoy-Herrera R: **The neuro-ecology of Drosophila pupation behavior**. *PLoS one* 2014, **17**(9(7)):e102159.
11. Grossfield J: **Non-sexual behavior of Drosophila**. In: *The genetics and biology of Drosophila.* Edited by Ashburner M, Wright TRF, vol. 2b. London

New York

San Francisco: Academic Press; 1978: 3-126.

12. Vandal NB, Siddalingamurthy GS, Shivanna N: **Larval pupation site preference on fruit in different species of Drosophila**. *Entomological Research* 2008, **38**:188-194.

13. Ashburner M, Richards G: **Sequential gene activation by ecdysone in polytene chromosomes of *Drosophila melanogaster*. III. Consequences of ecdysone withdrawal.** *Dev Biol* 1976, **54**:241-255.

14. Lehmann M: **Drosophila Sgs genes: stage and tissue specificity of hormone responsiveness**. *Bioessays* 1996, **18**(1):47-54.

15. Farkaš R: **The complex secretions of the salivary glands of *Drosophila melanogaster*, a model system**. In: *Extracellular composite matrices in Arthropods.* Edited by Cohen E, Moussian B. Switzerland: Springer International Publishing; 2016: 557-599.

16. Korge G: **Chromosome puff activity and protein synthesis in larval salivary glands of *Drosophila melanogaster***. *Proceedings of the National Academy of Sciences of the United States of America* 1975, **72**:4550-4554.

17. Akam ME, Roberts DB, Richards GP, Ashburner M: **Drosophila: the genetics of two major larval proteins**. *Cell* 1978, **13**(2):215-225.

18. Crosby MA, Meyerowitz EM: **Drosophila glue gene Sgs-3: sequences required for puffing and transcriptional regulation**. *Dev Biol* 1986, **118**:593-607.

19. Garfinkel MD, Pruitt RE, Meyerowitz EM: **DNA sequences, gene regulation and modular protein evolution in the Drosophila 68C glue gene cluster**. *J Mol Biol* 1983, **168**:765-789.

20. Guild GM, Shore EM: **Larval salivary glande secretion proteins in *Drosophila*. Identification and characterization of the *Sgs-5* structural gene**. *J Mol Biol* 1984, **179**:289-314.

21. Hofmann A, Garfinkel MD, Meyerowitz EM: ***cis*-acting sequences required for expression of the divergently transcribed *Drosophila melanogaster Sgs-7* and *Sgs-8* glue protein genes**. *Mol Cell Biol* 1991, **11**(6):2971-2979.

22. Hofmann A, Korge G: **Upstream sequences of dosage-compensated and non-compensated alleles of the larval secretion protein gene Sgs-4 in Drosophila**. *Chromosoma* 1987, **96**:1-7.

23. Lehmann M, Korge G: **The fork head product directly specifies the tissue-specific hormone responsiveness of the *Drosophila Sgs-4* gene**. *EMBO J* 1996, **15**(18):4825-4834.

24. Martin M, Giangrande A, Ruiz C, Richards G: **Induction and repression of the Drosophila Sgs-3 glue gene are mediated by distinct sequences in the proximal promoter**. *EMBO J* 1989, **8**(2):561-568.

25. Roth GE, Wattler S, Bornschein H, Lehmann M, Korge G: **Structure and regulation of the salivary gland secretion protein gene Sgs-1 of *Drosophila melanogaster***. *Genetics* 1999, **153**:753-762.

26. Shore EM, Guild GM: **Larval salivary gland secretion proteins in Drosophila structural analysis of the Sgs-5 gene**. *J Mol Biol* 1986, **190**:149-158.

27. Shore EM, Guild GM: **Closely linked DNA elements control the expression of the *Sgs-5* glue protein gene in *Drosophila***. *Genes Dev* 1987, **1**:829-839.

28. Restifo LL, Guild GM: **An ecdysterone-responsive puff site in *Drosophila* contains a cluster of seven differentially regulated genes**. *J Mol Biol* 1986, **1986**(188).

29. Wright LG, Chen T, Thummel CS, Guild GM: **Molecular characterization of the 71E late puff in Drosophila melanogaster reveals a family of novel genes**. *J Mol Biol* 1996, **255**:387-400.

30. Korayem AM, Fabbri M, Takahashi K, Scherfer C, Lindgren M, Schmidt O, Ueda R, Dushay MS, Theopold U: **A *Drosophila* salivary gland mucin is also expressed in immune tissues: evidence for a function in coagulation and the entrapment of bacteria**. *Insect Biochem Molec Biol* 2004, **34**:1297-1304.

31. Velissariou V, Ashburner M: **The secretory proteins of the larval salivary gland of *Drosophila melanogaster:* Cytogenetic correlation of a protein and a puff**. *Chromosoma* 1980, **77**(1):13-27.

32. Velissariou V, Ashburner M: **Cytogenetic and genetic mapping of a salivary gland secretion protein in *Drosophila melanogaster***. *Chromosoma* 1981, **84**:173-185.

33. Crowley TE, Bond MW, Meyerowitz EM: **The structural genes for three *Drosophila* glue proteins reside at a single polytene chromosome puff locus**. *Mol Cell Biol* 1983, **3**(4):623-634.

34. Meyerowitz EM, Hogness DS: **Molecular organization of a Drosophila puff site that responds to ecdysone**. *Cell* 1982, **28**:165-176.

35. Muskavitch MAT, Hogness DS: **An expandable gene that encodes a Drosophila glue protein is not expressed in variants lacking remote upstream sequences**. *Cell* 1982, **29**:1041-1051.

36. Mettling C, Bourouis M, Richards G: **Allelic variation at the nucleotide level in *Drosophila* glue genes**. *Mol Gen Genet* 1985, **201**:265-268.

37. Muskavitch MA, Hogness DS: **Molecular analysis of a gene in a developmentally regulated puff of *Drosophila melanogaster***. *Proceedings of the National Academy of Sciences of the United States of America* 1980, **77**(12):7362-7366.

38. Ometto L, Cestaro A, Ramasamy S, Grassi A, Revadi S, Siozios S, Moretto M, Fontana P, Varotto C, Pisani D *et al*: **Linking Genomics and Ecology to investigate the complex evolution of an invasive *Drosophila* pest**. *Genome Biology and Evolution* 2013, **5**(4):745-757.

39. Han MV, Thomas GW, Lugo-Martinez J, Hahn MW: **Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3**. *Mol Biol Evol* 2013, **30**(8):1987-1997.

40. Garrigan D, Kingan SB, Geneva AJ, Andolfatto P, Clark AG, Thornton KR, Presgraves DC: **Genome sequencing reveals complex speciation in the *Drosophila simulans* clade**. *Genome Res* 2012, **22**:1499-1511.

41. Cariou M-L, Silvain J-F, Daubin V, Da Lage J-L, Lachaise D: **Divergence between *Drosophila santomea* and allopatric or sympatric populations of *D. yakuba* using paralogous amylase genes and migration scenarios along the volcanic line.** *Mol Ecol* 2001, **10**(3):649-660.

42. Llopart A, Lachaise D: **An anomalous hybrid zone in Drosophila**. *Evolution* 2005, **59**(12):2602-2607.

43. Hickner PV, Rivaldi CL, Johnson CM, Siddappaji M, Raster GJ, Syed Z: **The making of a pest: insights from the evolution of chemosensory receptor families in a pestiferous and invasive fly, *Drosophila suzukii***. *BMC Genomics* 2016, **17**:DOI10.1186/s12864-12016-12983-12869.

44. Dosztányi Z, Csizmók V, Tompa P, Simon I: **IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content**. *Bioinformatics* 2005, **21**:3433-3434.

45. Ishida T, Kinoshita K:
PrDOS: prediction of disordered protein regions from amino acid sequence. *Nucl Ac Res* 2007, **35**:W460-464.

46. Linding R, Jensen LJ, Diella F, Bork P, Gibson TJ, Russell RB: **Protein disorder prediction: implications for structural proteomics**. *Structure* 2003, **11**(11):1453-1459.

47. Bomma R, Venkatesh P, Kumar A, Babu AY, Rao SK: **PONDR (Predicators of Natural Disorder Regions)**. *International Journal of Computer Technology and Electronics Engineering* 2012, **21**(4):61-70.

48. Stanley Jr CE, Kulathinal RJ: **flyDIVaS: A comparative genomics resource for Drosophila divergence and selection**. *Genes Genomes Genetics* 2016, **6**:2355-2363.

49. Lachaise D, Cariou M-L, David JR, Lemeunier F, Tsacas L, Ashburner M: **Historical biogeography of the *Drosophila melanogaster* species subgroup**. *Evol Biol* 1988, **22**:159-225.

50. Lack JB, Lange JD, Tang AD, Corbett-Detig RB, Pool JE: **A thousand fly genomes: an expanded Drosophila genome nexus**. *Mol Biol Evol* 2016, **33**(12):3308-3313.

51. Hollingsworth MA, Swanson BJ: **Mucins in cancer: protection and control of the cell surface**. *Nat Rev Cancer* 2004, **4**:45-60.

52. Syed ZA, Härd T, Uv A, van Dijk-Härd IF: **A potential role for Drosophila mucins in development and physiology**. *PLoS one* 2008, **3(8):e3041. doi: 10.1371/journal.pone.0003041**.

