Genomic signatures of globally enhanced gene duplicate accumulation in the megadiverse higher Diptera fueling intralocus sexual conflict resolution

Riyue Bao[1,2] and Markus Friedrich[3,4]

[1] Hillman Cancer Center, University of Pittsburgh, Pittsburgh, PA, USA
[2] Department of Medicine, University of Pittsburgh, Pittsburgh, PA, USA
[3] Department of Biological Sciences, Wayne State University, Detroit, MI, USA
[4] School of Medicine, Department of Anatomy and Cell Biology, Wayne State University, Detroit, MI, USA

ABSTRACT

Gene duplication is an important source of evolutionary innovation. To explore the relative impact of gene duplication during the diversification of major insect model system lineages, we performed a comparative analysis of lineage-specific gene duplications in the fruit fly *Drosophila melanogaster* (Diptera: Brachycera), the mosquito *Anopheles gambi*ae (Diptera: Culicomorpha), the red flour beetle *Tribolium castaneum* (Coleoptera), and the honeybee *Apis mellifera* (Hymenoptera). Focusing on close to 6,000 insect core gene families containing maximally six paralogs, we detected a conspicuously higher number of lineage-specific duplications in *Drosophila* (689) compared to *Anopheles* (315), *Tribolium* (386), and *Apis* (223). Based on analyses of sequence divergence, phylogenetic distribution, and gene ontology information, we present evidence that an increased background rate of gene duplicate accumulation played an exceptional role during the diversification of the higher Diptera (Brachycera), in part by providing enriched opportunities for intralocus sexual conflict resolution, which may have boosted speciation rates during the early radiation of the megadiverse brachyceran subclade Schizophora.

Subjects Biodiversity, Entomology, Genetics, Genomics, Zoology
Keywords Gene duplication, *Drosophila*, Brachycera, Genome evolution, Energy metabolism, Sexual conflict resolution, Speciation, Hexokinase, Hsp60, Thioredoxin

Submitted 15 January 2020
Accepted 31 August 2020
Published 12 October 2020

Corresponding author
Markus Friedrich,
ag7274@wayne.edu

Academic editor
Rodrigo Nunes Fonseca

Additional Information and Declarations can be found on page 28

DOI **10.7717/peerj.10012**

INTRODUCTION

A total of 50 years ago now, gene duplications became recognized as a major source of evolutionary innovation at the genetic level (*Ohno, 1970*). One hallmark validation of this conceptual advancement was the discovery of the Hox gene cluster, the deeply conserved string of tandem-duplicated transcription factor genes, which regulate the patterning of the longitudinal animal body axis (*Lewis, 1992*; *Carroll, 1995*; *Ryan et al., 2007*; *He et al., 2018*). Today, genome sequencing studies continue to uncover adaptive gene family expansions facilitated by gene duplication, which have affected a wide range of phenotype dimensions such as body plan innovation (*Cañestro, Yokoi & Postlethwait, 2007*; *Holland et al., 2017*; *Sakuma et al., 2019*), sensory reception (*Smadja et al., 2009*), sexual

reproduction (*Connallon & Clark, 2011*), or diet range (*Zhao et al., 2015*; *Pajic et al., 2019*), to list a few. The availability of whole genome sequences has made it also possible to quantify the global impact of gene duplication on the genetic evolution of organismal lineages. The first comprehensive study in this direction revealed that gene duplications occur at a frequency of close to 0.01 per gene per million years (*Lynch & Conery, 2000*), thus ranking equivalent with other mutational mechanisms in generating genetic variation in natural populations. The same work confirmed the expectation that most newly born gene duplicates experience rapid decay into pseudogenes. Recent studies, however, also produced evidence for a generic advantage of nascent gene duplicates in buffering gene expression noise (*Rodrigo & Fares, 2018*). Notwithstanding the nature of forces acting upon new gene duplicates, the opposing effects of gene duplication and loss events have been found to lead to rapid rates of gene turnover, which can translate into dynamic gene family size evolution(*Hahn, Han & Han, 2007*). There is, however, also a long trail of gene duplicates, which become long-term preserved due to either the complementary partitioning of ancestral pleiotropy (subfunctionalization) between sister paralogous gene duplicates or the acquisition of novel functions (neofunctionalization) by one of them (*Force et al., 1999*; *Innan & Kondrashov, 2010*). In addition, also benefits arising from genetic redundancy such as the suppression of regulatory noise can lead to hundreds of millions of years of conservation of overlapping ancestral functions in duplicated genes (*Conant & Wagner, 2004*; *Gu et al., 2003*; *Hanada et al., 2009*; *Hsiao & Vitkup, 2008*; *Tischler et al., 2006*; *Vavouri, Semple & Lehner, 2008*; *Friedrich, 2017*).

Despite these fundamental insights, much remains to be learned about the processes and mechanisms that lead to gene duplicate fixation and long-term conservation. Also the diversity, frequencies, and adaptive significance of gene duplication outcomes remain an area of continued progress in comparative genomics. In this context, it is the dichotomy of conservative or neutral vs innovative gene duplication outcomes, such as sub- vs neofunctionalization, which remains of fundamental interest in molecular evolution (*Van Hoof, 2005*; *Dean et al., 2008*; *Kondrashov, 2012*; *Qian & Zhang, 2014*; *Wang et al., 2016*; *Lan & Pritchard, 2016*; *Holland et al., 2017*; *MacKintosh & Ferrier, 2017*; *Marlétaz et al., 2018*; *Sandve, Rohlfs & Hvidsten, 2018*).

Arguably the most dramatic examples of how gene duplications provided the genetic substrate for expansions of organismal complexity and diversification are the gene duplicate enriched genomes of angiosperms in plants and of vertebrates in animals, both of which date back to several rounds of whole genome duplications (*Jiao et al., 2011*; *Cañestro et al., 2013*). A whole genome duplication has also been discovered at the base of the highly diversified arthropod subphylum Chelicerata followed by additional whole genome duplications in younger clades of this group (*Schwager et al., 2017*; *Nong et al., 2020*). At the same time, current data do not speak for an obligatory connection between gene duplication and taxonomic diversification. With close to one million documented species, insects stand at the forefront of understanding the origins of organismal diversity (*Labandeira & Sepkoski, 1993*; *Grimaldi & Engel, 2005*). However, comparative genomic evidence now solidly rules out that whole genome duplications preceded the unparalleled expansion of this group. Seeded by the *Drosophila* genome project (*Myers et al., 2000*),

the genomic exploration of insect diversity has amounted to over 100 sequenced genomes within the last 15 years (*Yin et al., 2016*). This progress is the result of targeted efforts such as i5k initiative, which strives for the completion of 5,000 high-priority arthropod genomes (*i5K Consortium, 2013*). The analysis of the first 28 genomes resulting from this effort together with 48 additional arthropod genomes suggests that gene duplicate accumulation rates remained remarkably steady over the about 450 million years of insect evolution (*Thomas et al., 2020*). The preceding analysis of the gene content in genome or transcriptome data sets of over 150 species representing all 41 insect orders, in contrast, reported evidence of substantial fluctuations in gene duplicate contents in over 15 insect orders (*Li et al., 2018*). While the strength of evidence for whole genome duplications has been refuted (*Nakatani & McLysaght, 2019*; *Roelofs et al., 2020*), these findings still leave room for the possibility that dramatically enhanced local gene duplicate retention rates played important roles in the exceptional diversification of insects.

Further noteworthy in this respect is that fact that the first reported insect genome with a dramatically higher number of lineage-specific gene duplications (close to 2,500), that of the pea aphid *Acyrthosiphon pisum* (*International Aphid Genomics Consortium, 2010*; *Julca et al., 2020*), failed to be detected in the recent study mentioned above (*Li et al., 2018*). This suggests that continued efforts are likely to refine our understanding of important gene family content changes in insect evolution. Studying the evolutionary histories of vision-related genes in insect genome models, we previously noted an unusually high number of duplicated genes in *Drosophila melanogaster* (*Bao & Friedrich, 2009*). In a followup study of over 350 developmental gene families, we discovered a preponderance of ancient, yet lineage-specific gene duplicates in *Drosophila* and the higher Diptera (Brachycera) (*Bao et al., 2018*). Moreover, more than 50% of these lineage-specific developmental gene duplications retained partial or complete genetic redundancy despite their ancient separation. This led us to hypothesize that the exceptional accumulation of developmental gene duplicates in *Drosophila* and the higher Diptera was of adaptive nature by increasing genetic robustness as a requirement for the fast development of brachyceran Diptera such as *Drosophila*. At the same time, the similarly higher number of structural vision genes gave reason to suspect the possibility of a genome-wide increase of gene duplicate accumulation in the higher Diptera (*Bao & Friedrich, 2009*; *Bao et al., 2018*).

As global studies of insect gene family contents did not produce evidence of an overall higher gene content in *Drosophila* vs non-dipteran insects (*Honeybee Genome Sequencing Consortium, 2006*; *Richards et al., 2008*; *Thomas et al., 2020*), we compared the numbers of lineage-specific gene duplicates in insect core gene families of small to moderate size (<6 paralogs). This approach was meant to eliminate the effect of adaptive gene family expansions in order to quantify the relative amounts of gene duplicate accumulation resulting from the average background rate of physical gene duplication events followed by successful fixation and longterm conservation. As a first step in this direction, we analyzed the gene contents of three distantly related, well-curated holometabolous insect genome species in comparison to *D. melanogaster*. The results from this four species comparison indicate that the megadiverse dipteran infraorder Brachycera is characterized by a genome-wide higher rate of gene duplicate accumulation.

Gene ontology analysis, however, further indicates that energy metabolism genes were exceptionally affected by this trend during the diversification of schizophoran Diptera, that is, the most recent of three major radiations in this insect order, between 40 and 60 million years ago (*Wiegmann et al., 2011*). Almost invariably, these lineage-specific energy metabolism gene duplications spawned germline-specific paralogs thereby resolving conflicting selection pressures on their ancestral singleton loci. Given the theoretical and empirical evidence that the emergence of germline-specific gene duplicates enforces species barriers, our findings point to a potentially important link between gene duplication and speciation rates in the higher Diptera in addition to documenting a higher global gene duplicate accumulation rate in this clade.

METHODS AND MATERIALS

Genome and sequence datasets

The genome assemblies used in this study were *Drosophila melanogaster* genome assembly 5.3 (*Myers et al., 2000*), *Mayetiola destructo*r genome assembly 1.0 (*Zhao et al., 2015*), *Anopheles gambiae* str. PEST genome database version 3.0 (*Sharakhova et al., 2007*), *Tribolium castaneum* Georgia GA2 genome database version 3.0 (*Richards et al., 2008*), and *Apis mellifera* DH4 genome database version 4.0 (*Honeybee Genome Sequencing Consortium, 2006*). The protein databases used in this study were GenBank RefSeq protein databases of *D. melanogaster*, *A. gambiae*, *T. castaneum*, and *A. mellifera* (*Pruitt, Tatusova & Maglott, 2005*) and Official gene set (OGS) protein databases of *M. destructor* (version 1.0) (*Zhao et al., 2015*), *T. castaneum* (version 1.0) (*Richards et al., 2008*), and *A. mellifera* (version 2.0) (*Honeybee Genome Sequencing Consortium, 2006*). The NCBI RefSeq and OGS protein sequence databases of *T. castaneum* and *A. mellifera* each were merged to create more comprehensive protein datasets. Four-way pairwise BLASTP searches were performed between the protein databases of *D. melanogaster*, *A. gambiae*, *T. castaneum*, and *A. mellifera*. *M. destructor* homologs were retrieved by searching *D. melanogaster* query proteins against *M. destructor* OGS protein database. In addition, *A. gambiae* genome sequences were downloaded from the GenBank RefSeq nucleotide database (*Pruitt, Tatusova & Maglott, 2005*) and searched against using *D. melanogaster* proteins as the query by TBLASTN (*Altschul et al., 1990*) to retrieve possible mosquito homologs not annotated in its RefSeq protein database. Additional genome databases interrogated in the analysis of the brachyceran enriched metabolism genes included that of the Mediterranean fruit fly *Ceratitis capitata* (*Papanicolaou et al., 2016*), the stalk-eyed fly *Teleopsis dalmanni* (*Vicoso & Bachtrog, 2015*), the calyptrate Diptera *Musca domestica* (common house fly) (*Scott et al., 2014*) and *Glossina morsitans* (tsetse fly) (*International Glossina Genome Initiative, 2014*), the onion fly *Delia antiqua* (*Zhang et al., 2014*), the robber fly species *Proctacanthus coquilletti* (*Dikow et al., 2017*), the bibionomorph species *Contarinia nasturtii* (swede midge) (GenBank assembly GCA_009176525.2) and *Sitodiplosis mosellana* (orange wheat blossom midge) (GenBank assembly GCA_009176505.1), the mosquito species *Aedes aegypti* (yellow fever mosquito) (*Nene et al., 2007*) and *Culex quinquefasciatus* (mosquito species) (*Pelletier & Leal, 2009*), and the

sandfly species *Lutzomyia longipalpis* (GCA_000265325.1) and *Phlebotomus papatasi* (GCA_000262795.1).

Duplicate detection and classification pipeline

Gene duplication and classification were conducted by adopting the pipeline developed in our previous analysis of developmental gene duplicates (*Bao et al., 2018*). In short, species-specific protein databases (RefSeq and OGS when available) were downloaded from GenBank or respective genome project websites. For each species, databases from different sources were merged based on identical associated GeneID suggested by the Gbrowser to obtain a final protein database void of redundant sequences. Inter- and intra-species BLAST searches were performed with protein sequences of each species as queries against the database of itself or from the other species, with *E*-value cutoff set as $1.0e^{-4}$ (*Altschul et al., 1990*).

Next, gene families were sorted into three classes, (1) 1:1 orthology if the gene had 1:1 orthologs in at least two additional species, (2) ancient duplication if it resulted from a duplication that happened before the insect diversification, (3) lineage-specific duplication if this gene has independent duplication(s) occurred in any of the four species. This classification was achieved in two steps. In the first step, ortholog numbers for a given gene family of each species were determined by the following logic. Assuming there are a, b, c, and d numbers of orthologs corresponding to 1 query gene in each of the four species, $a = b = c \geq d$ defined 1:1 orthology if each a, b, and c equaled 1 and as ancient duplication if a, b, and c were larger than 1. The condition $a > b = c \geq d$, by contrast, defined a family associated lineage-specific duplication in the species with more than one ortholog. The approximately 10% of gene families that did not fall into any of the three categories above were classified as unresolved groups and further analyzed in the second part of the classification analysis, which involved manual assignment of duplication labels to each gene family based on gene tree analysis results.

Gene family validation

In order to compare the validation data from our previous analysis of a manually curated subset of 377 validation gene families (*Drăghici, 2011*; *Bao et al., 2018*) with the genome-wide duplication identification output (Supplemental Data File 2), every individual gene in the two data sets was assigned with a with a six-digit code, which reflected its inferred gene duplication history. Genes with 1:1 orthologs in all four lineages were assigned with the code 100000. Genes associated with duplication events that were shared by two or three lineages were assigned with code 010000. Code positions 3–6 indicated lineage-specific duplication events in one of the four lineages. As an example, code variant 001010 represented a gene that had independently duplicated in the lineages to *Drosophila* and *Tribolium*. In the next step, genes present in both datasets were compared directly based on their assigned duplication labels and classified into four comparison results based on each position of the six digits: (1) False positive, if the gene appeared negative in the validation dataset but positive in the genome-wide analysis; (2) False negative for genes that were positive in the validation dataset but negative in the

study; (3) True positive for genes that were positive in both results; (4) True negative for genes associated with negative codes in both datasets. Each gene was calculated redundantly in each position of the six digits, meaning that if one gene was positive for 1:1 orthology, it was counted as zero for the other five positions.

Calculation of dS values

Protein sequences of duplicated genes were aligned with ClustalW2 (*Larkin et al., 2007*) and the corresponding cDNA alignments were generated by PAL2NAL version 13 (*Suyama, Torrents & Bork, 2006*) using the aligned protein sequences as input. Ambiguously aligned regions were removed by Gblocks 0.91b (*Talavera & Castresana, 2007*) with the block parameter "Allowed gap positions" set to "None". The gap-filtered cDNA alignments were used to calculate synonymous substitution divergence values (dS) with the yn00 algorithm of PAML version 4.4 (*Yang, 2007*). In the case of multiple duplications within the same gene family, dS values were determined for all paralog combinations and the smallest and largest values were selected for the analysis of gene duplicate age distribution.

Phylogenetic tree reconstruction

For the genome-wide survey, ClustalW2 (*Larkin et al., 2007*) was used to generate the multiple sequence alignments, which were subsequently purged of ambiguously aligned sites and divergent regions with Gblocks (*Castresana, 2000*) applying least stringency settings, before neighbor gene gene tree reconstruction with JTT protein substitution model distances executed in Phylip package version 3.69 on Wayne State Grid supercomputing cluster (*Saitou & Nei, 1987*; *Whelan & Goldman, 2001*; *Felsenstein, 2005*). Specifically, the Seqboot module created 100 bootstrap replicates from the protein alignment input, followed by "Protdist" calculation of protein distance matrices for each bootstrap dataset applying JTT model, followed by neighbor joining tree generation for each dataset with the "Neighbor" module, and consensus tree calculation through the "Consense" module.

For the expanded analyses of gene family evolution, we collected homologs via reciprocal BLAST in the species-specific protein sequence, whole genome sequence, and TSA databases at NCBI. Gene trees were estimated with RAxML as implemented in the CIPRES Science Gateway environment from multiple protein sequence alignments generated with Clustal followed by ambiguous site removal with Gblocks at least stringent settings (*Miller, Pfeiffer & Schwartz, 2010*; *Sievers et al., 2011*; *Stamatakis, 2014*). Tree rendering and annotation was conducted in the iTOL environment (*Letunic & Bork, 2016*).

