Positive diversifying selection is a pervasive adaptive force throughout the Drosophila radiation

Cicconardi, Francesco; Marcatili, Paolo; Arthofer, Wolfgang; Schlick-Steiner, Birgit C; Steiner, Florian M

Published in:
Molecular Phylogenetics and Evolution

Link to article, DOI:
10.1016/j.ympev.2017.04.023

Publication date:
2017

Document Version
Publisher's PDF, also known as Version of record

Molecular Phylogenetics and Evolution 112 (2017) 230–243

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

Positive diversifying selection is a pervasive adaptive force throughout the *Drosophila* radiation

Francesco Cicconardi [a,*], Paolo Marcatili [b], Wolfgang Arthofer [a], Birgit C. Schlick-Steiner [a,1], Florian M. Steiner [a,1]

[a] Institute of Ecology, University of Innsbruck, Technikerstr. 25, 6020 Innsbruck, Austria
[b] Center for Biological Sequence Analysis, Technical University of Denmark, Kgs. Lyngby, Denmark

ARTICLE INFO

Article history:
Received 25 November 2016
Revised 26 April 2017
Accepted 26 April 2017
Available online 27 April 2017

Keywords:
Adaptation
Development and evolution
Ecological genetics
Genomics/proteomics
Molecular evolution
Positive selection

ABSTRACT

The growing genomic information on non-model organisms eases exploring the evolutionary history of biodiversity. This is particularly true for *Drosophila* flies, in which the number of sequenced species doubled recently. Because of its outstanding diversity of species, *Drosophila* has become one of the most important systems to study adaptive radiation. In this study, we performed a genome-wide analysis of positive diversifying selection on more than 2000 single-copy orthologous groups in 25 species using a recent method of increased accuracy for detecting positive diversifying selection. Adopting this novel approach enabled us to find a consistent selection signal throughout the genus *Drosophila*, and a total of 1342 single-copy orthologous groups were identified with a putative signal of positive diversifying selection, corresponding to 1.9% of all loci. Specifically, in lineages leading to *D. grimshawi*, a strong putative signal of positive diversifying selection was found related to cell, morphological, neuronal, and sensorial development and function. A recurrent signal of positive diversifying selection was found on genes related to aging and lifespan, suggesting that selection had shaped lifespan diversity in *Drosophila*, including extreme longevity. Our study, one of the largest and most comprehensive ones on genome-wide positive diversifying selection to date, shows that positive diversifying selection has promoted species-specific differentiation among evolutionary lineages throughout the *Drosophila* radiation. Acting on the same biological processes via different routes, positive diversifying selection has promoted diversity of functions and adaptive divergence.

1. Introduction

Biologists have been passionately interested in deciphering how the variety of life forms was shaped by evolution. The neutral theory of molecular evolution (Kimura, 1968; King and Jukes, 1969) challenged the Darwinian concept of natural selection, stating that the main cause of evolution and variability at the molecular level was random fixation of selectively neutral mutations and that the effect of natural selection was insignificant. Today, many aspects of the neutralist school are accepted, and most scientists agree that both weak deleterious selection and occasional positive diversifying selection can be identified and represent important evolutionary forces (Fay et al., 2002). One of the main effects of selection is a change in the level of variability (Nielsen, 2005). For instance, at the population level (within species), selective sweeps drastically reduce variation, while purifying selection tends to reduce variability between species more drastically than within species (Nielsen, 2005). While population genetic approaches aim at detecting ongoing selection in a population, comparative genomic approaches, involving data from multiple species, are more suitable to detect past selection (Nielsen, 2005). Here, we refer to positive diversifying selection (by some authors also referred to as episodic diversifying selection or positive selection) as the selective force that increases amino-acid diversity in a gene at various phylogenetic levels (within and among species), promoting innovation and therefore adaptation (Yang et al., 2000). It can vary in a phylogeny over sites (site-to-site) and time (branch-site). Branch-site tests (Yang et al., 2000) measure selective pressure under specific phylogenetic hypotheses by ω, the ratio of non-synonymous to synonymous substitution rates (d_N/d_S) along lineages. If sites are statistically significant for a positive value of $\omega > 1$, positive diversifying selection is inferred, while purifying selection is inferred for $\omega < 1$ and neutrality for $\omega = 1$ (Zhang et al., 2005). Although extensively used (Clark et al.,