53. Feller G, Dehareng D, Da Lage J-L: **How to remain non-folded and pliable: the linkers in modular α-amylases as a case study**. *FEBS Journal* 2011, **278**:2333-2340.

54. Schüler A, Bornberg-Bauer E: **Evolution of protein domain repeats in Metazoa**. *Mol Biol Evol* 2016, **33**(12):3170-3182.

55. Tompa P: **Intrinsically unstructured proteins evolve by repeat expansion**. *Bioessays* 2003, **25**:847-855.

56. Shivanna N, Siddalinga Murthy GS, Ramesh SR: **Larval pupation site preference and its relationship to the glue proteins in a few species of *Drosophila***. *Genome* 1996, **39**:105-111.

57. Erezyilmaz DF, Stern DL: **Pupariation site preference within and between *Drosophila* sibling species**. *Evolution* 2013, **67**(9):2714-2727.

58. Marygold SJ, Crosby MA, Goodman JL, FlyBase C: **Using FlyBase, a Database of Drosophila Genes and Genomes**. In: *Methods Mol Biol.* vol. 1478; 2016: 1-31.

59. Chiu JC, Jiang X, Zhao L, Hamm CA, Cridland JM, Saelao P, Hamby KA, Lee EK, Kwok RS, Zhang G *et al*: **Genome of *Drosophila suzukii*, the Spotted Wing Drosophila**. *G3* 2013, **3**(12):2257-2271.

60. Andolfatto P, Hu T, Thornton K: **The Drosophila santomea genome - release 1.0**. In.; 2016.

61. Nolte V, Pandey RV, Kofler R, Schlötterer C: **Genome-wide patterns of natural variation reveal strong selective sweeps and ongoing genomic conflict in *Drosophila mauritiana***. *Genome Res* 2013, **23**(1):99-110.

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

63. Yin C, Shen G, Guo D, Wang S, Ma X, Xiao H, Liu J, Zhang Z, Liu Y, Zhang Y *et al*: **InsectBase: a resource for insect genomes and transcriptomes**. *Nucl Ac Res* 2016, **44**(Database issue):D801-D807.

64. Edgar RC: **MUSCLE: multiple sequence alignment with high accuracy and high throughput**. *Nucl Ac Res* 2004, **32**(5):1792-1797.

65. Dereeper A, Guignon V, Blanc G, Audic S, Buffet S, Chevenet F, Dufayard J-F, Guindon S, Lefort V, Lescot M *et al*: **Phylogeny.fr: robust phylogenetic analysis for the non-specialist**. *Nucl Ac Res* 2008, **36(Web Server Issue)**:W465-469.

66. Letunic I, Bork P: **Interactive tree of life(iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees**. *Nucl Ac Res* 2016, **44**(W1):W242-245.

67. Yang Z: **PAML4: plylogenetic analysis by maximum likelihood**. *Mol Biol Evol* 2007, **24**(8):1586-1591.

68. Kosakovsky Pond SL, Frost SD, Muse SV: **HyPhy: hypothesis testing using phylogenies**. *Bioinformatics* 2005, **21**(5):676-679.

69. Kosakovsky Pond SL, Murrell B, Fourment M, Frost SD, Delport W, Scheffler K: **A random effects branch-site model for detecting episodic diversifying selection**. *Mol Biol Evol* 2011, **28**(11):3033-3043.

70. Delport W, Poon AF, Frost SD, Kosakovski Pond SL: **Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology**. *Bioinformatics* 2010, **21**(10):2531-2533.

71. Gramates LS, Marygold SJ, Santos GD, Urbano JM, Antonazzo G, Matthews BB, Rey AJ, Tabone CJ, Crosby MA, Emmert DB *et al*: **FlyBase at 25: looking to the future**. *Nucl Ac Res* 2017, **45(D1)**:D663-D671.

72. Geer LY, Marchler-Bauer A, Geer RC, Han L, He J, Liu C, Shi W, Bryant SH: **The NCBI BioSystems database**. *Nucl Ac Res* 2010, **38**:D492-D496.

73. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs**. *Nucl Ac Res* 1997, **25**:3389-3402.

74. Enright AJ, Van Dongen S, Ouzounis CA: **An efficient algorithm for large-scale detection of protein families**. *Nucl Ac Res* 2002, **30**(7):1575-1584.

75. Thomas GWC, Hahn MW: **Drosophila 25 species phylogeny.** *FigShare* 2017:10.6084/m6089.figshare.5450602.

76. Lack JB, Cardeno CM, Crepeau MW, Taylor W, Corbett-Detig RB, Stevens KA, Langley CH, Pool JE: **The Drosophila genome Nexus: a population genomic resource of 623 Drosophila melanogaster genomes, including 197 from a single ancestral range population**. *Genetics* 2015, **199**:1229-1241.

77. Hervas S, Sanz E, XCasillas S, Pool JE, Barbadilla A: **PopFly: the Drosophila population genomics browser**. *Bioinformatics* 2017, **33**:2779-2780.

78. Rozas J: **DNA sequence polymorphism analysis using DnaSP**. *Methods in molecular biology* 2009, **537**:337-350.

79. Furia M, Digilio FA, Artiaco D, Favia G, Polito LC: **Molecular characterization of a Drosophila melanogaster variant strain defective in the Sgs-4 gene dosage compensation**. *Bioch Biophy Acta* 1992, **1130**:314-316.

80. Jukes TH, Cantor CR: **Evolution of protein molecules**. In: *Mammalian protein metabolism.* Edited by Munro HN. New York: Academic Press; 1969: 21-132.

Legends of Figures

Figure 1 : Schematic species tree showing glue gene distribution and the most parsimonious scenario for gene gains and losses. Gains are indicated by "+" and losses by "-". Numbers correspond to the glue gene name (eg. "3" for Sgs3). An inferred distribution of glue genes in the last common ancestor is shown at the bottom. The tree is from Thomas, G.W.C. and Hahn M.W. (2017) http://dx.doi.org/10.6084/m9.figshare.5450602. Pink is for *Sgs1*, yellow is for *Sgs3*, dark blue is for *Sgs7*, light blue is for *Sgs8*, green is for *Sgs4*, orange is for *Sgs5-5bis*, purple is for *Eig71Ee*. Along with each species is a schematic representation of the organization of the glue gene cluster, with relative position and orientation for the species with confirmed synteny information. Gene sizes and distances are not to scale. "R" means that internal repeats are present. "R?" means that no clear repeats were identified. In *D. pseudoobscura*, the relative orientation of the three clustered *Sgs3*-like sequences GA25425, GA23426, GA23878 suggested that GA23426 could be orthologous to *Sgs3* (it is inside an intron of GA11155, homologue of Mob2, which is close to *Sgs3* in *D. melanogaster*),

GA23425 to *Sgs7* and GA23878 to *Sgs8*. The last two had more similar sequences compared to GA23426, including the repeat region. Furthermore, the latter was neighbor to GA20420, a homologue of *chrb-PC*, a gene adjacent to *Sgs8* in *D. melanogaster*.

Figure 2 : Maximum likelihood (ML) unrooted tree of aligned Sgs5 and Sgs5bis amino acid sequences (repeated parts removed when present). Numbers along branches are the posterior probabilities.

Figure 3 : Unrooted ML tree of aligned Sgs3 (repeats removed), Sgs7 and Sgs8 amino acid sequences. Numbers along branches are the posterior probabilities.

Figure 4 : Example of predictions for disordered regions by PONDR. A: The glue proteins with internal repeats of *D. simulans,* except Sgs5, and Eig71Ee; B: example of Sgs5 protein with large internal repeats (*D. kikkawai*) compared to the one of *D. simulans*.

Figure 5 : Distribution of dN, dS and dN/dS for the pair *D. melanogaster*/*D. simulans* from the flyDIVaS database with the position of glue genes. Blue dots and arrows : dN ; red dots : dS ; green dots and arrows: dN/dS. Vertical axis : number of genes. Genes are binned into rate value categories with increment of 0.005.

Figure 6: Output trees of BS-REL analyses at Datamonkey website (classic.datamonkey.org). MEL: melanogaster, SIM: simulans, SECH: sechellia, SAN: santomea, YAK: yakuba, ERE: erecta, TAK: takahashii, SUZ: suzukii, BIAR: biarmipes, FIC: ficusphila, KIK: kikkawai, ANA: ananassae, BIP: bipectinata. “The hue of each color indicates strength of selection, with primary red corresponding to $\omega > 5$, primary blue to $w = 0$ and grey to $w = 1$. The width of each color component represent the proportion of sites in the corresponding class. Thicker branches have been classified as undergoing episodic diversifying selection by the sequential likelihood ratio test at corrected $p \leq 0.05$”.

Table 1: List of species and databases used in this study.