Gene ontology analysis

The distribution of functional categories associated with lineage-specific duplication gene sets was evaluated by analysis of Gene Ontology (GO) terms (*Ashburner et al., 2000*) using the BiNGO tool (*Maere, Heymans & Kuiper, 2005*) from Cytoscape package v. 2.8 (*Smoot et al., 2011*) with customized settings: study set and reference set as listed in (Supplemental Data File 5), annotation file as the FlyBase version downloaded from Gene

Ontology official website (http://www.geneontology.org/GO.downloads.annotations.shtml, updated 08/04/2011), and ontology file as OBO version 2.0 downloaded from Gene Ontology official website (http://www.geneontology.org/GO.downloads.ontology.shtml, updated 08/04/2011), with namespace chosen as Biological Process. The statistical test was set to hypergeometric test, and multiple testing correction set to Benjamini & Hochberg False Discovery Rate (FDR) correction (*Benjamini & Hochberg, 1995*). The significance level was chosen as 0.05. GO annotations of non-*Drosophila* genes were assigned based on their orthologs in *Drosophila*. Afterwards, representative GO terms were detected and visualized by removing redundant GO terms using REViGO (*Supek et al., 2011*) with the list of significant GO terms with their associated FDR corrected *p*-values as the input, the resulting list set to small (cutoff clustering score = 0.5), the database with GO term sizes set as *Drosophila melanogaster* UniProt (*McGarvey et al., 2019*), and the semantic similarity measure method set as SimRel (*Schlicker et al., 2006*).

RESULTS

Drosophila is two-fold richer in duplicated insect core genes compared to mosquito, flour beetle, and honeybee

To probe for a genome-wide increase in gene duplicate accumulation in the lineage leading to *Drosophila*, we compared the gene family content of *D. melanogaster* with three additional insect genome model species: the mosquito *Anopheles gambiae* (Diptera: Culicomorpha), the red flour beetle *Tribolium castaneum* (Coleoptera), and the honeybee *Apis mellifera* (Hymenoptera). To normalize the data sets, we excluded species-specific orphan gene families and restricted the analysis to gene families that were conserved in at least three of the four species. To compare base rates of gene duplicate accumulation, we excluded gene families with more than six members to eliminate the effect of massively adaptive gene family expansions. In combination, these criteria narrowed the comparison to 5,983 insect core gene families conserved in *Drosophila*. Of these, 5,581, 5,505, and 5,448 were recovered in mosquito, red flour beetle, and the honeybee, respectively (Supplemental Data File 1).

Lineage-specific gene family expansions were inferred via a previously described bioinformatic pipeline involving sequence similarity threshold filtering, reciprocal BLAST tests, and, in a subset of cases, gene tree reconstruction (*Bao et al., 2018*). Running a comparison to our previously manually curated sample of 377 developmental gene families for validation analysis (*Drăghici, 2011*; *Bao et al., 2018*), we obtained evidence of 90% or better accuracy (percentage of true positives) and 85% or better specificity (percentage of true negatives) for our automated gene family analysis pipeline (Supplemental Data File 2).

In total, our bioinformatic approach detected 1,211 gene families with high confidence lineage-specific gene duplications (Supplemental Data File 1). A total of 190 of these were characterized by independent gene duplication events in more than one of the four lineages. At the species level, we found 698 gene families with lineage-specific duplications in *Drosophila*, 315 in *Anopheles*, 386 in *Tribolium*, and 223 in *Apis* (Fig. 1). Normalized for the species differences in homolog contents, 11.7% of insect core gene families were

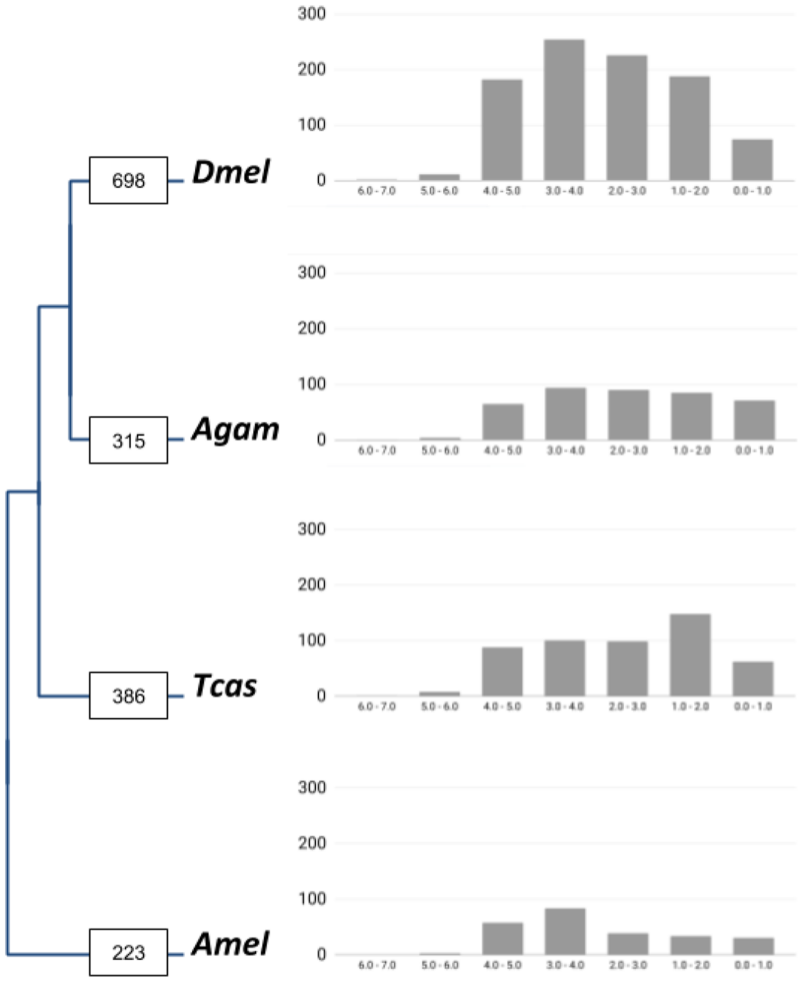

Figure 1 Lineage-specific core gene duplicate numbers across holometabolous insect genome model species. Boxes on terminal branches in the species tree report the numbers of gene families with lineage-specific gene duplications: Dmel, *Drosophila melanogaster*; Agam, *Anopheles gambiae*; Tcas, *Tribolium castaneum*; Amel, *Apis mellifera*. Bar charts depict distributions of lineage-specific gene duplicates by age based on neutral substitution divergence (dS) estimates with *Y*-axis representing the number of duplicates associated with each dS value bin. Full-size DOI: 10.7717/peerj.10012/fig-1

identified as expanded via lineage-specific gene duplications in the *Drosophila* lineage, which compared to 5.6%, 6.7%, and 4.1% in *Anopheles*, *Tribolium*, and *Apis*, respectively (Table 1). This on average 2-fold higher number of lineage-specific gene family expansions in *Drosophila* was consistently reflected in the large category of post-duplication 2-members large gene families and even more strongly in the gene families with 3 up to 6 extant paralogs (Table 1). Falling notably in line with the previously detected two times higher number of developmental gene duplications in the *Drosophila* lineage in the same comparative framework (*Bao et al., 2018*), these findings supported a genome-wide increase in gene duplicate accumulation as explanation for the previously detected higher numbers of both vision and developmental gene duplications in *Drosophila* vs *Anopheles*, *Tribolium*, and *Apis* (*Bao & Friedrich, 2009*; *Bao et al., 2018*).

Table 1 Lineage-specific gene duplications in insect core gene families. Species name abbreviations as in Fig. 1. Species-specific cell entries represent percentages of lineage-specific expanded gene families sorted by post-duplication size (2–6). Total percentages of lineage-specific expanded gene families given with absolute numbers and number of sampled gene families in parentheses.

Gene family sizes	Dmel	Agam	Tcas	Amel	Dmel/Others
2	7.9% (478)	4.4% (246)	4.9% (285)	3.4% (185)	1.9
3	2.4% (144)	0.8% (42)	1.0% (59)	0.5% (29)	3.1
4	0.9% (51)	0.3% (19)	0.3% (20)	0.1% (6)	3.2
5	0.3% (18)	0.1% (5)	0.1% (10)	0.0001% (2)	3.0
6	0.1% (7)	0.1% (3)	0.2% (12)	0.02% (1)	1.3
Total	11.7% (698/5,983)	5.6% (315/5,581)	6.7% (386/5,505)	4.1% (223/5,448)	2.1

The majority of the *Drosophila*-lineage gene duplicates are old

To gain insight into the time dimensions of insect core gene duplicate accumulation in the four insect lineages, we surveyed the synonymous substitution differences (dS) between gene family paralogs as proxies of gene duplicate ages (Supplemental Data File 3). The majority of the duplicate pairs in all four lineages were associated with dS values higher than 2.0, indicative of relatively ancient origins (Fig. 1). This trend was most pronounced in the *Drosophila* lineage where 92.2% of the gene duplicate pairs were characterized by dS values higher than 2.0 compared to 82.7%, 87.9%, and 88.2% in *Anopheles*, *Tribolium*, and *Apis*, respectively.

While straightforward to compute, dS values are coarse estimates of evolutionary time dimensions, especially in the case of short gene sequences due to sample size errors. We therefore also probed for the conservation of the *Drosophila* lineage-specific gene duplications in the genome of the Hessian fly *Mayetiola destructor* (*Zhao et al., 2015*). This pest species is a representative of the dipteran infraorder Bibionomorpha, the now well established sister clade of the Brachycera (*Wiegmann et al., 2011*). Thus, duplications shared by *M. destructor* and *D. melanogaster* to the exclusion of *A. gambiae* would be diagnosed to be of pre-Brachyceran origin, while duplications unique to *D. melanogaster* to the exclusion of both *M. destructor* and *A. gambiae*, were more likely to have occurred at a later time point, that is, during the diversification of brachyceran flies.

Reciprocal BLAST searches recovered 385 (55.1%) of the 698 gene families with *Drosophila*-lineage specific duplications in *M. destructor* (Supplemental Data File 4). A total of 75 (18.5%) of these shared two or more 1:1 orthologs in the Hessian fly, implying a pre-brachyceran origin. For the remaining 185 gene families, we only detected singleton orthologs in the Hessian fly genome, characterizing them as Brachycera-specific. Thus taken together, interparalog dS divergences and gene duplicate conservation in the Hessian fly suggested that only about 20% of the close to 700 *Drosophila*-specific insect core gene duplications detected in our initial four-species comparison had accumulated before the split of the last ancestor of Hessian fly and *Drosophila*, while the majority originated later, during the expansion of the megadiverse Brachycera.

Enrichment of energy metabolism functions in the Brachycera-specific gene duplicates

To gain insights into possible phenotypic corollaries of the heightened number of duplicated insect core genes in brachyceran Diptera, we tested for enrichment of biological processes (BP) using GO analysis tools (*Ashburner et al., 2000*; *Dennis et al., 2003*; *The Gene Ontology Consortium, 2015*). We were able to apply this approach to 1,403, 841, 860, and 426 gene duplicates of *Drosophila*, *Anopheles*, *Tribolium*, and *Apis*, respectively, based on the BP information available in FlyBase at the time of the analysis (*Drysdale & Crosby, 2005*).

Biologically meaningful GO term enrichment signals indicated informativeness of the approach (Fig. 2; Supplemental Data File 5). The GO-term "chitin metabolic processes" (GO:0006030), for instance, was exclusively enriched in the gene duplications specific to the *Tribolium*-lineage, consistent with previously reported expansions of chitin metabolism gene families in the Coleoptera (*Arakane et al., 2005*; *Dixit et al., 2008*) and the generally chitin-enriched cuticle of darkling beetles (Tenebrionidae) like *Tribolium* (*Finke, 2007*). Furthermore, the GO terms "generation of precursor metabolites and energy" (GO:0006091) and "cellular carbohydrate metabolic process" (GO:0044262) were significantly enriched in all species except *Anopheles* (Fig. 2A; Supplemental Data File 5). This genomic signal boded well with the fact that mosquitoes, in many cases, subsist on a carbohydrate-poor diet of vertebrate blood and plant pollen in contrast to the generally carbohydrate-rich diets in the clades represented by *Drosophila*, *Apis*, and *Tribolium* (*Foster, 1995*).

Interestingly, the population of *Drosophila*-lineage gene duplicates was characterized by the lowest number of enriched BP GO-terms (43) but the highest number of underrepresented BP GO-terms (206) (Fig. 2; Supplemental Data 5). Even more remarkably, the latter category included many developmental GO terms (Fig. 2; Supplemental Data 5), despite our previous finding that *Drosophila* possesses a two-fold higher number of duplicated developmental genes compared to *Anopheles*, *Tribolium*, and *Apis* (*Bao et al., 2018*). However, developmental GO terms were generally underrepresented in the lineage-specific gene duplications (Fig. 2), consistent with evidence that developmental gene regulatory networks are less tolerant to gene duplication (*Davidson & Erwin, 2006*).

In the *Drosophila* population of lineage-specific gene duplicates, significantly enriched BP GO-terms were predominantly related to energy metabolism (Fig. 2; Supplemental Data 5). Given the distinct structural and physiological features of brachyceran Diptera (*Wiegmann et al., 2011*), we also analyzed BP GO-term enrichment separately for the subsets of 234 pre-brachyceran vs 1,167 identified Brachycera-specific gene duplications. This approach detected 39 significantly enriched GO terms in the population of Brachycera-specific gene duplications all of which were related to energy metabolism. The 25 significantly enriched BP GO-terms in the population of pre-brachyceran gene duplications, in contrast, represented different categories of biological function (Fig. 2; Supplemental Data File 5).

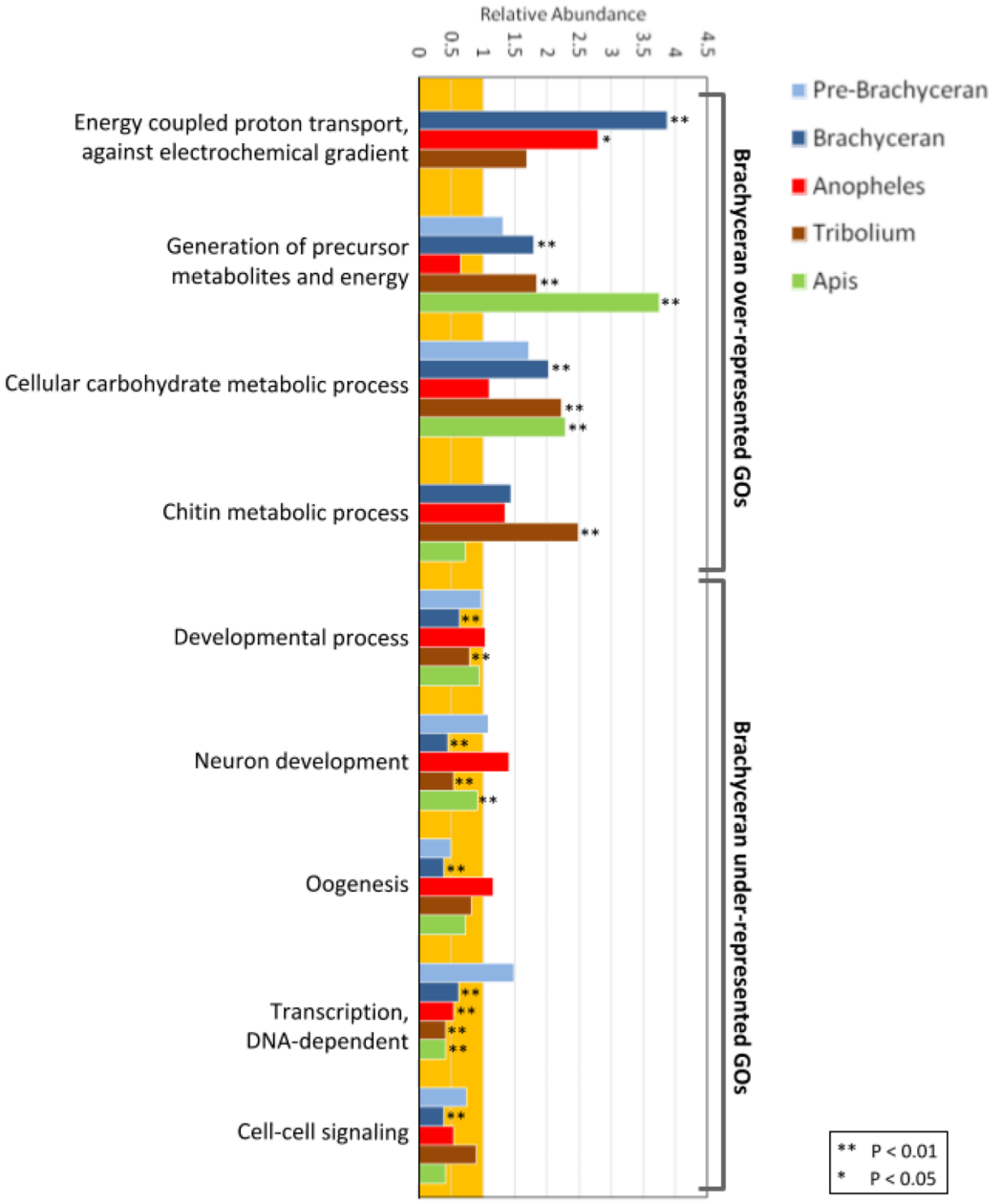

Figure 2 Functional enrichment analysis of lineage-specific gene duplicates. Distribution of GO functional categories in the duplicates of *Drosophila melanogaster* (Pre-Brachyceran and Brachyceran), *Anopheles gambiae*, *Tribolium castaneum*, and *Apis mellifera*. Bars represent abundance of genes associated with each listed individual GO term in the duplicates relative to those among all investigated genes (both singletons and duplicates) from each species. Relative abundance equal to one (background orange range) indicates that the GO term abundance in the duplicated genes is comparable to that in all genes, while a value lower than or higher than one indicates under- or over-representation, respectively. Brachyceran over-represented GO terms refer to the group of GO terms enriched in Brachycera-specific duplicates based on analysis of homolog conservation in *M. destructor*.