* Corresponding author.
E-mail address: francicco@gmail.com (F. Cicconardi).
[1] These authors contributed equally to this work.

http://dx.doi.org/10.1016/j.ympev.2017.04.023
1055-7903/

2007; Roux et al., 2014), these inferences lack precision under some scenarios and are not applicable under others (Kosakovsky Pond et al., 2011). Precision is reduced, for example, when selection strength varies in both foreground (the lineage tested for selection) and background (all other lineages in the phylogeny) or when positive diversifying selection additionally occurs in the background. Scenarios to which these methods are not applicable include backgrounds under purifying selection, which result in falsely identified positive diversifying selection on a neutrally evolving lineage (Kosakovsky Pond et al., 2011). Thus, a new class of models has been developed in which substitution rates may vary branch-to-branch and site-to-site, incorporating these variations via a "random effect" (Kosakovsky Pond et al., 2011; Smith et al., 2015). One of these methods, the adaptive branch-site random effects likelihood (aBSREL) method, provides three ω states to each branch and allows each site to evolve under any kind of ω value (<1; 1; >1) (Smith et al., 2015). Furthermore, like previous methods, aBSREL incorporates different amino acid substitution rates across lineages and is more sensitive than other branch-site methods (Smith et al., 2015).

The genus *Drosophila* (Diptera: Drosophilidae) started to diversify during the Cretaceous (~112 ± 28 million years ago) (Wheat and Wahlberg 2013a), occurs worldwide, and its four subgenera, *Dorsilopha*, *Drosophila*, *Siphlodora*, and *Sophophora*, comprise more than 2000 species (O'Grady and Markow 2009). These species vary starkly in geographic range and ecological niche and include, for example, climate specialists such as *Drosophila nigrosparsa* and *D. takahashii*, which are adapted to the mountainous and alpine zones, living at up to 2200 m above sea level (Bächli et al., 2004; Parkash et al., 2012), and nutritional specialists such as the cactophilic *D. mojavensis* and *D. erecta*, specialized on *Pandanus* (Gardiner et al., 2008), or *D. sechellia*, an extreme specialist of the *Morinda citrifolia* fruit, which is toxic to all other *Drosophila* species (Shiao et al., 2015). *Drosophila* diversity has been explored in multiple studies (Heger and Ponting, 2007; Jiggins and Kim 2007; Clark et al., 2007; Larracuente et al., 2008; Markova-Raina and Petrov, 2011; Roux et al., 2014) but always using the same 12 species and branch-site model implementations prone to false positives (Kosakovsky Pond et al., 2011). Here, we perform a genome-wide search of single-copy orthologous groups (scOGs) and infer the magnitude of the positive diversifying selection in the *Drosophila* radiation. We tried to infer the putative lineage-specific functional consequences of selection, focusing also on specific adaptation aspects such as thermal and lifespan adaptation. For the first aspect, thermal limits define the distribution and abundance of ectotherms through physiological tolerance, phenotypic plasticity, and their evolutionary potential. Because many *Drosophila* species have invaded various climatic regions, positively selected genes could have supported adaptation towards more thermally extreme environments. Finally, because *Drosophila* lifespans range from 50 days (*D. mojavensis*) to nine months (Hawaiian picture-wing *Drosophila*, e.g., *D. grimshawi*), we analysed candidate genes associated with different molecular mechanisms responsible for homeostasis, harmonizing the distribution of energy between body maintenance, growth, development, and reproduction, reducing age-associated diseases and the risk of death (Moskalev et al., 2014).

2. Methods

2.1. Mitogenomic assembly and phylogenetic analysis

All complete *Drosophila* mitochondrial genomes (mtDNA) were downloaded from MetAMiGA (Feijão et al., 2006) and GenBank (Benson et al., 2014) (Table S1). Raw SRA data (Leinonen et al., 2011) were filtered for PCR duplication, adaptor contamination, and low quality and assembled using IDBA-UD v1.1.1 (Peng et al., 2012) and SPAdes v3.5.0 (Bankevich et al., 2012). The best scaffold was selected using BLASTn v2.2.29+. Annotation was performed using MITOS (Bernt et al., 2013), checking start and stop codon. Protein-coding and ribosomal genes were aligned using MACSE v1.01b (Ranwez et al., 2011) and ClustalW v2 (Larkin et al., 2007). Genes shorter than a third of the whole-locus alignment were removed.