Species	Database	Version	URL	Date of access	reference
melanogaster	FlyBase	FB2015_02	flybase.org	06/2016	[58]
simulans	FlyBase	FB2015_02	flybase.org	02/2017	[58]
sechellia	FlyBase	FB2015_02	flybase.org	02/2017	[58]
mauritiana		v1.0	www.popoolation.at/mauritiana_genome/	12/2016	[61]
yakuba	FlyBase	FB2015_02	flybase.org	02/2017	[58]
santomea		v1.0	genomics.princeton.edu/AndolfattoLab/Dsantomea_genome.html	11/2016	[60]
erecta	FlyBase	FB2015_02	flybase.org	02/2017	[58]
takahashii	FlyBase	FB2015_02	flybase.org	02/2017	[58]
ficusphila	FlyBase	FB2015_02	flybase.org	02/2017	[58]
biarmipes	FlyBase	FB2015_02	flybase.org	02/2017	[58]
suzukii	SpottingWingFlybase	v1	http://spottedwingflybase.org/	02/2017	[59]
eugracilis	FlyBase	FB2015_02	flybase.org	02/2017	[58]
elegans	FlyBase	FB2015_02	flybase.org	02/2017	[58]
rhopaloa	FlyBase	FB2015_02	flybase.org	02/2017	[58]
kikkawai	FlyBase	FB2015_02	flybase.org	02/2017	[58]
ananassae	FlyBase	FB2015_02	flybase.org	02/2017	[58]
bipectinata	FlyBase	FB2015_02	flybase.org	02/2017	[58]
willistoni	FlyBase	FB2015_02	flybase.org	02/2017	[58]

Table 2 : Genomic coordinates of the glue genes in 20 Drosophila species. * indicates annotations and coordinates of the *Sgs5bis* gene; "M" indicates that part of the coding sequence was inferred manually by sequencing of PCR amplicons of relevant regions; "no" means that the gene sequence was not found by BLAST searches; Nterm and Cterm mean N-terminal and C-terminal region, respectively. **: this contig probably contains two paralogs of *Sgs3* with uncertain sequences.

Species	***Sgs1***	***Sgs3***	***Sgs4***	***Sgs5 Sgs5bis****	***Sgs7***	***Sgs8***	***Eig71Ee***
D. melanogaster	CG3047	CG11720	CG12181	CG7596 CG7587*	CG18087	CG6132	CG7604
D. simulans	GB:CM002910 4752550- 4754973	Dsim\GD14311	Dsim\GD16637	Dsim\GD19170 Dsim\GD19169*	Dsim\GD17634	Dsim\GD28639	Dsim\GD12546
D. sechellia	Dsec\GM18501 (M)	Dsec\GM25279 (M)	GB:CH480825 2852711-2853386 (M)	Dsec\GM15245 Dsec\GM15244*	Dsec\GM25278	Dsec\GM24748	NW_001999689 7761215-7759941
D. mauritiana	2L : 4721427-4722731	3L : 11002313-11003109	X : 2864998-2865616 (M)	3R : 7695225-7694660 relictual Sgs5bis 3R :7696600-7695629	3L : 10999955-11000249	no	3L : 15018149-15017249
D. yakuba	NT_167062 10588365-10585585	Dyak\Sgs3	Dyak\GE28681	Dyak\GE25481 Dyak\GE25480*	Dyak\GE20214 Dyak\GE21218	Dyak\Sgs8	Dyak\GE19823
D. santomea	2L : 10595909-10588129	3L : 11541799-11542678 (M)	X : 5242740-5241688 (M)	3R : 1975190-1975883 3R : 1974195-1974756*	3L : 11539572-11539861 3L : 11536774-11536485	3L : 11537383-11537681	3L : 18202978-18201736
D. erecta	no	Dere\Sgs3	Dere\GG27095	no Sgs5 Dere\GG22329*	Dere\GG13918	Dere\Sgs8	Dere\GG13528
D. eugracilis	AFPQ02004874 817906- 819883	KB465257 3401691- 3402412 3385186-3386300	no	KB464468 62658- 63338 61657-62202*	KB465257 3378701- 3378995	KB465257 3378110- 3377822	KB464880 383836-382228 (XM_017230731)
D. takahashii	KB461520 248469-250276	KB460792 317161-317949	no	KB461611 188299-187637 189545-188599*	KB461234 120246-120467	KB461234 119117-118896	XM_017142344
D. ficusphila	KB457325 1315471-1313145	KB457563 3180441-3179541	no	KB457381 2059719- 2058971	no	no	KB457515 1660700-1661809

		KB457373 332100-331262 3199436-3198351		2061615-2060148*			(XM_017197540)
D. biarmipes	KB462641 1521394-1523538	KB462590 1536842-1537624 (M) KB462646 54238-53374 (M)	no	KB462814 8082338-8083047 8081336-8081891*	KB462646 76095-75801	KB462646 77216-77501	KB462754 733209-734564
D. suzukii	KI419149 6645021-6638237	no	no	KI420542 10372-9639 11441-10912*	KI419359 22757-22464 KI420769 54293-54584 KI420610 25121-25412 55385-55094	KI420769 53260-52976	XM_017082231
D. elegans	KB458429 2603084- 2605600	KB458268 2467758- 2468497 KB458387 820622-819957 KB458387 18429-17499	no	KB458458 2864199- 2863401 no Sgs5bis	no	no	no
D. rhopaloa	KB450401 (Nterm) KB452165 (Cterm)	KB450817 117692-118515 KB452471 215593-216424 KB451944**	no	KB451039 15186-16018 no Sgs5bis	no	no	no
D. kikkawai	no	KB459615 1331679-1331220 KB459522 291906-292542	no	KB459676 1112222-1111011 1113233-1112671*	no	no	KB459876 1106397-1107027 (Nterm)
D. ananassae	no	NW_001939300 3959435-3957637 NW_001939293 5806878-5808646	no	NW_001939291 17741832-17741201 17742892-17742284*	no	no	GF10382(Nterm): NW_001939293 11506744-11507112
D. bipectinata	no	KB464001 557673-558039 KB464098 1120437-1121198	no	KB464382 185749-186362 184743-185354*	KB464098 1109828-1110127	KB464098 1109077-1108802	KB464259 2466431-2466234 (ortholog of GF10382)
D.	no	GA23425, GA23426,	no	no Sgs5	no	no	no

pseudoobscura		GA23878		Dpse\GA20459 *			
D. willistoni	no	NW_002032853 3296683-3295766 NW_002032860 11643758-11641972	no	no	NW_002032853 2792051-2792347 2793811-2794107	no	no
D. virilis	no	NW_002014431 6839085-6838999 (GJ27025) 6841799-6840888(GJ26085)	no	no Sgs5 NW_002014424 14511530-14512000*(GJ24445)	no	no	no

Table 3 : Characteristics of glue proteins in the species studied (except Sgs7 and Sgs8). Glycosylation sites were predicted from http://www.cbs.dtu.dk/services/NetNGlyc/ and http://www.cbs.dtu.dk/services/NetOGlyc/ for N glycosylation and O glycosylation, respectively. *: except for IUPred and PrDOS

protein	**species**	**length (aa)**	**kind of repeat**	**approx. nr of repeats**	**N glyc**	**O glyc**	**disoredered repeats**
Sgs1	*melanogaster*	1286	PTTTTPR/STTTTSTSR	ca 85	2	>25	yes
	simulans	785	CAPTTTTPR	ca 40	1	>25	yes
	mauritiana	412	CAPTTTTPR	ca 13	1	>25	yes
	sechellia	492	CAPTTTTPR	ca 22	1	>25	yes
	santomea		uncertain sequence				
	yakuba	619?	RPPTTSPSC	uncertain		>25	
	elegans	838	T rich stretches		0	>25	yes
	rhopaloa	ca. 624	T rich stretches		1	>25	yes
	ficusphila	758	CAPTTTPST	ca 59	0	>25	yes
	takahashii	585	TSTTTTPR	ca 25	1	>25	yes
	eugracilis	635	PRCTTTTT	ca 39	0	>25	yes
	biarmipes	696	VPTT/KCQMTTSSSAPTTAAPTATSTTAATTSTP	3/ca 12	1	>25	yes
	suzukii	2245	VPTT/RCPITTSTSAPTTTTATTTSTSTSTTSTP	8/ca 63	1	>25	yes
Sgs3	*melanogaster*	307	KPTTT	ca 31	0	>25	yes
	simulans	188	a few T rich stretches		0	>25	yes
	mauritiana	183	CAPPTRPPCTSPTTTTTTTTTT	ca 5	1	>25	yes
	sechellia	172	CKPTTTTTT	ca 8	0	>25	yes
	santomea	273	PTTTTTTTRR	ca 6	0	>25	yes
	yakuba	273	PTTTTTTTRR	ca 6	0	>25	yes
	erecta	333	TTRR	ca 35	3	>25	yes
	elegans	216	CAPTTTTTTTQR	ca 7	0	>25	yes

	elegans bis	202	KATT	ca 24	0	>25	yes
	elegans ter	287	PTTTTTKK	ca 23	1	>25	yes
	ficusphila	266	CAPTTTTTT	ca 12	0	>25	yes
	ficusphila bis	259	T rich stretches		0	>25	yes
	ficusphila ter	335	CKPPTTS/KPSKPT	ca 10/ca 28	1	>25	yes
	takahashii	585	PTTTSTTR	ca 27	1	>25	yes
	eugracilis	214	CAPTTTTTTTTT	ca 7	0	>25	yes
	eugracilis bis	348	PTK	ca 65	2	>25	yes
	biarmipes	244	KKPXTT	ca 21	0	>25	yes
	biarmipes bis	302	T rich stretches		0	>25	yes
	rhopaloa	254	ATTK	ca 21	0	>25	yes
	rhopaloa bis	256	T rich stretches		0	>25	yes
	rhopaloa ter	253	CAPTTTTTT	ca 12	0	>25	yes
	rhopaloa 4°	incomplete 5'	CAPTTTTTT	ca 9	0	>25	yes
	kikkawai	129	KPQP	ca 10	0	2	yes
	kikkawai bis	190	KPQPP	ca 16	0	6	yes
	ananassae	579	KPTTP	ca 55	1	>25	yes
	ananassae bis	566	PTR/PTE/PTV	ca 71/42/22	2	>25	yes
	bipectinata	272	T rich stretches/PTKSTR	ca 8	0	>25	yes
	bipectinata bis	254	QPPTKSTPKPT	ca 8	0	>25	yes
	pseudoobscura	207	KPT	ca 23	0	>25	yes
	pseudoobscura bis	229	KPTTTP	ca 14	0	>25	yes
	pseudoobscura ter	224	KPT	ca 33	0	>25	yes
	willistoni	283	P/T-rich stretch		0	>25	yes
	willistoni sgs3-like	546	CVTTRSSTPTP/CGPTPSPSPT	ca. 15/17	0	>25	yes
	virilis	242	RTTTTPTTTT	ca 12	0	>25	yes
	virilis bis	283	KPTTTRRT/KTIPTTTP	ca 11/9	2	>25	yes
Sgs4	*melanogaster*	287	CRTEPPT	ca 19	0	>25	yes*