Full-size DOI: 10.7717/peerj.10012/fig-2

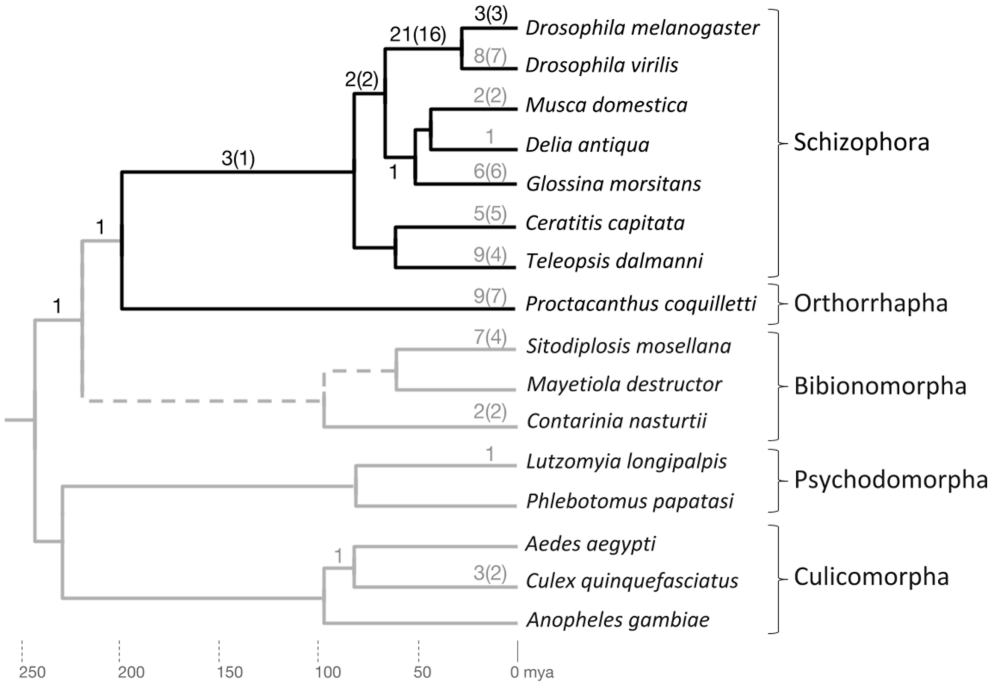

Figure 3 Phylogenetic accumulation of brachyceran energy metabolism gene duplicates. Summary of gene duplication time points based on homolog conservation in the species included in gene tree analysis. The Brachycera portion of the tree is indicated by dark gray branches. Dark gray numbers at branches indicate duplication events that generated conserved duplicates in the lineage to *D. melanogaster*. Light gray numbers at branches indicate parallel gene duplication events. Topology and branching time points based on *Wiegmann et al. (2011)*. mya, million years ago. Full-size DOI: 10.7717/peerj.10012/fig-3

Schizophora-concentrated origins of the *Drosophila* lineage energy metabolism-related gene duplicates

To explore the biological significance of energy metabolism-related gene duplications in brachyceran flies, we focused on gene duplicates in BP GO term categories that were significantly enriched in the Brachycera-specific gene duplications but not in any of the other investigated lineages (Supplemental Data 5). This condition was met for the BP GO terms "cell redox homeostasis" (GO:0045454), "protein targeting to mitochondrion" (GO:0006626), "protein localization in mitochondrion" (GO:0070585), "establishment of protein localization in mitochondrion" (GO:0072655), "carbohydrate phosphorylation" (GO:0046835), "dicarboxylic acid metabolic process" (GO:0043648), and "polyol metabolic process" (GO:0019751). Combined, these gene function populations constituted 189 *Drosophila* genes in 111 gene families of which, based on our initial pipeline results, 17 gene families had expanded due to Brachycera-specific gene duplications.

To scrutinize the predicted absence of these duplications outside Brachycera, we searched the genomes of additional mosquito (Suborder Culicomorpha), sandfly (Suborder Psychodomorpha), and gall midge (Suborder Bibionomorpha) species for homologs (Fig. 3). Most sampled gene families were only represented by singleton orthologs in these non-brachyceran species (Supplemental File Data 6). In two cases, however, that is, the *Thioredoxin* and *Malic enzyme* gene families, the manual homolog search raised

the initial gene family count from one to two with separate orthologs in the non-dipteran outgroups (Supplemental Data 6). Both of the *Malic enzyme* gene subfamilies contained *Drosophila*-lineage duplicates, thus increasing the number of de facto investigated gene family expansions to 18.

Only one duplication event mapped outside Brachycera to the last common ancestor (LCA) of Brachycera and Bibionomorpha (see below) (Supplemental Data File 6). Overall, the manual homolog conservation analysis confirmed the Brachycera-specificity of the energy-metabolism related gene duplications.

To explore the taxonomic depths of the metabolic gene duplications within brachyceran Diptera, we searched for homologs in a broader sample of schizophoran Diptera (Fig. 3; Supplemental Data File 6). Moreover, to differentiate between gene duplicates that dated back to the stem lineage of brachyceran Diptera vs duplicates that originated subsequently during brachyceran diversification, we searched the high coverage genome of the robber fly species *Proctacanthus coquilletti* as a representative of orthorrhaphan Brachycera (Fig. 3; Supplemental Data File 6) (*Dikow et al., 2017*).

Mapping the taxonomic distribution of gene duplicate conservation on this sampled framework of bachyceran phylogeny (*Bao et al., 2018*), we found that the majority of the energy metabolism-related gene duplications in the lineage to *Drosophila*, that is, 21 out of 32 (66%), originated in the lineage from the LCA of calyptrate + drosophilid Diptera to that of drosophilid Diptera (*D. melanogaster* and *D. virilis*) (Fig. 3). In addition, five further duplications mapped into the schizophoran clade of brachyceran Diptera: Three to the terminal branch of *D. melanogaster* representing the Sophophora subgroup of drosophilid Diptera and two back deeper to the lineage from the LCA of schizophoran Diptera to the LCA of calyptrate + drosophilid Diptera (Fig. 3).

Five duplications, finally, preceded the radiation of schizophoran Diptera (Fig. 3; Supplemental Data 6). Three of these stemmed from duplication events in a single gene family, generating the four *Thioredoxin* gene family paralogs of *D. melanogaster* that mapped to the relatively long, taxonomically still undersampled branch linking the LCAs of schizophoran and brachyceran Diptera (see below and Supplemental Data File 6). Only one duplication mapped outside Brachycera to the LCA of Brachycera and their sister clade, the Bibionomorpha (see below, Fig. 3; Supplemental Data 6). Overall, the taxonomic distribution of gene duplicate conservation was in line with the general ancientness of the Brachycera-specific gene duplications indicated by the interparalog dS divergences (Fig. 1) but also revealed the mostly schizophoran origins of the analyzed *D. melanogaster* energy metabolism gene duplicates.

Increased duplication accumulation rate in the energy metabolism gene population during schizophoran diversification

The concentration of energy metabolism gene duplications in the schizophoran lineage to the LCA of Drosophilidae differed from the previously reported, more evenly distributed accumulation of developmental gene duplications in brachyceran Diptera. Specifically, close to 85% of the duplications in energy metabolism-related gene families mapped into the schizophoran clade of dipteran phylogeny (Fig. 3) in contrast to close to 55% of

Table 2 Comparison of duplicate accumulation rates in developmental and energy metabolism gene families. Rates calculated as average numbers of duplications per gene family per million years. See Fig. 4 for branch definitions. Time intervals based on *Wiegmann et al. (2011)* and *Obbard et al. (2012)*.

Branches	Energy	Development	Time (My)
Drosophilidae LCA to present:	0.0009	0.0007	30
Drosophilidae + Calyptratae to Drosophilidae LCA:	0.0054	0.0011	35
Schizophora to Drosophilidae + Calyptratae LCA:	0.0012	0.0021	15
Brachycera to Schizophora LCA:	0.0003	0.0005	100
Brachycera + Bibionomorpha -> Brachycera LCA:	0.0002	0.0004	50
Average:	0.0016	0.001	

developmental gene duplications (*Bao et al., 2018*). One difference between the current analysis and the previous study of developmental gene duplications was the inclusion of the robber fly *P. coquilletti* to differentiate between gene duplicates that originated in the stem lineage of brachyceran Diptera vs the lineage preceding the diversification of schizophoran species (*Bao et al., 2018*). To make the data sets more comparable, we explored the conservation of 31 duplications in 25 developmental gene families that we had previously determined to have originated prior to the diversification of schizophoran Diptera. This was accomplished by probing for evidence of homolog conservation in the robber fly *P. coquilletti*. In addition, we scrutinized for the Brachycera-specificity of these duplications by searching for homologs in the gall midge *C. nasturtii* (Supplemental Data File 7; Fig. 3). This effort mapped 18 duplications to the lineage from the LCA of brachyceran Diptera to that of Schizophora, while eight duplications mapped to the brachyceran stem lineage and three to the even older LCA of Bibionomorpha and Brachycera (Supplemental Data File 7).

Normalizing the proportions of gene duplicates per total numbers of gene families sampled in the developmental and energy-related gene populations, that is, 377 vs 111 respectively, we generated estimates of duplicate accumulation rates in the two gene populations along the brachyceran lineages leading to *Drosophila* (Table 2). Along most branches, the accumulation rates appeared similar, not exceeding 2-fold differences and averaging 0.0016 vs 0.001 duplications per gene family per million years for energy metabolism vs developmental genes, respectively. Moreover, both gene populations were characterized by a 4-fold increase in gene duplicate accumulation rate in the lineage from the LCA of schizophoran Diptera to that of Drosophilidae and calyptrate Diptera (Table 2). Most notable, however, in the schizophoran lineage connecting the LCA of Drosophilidae + Calyptratae to that of Drosophilidae the gene duplicate accumulation rate in the energy metabolism gene population peaked even further to 0.0054, exceeding that of the developmental gene population (0.0011) by a factor of 5 (Table 2). These findings suggested that duplications accumulated at higher rate during early schizophoran evolution in both gene populations except for an additional spike in the energy metabolism population in the schizophoran lineage to Drosophilidae.

Table 3 Paralog expression specificities in the expanded energy metabolism gene families.

Gene families	Members	Cellular localization	Testis biased	Ovary biased	Long branch
Glutaredoxin	Grx1t	undefined	1		0
	Grx1		digestive system		0
Thioredoxin reductase	Trxr-1	cytosol/mitochondrial	digestive system/testis		0
	Trxr-2	mitochondrial	1		1
Thioredoxin	Trx-2	nuclear	0		0
	Trx-1 (dhd)	nuclear	0	1	1
	TrxT/1	chromosome	1		1
	CG13473	–	1		1
Heat Shock Protein 60	Hsp60	cytosol/mitochondrial	0		0
	Hsp60B		1		1
	Hsp60C		1	0	1
	Hsp60D		1		1
Mitochondrial inner membrane translocases	Tim13	mitochondrial	1		1
	CG34132		0		0
	CG42302		1		1
Mitochondrial inner membrane translocases	Tim17b1	mitochondrial	1		1
	Tim17b2		1		1
	CG1724		1		1
	Tim17b		0		0
P-P-bond-hydrolysis-driven protein transmembrane transporter 20	Tom20	mitochondrial	1	1	0
	tomboy20		1	0	1
P-P-bond-hydrolysis-driven protein transmembrane transporter 40	Tom40	mitochondrial	0	1	1
	tomboy40		1	0	1
Glycerol 3 phosphate dehydrogenase	Gpdh1	cytosolic	muscle specific		0
	Gpdh2		1		1
	Gpdh3		1		1
Glycerophosphate oxidase-1	Gpo-1	mitochondrial	muscle specific		0
	Gpo-3		1		1
	Gpo-2		1		1
Mitochondrial anion carrier protein (MACP) gene family	Ucp4A	mitochondrial	0		0
	Ucp4B		1		1
	Ucp4C		1		1
Mitochondrial carrier (TC 2.A.29) family	MME1	mitochondrial	1		1
	colt		0		0
Eukaryotic mitochondrial porin family	porin	mitochondrial	0		0
	Porin2		1		1
Hexokinase	Hex-A	cytosolic	0		0
	Hex-C		0		1
	Hex-t1		1		1
	Hex-t2		1		1

(Continued)

Table 3 (continued)

Gene families	Members	Cellular localization	Testis biased	Ovary biased	Long branch
Malate dehydrogenase 2	Mdh2	mitochondrial	0		0
	CG10748		1		1
	CG10749		1		1
Succinate dehydrogenase, subunit C	SdhC	mitochondrial	0		0
	CG6629		1		1
Malic enzyme like-1	Menl-1	mitochondrial	1		1
	Menl-2		1		1
	Men-b		0		0
Malic enzyme	CG7848	mitochondrial	1		1
	Men	mitochondrial/cytosol	0		0

Pervasive germline subfunctionalization in the enriched energy metabolism-related gene duplicate population

To gain insights into the biological significance of the energy metabolism-related gene duplications, we mined *D. melanogaster* gene expression data available through the modENCODE database (*Chen et al., 2014*). This effort revealed that all of the 18 energy metabolism-related gene families included germline-specific paralogs (Table 3). In most cases, only one paralog was documented to be expressed in a broader range of tissues. Moreover, most of the energy metabolism-related gene families (13/18) represented nuclearly encoded mitochondrial proteins (Table 3), many of which had been previously reported in testes-, or, to a much lesser extent, ovary-specific expression datasets (*Haerty et al. (2007)*: 18; *Mikhaylova, Nguyen & Nurminsky (2008)*: 6; *Wasbrough et al. (2010)*: 14; *Gallach, Chandrasekaran & Betrán (2010)*: 29).

Mapping the paralog gene expression characteristics onto gene tree topologies further revealed that germline-specificity was associated with pronounced protein sequence divergence compared to broadly expressed paralogs in the majority of cases (Supplemental Data File 8). As a paradigmatic example, the *Heat shock protein 60* (*Hsp60*) family is represented by the uniformly expressed paralog *Hsp60* and three testis-biased paralogs in *D. melanogaster*, that is, *Hsp60B*, *Hsp60C*, and *Hsp60D* (Fig. 4). Maximum likelihood gene tree estimation revealed that the uniformly expressed *D. melanogaster Hsp60* paralog is characterized by a terminal branch length that falls well within the range of that of singleton homologs of other both brachyceran and non-brachyceran species. The testis-biased paralogs, by contrast, which were unique to drosophilid species, were characterized by moderately (*Hsp60C*) to extremely extended terminal branch lengths (*Hsp60B* and *Hsp60D*) (Fig. 4).

Germline subfunctionalization combined with extremely asymmetric protein sequence evolution of sister paralog gene duplicates has been recognized to constitute a signature outcome of intralocus sexual conflict resolution (ISCR) (*Haerty et al., 2007*; *Gallach, Chandrasekaran & Betrán, 2010*). The combined evidence from gene expression and gene

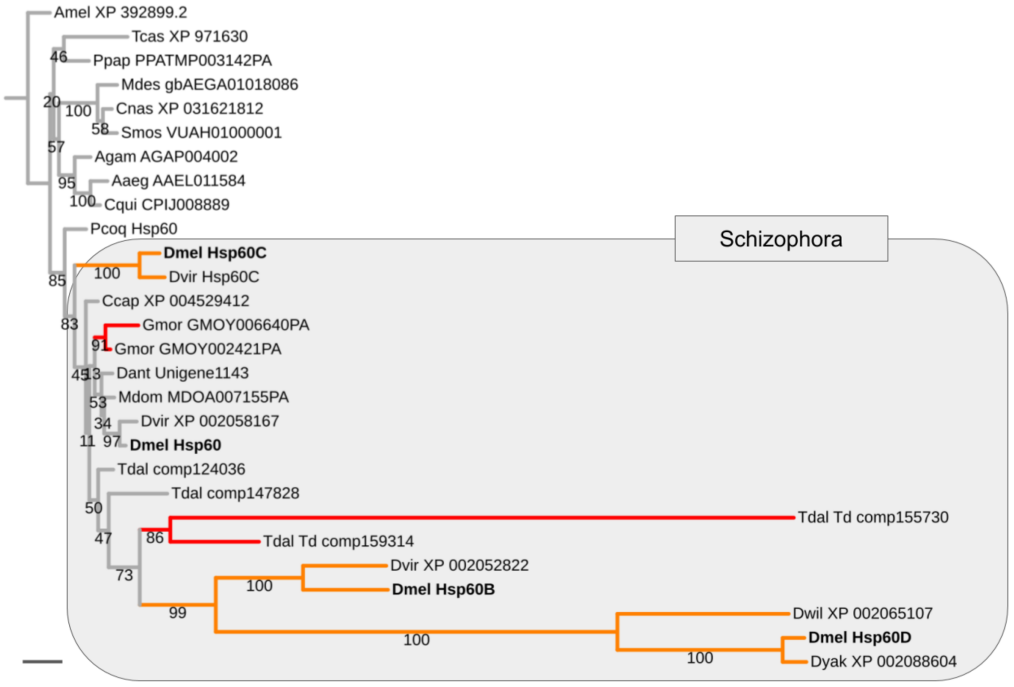

Figure 4 Expansion and diversification of the *Hsp60* gene family in schizophoran Diptera. Maximum likelihood tree of *Hsp60* gene family homologs compiled from dipteran species and select non-dipteran outgroup species (Amel, Tcas). Numbers at branches represent branch support from non-parametric bootstrap analysis. Branch support values lower than 50% not shown. The four *Hsp60* paralogs of *D. melanogaster* are highlighted in bold font. Orange branches indicate germline-specific paralogs. Red branches indicate parallel duplications. Species abbreviations: Aaed, *Aedes aegypti*; Agam, *Anopheles gambiae*; Ccap, *Ceratitis capitata*; Cnas, *Contarinia nasturtii*; Cqui, *Culex quinquefasciatus*; Dant, *Delia antiqua*; Dmel, *Drosophila melanogaster*; Dwil, *Drosophila willistoni*; Gmor, *Glossina morsitans*; Llon, *Lutzomyia longipalpis*; Mdes, *Mayetiola destructor*; Mdom, *Musca domestica*; Ppap, *Phlebotomus papatasi*; Pcoq, *Proctacanthus coquilletti*; Smos, *Sitodiplosis mosellana*; Tdal, *Teleopsis dalmanni*. Note: We failed to detect *Hsp60D* in *D. virilis* but found *Hsp60D* in the closely related *D. willistoni* implying that *Hsp60D* originated through a duplication that preceded the split of the Sophophora and Drosophila subgroups, which are represented by *D. melanogaster* vs *D. virilis* and *D. willistoni*, respectively. Scale bar corresponds to 0.1 substitutions per site. Full-size DOI: 10.7717/peerj.10012/fig-4

tree analyses therefore suggested that the enriched population of duplicated energy metabolism genes had been primarily deployed in ISCR events.

Molecular signatures of intralocus sexual conflict resolution events throughout schizophoran lineages

Asking whether the duplication events in energy metabolism gene families had been of general impact in schizophoran Diptera diversity as opposed to specifically the lineage leading to *Drosophila*, we compared the numbers of independent gene duplications in the 18 investigated gene families that were detectable in the terminal branches to schizophoran and non-schizophoran species. As an example, in the *Hsp60* gene family two parallel gene duplications mapped to terminal branches of schizophoran species (*G. morsitans*, *T. dalmanni*) but no parallel duplications were detectable in the nine sampled non-schizophoran lineages (Fig. 4).