The best partition and model of evolution of the concatenated mtDNA alignment were found with PartitionFinder v1.1.1_Mac (Lanfear et al., 2014). MtDNA trees were searched with Maximum Likelihood (ML) and Bayesian Inference (BI) algorithms. For ML, GARLI v2.01.1067 (http://code.google.com/p/garli) was used performing 20 + 5 runs from random starting trees. Runs were continued until no further improvement in log-likelihood was found. After the best tree had been found, 1000 ML nonparametric bootstrap pseudoreplicates were performed. The results were summarized using SumTrees v3.3.1 (http://bit.ly/DendroPy). With RAxML v8.2.3 (http://bit.ly/RAxMLv8), rapid bootstrap search was done with the MRE-based bootstopping criterion.

For BI, BEAST v2 (Bouckaert et al., 2014) was used, modeling population and the speciation tree (template *BEAST (Heled and Drummond 2010)). Five runs of 5×10^7 generations were sampled every 5000th generation. Each partition was modeled with an uncorrelated relaxed clock and its best substitution model. Models not implemented in BEAUTi v2 (Bouckaert et al., 2014) were manually edited in the xml file. Tracer v1.6 (http://beast.bio.ed.ac.uk/Tracer) was used to evaluate convergence and the parameters' effective sampling size (ESS) and to define the burn-in. LogCombiner (Bouckaert et al., 2014) and TreeAnnotator (Bouckaert et al., 2014) were used to summarize the results in a single consensus tree. In ML and BI, *Phortica variegata* (Diptera, Drosophilidae, Steganinae) was used as outgroup. Nodes were considered supported if ML and BI analysis gave a bootstrap support (Bs) $\geq$0.70 and a posterior probability (Pp) $\geq$0.95, respectively.

To compute the species tree from nuclear DNA (nuDNA), scOGs present in all species were used. Single-locus trees were estimated using ML search as implemented in FastTree v2.1.8 SSE3 (Price et al., 2010). Trees were summarized with MP-EST (Liu et al., 2010) using the STRAW web-server (Shaw et al., 2013) and considered together with the mtDNA tree to define the final species tree. Single-gene trees may differ in topology from the species tree. Therefore, the species tree was not used for the diversifying selection test; instead, it was only used for grouping orthologous groups (OGs) with a diversifying selection signal on internal branches.

2.2. Transcriptome datasets

Twenty-three *Drosophila* transcriptomes were downloaded from the *Drosophila* 12 Genomes Consortium (Clark et al., 2007), modENCODE (http://bit.ly/modENCODE) (Chen et al., 2014), and GenBank (Benson et al., 2014) (Table S1). *Drosophila albomicans* and *D. mauritiana* SRA raw reads (Leinonen et al., 2011) were quality filtered and *de novo* assembled using Trinity v2.0.6 (Iyer and Chinnaiyan, 2011).

For 12 unannotated transcriptomes, non-coding transcripts were filtered identifying homologs with BLASTn v2.2.29 + ($e < 1 \times 10^{-10}$) and the best putative protein-coding sequences inferred using TransDecoder v2.0.1 (http://transdecoder.github.io/) and HMMscan (Eddy, 1998) (aa > 100). Best open reading frames (ORFs) for each transcript were defined using BLASTp, considering the lowest *e*-value and the highest *p*-identity.