	simulans	266	CDTEPPT	ca 8	0	>25	yes*
	mauritiana	360	CNTEPPT	ca 31	0	>25	yes*
	sechellia	255	CNTEPPT/CDTEPPT	ca5/4	0	>25	yes*
	santomea	351	C(K/R)T(E/T)PPT / CKTKPPCTTV	ca 14/9	0	>25	yes*
	yakuba	361	C(K/R)T(E/T)PPT	ca 23	0	>25	yes*
	erecta	280	CRTEPPT/NAPTRRT	ca 8/7	1	>25	yes*
Sgs5 and 5bis	*melanogaster*	163	no repeats		0	2	NA
	melanogaster bis	142	no repeats		0	0	NA
	simulans	169	PE/TE	ca 6	0	8	yes
	simulans bis	142	no repeats		0	0	NA
	mauritiana	169	PE/TE	ca 6	0	10	yes
	sechellia	169	PE/TE	ca 6	0	10	yes
	sechellia bis	142	no repeats		0	0	NA
	santomea	192	TE	ca 7	0	8	yes
	santomea bis	142	no repeats		0	0	NA
	yakuba	192	TE	ca 7	0	12	yes
	erecta bis	142	no repeats		0	0	NA
	ficusphila	208	DP or EP, ES, ET	ca 28	0	22	yes
	ficusphila bis	142	no repeats		0	0	NA
	takahashii	217	EP or EE	ca 12	0	19	yes
	takahashii bis	161	no repeats		0	3	NA
	biarmipes	190	PED or PET	ca 10	0	17	yes
	biarmipes bis	143	no repeats		0	1	NA
	elegans	223	EP	ca 27	0	11	yes
	eugracilis	187	PE	ca 16	0	14	yes
	eugracilis bis	142	no repeats		0	0	NA
	suzukii	203	PETE	ca 11	0	23	yes
	suzukii bis	142?	no repeats		0	1	NA

	kikkawai	362	PEDEED	ca 37	0	11	yes
	kikkawai bis	146	no repeats		0	2	NA
	rhopaloa	236	EP	ca 38	0	9	yes
	ananassae	172	almost no repeats		0	2	NA
	ananassae bis	146	no repeats		0	0	NA
	bipectinata	162	almost no repeats		0	3	NA
	bipectinata bis	146	no repeats		0	1	NA
	pseudoobscura bis	144	no repeats		0	0	NA
	virilis	143	no repeats		0	0	NA
Eig71Ee	*melanogaster*	445	CTCTESTT/(R/K)TNPT	ca 9/ca 7	8	>25	yes
	simulans	321	CTCTDSTT(R/K)KTNPT	ca 4/ca 2	2	>25	yes
	sechellia	408	CTDSTTKTTNPPCT	ca 8	3	>25	yes
	mauritiana	284	no clear repeats		0	>25	yes
	yakuba	417	CTESTTQKPNPPSTQKTRPPCG	ca 5	1	>25	yes
	santomea	394	CTESTTQKPNPPSTEKTRPPCG	ca 3	1	>25	yes
	erecta	454	CTESTTRRTKPPSTRKTRPP	ca 5	0	>25	yes
	ficusphila	384	TE(K/R)T	ca 11	1	>25	yes
	takahashii	302	CTEKTTQKPEPP	ca 7	0	>25	yes
	biarmipes	434	no clear repeats		6	>25	yes
	suzukii	346	no clear repeats		0	>25	yes
	eugracilis	447	CTETTTQKTNPP	ca 5	0	>25	yes

Table 4: List of strains used for PCR amplification. Number of repeats and repeat motifs in Sgs3 and Sgs4 in populations of *D. melanogaster* and *D. mauritiana*. Sequences of Sgs4 for Oregon R and Samarkand strains are from [79]. * indicate lines also used in the Drosophila Nexus project. @ indicate suspected artifactual repeat losses during cloning. PSC indicates the presence of a premature stop codon.

protein	species	sample	Origin	nr of repeats	type of repeat	remarks
Sgs3	*D. melanogaster*	Cayenne	French Guyana	29	(K/N)(P/Q/A)TTT	
		Chavroche	France	29	(K/N)(P/Q/A)TTT	
		Chavroche2	France	29	(K/N)(P/Q/A)TTT	
		Chavroche3	France	30	(K/N)(P/Q/A)TTT	
		Cotonou	Benin	31	(K/N)(P/Q/A)TTT	
		Delhi1	India	27	(K/N)(P/Q/A)TTT	
		Delhi2	India	29	(K/N)(P/Q/A)TTT	
		Delhi B	India	27	(K/N)(P/Q/A)TTT	
		Gally A	France	29	(K/N)(P/Q/A)TTT	
		Gally B	France	29	(K/N)(P/Q/A)TTT	
		Gally C	France	29	(K/N)(P/Q/A)TTT	
		Gally D	France	29	(K/N)(P/Q/A)TTT	
		EF1 B	Ethiopia*	24	(K/N)(P/Q/A)TTT	
		EF1 3	Ethiopia*	29	(K/N)(P/Q/A)TTT	
		EG15N	Cairo, Egypt*	30	(K/N)(P/Q/A)TTT	
		EG16N	Cairo, Egypt*	>25	(K/N)(P/Q/A)TTT	
		EG25N	Cairo, Egypt*	29	(K/N)(P/Q/A)TTT	
		EG28N	Cairo, Egypt*	>29	(K/N)(P/Q/A)TTT	
		EG33N a	Cairo, Egypt*	12@	(K/N)(P/Q/A)TTT	
		EG33N c	Cairo, Egypt*	31	(K/N)(P/Q/A)TTT	
		EG34N	Cairo, Egypt*	7@	(K/N)(P/Q/A)TTT	
		EG55N	Cairo, Egypt*	23	(K/N)(P/Q/A)TTT	

		EG59N	Cairo, Egypt*	22	(K/N)(P/Q/A)TTT
		EG74N	Cairo, Egypt*	23	(K/N)(P/Q/A)TTT
	D. mauritiana	GM21	Grande Montagne (Rodrigues Island)	5	CAPPTRPP(T)n
		GM23a	Grande Montagne (Rodrigues Island)	5	CAPPTRPP(T)n
		GM23b	Grande Montagne (Rodrigues Island)	3	CAPPTRPP(T)n
		GM24	Grande Montagne (Rodrigues Island)	4	CAPPTRPP(T)n
		GM25	Grande Montagne (Rodrigues Island)	5	CAPPTRPP(T)n
		GRNM1	Gorges de la Rivière Noire (Mauritius)	5	CAPPTRPP(T)n
		MaurII-704	Mauritius	5	CAPPTRPP(T)n
		MaurII-a	Mauritius	5	CAPPTRPP(T)n
Sgs4	*D. melanogaster*	CG12181	reference strain Iso1	20	C(K/R/E)TEPP(R/T)
		OregonR	lab strain (from [79])	22	C(K/R/E)TEPP(R/T)
		Samarkand	[79]	21	C(K/R/E)TEPP(R/T)
		Canton S	Lab strain	>21	C(K/R/E)TEPP(R/T)
		Cayenne1	French Guyana	>21	C(K/R/E)TEPP(R/T)
		Cayenne2	French Guyana	>22	C(K/R/E)TEPP(R/T)
		Cayenne3	French Guyana	>21	C(K/R/E)TEPP(R/T)
		Chavroche1	France	>22	C(K/R/E)TEPP(R/T)
		Chavroche3	France	>22	C(K/R/E)TEPP(R/T)
		Comores1	Comores	>22	C(K/R/E)TEPP(R/T)
		Comores2	Comores	>22	C(K/R/E)TEPP(R/T)
		Cotonou	Benin	>22	C(K/R/E)TEPP(R/T)
		Delhi1	India	>21	C(K/R/E)TEPP(R/T)
		Delhi2	India	>21	C(K/R/E)TEPP(R/T)
		Gally1	France	>20	C(K/R/E)TEPP(R/T)
		Gally2	France	>20	C(K/R/E)TEPP(R/T)
		EF1	Ethiopia*	>22	C(K/R/E)TEPP(R/T)
		Tai1	Ivory Coast	>20	C(K/R/E)TEPP(R/T)