Overall, we detected 31 independent energy metabolism gene duplications in the terminal branches to the six schizophoran lineages sampled in addition to *D. melanogaster*. This compared to 9 in the robber fly *P. coquilletti* and 13 in the 8 non-brachyceran sampled dipteran species (Fig. 3). Outside Diptera, we found three parallel duplications in each *Apis* and *Tribolium* (Supplemental Data File 6). Taken together, these numbers constituted evidence in support of a Schizophora-wide increase in energy metabolism gene duplicate accumulation rate. Moreover, based on the *P. coquilleti* results, also orthorrphaphan Diptera might have experienced a relative increase in energy metabolism gene duplications. Of note, our finding of independent duplications in the *Thioredoxin*, *Hsp60*, *MME1/colt*, and *Hexokinase* gene families in the stalk eyed fly *T. dalmanni* was consistent with the results of the original study of sperm-enriched genes in *T. dalmanni* (*Baker et al., 2016*). Some orthology assignments differed between the two studies most likely reflecting the uncertainties coming along with low branch support values in some of the respective gene trees (Fig. 4; Supplemental Data File 8, and below).

Asking whether there was also evidence of ISCR-related trajectories in the population of independent energy metabolism gene duplication events, we tallied the fractions of parallel gene duplications that generated paralogs with pronounced asymmetric, that is, more than 2-fold, branch length differences based on maximum likelihood tree estimation results. Both parallel brachyceran gene duplications in the *Hsp60* gene family, for example, were associated with pronounced asymmetric terminal branch lengths (Fig. 4). Surveying across all energy metabolism related gene families, we found that an average of 3.0 parallel duplications with asymmetrically diverged paralogs in the schizophoran species compared to an average of 1.1 among the non-brachyceran dipteran species (Supplemental Data 6). Most of the independent gene duplications in the robber fly, however, also produced asymmetrically diverged paralogs (Fig. 3).

Comparing the average numbers of gene families in which parallel asymmetric gene duplications were detected indicated a similarly pronounced difference between the schizophoran and non-schizophoran terminal lineages (Fig. 5; Supplemental Data 6). With the caveat of lacking data on the tissue specificities for most of the of the non-*Drosophila* gene duplicates except for the stalk eyed fly *T. dalmanni* (*Baker et al., 2012*, *2016*), these findings did amount to evidence that ISCR via gene duplication has been more common in schizophoran Diptera than in other dipteran lineages with the possible exception of orthorraphan Diptera represented by *P. coquilleti*.

Pre-schizophoran expansion of the Drosophila Thioredoxin gene family

While our analyses indicated a the prevalence of ISCR-related duplications in the energy metabolism gene population during the diversification of schizophoran Diptera, five duplications in three energy-metabolism gene families predated this time window, based on the compilation of gene family homologs and phylogenetic gene tree analyses (Fig. 3; Supplemental Data File 6). It was therefore of interest to examine whether the protein sequence evolution characteristics of the paralogs resulting from these duplications were

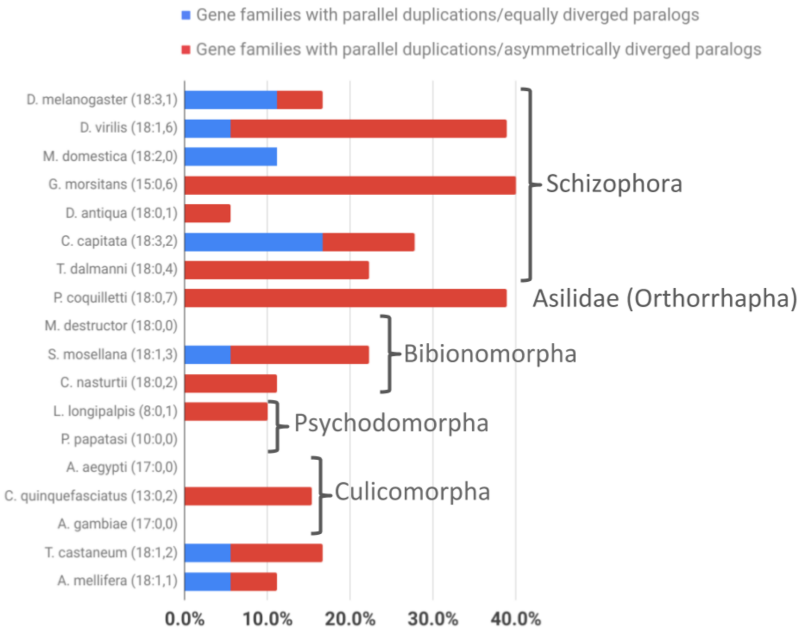

Figure 5 Taxonomic survey of parallel duplications in the overrepresented energy metabolism gene family populations of Brachycera-specific gene duplicates. Numbers in parentheses next to species names indicate numbers of gene families sampled followed by the number of parallel gene duplications that produced equally diverged paralogs (represented as a percentage of all gene families sampled by blue bar portions) and the number of parallel duplications with strong branch asymmetry (>2) between sister paralogs (represented as a percentage of all gene families sampled by red bar portions).

Full-size DOI: 10.7717/peerj.10012/fig-5

likewise indicative of ISCR trajectories despite their considerably more ancient time points of occurrence.

Three of the five pre-schizophoran gene duplications were detected in the disulfide oxidoreductase encoding *Thioredoxin* gene family, which is represented by four paralogs in *D. melanogaster*: *Trx-2* (CG31884), *Trx-1* (*deadhead*) (CG4193), *TrxT/1* (CG3315), and the yet uncharacterized locus *CG13473* (Supplemental Data File 6). While *Trx-2* is broadly expressed throughout tissues and life cycle in *D. melanogaster*, *TrxT/1* and *CG13473* transcripts are testis-enriched and *Trx-1* (*dhd*) is ovary-enriched (*Svensson et al., 2003*; *Svensson & Larsson, 2007*).

In most non-schizophoran species, including the robber fly, we only detected singleton *Thioredoxin* homologs (Fig. 6; Supplemental Data File 6). Multiple homologs were identified in the gall midge species *S. mosellana* and *C. nasturtii* as products of parallel gene duplications based on gene tree analysis results (Fig. 6). Most importantly, in the Mediterranean fruit fly *C. capitata*, we found candidate 1:1 orthologs for each of the *D. melanogaster Thioredoxin* gene family members (contig5575_1, contig5575_2, comp62603, comp61885). While branch support values were too low to draw conclusions with high confidence (number of homologous alignment sites: 75; see Supplemental Data File 7), the distribution of the *C. ceratitis Thioredoxin* homologs was best compatible with a pre-schizophoran expansion of the *D. melanogaster Thioredoxin* gene family in light of the fact that the LCA of *D. melanogaster* and *C. ceratitis* dates back to the root of

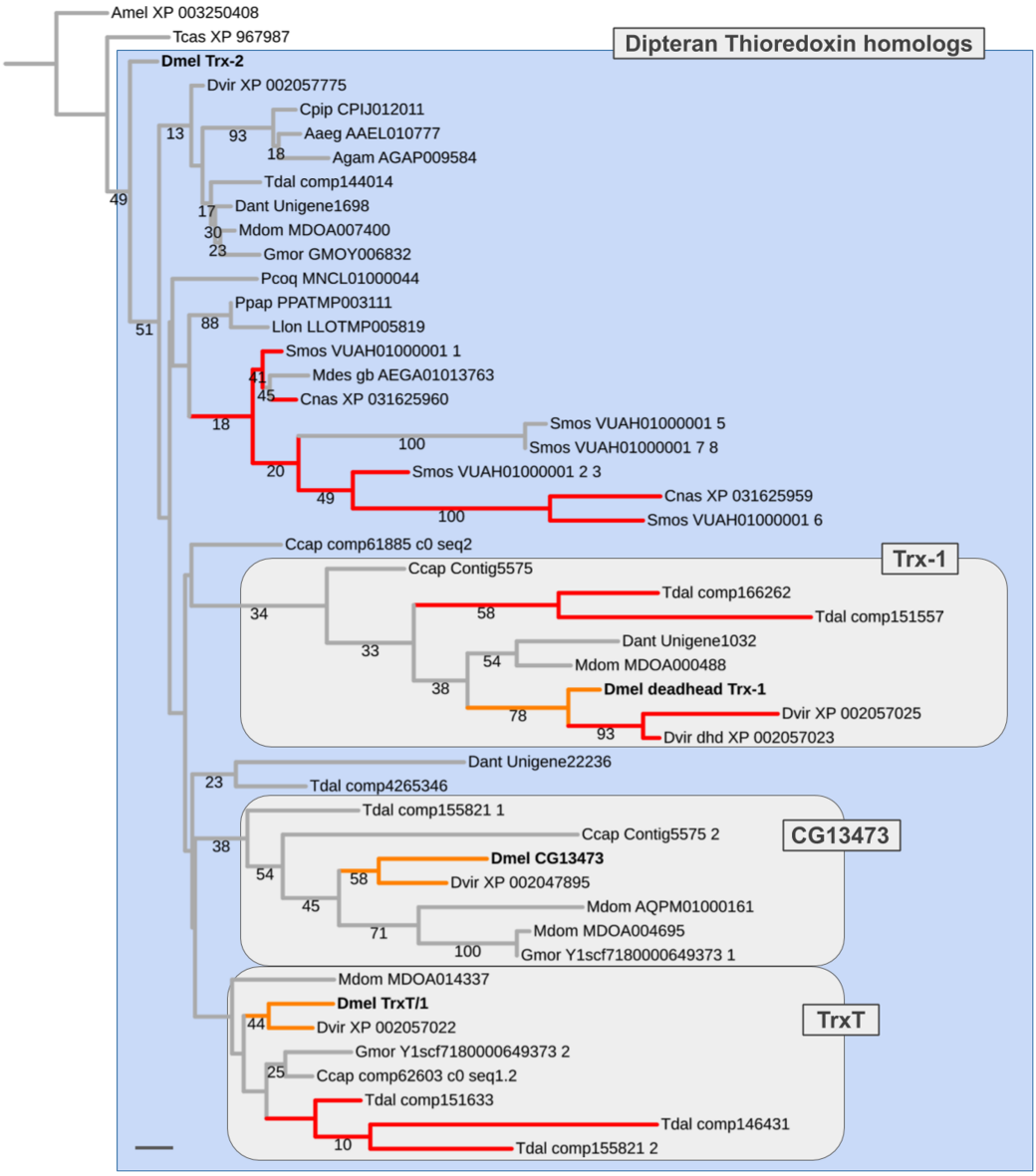

Figure 6 Dipteran *Thioredoxin* gene family tree. Maximum likelihood tree of *Thioredoxin* homologs compiled from dipteran species and select outgroup species. Numbers at branches represent branch support from non-parametric bootstrap analysis. Branch support values lower than 10% not shown. The four *Thioredoxin* gene family paralogs of *D. melanogaster* are highlighted in bold font. Orange branches indicate germline-specific paralogs in *Drosophila*. Red branches indicate originated asymmetric gene duplicates that originated in parallel to the duplications in the *Drosophila* lineage. Purple branches lead to robber fly homologs. Species abbreviations same as in Fig. 4. Scale bar corresponds to 0.1 substitutions per site. Full-size DOI: 10.7717/peerj.10012/fig-6

schizophoran Diptera (Figs. 3 and 5) (*Wiegmann et al., 2011*). Further, consistent with the somatic, that is, likely ancestral, requirement of the *D. melanogaster Trx-2* paralog, all non-schizophoran homologs clustered with this member of the *D. melanogaster Thioredoxin* gene family (Fig. 6). The germline-specific *TrxT/1*, *CG13473*, and *Trx-1* paralogs, in contrast, each clustered with a large number of bona fide orthologs from

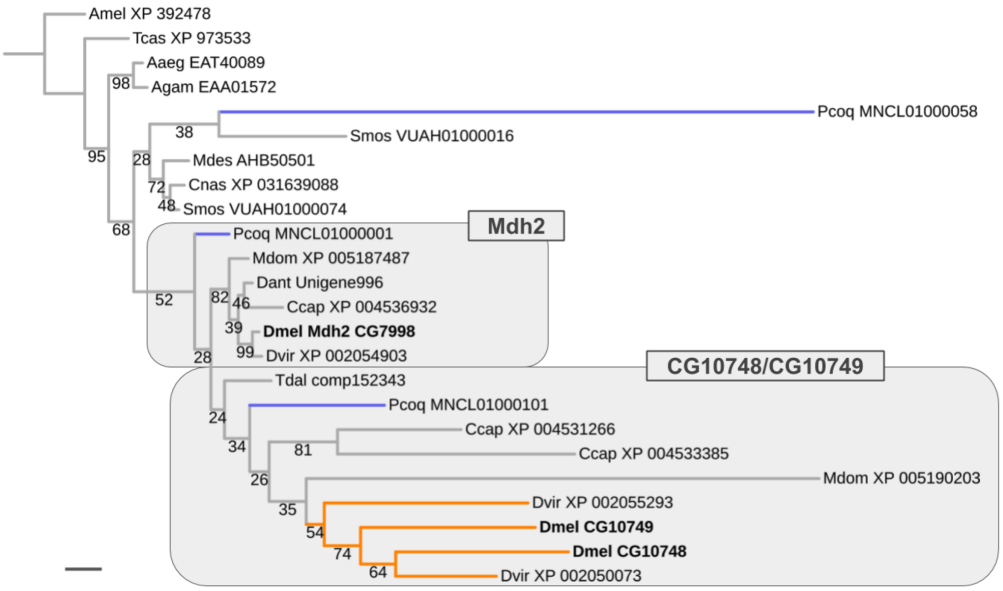

Figure 7 Dipteran Mdh2 gene family tree. Maximum likelihood tree of *Mdh2* homologs compiled from dipteran species and select outgroup species. Numbers at branches represent branch support from non-parametric bootstrap analysis. Branch support values lower than 50% not shown. Orange branches indicate germline-specific paralogs in the *Drosophila* lineage. Blue branches lead to robber fly homologs. Species abbreviations as in Fig. 4. Scale bar corresponds to 0.1 substitutions per site.

Full-size DOI: 10.7717/peerj.10012/fig-7

schizophoran species (Fig. 6). Thus overall, the gene tree analysis results suggested a pre-schizophoran ISCR-related expansion of the *Thioredoxin* gene family.

Ancient ISCR-related expansion of the Malate dehydrogenase 2 gene family

Results similar to that obtained for the *Thioredoxin* gene family were encountered for the *Malate dehydrogenase 2* (*Mdh2*) gene family, which comprises two male germline-specific paralogs in *D. melanogaster* (CG10748, CG10749) besides *Mdh2* (CG7998) (Fig. 7). However, unlike in the case of the *Thioredoxin* gene family, evidence for the existence of not only one but three *Mdh2* gene family homologs was found in the robber fly *P. coquilletti* (Fig. 7; Supplemental Data File 6). One of them, located on genome assembly contig MNCL01000001, grouped basally with the *Mdh2* homolog cluster while the second, located on genome assembly contig MNCL01000101, branched out basally in the cluster that included the germline-specific *D. melanogaster* homologs *CG10748* and *CG10749* (Fig. 7; Supplemental Data File 6). Equivalently placed homologs were also recovered from the Mediterranean fruit fly *C. capitata* (XP 004531266, XP 004536932).

The third *P. coquilletti Mdh2* gene family homolog located on contig MNCL01000058 appeared extremely derived, branching out in the cluster of gall midge (Bibionomorpha) species homologs (Fig. 7). This was likely an artifact due to extreme amino acid substitution differences and limited number of multiple sequence alignments sites (290) for accurate gene tree reconstruction. Notwithstanding these limitations, the existence of reasonably well supported separate orthologs to the somatic and germline-specific

paralogs of the *D. melanogaster Mdh2* gene family in the robber fly *P. coquilletti* provided compelling evidence of an ISCR-related expansion of this gene family in the stem lineage to brachyceran Diptera, thus over 180 million years ago.

Early metabolic and late germline-specific expansions in the Hexokinase gene family

The oldest duplication event in the energy metabolism gene population was detected in the Hexokinase gene family, which comprises four paralogs in *D. melanogaster*: *Hex-A* (CG3001), *Hex-C* (CG8094), *Hex-t1* (CG33102), and *Hex-t2* (CG32849) (Fig. 8; Supplemental Data File 6). While *Hex-t1* and *Hex-t2* are germline-specific paralogs, both *Hex-A* and *Hex-C* are characterized by high expression levels in a wide number of body regions based on modENCODE data and previous studies (*Moser, Johnson & Lee, 1980*; *Bourbon et al., 2002*; *Chen et al., 2014*). Moreover, *Hex-A* is expressed at higher levels than *Hex-C* in most cases except for the digestive system, head, and male testis (*Chen et al., 2014*). This correlates with an ancestral functionality of *Hex-A* in glucose metabolism in contrast to the derived affinity of *Hex-C* to fructose (*Moser, Johnson & Lee, 1980*).

Our homolog searches and gene tree analyses revealed that the broadly expressed *Hex-A* and *Hex-C* paralogs were the products of a gene duplication in the LCA of Brachycera and Bibionomorpha (Fig. 8; Supplemental Data File 6). High confidence orthologs of both *Hex-A* and *Hex-C* were detected in all dipteran species sampled with parallel duplications of *Hex-A* in the robber fly (MNCL01000216, MNCL01000057, MNCL01005250_1) and the orange wheat blossom midge *S. mosellana* (VUAH01006190, VUAH01003649) and parallel duplications of *Hex-C* in the stalk-eyed fly *T. dalmanni* (comp147884, comp160205, comp157604) (Fig. 8).

Moreover, the Hexokinase gene family tree produced strong support that the germline-specific *Hex-t1* and *Hex-t2* paralogs were born through additional duplications in the Hex-C cluster (Fig. 8). 1:1 orthologs of each germline-specific paralog, however, were only detectable in *D. virilis*, thus dating their origin prior to the diversification of drosophilid Diptera. A second cluster of independently duplicated paralogs was detected in the calyptrate species *M. domestica* and *G. morsitans* (Fig. 8). Only low support, however, was recovered for a closer relationship between these calyptrate *Hex-C* homologs and the *Drosophila Hex-t1/t2* homolog cluster.

In summary, the reconstructed sequence of gene duplication events in the Hexokinase gene family corroborated the strong association of ISCR-related gene duplication events with the schizophoran species radiation in addition to revealing an ancient duplication in the LCA of Brachycera and Bibionomorpha that most likely resulted in an expansion of metabolite usage for energy production in these clades.