2.3. Single-copy orthologous group search

For each transcript, the longest ORFs were used for OG search. InParanoid v8 (Sonnhammer and Östlund, 2015) was used, the algorithm implemented in Hieranoid v2 (Schreiber and Sonnhammer, 2013), adopting the mtDNA tree as guide. Hieranoid implements the best reciprocal hit clustering algorithm, while the hierarchical approach through a guide tree reduces total-runtime complexity to a linear function proportional to the species number. At each node, the algorithm aggregates multiple sequences used for similarity searching techniques, yielding more accurate OGs. In the recently published OrthoBench benchmark, a test suite used to evaluate the quality of OGs (Trachana et al., 2011; Trachana et al., 2014), Hieranoid is among the more balanced approaches reducing both false positive and false negative results in most benchmarks (Hulsen et al., 2006; Chen et al., 2007; Altenhoff and Dessimoz, 2009; Schreiber and Sonnhammer, 2013; Lechner et al., 2014) and showing solid performance in a very recent standardized benchmarking among 15 well-established inference methods and resources on a battery of 20 different benchmarks (Altenhoff et al., 2016). To further improve the accuracy and limit false-positive results, BLASTp was used rather than other computationally efficient search tools because it showed better sensitivity (Edgar, 2010). Furthermore, to facilitate the branch-site test and avoid biases related to duplication among lineages and out-paralog genes (Roux et al., 2014), only scOGs were retained and analysed.

There were only 43 single-copy gene families including exactly one ortholog for each of the 25 species. Because annotations and transcriptomes are likely incomplete (see Results), OGs with a few missing genes (gene losses or unannotated genes) were included, retaining a minimal number of species (16, i.e. 64%) and keeping the *Drosophila* 12 Genomes Consortium phylogeny unimpaired.

2.4. Diversifying selection analysis

Signatures of positive diversifying selection were searched for each OG as follows:

(1) Alignment: CDS without internal stop codons were aligned using MACSE v1.01b (Ranwez et al., 2011), which implements a pairwise CDS alignment method detecting CDS preserving codon structure. From each gene alignment, gaps and ambiguously aligned sites were removed with Gblocks v0.91b (Castresana 2000) under a "relaxed" condition (settings: -t = c -b1 = $b -b2 = $b -b3 = 1 -b4 = 6 -b5 = h; where $b is the number of sequences divided by two plus one), as applied by Parker et al. (2013) and Cicconardi et al. (2017), and comparable with the protocol of Talavera and Castresana (2007). A subset of alignments was randomly selected to check whether the alignment and subsequent filtering were effective.

(2) *De novo* phylogenetic gene tree estimation: Incorrect results due to gene tree / mitochondrial tree incongruences were avoided reconstructing a phylogenetic tree for each scOG, as potential incomplete lineage sorting is more likely to affect short branches, where selection signal is harder to detect due to statistical power. ML search was performed with FastTree v2.1.8 SSE3 (Price et al., 2010), reducing computational time without accuracy loss.

(3a) aBSREL analysis: Signatures of positive diversifying selection were searched using the aBSREL algorithm as implemented in the HyPhy batch language (Kosakovsky Pond et al., 2005), performing all likelihood calculations and parameter optimizations using a batch script (BranchSiteREL) in HyPhy (http://github.com/veg/hyphy). Because for each tested OG more than one branch is being tested, the Bonferroni-Holm sequential rejection procedure (Abdi 2010), more balanced than Bonferroni correction, was used to control the family-wise error rate and the probability of making one or more false discoveries, as implemented in HyPhy. Tests with adjusted p-values < 0.05 were considered significant.

(3b) MEME analysis: A site-wise positive diversifying selection test was performed on scOGs under positive diversifying selection (thereafter pOGs) for which we discuss amino acid structural implications and the signal of selection in Gene Ontology (GO) terms enriched in multiple lineages (henceforth multilineage GO terms; also see above) to detect the distribution of synonymous (α) and non-synonymous (β) substitution rates over sites in the branch under selection. Sites were selected with a significantly higher evolutionary rate for which $\beta > \alpha > 0$. For each alignment, the best substitution model was determined (HyPhy: CodonModelCompare), and a posterior probability threshold of 0.95 was adopted to identify the most relevant non-synonymous (β) substitutions for the sequence under positive selection.

After the selection test, all possible sources of bias were considered. These were codon usage, which in principle could cause deviations of the distribution of nucleotide substitutions, and the presence of false positives due to long branches in the phylogenetic trees (see Supplementary materials).

As misalignment is a relevant problem with positive diversifying selection tests (Markova-Raina and Petrov 2011), the effect of the common strategy of filtering (Gblock) the alignment to limit false positive and false negative results (Jordan and Goldman 2012) was studied. The filtering procedure was checked focusing on how efficient it was in removing this type of error. This evaluation was performed re-analyzing all scOGs without the alignment filtering, focusing for simplicity on terminal branches (see Supplementary materials).