		Tai2	Ivory Coast	>20	C(K/R/E)TEPP(R/T)	
		EG15N	Cairo, Egypt*	>26	C(K/R/E)TEPP(R/T)	PSC
		EG16N	Cairo, Egypt*	22	C(K/R/E)TEPP(R/T)	PSC
		EG25N	Cairo, Egypt*	20	C(K/R/E)TEPP(R/T)	PSC
		EG28N	Cairo, Egypt*	20	C(K/R/E)TEPP(R/T)	PSC
		EG33N	Cairo, Egypt*	20	C(K/R/E)TEPP(R/T)	PSC
		EG34N	Cairo, Egypt*	22	C(K/R/E)TEPP(R/T)	PSC
		EG36N	Cairo, Egypt*	22	C(K/R/E)TEPP(R/T)	PSC
		EG44N	Cairo, Egypt*	>26	C(K/R/E)TEPP(R/T)	PSC
		EG55N	Cairo, Egypt*	>26	C(K/R/E)TEPP(R/T)	PSC
		EG59N	Cairo, Egypt*	>26	C(K/R/E)TEPP(R/T)	PSC
		EG74N	Cairo, Egypt*	>26	C(K/R/E)TEPP(R/T)	
		ZI395	Zambia*	25	C(K/R/E)TEPP(R/T)	
		ZI420	Zambia*	18	C(K/R/E)TEPP(R/T)	
	D.mauritiana	GM22	Grande Montagne (Rodrigues Island)	>30	C(N/D)TEPP	
		GM23	Grande Montagne (Rodrigues Island)	>31	C(N/D)TEPP	
		GM25	Grande Montagne (Rodrigues Island)	>30	C(N/D)TEPP	
		GRNM1	Gorges de la Rivière Noire (Mauritius)	>27	C(N/D)TEPP	
		GRNM2	Gorges de la Rivière Noire (Mauritius)	>32	C(N/D)TEPP	
		GRNM3	Gorges de la Rivière Noire (Mauritius)	>27	C(N/D)TEPP	
		GRNM6	Gorges de la Rivière Noire (Mauritius)	>24	C(N/D)TEPP	
		MaurII-a	Mauritius	>28	C(N/D)TEPP	
		MaurII-704	Mauritius	>28	C(N/D)TEPP	
Sequence checking	*D. sechellia*		Praslin Island			
	D. santomea	STO3	Sao Tomé			
	D. virilis		Spain			
	D. biarmipes		India			

Table 5 : Non-synonymous (dN) and synonymous (dS) substitution rates, and the dN/dS ratio for glue genes between *D. melanogaster* and *D. simulans* in pairwise alignments. *Sgs3* was not included, and unalignable regions were removed.

	Sgs1	*Sgs4*	*Sgs5*	*Sgs5bis*	*Sgs7*	*Sgs8*	*Eig71Ee*
dN	0.110	0.183	0.034	0.029	0.047	0.179	0.0678
dS	0.079	0.334	0.084	0.067	0.146	0.146	0.110
dN/dS	1.393	0.547	0.405	0.430	0.323	1.259	0.616

Table 6 : Non-synonymous (dN) and synonymous (dS) substitution rates and the ratio dN/dS for *Sgs7* and *Sgs8* between related species pairs in pairwise alignments.

Species pair	Gene	dN	dS	dN/dS
melanogaster/simulans	*Sgs7*	0.0475	0.1459	0.323
	Sgs8	0.1789	0.1420	1.259
melanogaster/sechellia	*Sgs7*	0.0990	0.1339	0.739
	Sgs8	0.1866	0.1216	1.534
simulans/sechellia	*Sgs7*	0.0696	0.0559	1.245
	Sgs8	0.0060	0.0564	0.106
yakuba/erecta	*Sgs7*	0.1780	0.2235	0.796
	Sgs8	0.1623	0.2164	0.750
biarmipes/suzukii	*Sgs7*	0.0592	0.4329	0.137
	Sgs8	0.0565	0.4533	0.125

Table 7 : A) Nucleotide diversity π of *Sgs5* and *Sgs5bis* in four populations, computed from Jukes and Cantor [80] using DnaSP. **B)** Nucleotide divergence between populations D_{xy} computed from Jukes and Cantor in DnaSP. EF: Ethiopia, FR : France, ZI : Zambia, RAL : Raleigh. N : number of lines, n : number of sites, S : number of segregating sites, S.D. : standard deviation, π_{global} and $D_{xy\ global}$: nucleotide diversity and nucleotide divergence across the genomes, respectively, from [50].

A)

Sgs5	**N**	**n**	**S**	**π (S.D.)**	**π_{global}**
EF	35	467	11	0.00450 (0.00106)	0.00622
FR	45	476	5	0.00423 (0.00023)	0.00471
ZI	183	489	38	0.00998 (0.00030)	0.00843
RAL	153	386	8	0.00257 (0.00015)	0.00569
Sgs5bis	**N**	**n**	**S**	**π (S.E.)**	**π_{global}**
EF	35	406	3	0.00267 (0.00024)	0.00622
FR	45	422	8	0.00460 (0.00029)	0.00471
ZI	201	426	37	0.00614 (0.00034)	0.00843
RAL	172	278	5	0.00322 (0.00018)	0.00569

B)

Sgs5	**N**	**n**	**S**	**D_{xy} (S.D.)**	**$D_{xy\ global}$**
ZI/EF	183/35	467	37/11	0.01197 (0.00082)	0.00855
ZI/FR	183/45	476	33/5	0.00685 (0.00046)	0.00868
ZI/RAL	183/153	386	25/8	0.00488 (0.00036)	0.00864
EF/FR	35/45	454	8/5	0.00810 (0.00128)	0.00795
EF/RAL	35/153	373	6/8	0.00705 (0.00093)	0.00790
FR/RAL	45/153	379	7/2	0.00162 (0.00025)	0.00546

Sgs5bis	**N**	**n**	**S**	**D_{xy} (S.D.)**	**$D_{xy\ global}$**
ZI/EF	201/35	406	35/3	0.00506 (0.00055)	0.00855
ZI/FR	201/45	422	36/8	0.00639 (0.00057)	0.00868
ZI/RAL	201/172	278	23/5	0.00423 (0.00033)	0.00864
EF/FR	35/45	402	3/6	0.00477 (0.00091)	0.00795
EF/RAL	35/172	263	3/5	0.00551 (0.00090)	0.00790
FR/RAL	45/172	276	6/5	0.00289 (0.00035)	0.00546

Legends of Supplementary Materials

Figure S1: Ancestral states for the *Sgs1-3-7-8* gene family inferred by CAFE. Species tips are labeled with the observed gene count and internal nodes are labeled with inferred gene counts. Orange branches represent gene losses, blue branches represent gene gains, while black branches represent lineages in which no change in gene copy number is observed. Branches marked with asterisks have marginally significant p-values (< 0.05).

Figure S2 : Partial alignment of *Sgs3* sequences with translation in *D. melanogaster* individuals. EF : Ethiopia ; Chavroche and Gally : France ; Cotonou : Benin; Delhi : India; Cayenne : French Guyana.

Figure S3: Partial alignment of *Sgs3* sequences with translation in the EG population (Cairo) of *D. melanogaster*.

Figure S4: Partial alignment of *Sgs4* sequences with translation in *D. melanogaster* individuals. EF : Ethiopia ; Chavroche and Gally : France ; Cotonou : Benin; Delhi : India; Cayenne : French Guyana ; Tai : Ivory Coast.

Figure S5: Partial alignment of Sgs4 protein sequences in the EG population (Cairo) and ZI (Zambia) of *D. melanogaster*. The reference sequence is shown. Asterisks indicate premature stop codons.

Figure S6 : Partial alignment of Sgs3 (A) and Sgs4 (B) amino acid sequences in *D. mauritiana* individuals. Sgs3 mau and Sgs4 mau are the sequences from the online genome. Sgs4 mau has been corrected with our resequencing. Xs are undetermined amino acids.