DISCUSSION

Elevated background accumulation of insect core gene duplicates in the higher Diptera

Studying the impact of gene duplication on phenotypic evolution is of particular interest for understanding the origins of highly diverse organismal groups such as the true flies

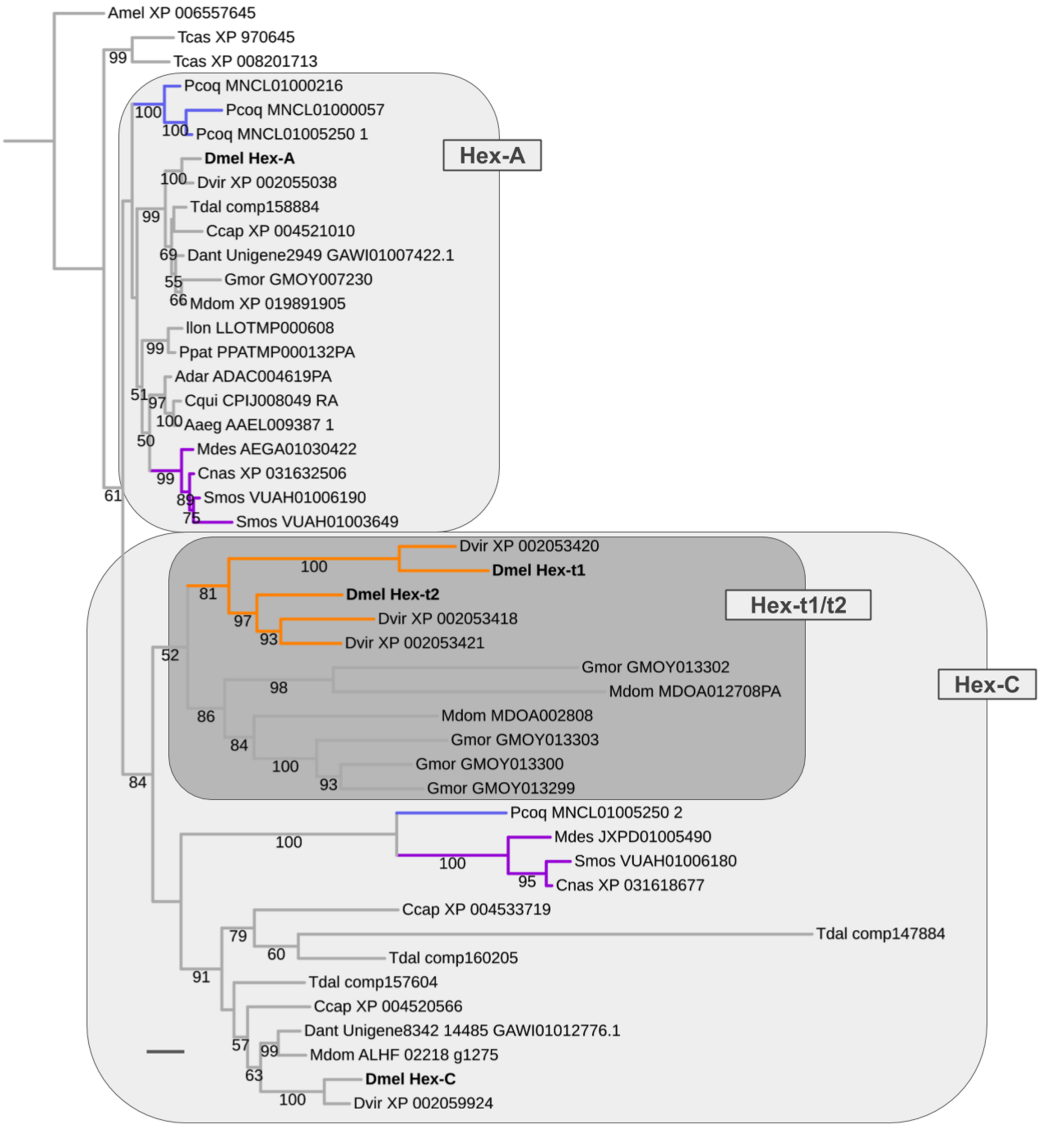

Figure 8 Dipteran *Hexokinase* gene family tree. Maximum likelihood tree of *Hexokinase* homologs compiled from dipteran and select outgroup species. Numbers at branches represent branch support from non-parametric bootstrap analysis. Branch support values lower than 50% not shown. Orange branches indicate germline-specific paralogs in the *Drosophila* lineage. Blue branches lead to robber fly homologs. Turquoise branches indicate gall midge (Bibionomorpha) homolog clusters. Species abbreviations same as in Fig. 4. Scale bar corresponds to 0.1 substitutions per site.

Full-size DOI: 10.7717/peerj.10012/fig-8

which encompass over 150,000 described species (*Yeates et al., 2007*; *Wiegmann et al., 2011*). Three major radiations have been recognized in the expansion of this megadiverse animal clade. Besides an initial rapid diversification into seven separate lineages approximately 180 million years ago, followed by the radiation of brachyceran Diptera into about 100,000 contemporary species (*Wiegmann et al., 2011*), roughly 50% of this diversity was the result of a third radiation, that of schizophoran Brachycera into over 150 families starting only about 65 million years ago (*Wiegmann et al., 2011*).

Our study was motivated by preliminary evidence that the genome of a premier representative of the Brachycera and Schizophora, i.e., *D. melanogaster*, was notably richer in lineage-specific gene duplicates compared to other equally old or older insect lineages leading to mosquito, beetle, and hymenopteran genome model species (*Bao & Friedrich, 2009*; *Bao et al., 2018*). Our findings at the genome-wide scale presented here point at an approximately 2-fold higher number of duplicated insect core gene in the lineage to *Drosophila* compared to non-brachyceran Diptera, that is, mosquitoes (Culicomorpha), honeybee (Hymenoptera), and the red flour beetle (Coleoptera). In addition to our own earlier studies, this result is consistent with multiple notions of *Drosophila*-specific gene duplications in earlier gene family studies (*Svensson, Stenberg & Larsson, 2007*; *Porcelli et al., 2007*; *Bao & Friedrich, 2009*; *Carmon & MacIntyre, 2010*; *Jiménez-Guri et al., 2013*; *Fraga et al., 2013*; *Lewis, Salmela & Obbard, 2016*; *Bao et al., 2018*; *Helleu & Levine, 2018*). At the same time, however, recent large scale analyses of gene duplication rates in insects did not detect evidence of exceptional gene gain in the *Drosophila* lineage (*Li et al., 2018*; *Thomas et al., 2020*). To the contrary, *Neafsey et al. (2015)* reported a higher gene gain rate in mosquito species compared to drosophilid Diptera. Comparing a similarly small sample of holometabolan insect lineages, *Roelofs et al. (2020)*, however, encountered evidence of a higher genome-wide gene duplication rate in the lineage to *Drosophila* in comparison to Lepidoptera, Coleoptera, and Hymenoptera.

While many high quality insect genome drafts have been generated by now, it is still reasonable to assume that the *Drosophila* genome continues to be the most comprehensively annotated and best curated one. Therefore, an alternative explanation for the larger number of detected lineage-specific gene duplicates in *Drosophila* could be lower quality of sequence coverage and gene annotation sensitivity in the genome assemblies of the non-*Drosophila* species we sampled. Three lines of evidence speak against this possible ascertainment bias: (1) The above mentioned congruence with previous gene-specific studies, (2) the high level of accuracy in our validation test, and (3) the equally high validation in our deeper analyses of gene duplications in energy metabolism gene families. Combined with our analysis of developmental gene families, the same analysis further suggests that the *Drosophila*-lineage increase of gene duplicate accumulation occurred for the most part during the early stage of the schizophoran radiation, extending from approximately 65 to 30 million years ago.

Causation scenarios

One obvious followup question arising from our findings is which processes might have been responsible for the enhanced accumulation of duplicated insect core genes in the higher Diptera. The exclusion of adaptive gene family expansions through our focus on small size gene families and the genome-wide dimension of gene duplicate increase makes non-adaptive mechanisms appear more likely. Also the fact that the *Drosophila*-lineage gene duplicate population is characterized by the lowest number of enriched BP GO-terms (centered around a single context: energy metabolism), but the highest number of underrepresented BP GO-terms compared to the less core gene duplicate rich lineages

could be interpreted in this direction. This finding seems best explained by a gene-function independent origination mechanism that is tolerated to a similar degree by most functional contexts thus explaining the scarcity of enriched biological functions. The relatively high number of underrepresented biological functions could be envisioned to represent functional contexts that are more sensitive to the immediate consequences of gene duplications such as dosage increase. As a point in case, the underrepresentation of developmental functions in the *Drosophila* population of lineage-specific gene duplicates is consistent with the in most cases highly pleiotropic nature and dosage-sensitive action of developmental regulators, as has been noted in other cases as well (*Conant, 2020*). And yet, in the comparison between species *Drosophila* still stands out with a higher number of developmental gene duplicates (*Bao et al., 2018*), thus documenting a measurable impact of heightened gene duplicate accumulation even on this exceptionally sensitive class of genes.

Discounting adaptive forces, the higher proportion of duplicated insect core genes in the *Drosophila* lineage could be due to an increase in the rate of nonhomologous recombination, an increase in the fixation rate of nascent gene duplicates, or a decrease in gene duplicate loss rates. In principle, all of these candidate variables can be tested through comparative population genomic approaches. Sample sizes will, however, likely need to be considerable in light of the fact that our findings suggest that the 2-fold higher number of insect core gene duplicates in modern Brachycera built over a long evolutionary time span, that is, up to 65 million years, since the origin of the schizophoran stem lineage and left traces in less than 10% of the insect core gene repertoire.

In this context, it is informative to relate our findings to the more recent and dramatic expansion of the pea aphid genee content via local duplications (*International Aphid Genomics Consortium, 2010*; *Armisen et al., 2018*; *Panfilio et al., 2019*). While examples of adaptive gene family expansions have been detected for the pea aphid (*Smadja et al., 2009*), it is intriguing to note that pea aphids are characterized by cyclical parthenogenesis, that is, a mode of asexual evolution (*Miura et al., 2003*). The latter in turn implies relaxed purifying selection due to reduced effective population size, which can be hypothesized to increase the survival probability of nascent gene duplicates via genetic drift (*Lynch & Conery, 2000*). While this reference point lends support to effective population size based mechanisms as candidate explanations of the elevated gene duplicate accumulation in the higher Diptera, population genomic evidence suggests that positive selection has been the stronger effector in the fixation of gene duplicates during the more recent evolutionary history of the genus *Drosophila* (*Cardoso-Moreira et al., 2016*). Moreover, while a considerable number of parthenogenesis-capable species have been documented in flies including drosophilids (*Meyer et al., 2010*; *Gokhman & Kuznetsova, 2018*), their overall rarity gives little reason to suspect a broader impact of parthenogenesis during the early schizophoran radiation.

For these reasons, the recently identified immediate fitness benefit of gene regulatory noise suppression through gene duplications deserves equal consideration as a possible mechanism underlying the schizophoran gene duplicate accumulation increase (*Rodrigo & Fares, 2018*). Intriguingly, our study of developmental gene duplications produced

evidence for a larger amount of long-term conserved, genetically redundant paralogs in the higher Diptera compared to other insect genome models (*Bao et al., 2018*). This led us to the prediction of a higher level of genetic robustness during development, which might have benefited the comparatively fast speed of embryonic and postembryonic development in the higher Diptera. As noted above, although richer in gene duplicates compared to other insect lineages, developmental genes do not represent a significantly enriched fraction in the population of lineage-specific gene duplicates in *Drosophila*. It seems therefore reasonable to speculate that much of the functional spectrum of gene duplications in the *Drosophila* lineage produced a similar blend of conservative vs innovative functionalization outcomes as found for the developmental gene cohort. This prediction can be assessed in future studies through comprehensive analysis of expression and gene function data available for *D. melanogaster*. Even more decisive would be the integration of tissue specific expression data from a number of dipteran key species, an approach which proved highly informative in recent studies of gene duplication outcomes in vertebrates (*Marlétaz et al., 2018*).

Possible links between enhanced gene duplicate accumulation, intralocus sexual conflict resolution, and speciation rates in the higher Diptera

Energy metabolism-related genes emerged as the only overrepresented functional category in the brachyceran gene duplicates. As previous studies noted evidence of adaptive evolution of mitochondrially encoded energy metabolism genes associated with the transition to flight in insects (*Yang et al., 2014*), we initially suspected the emergence of exceptionally fast flight capacities in the brachyceran Diptera as a possible adaptive outcome of the enriched proportion of duplicated energy metabolism genes. Our detailed analysis, however, paints a different picture. With a few notable exceptions in the *Thioredoxin* and *Mdh2* gene families, the energy metabolism gene duplications date to the early diversification of schizophoran Diptera, that is, over 100 million of years later than the primary radiation of brachyceran Diptera. Moreover, tissue-specific expression data and asymmetric paralog divergencies identified the great majority of lineage-specific energy metabolism gene duplicates as facilitators of ISCR. Consistent with this, most of our energy metabolism duplicates were previously identified as the genomic products of ISCR (*Rand, Clark & Kann, 2001*; *Gallach, Chandrasekaran & Betrán, 2010*; *Connallon & Clark, 2011*; *Wyman, Cutter & Rowe, 2012*; *Chakraborty & Fry, 2015*). In this process, the faster sequence divergence of male germline-specific paralogs is thought to be driven by higher energy requirements of competing sperm, release from conflicting functional constraints, and the higher mutagenic physiology of sperm cells due to higher radical oxygen species production (*Rettie & Dorus, 2012*; *Patel et al., 2016*; *Jiang & Assis, 2017*). Of note, ISCR by gene duplication is not considered subfunctionalization in the strict sense of leading to an adaptively neutral breakup of ancestrally pleiotropic functionality (*Gallach & Betrán, 2011*). However, it seems commonly assumed, and is testable, that the precursor homologs of ISCR paralogs have been homogeneously expressed in germline and somatic cells, thus defining the

differential expression of ISCR paralogs in germline vs somatic tissues as subfunctions of the ancestral expression repertoire.

Overall, our finding of a large number of germline-biased paralogs in the germline-biased energy metabolism gene population is of little surprise in light of the fact that over 10% of the *Drosophila* coding genome is characterized by germline biased gene expression (*Chintapalli, Wang & Dow, 2007*). More remarkable may be our finding that the outcome of ISCR by gene duplication can persist over long evolutionary time spans. The majority of the ISCR-related gene duplications in the lineage to *Drosophila* mapped to the early radiation of schizophoran Diptera, about 40–60 million years ago. Moreover, the oldest ISCR-related gene duplications captured in our analysis occurred most likely at least 80 million years ago (*Mdh2* and *Thioredoxin* gene families) (Supplemental Data File 6). For comparison, the most thoroughly studied case of gene duplication facilitated ISCR studied in *Drosophila* thus far dates back about 200,000 years (*VanKuren & Long, 2018*).

Our genomic detection of ISCR associated gene duplicates in subclades of the *Drosophila* family, that is, the *Drosophila* and *Sophophora* groups, is consistent with the notion that ISCR has continued to be of relevance during the more recent diversification of drosophilid Diptera (*Mikhaylova, Nguyen & Nurminsky, 2008*; *Kondo et al., 2017*). In addition, we detected a considerable number of parallel gene duplications with ISCR signatures, that is, extremely asymmetrically diverged sister paralogs, in other schizophoran lineages. Taken together, these findings and that of others in tephritid Diptera indicate that ISCR has been of widespread occurrence and significance in schizophoran Diptera (*Baker et al., 2016*).

Two lines of evidence indicate a possible link between ISCR frequency and speciation rate in our findings. For one, our data suggest a higher frequency of ISCR in the overall largely expanded schiziphoran Diptera compared to other dipteran clades with exception of the robber fly lineage. Second, we detect evidence of a spike of ISCR related gene duplication events during early schizophoran radiation. Together, tese findings speak to the still open debate of whether ISCR impacts reproductive isolation and therefore ultimately speciation rates (*Coyne & Orr, 1989*; *Parker & Partridge, 1998*; *Gavrilets, 2014*). Experimental studies produced mixed signals. Consistent with our results for the higher Diptera, *Katzourakis et al. (2001)* found evidence for a correlation between sexual selection and species richness in hoverflies. Similar conclusions have been drawn in a study of tephritid species (*Congrains et al., 2018*), which aligns taxonomically more closely with our findings of parallel sex biased gene duplications in schizophoran lineages outside the Drosophilidae. Based on artificial breeding experiments, however, sexual conflict was concluded to play no role in reproductive isolation in *D. pseudoobscura* (*Bacigalupe et al., 2007*). At the same time, there is evidence for the action of sexual conflict in allopatric experimental populations of *D. melanogaster* (*Syed et al., 2017*). At this point, the available evidence may still be summed up to suggest that speciation rate increase is driven by sexual conflict in some but not all clades (*Gavrilets, 2014*).

In support of the latter notion, we not only detected substantially lower numbers of parallel candidate ISCR gene duplications in non-brachyceran Diptera but also in

Tribolium and the honeybee, representatives of two megadiverse insect groups. Given the relatively small number and biased selection of gene families sampled, future studies will be needed to scrutinize the preliminary evidence that ISCR may have been of exceptional importance during the radiation of schizophoran Diptera compared to other megadiverse insect groups such as the Hymenoptera and Coleoptera. At this point, however, it is tempting to speculate that a heightened gene duplication background rate fueled ISCR events during the massive radiation of schizophoran Diptera. Encouragingly, our finding that the ISCR promoted germline-specificity of gene duplicates can remain stable over long evolutionary time scales is also of practical significance as it opens up venues for studying the global significance of ISCR in species diversification via comparative genomics and transcriptomics. This promises more definitive answers to the questions raised above and may potentially deliver even generally deeper insights into the role of molecular germ line subfunctionalization in animal speciation (*White-Cooper & Bausek, 2010*).

CONCLUSIONS

Our comparative analysis of lineage-specific gene duplicates in four holometabolous insect genome species produced evidence of an enhanced gene duplicate accumulation rate in the lineage to *Drosophila*. Our phylogenetic surveys of developmental and energy metabolism gene duplicates suggest that this increase occurred largely during the diversification of schizophoran Diptera about 60 million years ago. Energy metabolism gene duplicates seem to have experienced an exceptional increase facilitating ISCR via gene duplication, which may also have impacted speciation rates during the early phase of the dramatic schizophoran radiation. Our study further shows that ISCR originated gene duplicates can remain conserved over considerable evolutionary time scales, which should facilitate broader genomic and transcriptomic studies of the relationship between ISCR and speciation rates.