2.5. Functional enrichment analysis

To understand the putative biological meaning of selection signatures, not the implication of single genes under positive diversifying selection was evaluated, but gene-set enrichments were tested. By doing that, we believe we diminished the false positive rate. Therefore, we only report and discuss enrichment test results.

Ontology enrichment analysis was performed using the *D. melanogaster* annotation and DAVID v6.7 (Huang et al., 2009) with the *all molecular function*, *all cellular component*, and *all biological process* terms. For all enrichment analysis tests, the complete list of scOGs was used as background instead of the whole genome to avoid potential bias due to some degree of enrichment in scOGs. DAVID bioinformatics resources (Huang et al., 2009) consist of an integrated biological knowledgebase and analytic tools aimed at systematically extracting biological meaning from large gene/protein lists. By following this protocol, extensively used in this field (Roux et al., 2014; Le Duc et al., 2015; Wang et al., 2015; Engel et al., 2016), investigators gain an in-depth understanding of the biological themes in lists of genes enriched in genome-scale studies (Huang et al., 2009). Enrichment hypotheses for temperature (38 loci (Hoffmann and Willi, 2008; Paaby et al., 2010; Morrow and Tanguay, 2015)) and lifespan (71 loci (Proshkina et al., 2015)) genes were tested using Fisher's Exact Test as implemented in the R function fisher.test (p-value < 0.05), using as background all genes in the scOG list. For both ontology enrichment analysis and custom gene lists (temperature and lifespan), the Bonferroni-Holm correction was used. Gene functions and interactions were inferred through physical and genetic interactions using a

network-inferring algorithm as implemented in GeneMANIA (Zuberi et al., 2013).

Gene enrichment was tested at two phylogenetic scales according to the topological localization. In the first strategy, pOGs were grouped according to the subgenus (*Sophophora*, *Drosophila*). In the second strategy, pOGs were grouped instead, according to the evolutionary lineage (internal and/or terminal branches) where a signal of selection was found.

2.6. Protein structure modeling

For two genes with relevant biological function, the effect of selective forces on possible structural modification was evaluated. For each gene, 3D models were generated with MPI toolkit (Biegert et al., 2006) using HHpred (Remmert et al., 2011) and MODELLER (Webb and Sali, 2014). N- and C-terminal protein regions for which no template was found were removed from the final models. Electrostatic potentials were generated using APBS (Baker, 2001).

3. Results

3.1. Mitochondrial genome assembly and phylogeny

To obtain a reliable and independent *Drosophila* phylogeny for our analyses (see below), without extensively using single-copy genes and avoiding circularity in the analysis, we used mitochondrial DNA (mtDNA) as a proxy to infer the actual phylogeny of *Drosophila*. A total of 109 mitogenomes was collected from databases or assembled *de novo* (Table S1). Genome size ranged from 11.3 kilobases (kb; *D. affinis*) to 16.7 kb (*D. ficusphila*), averaging 15.3 kb, giving a final concatenated matrix of 13,253 nucleotides, more than 1.4 million characters.

The phylogenetic relationships as reconstructed by two ML and BI approaches were almost completely identical (Fig. S1), and high support values were recovered in most tree nodes (Fig. 1). The topology of the mtDNA tree was then tested using the 43 scOGs recovered in all the 25 focus species. The topologies of the two phylogenies, mtDNA and nuDNA, were almost identical. Relevant exceptions were two nodes which both had low mtDNA support and low nuDNA concordance factor (CF) values (Figs. S1–S2), namely the phylogenetic position of *D. eugracilis* and the relationship between *D. virilis* and *D. mojavensis*.

3.2. De novo transcriptome assembly, annotation, and orthologous group search

The number of protein-coding genes per species ranged from 7026 (*D. nigrosparsa*) to 19,765 (*D. albomicans*), averaging 13,529 (Fig. 1, Table S1), and no correlation was found between the number of protein-coding loci and genome size (Spearman correlation: $\rho = 0.26$, p-value = 0.22, $R^2 = 0.04$) (Fig. S3a, Table S1).