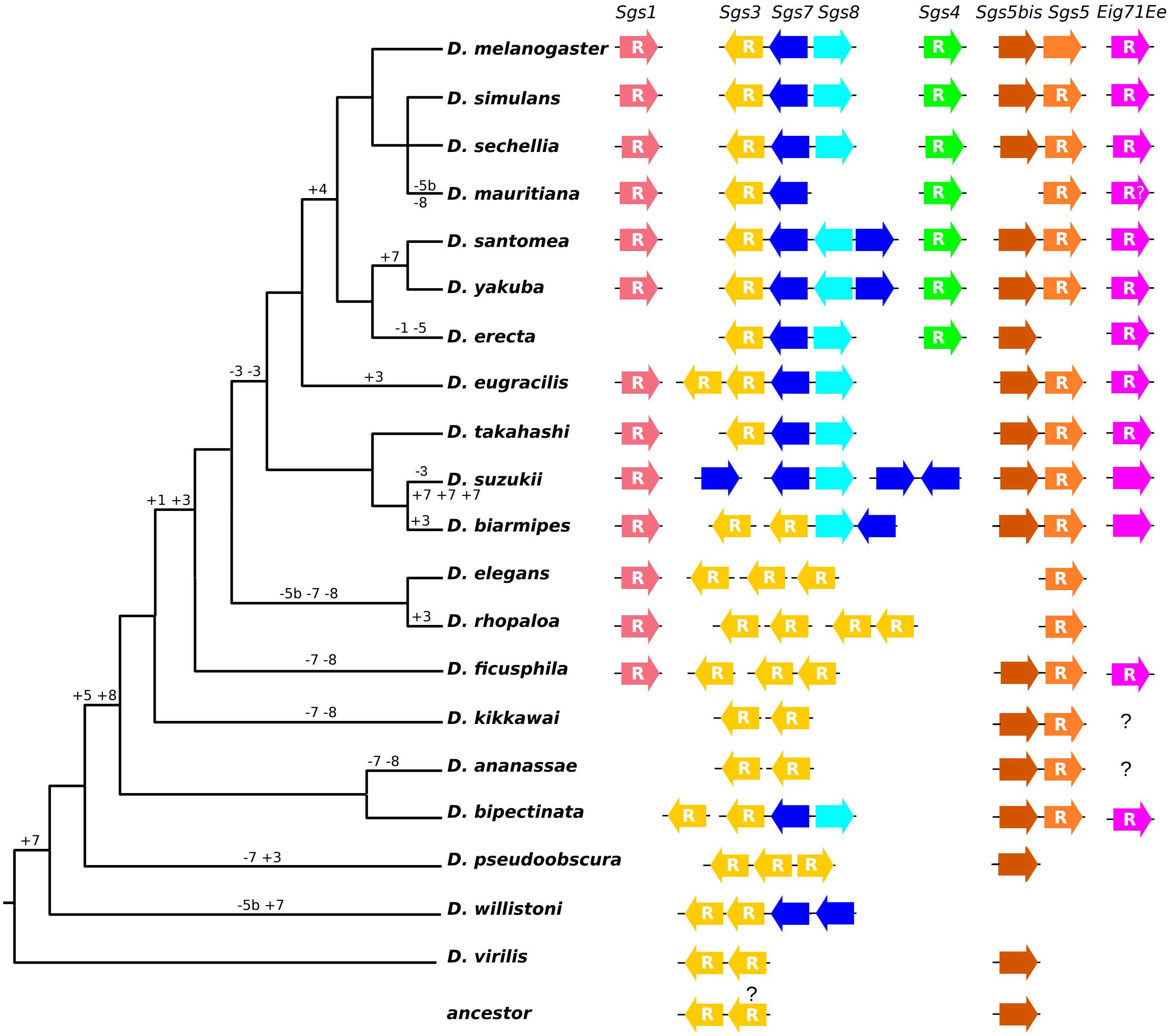
Table S1: Number of gene copies for each family, and results of CAFE analysis for the glue gene families.

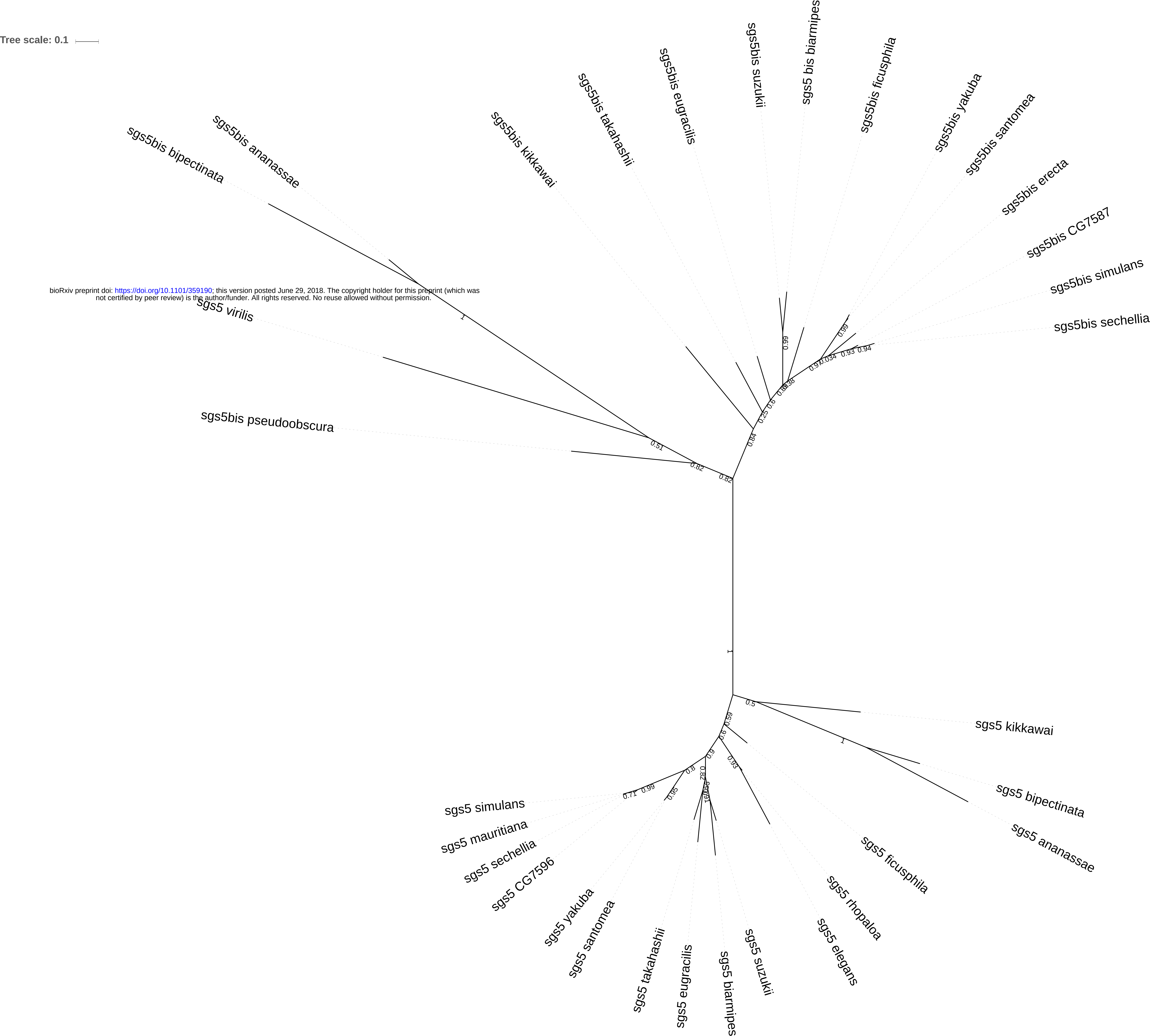
Table S2 : List of primers used for this study. Different combinations were used to amplify glue genes. All primers were chosen outside the repeated regions. *D. sechellia*, *D. santomea*, *D. virilis* and *D. biarmipes* were resequenced because of uncertainties or putative errors in the

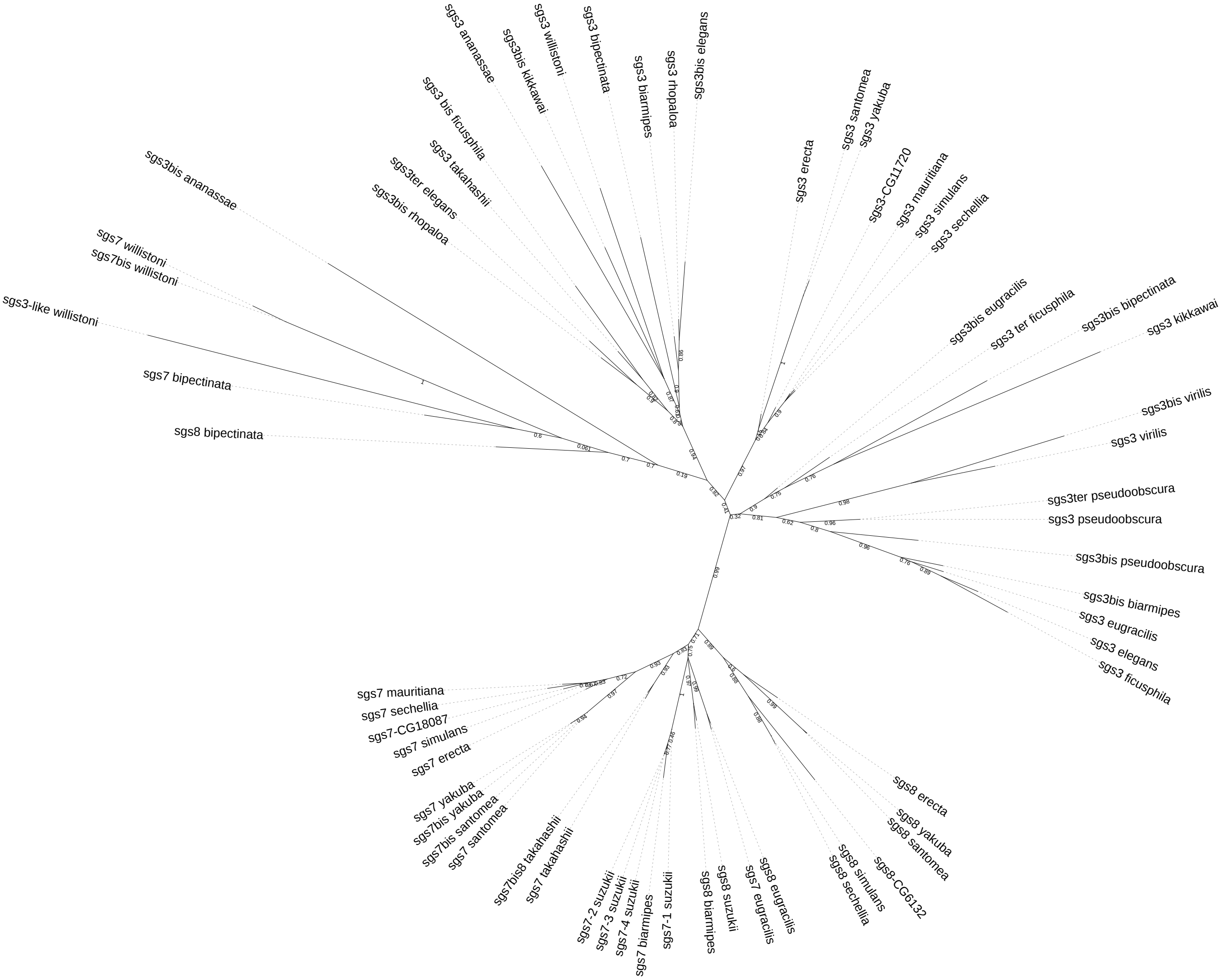
online sequences. *D. melanogaster* and *D. mauritiana* were resequenced for studying RNV in *Sgs3* and *Sgs4*.