ACKNOWLEDGEMENTS

We thank the anonymous reviewers for critical input, Sorin Draghici, Chuanzhou Fan, and Jason Caravas for comments and advice on the project, Wayne State University Scientific Computing Program for providing and maintaining the Grid supercomputing cluster service.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

Riyue Bao was supported by Thomas C. Competitive Rumble University Graduate Fellowship and College of Liberal Arts and Sciences Enhancement GRA Scholarship. This project was supported by NSF award EF-0334948. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
Thomas C. Competitive Rumble University Graduate Fellowship.
College of Liberal Arts and Sciences Enhancement GRA Scholarship.
NSF: EF-0334948.

Competing Interests

The authors declare that they have no competing interests.

Author Contributions

- Riyue Bao conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.
- Markus Friedrich conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

This study relies on public domain genomic and transcriptomic data sources available at the RefSeq, Official Gene Set (OGS), whole genome shotgun contigs (wgs), and Transcriptome Shotgun Assemblies (TSA) databases maintained by the National Center of Biotechnology Information (NCBI). All analyzed sequences are also available in the Supplemental File.

- RefSeq: XP_001120807, XP_001121522, XP_001123018, XP_001653595, XP_001664283, XP_001687881, XP_001841997, XP_001845276, XP_001850913, XP_001850913, XP_001859594, XP_001862425, XP_002047597, XP_002047895, XP_002047939, XP_002048011, XP_002048624, XP_002049040, XP_002049720, XP_002050073, XP_002050587, XP_002050844, XP_002050845, XP_002050869, XP_002051123, XP_002051268, XP_002051368, XP_002051379, XP_002051380, XP_002051381, XP_002051422, XP_002051833, XP_002051868, XP_002052164, XP_002052597, XP_002052598, XP_002052822, XP_002053418, XP_002053420, XP_002053421, XP_002053422, XP_002053434, XP_002053611, XP_002054324, XP_002054373, XP_002054903, XP_002055038, XP_002055293, XP_002055407, XP_002055451, XP_002056569, XP_002056572, XP_002056889, XP_002057022, XP_002057023, XP_002057025, XP_002057046, XP_002057047, XP_002057312, XP_002057775, XP_002057863, XP_002058101, XP_002058167, XP_002058764, XP_002059026, XP_002059558, XP_002059924, XP_002062810, XP_002065107, XP_003249498, XP_003250408, XP_003436731, XP_004518508, XP_004519116, XP_004519322, XP_004520566, XP_004521010, XP_004523446, XP_004523968, XP_004524766, XP_004526337, XP_004529383, XP_004529412, XP_004530158, XP_004531038, XP_004531266, XP_004533385, XP_004533719, XP_004535304, XP_004536134, XP_004536932, XP_004537692, XP_005187487, XP_005190203,

XP_005190609, XP_006557645, XP_006558427, XP_006560286, XP_006563718, XP_006564634, XP_006564913, XP_006565977, XP_008191172, XP_008194238, XP_008199565, XP_008201416, XP_008201713, XP_019891905, XP_020713613, XP_021693932, XP_310951, XP_311387, XP_316164, XP_391836, XP_391836, XP_392478, XP_392899.2, XP_395280, XP_562185, XP_002056217, XP_966771, XP_967309,XP_967987, XP_968248,XP_969151, XP_969226, XP_969619, XP_970645, XP_971201, XP_971630, XP_972413, XP_972464, XP_973533, XP_975253, XP_EFA08711, AAEL002886, AAEL007001, AAEL007235, AAEL007392, AAEL008128, AAEL009387, AAEL010777, AAEL013980, AEGA01006707, AEGA01013763, AEGA01014751, AEGA01022594, AEGA01022595, AEGA01030422, AGAP000565RA, AGAP002277, AGAP004002, AGAP004002, AGAP004657, AGAP004657, AGAP007871, AGAP009584, AGAP009833, AGAP011107, AGAP012339, AHB50501, AAF46272, ALHF_02218.g1275, ALHF_11471, AQPM01000161, CH477216, CL1299, CL1818, CPIJ002542, CPIJ007967, CPIJ008049, CPIJ008889, EAA01572, EAT40089, EAT42717, EAT42717, EW987750, EX212033, FK813889, AEGA01003790, AEGA01018086, gi_145648988, gi_158703262, gi_78216392, gi309241287, GL501425, GL501437, GL630235, JP550838, NP_001014994, NP_001171496

- Official Gene Set (OGS): CG7975, CG10120, CG10748, CG10749, CG11401, CG1158, CG11611, CG12101, CG12157, CG13473, CG14690, CG15257, CG16954, CG17137, CG1724, CG18340, CG2137, CG2151, CG2830, CG3001, CG3057, CG31884, CG3215, CG32849, CG33102, CG3315, CG34132, CG3476, CG40451, CG4193, CG42302, CG43343, CG5495, CG5889, CG6492, CG6629, CG6647, CG6666, CG6852, CG7235, CG7311, CG7654, CG7964, CG7969, CG7998, CG8094, CG8256, CG8330, CG9042, CG9064

- Whole genome shotgun contigs (wgs), and Transcriptome Shotgun Assemblies

TSA: GAWI01006724, GAWI01008942, GAMC01007766, GAWI01002350, GAWI01002349, GAWI01003186, GAWI01003819, GAWI01003871, GAWI01005176, GAWI01005478, GAWI01005502, GAWI01005514, GAWI01005625, GAWI01005678, GAWI01005947, GAWI01006090, GAWI01006179, GAWI01006311, GAWI01007175, GAWI01007422, GAWI01020148, GAWI01026563, GAMC01000264, GAMC01005605, GAMC01009657, GAMC01011211, GAMC01011214, GBBP01071272, GBBP01072246, GBBP01083257, GBBP01087487, GBBP01052257, GBBP01035041, GBBP01013511, GBBP01076789, GBBP01037147, GBBP01037184, GBBP01041269, GBBP01130520, GBBP01052076, GBBP01042577, GBBP01054104, GBBP01054167, GBBP01064369, GBBP01081694, GBBP01043905, GBBP01015449, GBBP01064676, GBBP01072051, GBBP01087855, GBBP01044033, GBBP01074354, GBBP01077960, GBBP01079055, GBBP01080199, GBBP01113631, GBBP01155502, GBBP01194611, GBBP01064676

- The National Center of Biotechnology Information (NCBI), and VectorBase: GMOY002421, GMOY002543, GMOY002684, GMOY003090, GMOY003575, GMOY004864, GMOY006103, GMOY006640, GMOY006832, GMOY007230, GMOY007568, GMOY007874, GMOY008210, GMOY009558, GMOY009911, GMOY010870, GMOY010871, GMOY011241, GMOY011667, GMOY011672, GMOY011701, GMOY012330, GMOY013299, GMOY013300, GMOY013302,

GMOY013303, LLOJ000608, LLOJ002548, LLOJ005819, LLOJ006374, LLOJ006852, LLOJ008094, LLOJ010089, LLOJ010089, MDOA000488, MDOA000548PA, MDOA001306, MDOA002306, MDOA002808, MDOA004695, MDOA006332PA, MDOA007103, MDOA007155PA, MDOA007400, MDOA007993, MDOA008701, MDOA009054, MDOA011431, MDOA011608, MDOA011810, MDOA012163, MDOA012708, MDOA013911, MDOA014337, MDOA014768, PPAIP000122, PPAIP000132, PPAIP003111, PPAIP003142PA, PPAIP003321, PPAIP006241, PPAIP006242, PPAIP010765.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.10012#supplemental-information.

REFERENCES

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *Journal of Molecular Biology* **215(3)**:403–410 DOI 10.1016/S0022-2836(05)80360-2.

Arakane Y, Muthukrishnan S, Kramer KJ, Specht CA, Tomoyasu Y, Lorenzen MD, Kanost M, Beeman RW. 2005. The *Tribolium* chitin synthase genes TcCHS1 and TcCHS2 are specialized for synthesis of epidermal cuticle and midgut peritrophic matrix. *Insect Molecular Biology* **14(5)**:453–463 DOI 10.1111/j.1365-2583.2005.00576.x.

Armisen D, Rajakumar R, Friedrich M, Benoit JB. 2018. The genome of the water strider *Gerris buenoi* reveals expansions of gene repertoires associated with adaptations to life on the water. *BMC Genomics* **19(1)**:832 DOI 10.1186/s12864-018-5163-2.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. 2000. Gene ontology: tool for the unification of biology. *Nature Genetics* **25(1)**:25–29 DOI 10.1038/75556.

Bacigalupe LD, Crudgington HS, Hunter F, Moore AJ, Snook RR. 2007. Sexual conflict does not drive reproductive isolation in experimental populations of *Drosophila pseudoobscura*. *Journal of Evolutionary Biology* **20(5)**:1763–1771 DOI 10.1111/j.1420-9101.2007.01389.x.

Baker RH, Narechania A, DeSalle R, Johns PM, Reinhardt JA, Wilkinson GS. 2016. Spermatogenesis drives rapid gene creation and masculinization of the X chromosome in stalk-eyed flies (Diopsidae). *Genome Biology and Evolution* **8(3)**:896–914 DOI 10.1093/gbe/evw043.

Baker RH, Narechania A, Johns PM, Wilkinson GS. 2012. Gene duplication, tissue-specific gene expression and sexual conflict in stalk-eyed flies (Diopsidae). *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* **367(1600)**:2357–2375 DOI 10.1098/rstb.2011.0287.

Bao R, Dia SE, Issa HA, Alhusein D, Friedrich M. 2018. Comparative evidence of an exceptional impact of gene duplication on the developmental evolution of *Drosophila* and the higher Diptera. *Frontiers in Ecology and Evolution* **6**:63 DOI 10.3389/fevo.2018.00063.

Bao R, Friedrich M. 2009. Molecular evolution of the *Drosophila* retinome: exceptional gene gain in the higher Diptera. *Molecular Biology and Evolution* **26(6)**:1273–1287 DOI 10.1093/molbev/msp039.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B: Statistical Methodology* **57**:289–300.

Bourbon H-M, Gonzy-Treboul G, Peronnet F, Alin M-F, Ardourel C, Benassayag C, Cribbs D, Deutsch J, Ferrer P, Haenlin M, Lepesant J-A, Noselli S, Vincent A. 2002. A P-insertion screen identifying novel X-linked essential genes in *Drosophila*. *Mechanisms of Development* **110(1–2)**:71–83 DOI 10.1016/S0925-4773(01)00566-4.

Cañestro C, Yokoi H, Postlethwait JH. 2007. Evolutionary developmental biology and genomics. *Nature Reviews Genetics* **8(12)**:932–942 DOI 10.1038/nrg2226.

Cañestro C, Albalat R, Irimia M, Garcia-Fernàndez J. 2013. Impact of gene gains, losses and duplication modes on the origin and diversification of vertebrates. *Seminars in Cell & Developmental Biology* **24(2)**:83–94 DOI 10.1016/j.semcdb.2012.12.008.

Cardoso-Moreira M, Arguello JR, Gottipati S, Harshman LG, Grenier JK, Clark AG. 2016. Evidence for the fixation of gene duplications by positive selection in *Drosophila*. *Genome Research* **26(6)**:787–798 DOI 10.1101/gr.199323.115.

Carmon A, MacIntyre R. 2010. The α glycerophosphate cycle in *Drosophila melanogaster* VI: structure and evolution of enzyme paralogs in the genus Drosophila. *Journal of Heredity* **101(2)**:225–234 DOI 10.1093/jhered/esp111.

Carroll SB. 1995. Homeotic genes and the evolution of arthropods and chordates. *Nature* **376(6540)**:479–485 DOI 10.1038/376479a0.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular Biology and Evolution* **17(4)**:540–552 DOI 10.1093/oxfordjournals.molbev.a026334.

Chakraborty M, Fry JD. 2015. Parallel functional changes in independent testis-specific duplicates of Aldehyde dehydrogenase in *Drosophila*. *Molecular Biology and Evolution* **32(4)**:1029–1038 DOI 10.1093/molbev/msu407.

Chen Z-X, Sturgill D, Qu J, Jiang H, Park S, Boley N, Suzuki AM, Fletcher AR, Plachetzki DC, FitzGerald PC, Artieri CG, Atallah J, Barmina O, Brown JB, Blankenburg KP, Clough E, Dasgupta A, Gubbala S, Han Y, Jayaseelan JC, Kalra D, Kim Y-A, Kovar CL, Lee SL, Li M, Malley JD, Malone JH, Mathew T, Mattiuzzo NR, Munidasa M, Muzny DM, Ongeri F, Perales L, Przytycka TM, Pu L-L, Robinson G, Thornton RL, Saada N, Scherer SE, Smith HE, Vinson C, Warner CB, Worley KC, Wu Y-Q, Zou X, Cherbas P, Kellis M, Eisen MB, Piano F, Kionte K, Fitch DH, Sternberg PW, Cutter AD, Duff MO, Hoskins RA, Graveley BR, Gibbs RA, Bickel PJ, Kopp A, Carninci P, Celniker SE, Oliver B, Richards S. 2014. Comparative validation of the D. melanogaster modENCODE transcriptome annotation. *Genome Research* **24(7)**:1209–1223 DOI 10.1101/gr.159384.113.

Chintapalli VR, Wang J, Dow JAT. 2007. Using FlyAtlas to identify better *Drosophila melanogaster* models of human disease. *Nature Genetics* **39(6)**:715–720 DOI 10.1038/ng2049.

Conant GC, Wagner A. 2004. Duplicate genes and robustness to transient gene knock-downs in Caenorhabditis elegans. *Proceedings of the Royal Society B: Biological Sciences* **271**:89–96.

Conant GC. 2020. The lasting after-effects of an ancient polyploidy on the genomes of teleosts. *PLOS ONE* **15**:e0231356.

Congrains C, Campanini EB, Torres FR, Rezende VB, Nakamura AM, De Oliveira JL, Lima ALA, Chahad-Ehlers S, Sobrinho IS Jr, De Brito RA. 2018. Evidence of adaptive evolution and relaxed constraints in sex-biased genes of South American and West Indies fruit flies (Diptera: Tephritidae). *Genome Biology and Evolution* **10**:380–395.

Connallon T, Clark AG. 2011. The resolution of sexual antagonism by gene duplication. *Genetics* **187(3)**:919–937 DOI 10.1534/genetics.110.123729.

Coyne JA, Orr HA. 1989. Patterns of speciation in *Drosophila*. *Evolution* **43(2)**:362–381 DOI 10.1111/j.1558-5646.1989.tb04233.x.

Davidson EH, Erwin DH. 2006. Gene regulatory networks and the evolution of animal body plans. *Science* **311(5762)**:796–800 DOI 10.1126/science.1113832.

Dean EJ, Davis JC, Davis RW, Petrov DA. 2008. Pervasive and persistent redundancy among duplicated genes in yeast. *PLOS Genetics* **4(7)**:e1000113 DOI 10.1371/journal.pgen.1000113.

Dennis G Jr, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA. 2003. DAVID: database for annotation, visualization, and integrated discovery. *Genome Biology* **4(5)**:P3 DOI 10.1186/gb-2003-4-5-p3.

Dikow RB, Frandsen PB, Turcatel M, Dikow T. 2017. Genomic and transcriptomic resources for assassin flies including the complete genome sequence of Proctacanthus coquilletti (*Insecta: Diptera: Asilidae*) and 16 representative transcriptomes. *PeerJ* **5**:e2951.

Dixit R, Arakane Y, Specht CA, Richard C, Kramer KJ, Beeman RW, Muthukrishnan S. 2008. Domain organization and phylogenetic analysis of proteins from the chitin deacetylase gene family of *Tribolium castaneum* and three other species of insects. *Insect Biochemistry and Molecular Biology* **38(4)**:440–451 DOI 10.1016/j.ibmb.2007.12.002.

Drăghici S. 2011. *Statistics and data analysis for microarrays using R and bioconductor*. Boca Raton: CRC Press.

Drysdale RA, Crosby MA, FlyBase Consortium. 2005. FlyBase: genes and gene models. *Nucleic Acids Research* **33**:D390–D395.

Felsenstein J. 2005. *PHYLIP*. Version 3.6. Software package, Department of Genome Sciences, University of Washington, Seattle, USA. *Available at https://evolution.genetics.washington.edu/phylip.html*.

Finke MD. 2007. Estimate of chitin in raw whole insects. *Zoo Biology* **26(2)**:105–115 DOI 10.1002/zoo.20123.

Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J. 1999. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151**:1531–1545.

Foster WA. 1995. Mosquito sugar feeding and reproductive energetics. *Annual Review of Entomology* **40(1)**:443–474 DOI 10.1146/annurev.en.40.010195.002303.

Fraga A, Ribeiro L, Lobato M, Santos V, Silva JR, Gomes H, Da Cunha Moraes JL, De Souza Menezes J, De Oliveira CJL, Campos E, Da Fonseca RN. 2013. Glycogen and glucose metabolism are essential for early embryonic development of the red flour beetle *Tribolium castaneum*. *PLOS ONE* **8(6)**:e65125 DOI 10.1371/journal.pone.0065125.

Friedrich M. 2017. Ancient genetic redundancy of eyeless and twin of eyeless in the arthropod ocular segment. *Developmental Biology* **432**:192–200.

Gallach M, Chandrasekaran C, Betrán E. 2010. Analyses of nuclearly encoded mitochondrial genes suggest gene duplication as a mechanism for resolving intralocus sexually antagonistic conflict in Drosophila. *Genome Biology and Evolution* **2(5)**:835–850 DOI 10.1093/gbe/evq069.

Gallach M, Betrán E. 2011. Intralocus sexual conflict resolved through gene duplication. *Trends in Ecology & Evolution* **26(5)**:222–228 DOI 10.1016/j.tree.2011.02.004.

Gavrilets S. 2014. Is sexual conflict an "engine of speciation"? *Cold Spring Harbor Perspectives in Biology* **6**:a0177231-13 DOI 10.1101/cshperspect.a017723.

Gokhman VE, Kuznetsova VG. 2018. Parthenogenesis in Hexapoda: holometabolous insects. *Journal of Zoological Systematics and Evolutionary Research* **56(1)**:23–34 DOI 10.1111/jzs.12183.

Grimaldi D, Engel MS. 2005. *Evolution of the insects*. Cambridge: Cambridge University Press.

Gu Z, Steinmetz LM, Gu X, Scharfe C, Davis RW, Li W-H. 2003. Role of duplicate genes in genetic robustness against null mutations. *Nature* **421**:63–66.

Haerty W, Jagadeeshan S, Kulathinal RJ, Wong A, Ravi Ram K, Sirot LK, Levesque L, Artieri CG, Wolfner MF, Civetta A, Singh RS. 2007. Evolution in the fast lane: rapidly evolving sex-related genes in *Drosophila*. *Genetics* **177(3)**:1321–1335 DOI 10.1534/genetics.107.078865.