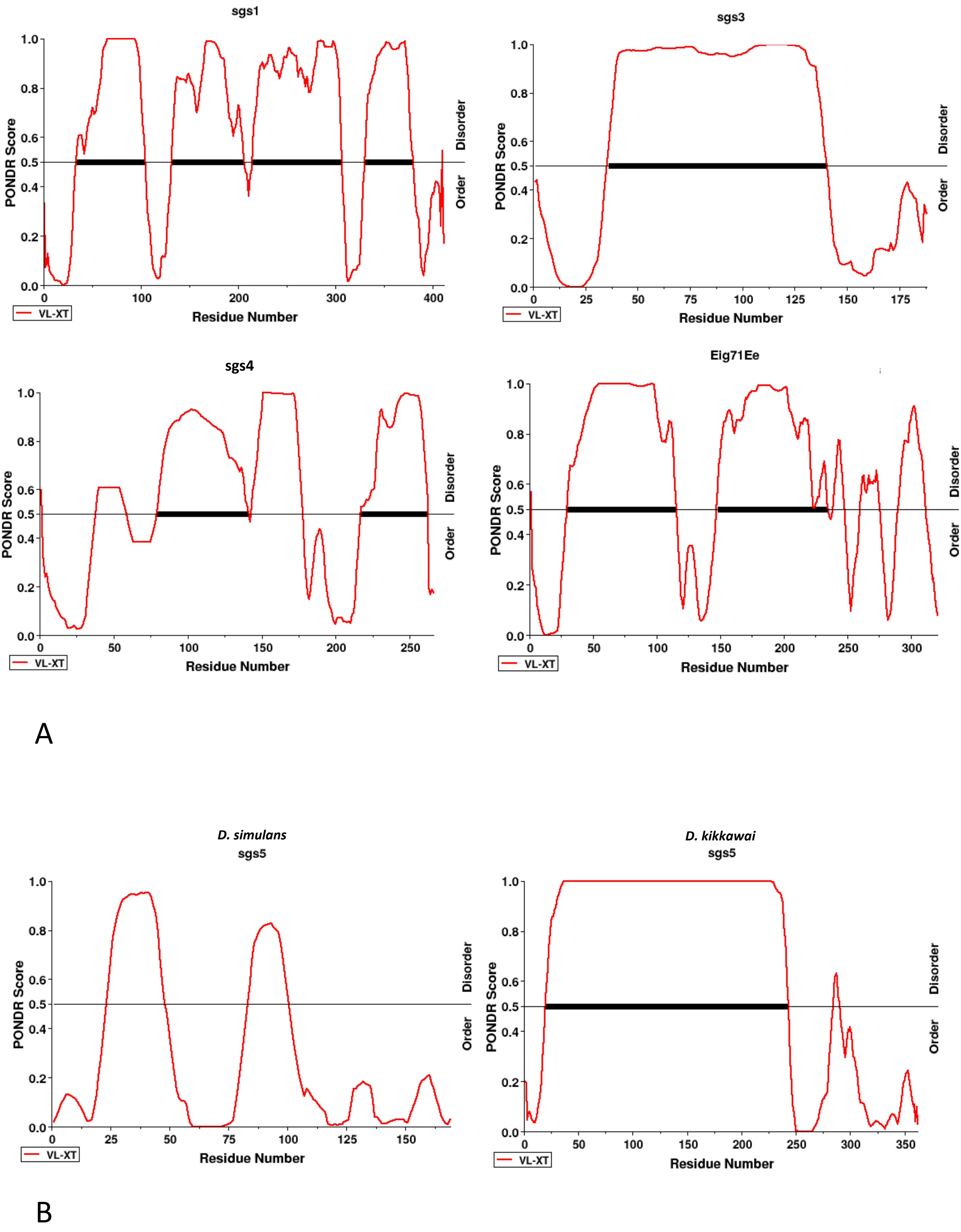
Table S3: Assembly/Annotation error estimation and gene gain/loss rates in a single λ model in the 25 *Drosophlia* species included in this study compared to previous studies using fewer species.

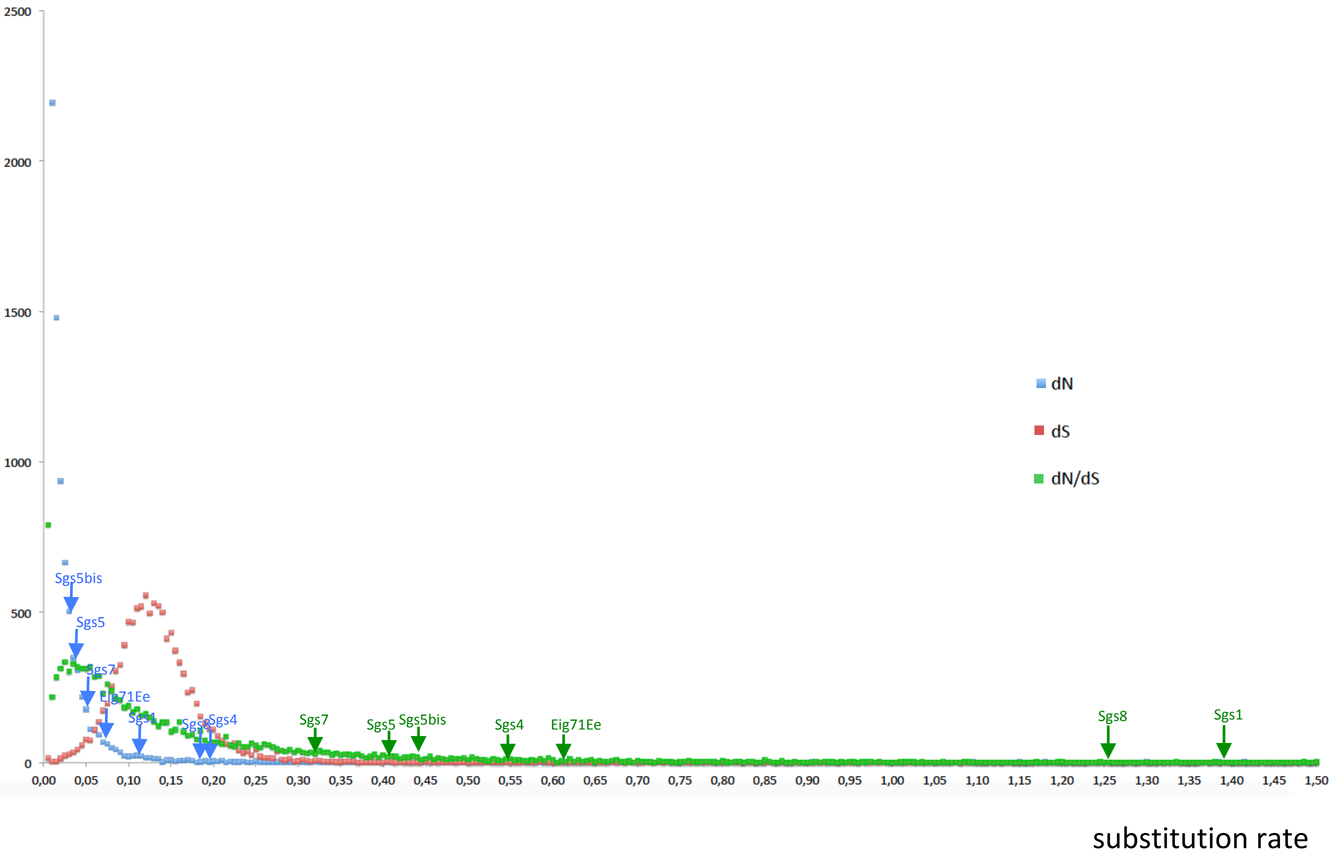
Table S4: Summary of gene gain and loss events inferred after correcting for annotation and assembly error across all 25 *Drosophila* species. The number of rapidly evolving families is shown in parentheses for each type of change.

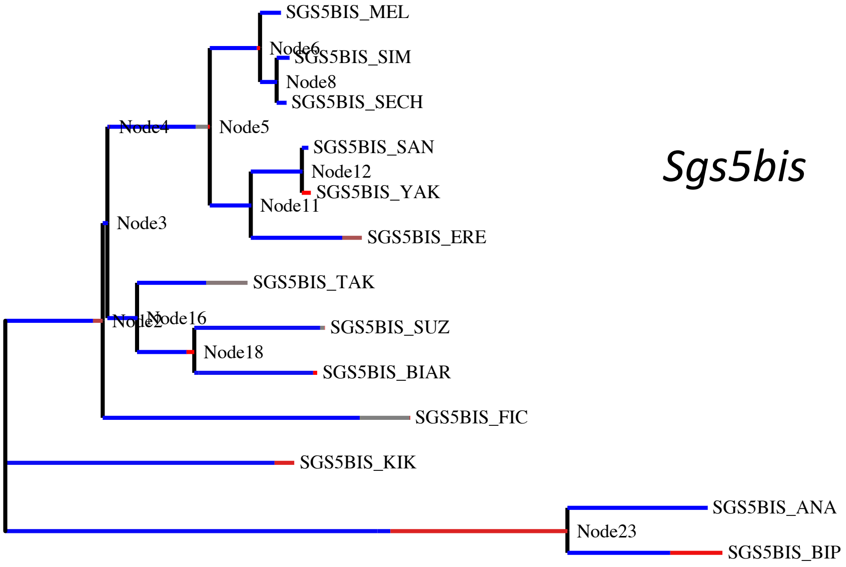

Sgs1
Sgs3
Sgs7
Sgs8
Sgs4
Sgs5bis
Sgs5
Eig71Ee
D. melanogaster
D. simulans
D. sechellia
D. mauritiana
D. santomea
D. yakuba
D. erecta
D. eugracilis
D. takahashi
D. suzukii
D. biarmipes
D. elegans
D. rhopaloa
D. ficusphila
D. kikkawai
D. ananassae
D. bipectinata
D. pseudoobscura
D. willistoni
D. virilis
ancestor
+4
-5b
-8
+7
-1 -5
+3
-3 -3
-3
+7 +7 +7
+3
+1 +3
-5b -7 -8
+3
-7 -8
-7 -8
-7 -8
+5 +8
-7 +3
+7
-5b +7
R
R?
?

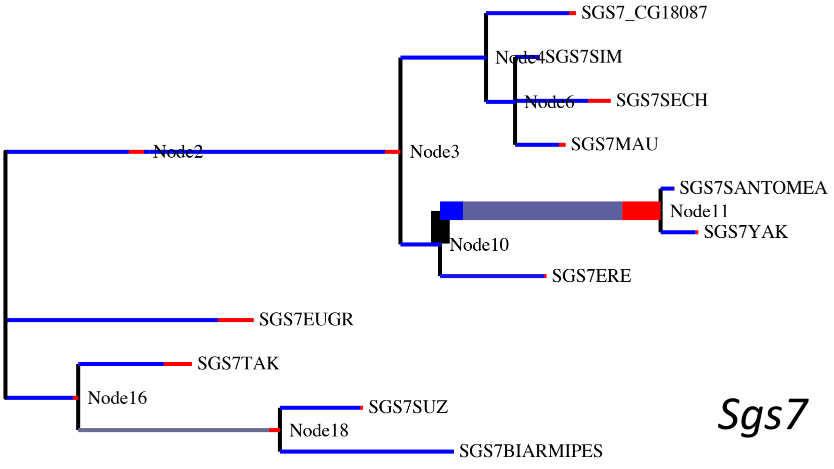

Tree scale: 0.1
sgs5bis bipectinata
sgs5bis ananassae
sgs5bis kikkawai
sgs5bis takahashii
sgs5bis eugracilis
sgs5bis suzukii
sgs5 bis biarmipes
sgs5bis ficusphila
sgs5bis yakuba
sgs5bis santomea
sgs5bis erecta
sgs5bis CG7587
sgs5bis simulans
sgs5bis sechellia
sgs5 virilis
sgs5bis pseudoobscura
sgs5 kikkawai
sgs5 bipectinata
sgs5 ananassae
sgs5 ficusphila
sgs5 rhopaloa
sgs5 elegans
sgs5 suzukii
sgs5 biarmipes
sgs5 eugracilis
sgs5 takahashii
sgs5 santomea
sgs5 yakuba
sgs5 CG7596
sgs5 sechellia
sgs5 mauritiana
sgs5 simulans
1
0.51
0.82
0.82
0.84
0.25
0.6
0.99
0.38
0.9
0.034
0.93
0.94
0.99
1
0.5
0.59
0.6
1
0.93
0.9
0.8
0.82
0.95
0.71
0.99

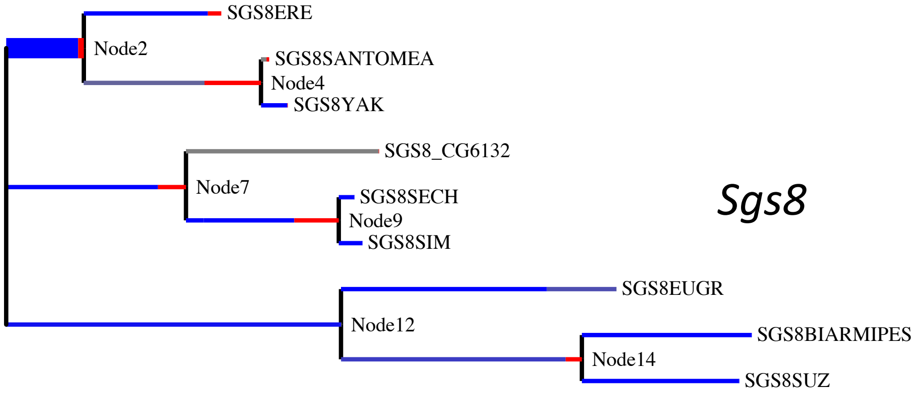

sgs3 ananassae
sgs3bis kikkawai
sgs3 willistoni
sgs3 bipectinata
sgs3 biarmipes
sgs3 rhopaloa
sgs3bis elegans
sgs3 bis ficusphila
sgs3 takahashii
sgs3ter elegans
sgs3bis rhopaloa
sgs3bis ananassae
sgs7 willistoni
sgs7bis willistoni
sgs3-like willistoni
sgs7 bipectinata
sgs8 bipectinata
sgs3 erecta
sgs3 santomea
sgs3 yakuba
sgs3-CG11720
sgs3 mauritiana
sgs3 simulans
sgs3 sechellia
sgs3bis eugracilis
sgs3 ter ficusphila
sgs3bis bipectinata
sgs3 kikkawai
sgs3bis virilis
sgs3 virilis
sgs3ter pseudoobscura
sgs3 pseudoobscura
sgs3bis pseudoobscura
sgs3bis biarmipes
sgs3 eugracilis
sgs3 elegans
sgs3 ficusphila
sgs7 mauritiana
sgs7 sechellia
sgs7-CG18087
sgs7 simulans
sgs7 erecta
sgs7 yakuba
sgs7bis yakuba
sgs7bis santomea
sgs7 santomea
sgs7bis8 takahashii
sgs7 takahashii
sgs7-2 suzukii
sgs7-3 suzukii
sgs7-4 suzukii
sgs7 biarmipes
sgs7-1 suzukii
sgs8 biarmipes
sgs8 suzukii
sgs7 eugracilis
sgs8 eugracilis
sgs8 sechellia
sgs8 simulans
sgs8-CG6132
sgs8 santomea
sgs8 yakuba
sgs8 erecta


sgs1
sgs3
sgs4
Eig71Ee
PONDR Score
Residue Number
Disorder
Order
VL-XT
0.0
0.2
0.4
0.5
0.6
0.8
1.0
0
100
200
300
400
25
50
75
125
150
175
250
A
D. simulans
sgs5
D. kikkawai
sgs5
350
B


2500
2000
1500
1000
500
0
🟦 dN
🟥 dS
🟩 dN/dS
Sgs5bis
Sgs5
Sgs7
Eig71Ee
Sgs1
Sgs8
Sgs4
Sgs7
Sgs5
Sgs5bis
Sgs4
Eig71Ee
Sgs8
Sgs1
0,00
0,05
0,10
0,15
0,20
0,25
0,30
0,35
0,40
0,45
0,50
0,55
0,60
0,65
0,70
0,75
0,80
0,85
0,90
0,95
1,00
1,05
1,10
1,15
1,20
1,25
1,30
1,35
1,40
1,45
1,50
substitution rate


Sgs5bis
SGS5BIS_MEL
Node6
SGS5BIS_SIM
Node8
SGS5BIS_SECH
Node4
Node5
SGS5BIS_SAN
Node12
SGS5BIS_YAK
Node11
SGS5BIS_ERE
Node3
SGS5BIS_TAK
Node2
Node16
SGS5BIS_SUZ
Node18
SGS5BIS_BIAR
SGS5BIS_FIC
SGS5BIS_KIK
SGS5BIS_ANA
Node23
SGS5BIS_BIP


SGS7_CG18087
Node4
SGS7SIM
Node6
SGS7SECH
SGS7MAU
Node2
Node3
SGS7SANTOMEA
Node11
SGS7YAK
Node10
SGS7ERE
SGS7EUGR
SGS7TAK
Node16
SGS7SUZ
Sgs7
Node18
SGS7BIARMIPES


SGS8ERE
Node2
SGS8SANTOMEA
Node4
SGS8YAK
SGS8_CG6132
Node7
SGS8SECH
Sgs8
Node9
SGS8SIM
SGS8EUGR
Node12
SGS8BIARMIPES
Node14
SGS8SUZ