Hahn MW, Han MV, Han S-G. 2007. Gene family evolution across 12 *Drosophila* genomes. *PLOS Genetics* **3(11)**:e197 DOI 10.1371/journal.pgen.0030197.

Hanada K, Kuromori T, Myouga F, Toyoda T, Li W-H, Shinozaki K. 2009. Evolutionary persistence of functional compensation by duplicate genes in Arabidopsis. *Genome Biology and Evolution* **1**:409–414.

He S, Del Viso F, Chen C-Y, Ikmi A, Kroesen AE, Gibson MC. 2018. An axial Hox code controls tissue segmentation and body patterning in *Nematostella vectensis*. *Science* **361(6409)**:1377–1380 DOI 10.1126/science.aar8384.

Helleu Q, Levine MT. 2018. Recurrent amplification of the heterochromatin protein 1 (HP1) gene family across Diptera. *Molecular Biology and Evolution* **35(10)**:2375–2389 DOI 10.1093/molbev/msy128.

Holland PWH, Marlétaz F, Maeso I, Dunwell TL, Paps J. 2017. New genes from old: asymmetric divergence of gene duplicates and the evolution of development. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* **372(1713)**:20150480 DOI 10.1098/rstb.2015.0480.

Honeybee Genome Sequencing Consortium. 2006. Insights into social insects from the genome of the honeybee *Apis mellifera*. *Nature* **443(7114)**:931–949 DOI 10.1038/nature05260.

Hsiao T-L, Vitkup D. 2008. Role of duplicate genes in robustness against deleterious human mutations. *PLOS Genetics* **4**:e1000014.

i5K Consortium. 2013. The i5K initiative: advancing arthropod genomics for knowledge, human health, agriculture, and the environment. *Journal of Heredity* **104**:595–600.

Innan H, Kondrashov F. 2010. The evolution of gene duplications: classifying and distinguishing between models. *Nature Reviews Genetics* **11(2)**:97–108 DOI 10.1038/nrg2689.

International Aphid Genomics Consortium. 2010. Genome sequence of the pea aphid *Acyrthosiphon pisum*. *PLOS Biology* **8(2)**:e1000313 DOI 10.1371/journal.pbio.1000313.

International Glossina Genome Initiative. 2014. Genome sequence of the tsetse fly (*Glossina morsitans*): vector of African trypanosomiasis. *Science* **344**:380–386.

Jiang X, Assis R. 2017. Natural selection drives rapid functional evolution of young *Drosophila* duplicate genes. *Molecular Biology and Evolution* **34(12)**:3089–3098 DOI 10.1093/molbev/msx230.

Jiménez-Guri E, Huerta-Cepas J, Cozzuto L, Wotton KR, Kang H, Himmelbauer H, Roma G, Gabaldón T, Jaeger J. 2013. Comparative transcriptomics of early dipteran development. *BMC Genomics* **14(1)**:123 DOI 10.1186/1471-2164-14-123.

Jiao Y, Wickett NJ, Ayyampalayam S, Chanderbali AS, Landherr L, Ralph PE, Tomsho LP, Hu Y, Liang H, Soltis PS, Soltis DE, Clifton SW, Schlarbaum SE, Schuster SC, Ma H, Leebens-Mack J, DePamphilis CW. 2011. Ancestral polyploidy in seed plants and angiosperms. *Nature* **473(7345)**:97–100 DOI 10.1038/nature09916.

Julca I, Marcet-Houben M, Cruz F, Vargas-Chavez C, Johnston JS, Gómez-Garrido J, Frias L, Corvelo A, Loska D, Cámara F, Gut M, Alioto T, Latorre A, Gabaldón T. 2020.

Phylogenomics identifies an ancestral burst of gene duplications predating the diversification of Aphidomorpha. *Molecular Biology and Evolution* **37**:730–756.

Katzourakis A, Purvis A, Azmeh S, Rotheray G, Gilbert F. 2001. Macroevolution of hoverflies (Diptera: Syrphidae): the effect of using higher-level taxa in studies of biodiversity, and correlates of species richness. *Journal of Evolutionary Biology* **14(2)**:219–227 DOI 10.1046/j.1420-9101.2001.00278.x.

Kondo S, Vedanayagam J, Mohammed J, Eizadshenass S, Kan L, Pang N, Aradhya R, Siepel A, Steinhauer J, Lai EC. 2017. New genes often acquire male-specific functions but rarely become essential in *Drosophila*. *Genes & Development* **31(18)**:1841–1846 DOI 10.1101/gad.303131.117.

Kondrashov FA. 2012. Gene duplication as a mechanism of genomic adaptation to a changing environment. *Proceedings of the Royal Society B* **279(1749)**:5048–5057 DOI 10.1098/rspb.2012.1108.

Labandeira CC, Sepkoski JJ Jr. 1993. Insect diversity in the fossil record. *Science* **261(5119)**:310–315 DOI 10.1126/science.11536548.

Lan X, Pritchard JK. 2016. Coregulation of tandem duplicate genes slows evolution of subfunctionalization in mammals. *Science* **352(6288)**:1009–1013 DOI 10.1126/science.aad8411.

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG. 2007. Clustal W and Clustal X version 2.0. *Bioinformatics* **23(21)**:2947–2948 DOI 10.1093/bioinformatics/btm404.

Letunic I, Bork P. 2016. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Research* **44(W1)**:W242–W245 DOI 10.1093/nar/gkw290.

Lewis EB. 1992. Clusters of master control genes regulate the development of higher organisms. *JAMA* **267(11)**:1524–1531 DOI 10.1001/jama.1992.03480110100042.

Lewis SH, Salmela H, Obbard DJ. 2016. Duplication and diversification of dipteran *Argonaute* genes, and the evolutionary divergence of *Piwi* and *Aubergine*. *Genome Biology and Evolution* **8(3)**:507–518 DOI 10.1093/gbe/evw018.

Li Z, Tiley G, Galuska S, Reardon C, Kidder T, Rundell R, Barker MS. 2018. Multiple large-scale gene and genome duplications during the evolution of hexapods. *Proceedings of the National Academy of Sciences* **115(18)**:4713–4718 DOI 10.1073/pnas.1710791115.

Lynch M, Conery JS. 2000. The evolutionary fate and consequences of duplicate genes. *Science* **290(5494)**:1151–1155 DOI 10.1126/science.290.5494.1151.

Maere S, Heymans K, Kuiper M. 2005. BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics* **21(16)**:3448–3449 DOI 10.1093/bioinformatics/bti551.

Marlétaz F, Firbas PN, Maeso I, Tena JJ, Bogdanovic O, Perry M, Wyatt CDR, De la Calle-Mustienes E, Bertrand S, Burguera D, Acemel RD, Van Heeringen SJ, Naranjo S, Herrera-Ubeda C, Skvortsova K, Jimenez-Gancedo S, Aldea D, Marquez Y, Buono L, Kozmikova I, Permanyer J, Louis A, Albuixech-Crespo B, Le Petillon Y, Leon A, Subirana L, Balwierz PJ, Duckett PE, Farahani E, Aury J-M, Mangenot S, Wincker P, Albalat R, Benito-Gutiérrez È, Cañestro C, Castro F, D'Aniello S, Ferrier DEK, Huang S, Laudet V, Marais GAB, Pontarotti P, Schubert M, Seitz H, Somorjai I, Takahashi T, Mirabeau O, Xu A, Yu J-K, Carninci P, Martinez-Morales JR, Crollius HR, Kozmik Z, Weirauch MT, Garcia-Fernàndez J, Lister R, Lenhard B, Holland PWH, Escriva H, Gómez-Skarmeta JL,

Irimia M. 2018. *Amphioxus* functional genomics and the origins of vertebrate gene regulation. *Nature* **564(7734)**:64–70 DOI 10.1038/s41586-018-0734-6.

MacKintosh C, Ferrier DEK. 2017. Recent advances in understanding the roles of whole genome duplications in evolution. *F1000Research* **6**:1623.

McGarvey PB, Nightingale A, Luo J, Huang H, Martin MJ, Wu C, the UniProt Consortium. 2019. UniProt genomic mapping for deciphering functional effects of missense variants. *Human Mutation* **40**:694–705.

Meyer RE, Delaage M, Rosset R, Capri M, Aït-Ahmed O. 2010. A single mutation results in diploid gamete formation and parthenogenesis in a *Drosophila yemanuclein-alpha meiosis I* defective mutant. *BMC Genetics* **11(1)**:104 DOI 10.1186/1471-2156-11-104.

Mikhaylova LM, Nguyen K, Nurminsky DI. 2008. Analysis of the *Drosophila melanogaster* testes transcriptome reveals coordinate regulation of paralogous genes. *Genetics* **179(1)**:305–315 DOI 10.1534/genetics.107.080267.

Miller MA, Pfeiffer W, Schwartz T. 2010. Creating the CIPRES science gateway for inference of large phylogenetic trees. In: *2010 Gateway Computing Environments Workshop (GCE).* Piscataway: IEEE, 1–8.

Miura T, Braendle C, Shingleton A, Sisk G, Kambhampati S, Stern DL. 2003. A comparison of parthenogenetic and sexual embryogenesis of the pea aphid *Acyrthosiphon pisum* (Hemiptera: Aphidoidea). *Journal of Experimental Zoology* **295B(1)**:59–81 DOI 10.1002/jez.b.3.

Moser D, Johnson L, Lee CY. 1980. Multiple forms of *Drosophila* hexokinase: purification, biochemical and immunological characterization. *Journal of Biological Chemistry* **255**:4673–4679.

Myers EW, Sutton GG, Delcher AL, Dew IM, Fasulo DP, Flanigan MJ, Kravitz SA, Mobarry CM, Reinert KH, Remington KA, Anson EL, Bolanos RA, Chou HH, Jordan CM, Halpern AL, Lonardi S, Beasley EM, Brandon RC, Chen L, Dunn PJ, Lai Z, Liang Y, Nusskern DR, Zhan M, Zhang Q, Zheng X, Rubin GM, Adams MD, Venter JC. 2000. A whole-genome assembly of *Drosophila*. *Science* **287(5461)**:2196–2204 DOI 10.1126/science.287.5461.2196.

Nakatani Y, McLysaght A. 2019. Macrosynteny analysis shows the absence of ancient whole-genome duplication in lepidopteran insects. *Proceedings of the National Academy of Sciences* **116(6)**:1816–1818 DOI 10.1073/pnas.1817937116.

Neafsey DE, Waterhouse RM, Abai MR, Aganezov SS, Alekseyev MA, Allen JE, Amon J, Arcà B, Arensburger P, Artemov G, Assour LA, Basseri H, Berlin A, Birren BW, Blandin SA, Brockman AI, Burkot TR, Burt A, Chan CS, Chauve C, Chiu JC, Christensen M, Costantini C, Davidson VLM, Deligianni E, Dottorini T, Dritsou V, Gabriel SB, Guelbeogo WM, Hall AB, Han MV, Hlaing T, Hughes DST, Jenkins AM, Jiang X, Jungreis I, Kakani EG, Kamali M, Kemppainen P, Kennedy RC, Kirmitzoglou IK, Koekemoer LL, Laban N, Langridge N, Lawniczak MKN, Lirakis M, Lobo NF, Lowy E, MacCallum RM, Mao C, Maslen G, Mbogo C, McCarthy J, Michel K, Mitchell SN, Moore W, Murphy KA, Naumenko AN, Nolan T, Novoa EM, O'Loughlin S, Oringanje C, Oshaghi MA, Pakpour N, Papathanos PA, Peery AN, Povelones M, Prakash A, Price DP, Rajaraman A, Reimer LJ, Rinker DC, Rokas A, Russell TL, Sagnon N, Sharakhova MV, Shea T, Simão FA, Simard F, Slotman MA, Somboon P, Stegniy V, Struchiner CJ, Thomas GWC, Tojo M, Topalis P, Tubio JMC, Unger MF, Vontas J, Walton C, Wilding CS, Willis JH, Wu Y-C, Yan G, Zdobnov EM, Zhou X, Catteruccia F, Christophides GK, Collins FH, Cornman RS, Crisanti A, Donnelly MJ, Emrich SJ, Fontaine MC, Gelbart W, Hahn MW, Hansen IA, Howell PI, Kafatos FC, Kellis M, Lawson D, Louis C, Luckhart S, Muskavitch MAT,

Ribeiro JM, Riehle MA, Sharakhov IV, Tu Z, Zwiebel LJ, Besansky NJ. 2015. Highly evolvable malaria vectors: the genomes of 16 *Anopheles* mosquitoes. *Science* **347(6217)**:1258522 DOI 10.1126/science.1258522.

Nene V, Wortman JR, Lawson D, Haas B, Kodira C, Tu ZJ, Loftus B, Xi Z, Megy K, Grabherr M, Ren Q, Zdobnov EM, Lobo NF, Campbell KS, Brown SE, Bonaldo MF, Zhu J, Sinkins SP, Hogenkamp DG, Amedeo P, Arensburger P, Atkinson PW, Bidwell S, Biedler J, Birney E, Bruggner RV, Costas J, Coy MR, Crabtree J, Crawford M, Debruyn B, Decaprio D, Eiglmeier K, Eisenstadt E, El-Dorry H, Gelbart WM, Gomes SL, Hammond M, Hannick LI, Hogan JR, Holmes MH, Jaffe D, Johnston JS, Kennedy RC, Koo H, Kravitz S, Kriventseva EV, Kulp D, Labutti K, Lee E, Li S, Lovin DD, Mao C, Mauceli E, Menck CFM, Miller JR, Montgomery P, Mori A, Nascimento AL, Naveira HF, Nusbaum C, O'leary S, Orvis J, Pertea M, Quesneville H, Reidenbach KR, Rogers Y-H, Roth CW, Schneider JR, Schatz M, Shumway M, Stanke M, Stinson EO, Tubio JMC, Vanzee JP, Verjovski-Almeida S, Werner D, White O, Wyder S, Zeng Q, Zhao Q, Zhao Y, Hill CA, Raikhel AS, Soares MB, Knudson DL, Lee NH, Galagan J, Salzberg SL, Paulsen IT, Dimopoulos G, Collins FH, Birren B, Fraser-Liggett CM, Severson DW. 2007. Genome sequence of *Aedes aegypti*, a major arbovirus vector. *Science* **316(5832)**:1718–1723 DOI 10.1126/science.1138878.

Nong W, Qu Z, Li Y, Barton-Owen T, Wong AYP, Yip HY, Lee HT, Narayana S, Baril T, Swale T, Cao J, Chan TF, Kwan HS, Ming NS, Panagiotou G, Qian P-Y, Qiu J-W, Yip KY, Ismail N, Pati S, John A, Tobe SS, Bendena WG, Cheung SG, Hayward A, Hui JHL. 2020. Horseshoe crab genomes reveal the evolutionary fates of genes and microRNAs after three rounds (3R) of whole genome duplication. *bioRxiv* DOI 10.1101/2020.04.16.045815.

Obbard DJ, Maclennan J, Kim K-W, Rambaut A, O'Grady PM, Jiggins FM. 2012. Estimating divergence dates and substitution rates in the *Drosophila* phylogeny. *Molecular Biology and Evolution* **29(11)**:3459–3473 DOI 10.1093/molbev/mss150.

Ohno S. 1970. *Evolution by gene duplication*. New York: Springer Verlag.

Pajic P, Pavlidis P, Dean K, Neznanova L, Romano R-A, Garneau D, Daugherity E, Globig A, Ruhl S, Gokcumen O. 2019. Independent amylase gene copy number bursts correlate with dietary preferences in mammals. *eLife* **8**:e44628 DOI 10.7554/eLife.44628.

Panfilio KA, Vargas Jentzsch IM, Benoit JB, Erezyilmaz D, Suzuki Y, Colella S, Robertson HM, Poelchau MF, Waterhouse RM, Ioannidis P, Weirauch MT, Hughes DST, Murali SC, Werren JH, Jacobs CGC, Duncan EJ, Armisén D, Vreede BMI, Baa-Puyoulet P, Berger CS, Chang C-C, Chao H, Chen M-JM, Chen Y-T, Childers CP, Chipman AD, Cridge AG, Crumière AJJ, Dearden PK, Didion EM, Dinh H, Doddapaneni HV, Dolan A, Dugan S, Extavour CG, Febvay G, Friedrich M, Ginzburg N, Han Y, Heger P, Holmes CJ, Horn T, Hsiao Y-M, Jennings EC, Johnston JS, Jones TE, Jones JW, Khila A, Koelzer S, Kovacova V, Leask M, Lee SL, Lee C-Y, Lovegrove MR, Lu H-L, Lu Y, Moore PJ, Munoz-Torres MC, Muzny DM, Palli SR, Parisot N, Pick L, Porter ML, Qu J, Refki PN, Richter R, Rivera-Pomar R, Rosendale AJ, Roth S, Sachs L, Santos ME, Seibert J, Sghaier E, Shukla JN, Stancliffe RJ, Tidswell O, Traverso L, Van der Zee M, Viala S, Worley KC, Zdobnov EM, Gibbs RA, Richards S. 2019. Molecular evolutionary trends and feeding ecology diversification in the Hemiptera, anchored by the milkweed bug genome. *Genome Biology* **20(1)**:64 DOI 10.1186/s13059-019-1660-0.

Papanicolaou A, Schetelig MF, Arensburger P, Atkinson PW, Benoit JB, Bourtzis K, Castañera P, Cavanaugh JP, Chao H, Childers C, Curril I, Dinh H, Doddapaneni H, Dolan A, Dugan S, Friedrich M, Gasperi G, Geib S, Georgakilas G, Gibbs RA, Giers SD, Gomulski LM, González-Guzmán M, Guillem-Amat A, Han Y, Hatzigeorgiou AG, Hernández-Crespo P, Hughes DST, Jones JW, Karagkouni D, Koskinioti P, Lee SL,

Malacrida AR, Manni M, Mathiopoulos K, Meccariello A, Murali SC, Murphy TD, Muzny DM, Oberhofer G, Ortego F, Paraskevopoulou MD, Poelchau M, Qu J, Reczko M, Robertson HM, Rosendale AJ, Rosselot AE, Saccone G, Salvemini M, Savini G, Schreiner P, Scolari F, Siciliano P, Sim SB, Tsiamis G, Ureña E, Vlachos IS, Werren JH, Wimmer EA, Worley KC, Zacharopoulou A, Richards S, Handler AM. 2016. The whole genome sequence of the Mediterranean fruit fly, *Ceratitis capitata* (Wiedemann), reveals insights into the b*iology* and adaptive evolution of a highly invasive pest species. *Genome Biology* **17(1)**:192 DOI 10.1186/s13059-016-1049-2.

Parker GA, Partridge L. 1998. Sexual conflict and speciation. *Philosophical Transactions of the Royal Society B: Biological Sciences* **353(1366)**:261–274 DOI 10.1098/rstb.1998.0208.

Patel MR, Miriyala GK, Littleton AJ, Yang H, Trinh K, Young JM, Kennedy SR, Yamashita YM, Pallanck LJ, Malik HS. 2016. A mitochondrial DNA hypomorph of cytochrome oxidase specifically impairs male fertility in *Drosophila melanogaster*. *eLife* **5**:e16923 DOI 10.7554/eLife.16923.030.

Pelletier J, Leal WS. 2009. Genome analysis and expression patterns of odorant-binding proteins from the Southern House mosquito *Culex pipiens quinquefasciatus*. *PLOS ONE* **4(7)**:e6237 DOI 10.1371/journal.pone.0006237.

Porcelli D, Barsanti P, Pesole G, Caggese C. 2007. The nuclear OXPHOS genes in insecta: a common evolutionary origin, a common cis-regulatory motif, a common destiny for gene duplicates. *BMC Evolutionary Biology* **7(1)**:215 DOI 10.1186/1471-2148-7-215.

Pruitt KD, Tatusova T, Maglott DR. 2005. NCBI reference sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research* **33**:D501–D504 DOI 10.1093/nar/gki025.

Qian W, Zhang J. 2014. Genomic evidence for adaptation by gene duplication. *Genome Research* **24(8)**:1356–1362 DOI 10.1101/gr.172098.114.

Rand DM, Clark AG, Kann LM. 2001. Sexually antagonistic cytonuclear fitness interactions in Drosophila melanogaster. *Genetics* **159**:173–187.

Rettie EC, Dorus S. 2012. *Drosophila* sperm proteome evolution: insights from comparative genomic approaches. *Spermatogenesis* **2(3)**:213–223 DOI 10.4161/spmg.21748.

Richards S, Gibbs RA, Weinstock GM, Brown SJ, Denell R, Beeman RW, Gibbs R, Beeman RW, Brown SJ, Bucher G, Friedrich M, Grimmelikhuijzen CJP, Klingler M, Lorenzen M, Richards S, Roth S, Schröder R, Tautz D, Zdobnov EM, Muzny D, Gibbs RA, Weinstock GM, Attaway T, Bell S, Buhay CJ, Chandrabose MN, Chavez D, Clerk-Blankenburg KP, Cree A, Dao M, Davis C, Chacko J, Dinh H, Dugan-Rocha S, Fowler G, Garner TT, Garnes J, Gnirke A, Hawes A, Hernandez J, Hines S, Holder M, Hume J, Jhangiani SN, Joshi V, Khan ZM, Jackson L, Kovar C, Kowis A, Lee S, Lewis LR, Margolis J, Morgan M, Nazareth LV, Nguyen N, Okwuonu G, Parker D, Richards S, Ruiz S-J, Santibanez J, Savard J, Scherer SE, Schneider B, Sodergren E, Tautz D, Vattahil S, Villasana D, White CS, Wright R, Park Y, Beeman RW, Lord J, Oppert B, Lorenzen M, Brown S, Wang L, Savard J, Tautz D, Richards S, Weinstock G, Gibbs RA, Liu Y, Worley K, Weinstock G, Elsik CG, Reese JT, Elhaik E, Landan G, Graur D, Arensburger P, Atkinson P, Beeman RW, Beidler J, Brown SJ, Demuth JP, Drury DW, Du Y-Z, Fujiwara H, Lorenzen M, Maselli V, Osanai M, Park Y, Robertson HM, Tu Z, Wang J-J, Wang S, Richards S, Song H, Zhang L, Sodergren E, Werner D, Stanke M, Morgenstern B, Solovyev V, Kosarev P, Brown G, Chen H-C, Ermolaeva O, Hlavina W, Kapustin Y, Kiryutin B, Kitts P, Maglott D, Pruitt K, Sapojnikov V, Souvorov A, Mackey AJ, Waterhouse RM, Wyder S, Zdobnov EM, Zdobnov EM, Wyder S, Kriventseva EV, Kadowaki T, Bork P, Aranda M, Bao R, Beermann A, Berns N, Bolognesi R, Bonneton F, Bopp D, Brown SJ, Bucher G, Butts T,

Chaumot A, Denell RE, Ferrier DEK, Friedrich M, Gordon CM, Jindra M, Klingler M, Lan Q, Lattorff HMG, Laudet V, Von Levetsow C, Liu Z, Lutz R, Lynch JA, Da Fonseca RN, Posnien N, Reuter R, Roth S, Savard J, Schinko JB, Schmitt C, Schoppmeier M, Schröder R, Shippy TD, Simonnet F, Marques-Souza H, Tautz D, Tomoyasu Y, Trauner J, Van der Zee M, Vervoort M, Wittkopp N, Wimmer EA, Yang X, Jones AK, Sattelle DB, Ebert PR, Nelson D, Scott JG, Beeman RW, Muthukrishnan S, Kramer KJ, Arakane Y, Beeman RW, Zhu Q, Hogenkamp D, Dixit R, Oppert B, Jiang H, Zou Z, Marshall J, Elpidina E, Vinokurov K, Oppert C, Zou Z, et al. 2008. The genome of the model beetle and pest *Tribolium castaneum*. *Nature* **452(7190)**:949–955 DOI 10.1038/nature06784.

Rodrigo G, Fares MA. 2018. Intrinsic adaptive value and early fate of gene duplication revealed by a bottom-up approach. *eLife* **7**:e29739 DOI 10.7554/eLife.29739.

Roelofs D, Zwaenepoel A, Sistermans T, Nap J, Kampfraath AA, Van de Peer Y, Ellers J, Kraaijeveld K. 2020. Multi-faceted analysis provides little evidence for recurrent whole-genome duplications during hexapod evolution. *BMC Biology* **18**:57.

Ryan JF, Mazza ME, Pang K, Matus DQ, Baxevanis AD, Martindale MQ, Finnerty JR. 2007. Pre-bilaterian origins of the Hox cluster and the Hox code: evidence from the sea anemone, *Nematostella vectensis*. *PLOS ONE* **2(1)**:e153 DOI 10.1371/journal.pone.0000153.

Saitou N, Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* **4**:406–425.

Sakuma S, Golan G, Guo Z, Ogawa T, Tagiri A, Sugimoto K, Bernhardt N, Brassac J, Mascher M, Hensel G, Ohnishi S, Jinno H, Yamashita Y, Ayalon I, Peleg Z, Schnurbusch T, Komatsuda T. 2019. Unleashing floret fertility in wheat through the mutation of a homeobox gene. *Proceedings of the National Academy of Sciences* **116(11)**:5182–5187 DOI 10.1073/pnas.1815465116.

Sandve SR, Rohlfs RV, Hvidsten TR. 2018. Subfunctionalization versus neofunctionalization after whole-genome duplication. *Nature Genetics* **50(7)**:908–909 DOI 10.1038/s41588-018-0162-4.

Schlicker A, Domingues FS, Rahnenführer J, Lengauer T. 2006. A new measure for functional similarity of gene products based on gene ontology. *BMC Bioinformatics* **7(1)**:302 DOI 10.1186/1471-2105-7-302.

Schwager EE, Sharma PP, Clarke T, Leite DJ, Wierschin T, Pechmann M, Akiyama-Oda Y, Esposito L, Bechsgaard J, Bilde T, Buffry AD, Chao H, Dinh H, Doddapaneni H, Dugan S, Eibner C, Extavour CG, Funch P, Garb J, Gonzalez LB, Gonzalez VL, Griffiths-Jones S, Han Y, Hayashi C, Hilbrant M, Hughes DST, Janssen R, Lee SL, Maeso I, Murali SC, Muzny DM, Nunes da Fonseca R, Paese CLB, Qu J, Ronshaugen M, Schomburg C, Schönauer A, Stollewerk A, Torres-Oliva M, Turetzek N, Vanthournout B, Werren JH, Wolff C, Worley KC, Bucher G, Gibbs RA, Coddington J, Oda H, Stanke M, Ayoub NA, Prpic N-M, Flot J-F, Posnien N, Richards S, McGregor AP. 2017. The house spider genome reveals an ancient whole-genome duplication during arachnid evolution. *BMC Biology* **15(1)**:62 DOI 10.1186/s12915-017-0399-x.

Scott JG, Warren WC, Beukeboom LW, Bopp D, Clark AG, Giers SD, Hediger M, Jones AK, Kasai S, Leichter CA, Li M, Meisel RP, Minx P, Murphy TD, Nelson DR, Reid WR, Rinkevich FD, Robertson HM, Sackton TB, Sattelle DB, Thibaud-Nissen F, Tomlinson C, Van de Zande L, Walden KKO, Wilson RK, Liu N. 2014. Genome of the house fly, *Musca domestica* L., a global vector of diseases with adaptations to a septic environment. *Genome Biology* **15(10)**:466 DOI 10.1186/s13059-014-0466-3.

Sharakhova MV, Hammond MP, Lobo NF, Krzywinski J, Unger MF, Hillenmeyer ME, Bruggner RV, Birney E, Collins FH. 2007. Update of the *Anopheles gambiae* PEST genome assembly. *Genome Biology* **8(1)**:R5 DOI 10.1186/gb-2007-8-1-r5.

Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, Thompson JD, Higgins DG. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega. *Molecular Systems Biology* **7(1)**:539 DOI 10.1038/msb.2011.75.

Smadja C, Shi P, Butlin RK, Robertson HM. 2009. Large gene family expansions and adaptive e*volution* for odorant and gustatory receptors in the pea aphid, *Acyrthosiphon pisum*. *Molecular Biology and Evolution* **26(9)**:2073–2086 DOI 10.1093/molbev/msp116.

Smoot ME, Ono K, Ruscheinski J, Wang P-L, Ideker T. 2011. Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics* **27(3)**:431–432 DOI 10.1093/bioinformatics/btq675.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30(9)**:1312–1313 DOI 10.1093/bioinformatics/btu033.

Supek F, Bošnjak M, Škunca N, Šmuc T. 2011. REVIGO summarizes and visualizes long lists of gene ontology terms. *PLOS ONE* **6(7)**:e21800 DOI 10.1371/journal.pone.0021800.

Suyama M, Torrents D, Bork P. 2006. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Research* **34**:W609–W612 DOI 10.1093/nar/gkl315.

Svensson MJ, Chen JD, Pirrotta V, Larsson J. 2003. The ThioredoxinT and deadhead gene pair encode testis- and ovary-specific thioredoxins in Drosophila melanogaster. *Chromosoma* **112**:133–143.

Svensson MJ, Stenberg P, Larsson J. 2007. Organization and regulation of sex-specific thioredoxin encoding genes in the genus *Drosophila*. *Development Genes and Evolution* **217(9)**:639–650 DOI 10.1007/s00427-007-0175-y.

Svensson MJ, Larsson J. 2007. Thioredoxin-2 affects lifespan and oxidative stress in Drosophila. *Hereditas*.

Syed ZA, Chatterjee M, Samant MA, Prasad NG. 2017. Reproductive isolation through experimental manipulation of sexually antagonistic coevolution in *Drosophila melanogaster*. *Scientific Reports* **7(1)**:3330 DOI 10.1038/s41598-017-03182-1.

Talavera G, Castresana J. 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Systematic Biology* **56(4)**:564–577 DOI 10.1080/10635150701472164.

The Gene Ontology Consortium. 2015. Gene ontology consortium: going forward. *Nucleic Acids Research* **43**:D1049–D1056.

Thomas GWC, Dohmen E, Hughes DST, Murali SC, Poelchau M, Glastad K, Anstead CA, Ayoub NA, Batterham P, Bellair M, Binford GJ, Chao H, Chen YH, Childers C, Dinh H, Doddapaneni HV, Duan JJ, Dugan S, Esposito LA, Friedrich M, Garb J, Gasser RB, Goodisman MAD, Gundersen-Rindal DE, Han Y, Handler AM, Hatakeyama M, Hering L, Hunter WB, Ioannidis P, Jayaseelan JC, Kalra D, Khila A, Korhonen PK, Lee CE, Lee SL, Li Y, Lindsey ARI, Mayer G, McGregor AP, McKenna DD, Misof B, Munidasa M, Munoz-Torres M, Muzny DM, Niehuis O, Osuji-Lacy N, Palli SR, Panfilio KA, Pechmann M, Perry T, Peters RS, Poynton HC, Prpic N-M, Qu J, Rotenberg D, Schal C, Schoville SD, Scully ED, Skinner E, Sloan DB, Stouthamer R, Strand MR, Szucsich NU, Wijeratne A, Young ND, Zattara EE, Benoit JB, Zdobnov EM, Pfrender ME, Hackett KJ, Werren JH, Worley KC, Gibbs RA, Chipman AD, Waterhouse RM, Bornberg-Bauer E, Hahn MW,

Richards S. 2020. Gene content evolution in the arthropods. *Genome Biology* **21(1)**:15 DOI 10.1186/s13059-019-1925-7.

Tischler J, Lehner B, Chen N, Fraser AG. 2006. Combinatorial RNA interference in Caenorhabditis elegans reveals that redundancy between gene duplicates can be maintained for more than 80 million years of evolution. *Genome Biology* **7**:R69.

Van Hoof A. 2005. Conserved functions of yeast genes support the duplication, degeneration and complementation model for gene duplication. *Genetics* **171(4)**:1455–1461 DOI 10.1534/genetics.105.044057.

VanKuren NW, Long M. 2018. Gene duplicates resolving sexual conflict rapidly evolved essential gametogenesis functions. *Nature Ecology & Evolution* **2(4)**:705–712 DOI 10.1038/s41559-018-0471-0.

Vavouri T, Semple JI, Lehner B. 2008. Widespread conservation of genetic redundancy during a billion years of eukaryotic evolution. *Trends in Genetics: TIG* **24**:485–488.

Vicoso B, Bachtrog D. 2015. Numerous transitions of sex chromosomes in Diptera. *PLOS Biology* **13(4)**:e1002078 DOI 10.1371/journal.pbio.1002078.

Wang J, Tao F, Marowsky NC, Fan C. 2016. Evolutionary fates and dynamic functionalization of young duplicate genes in *Arabidopsis* genomes. *Plant Physiology* **172(1)**:427–440 DOI 10.1104/pp.16.01177.

Wasbrough ER, Dorus S, Hester S, Howard-Murkin J, Lilley K, Wilkin E, Polpitiya A, Petritis K, Karr TL. 2010. The *Drosophila melanogaster* sperm proteome-II (DmSP-II). *Journal of Proteomics* **73(11)**:2171–2185 DOI 10.1016/j.jprot.2010.09.002.

Whelan S, Goldman N. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Molecular Biology and Evolution* **18(5)**:691–699 DOI 10.1093/oxfordjournals.molbev.a003851.

White-Cooper H, Bausek N. 2010. Evolution and spermatogenesis. *Philosophical Transactions of the Royal Society B: Biological Sciences* **365(1546)**:1465–1480 DOI 10.1098/rstb.2009.0323.

Wiegmann BM, Trautwein MD, Winkler IS, Barr NB, Kim J-W, Lambkin C, Bertone MA, Cassel BK, Bayless KM, Heimberg AM, Wheeler BM, Peterson KJ, Pape T, Sinclair BJ, Skevington JH, Blagoderov V, Caravas J, Kutty SN, Schmidt-Ott U, Kampmeier GE, Thompson FC, Grimaldi DA, Beckenbach AT, Courtney GW, Friedrich M, Meier R, Yeates DK. 2011. Episodic radiations in the fly tree of life. *Proceedings of the National Academy of Sciences* **108(14)**:5690–5695 DOI 10.1073/pnas.1012675108.

Wyman MJ, Cutter AD, Rowe L. 2012. Gene duplication in the evolution of sexual dimorphism. *Evolution* **66(5)**:1556–1566 DOI 10.1111/j.1558-5646.2011.01525.x.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution* **24(8)**:1586–1591 DOI 10.1093/molbev/msm088.

Yang Y, Xu S, Xu J, Guo Y, Yang G. 2014. Adaptive evolution of mitochondrial energy metabolism genes associated with increased energy demand in flying insects. *PLOS ONE* **9(6)**:e99120 DOI 10.1371/journal.pone.0099120.

Yeates DK, Wiegmann BM, Courtney GW, Meier R, Lambkin C, Pape T. 2007. Phylogeny and systematics of Diptera: two decades of progress and prospects. *Zootaxa* **1668(1)**:565–590 DOI 10.11646/zootaxa.1668.1.27.

Yin C, Shen G, Guo D, Wang S, Ma X, Xiao H, Liu J, Zhang Z, Liu Y, Zhang Y, Yu K, Huang S, Li F. 2016. InsectBase: a resource for insect genomes and transcriptomes. *Nucleic Acids Research* **44(D1)**:D801–D807 DOI 10.1093/nar/gkv1204.

Zhang Y-J, Hao Y, Si F, Ren S, Hu G, Shen L, Chen B. 2014. The *de novo* transcriptome and its analysis in the worldwide vegetable pest Delia antiqua (Diptera: Anthomyiidae). *G3 Genes Genomes Genetics* **4**:851–859.

Zhao C, Escalante LN, Chen H, Benatti TR, Qu J, Chellapilla S, Waterhouse RM, Wheeler D, Andersson MN, Bao R, Batterton M, Behura SK, Blankenburg KP, Caragea D, Carolan JC, Coyle M, El-Bouhssini M, Francisco L, Friedrich M, Gill N, Grace T, Grimmelikhuijzen CJP, Han Y, Hauser F, Herndon N, Holder M, Ioannidis P, Jackson L, Javaid M, Jhangiani SN, Johnson AJ, Kalra D, Korchina V, Kovar CL, Lara F, Lee SL, Liu X, Löfstedt C, Mata R, Mathew T, Muzny DM, Nagar S, Nazareth LV, Okwuonu G, Ongeri F, Perales L, Peterson BF, Pu L-L, Robertson HM, Schemerhorn BJ, Scherer SE, Shreve JT, Simmons D, Subramanyam S, Thornton RL, Xue K, Weissenberger GM, Williams CE, Worley KC, Zhu D, Zhu Y, Harris MO, Shukle RH, Werren JH, Zdobnov EM, Chen M-S, Brown SJ, Stuart JJ, Richards S. 2015. A massive expansion of effector genes underlies gall-formation in the wheat pest *Mayetiola destructor*. *Current Biology* **25(5)**:613–620 DOI 10.1016/j.cub.2014.12.057.