The search for gene orthology gave 31,003 OGs, of which ~ 70% were scOGs. This lack of completeness was mainly due to insufficient sequencing, as a strong positive correlation between scOG abundance and total number of loci per species was found (Spearman correlation: $\rho = 0.93$, p-value = 9.41e-12, $R^2 = 0.80$) (Fig. S3b). The 12 species of the *Drosophila* 12 Genomes Consortium (Clark et al., 2007) had significantly more orthologs than all other modEncode (Chen et al., 2014) and non-modEncode species (one-tailed Wilcoxon rank-sum test, p-adjusted = 0.035 and p-adjusted < 0.0004), and the other two groups did not significantly differ (two tailed Wilcoxon rank-sum test, p-adjusted = 1) (Fig. S3c). Once a complete list of scOGs had been established, we checked the degree of gene enrichment against the full list of genes in the *D. melanogaster* genome and found that only three biological processes were enriched (GO:0006643 ~ membrane lipid metabolic process; GO:0006350 ~ transcription; GO:0008610 ~ lipid biosynthetic process) (Table S2). This list was therefore used as background for the following GO enrichment test (see Section 2).

3.3. Signal of positive diversifying selection in the Drosophila radiation

Episodes of positive diversifying selection were tested on 2032 protein alignments of scOGs in at least 16 of the 25 *Drosophila* species (see Methods) (Fig. 1, Table S1). Potential bias in the signal of positive diversifying selection given by codon usage, d_S saturation, and alignment properties and filtering were also considered, and no evident effect was detected (see Section 2 and Supplementary Material, especially Figs. S5–S8). After correction for false discovery rate, a total of 1342 scOGs (66%) displayed a putative signal of positive diversifying selection in at least one of the branches tested (p-adjusted < 0.05) (Fig. 1). On average, 255 pOGs were detected per species, corresponding to 1.9% of all loci, very similar to a previous estimation of 2% (Clark et al., 2007). Of all gene-tree branches (176,160), 3587 showed positive diversifying selection, and the portion of sites under selection ($\omega > 1$) ranged from 0.04% to 94.98%. Although some genes had a high fraction of sites with ω^+, these were considered as outliers as the mean and median of the fraction of sites with ω^+ were much lower (mean = 5.96%, median = 3.22%; Fig. S4). On average, 64 pOGs were detected per branch; with 294 pOGs on its branch, *D. persimilis* scored highest (Fig. 1). In 535 branches under selection (15%), gene and mitochondrial trees differed in topology, probably due to homoplasious substitutions, neutral processes, or adaptive convergent evolution; to avoid misleading results, these pOGs were excluded from all subsequent analyses.

As the proportion of positively selected genes differed across branches, the correlation between branch lengths and p-values / log-likelihood ratio was assessed to evaluate the power of the branch-site detection. Similarly to previous analyses with experimental and simulated data sets, a significant negative correlation was found (Spearman correlation with p-values: $\rho = -0.57$ p-value = 0; and log-likelihood ratio: $\rho = -0.56$ p-value = 0). The shorter the branches, the smaller were the log-likelihood ratios and the larger the numerical values of p, suggesting that not all selection signals were detected (Fig. S6a). This was, at least in part, likely the result of the lower power of the branch-site test in shorter branches (Studer and Robinson-Rechavi, 2009; Fletcher and Yang, 2010; Yang and Dos Reis, 2011; Roux et al., 2014).

To gain insight into the putative biological meaning of selection signatures, gene-set enrichment tests were twice performed using GO: (i) once using two subgenus-specific sets of pOGs, corresponding to the *Sophophora* and *Drosophila* subgenera, and (ii) once using *Drosophila* lineage-specific sets (internal and external branches) of pOGs.

3.4. Proportion of positive diversifying selection in the subgenera Sophophora and Drosophila

To establish common and specific traits under selection along the *Drosophila* phylogeny, GO term enrichment was tested on pOGs, and an enrichment of 48 GO terms was found in the two subgenera (Fig. 2, Table S3). Three and 37 were enriched exclusively in the *Sophophora* and *Drosophila* subgenus, respectively, and eight were common to both. The enriched terms were related to cell biology, regulation of gene expression and metabolic processes, and eye, nervous system, and anatomy development. Cell terms were generally common to both subgenera, while the terms related to anatomy, neurons, eye, and regulation were mostly enriched in the *Drosophila* subgenus only. When terms were

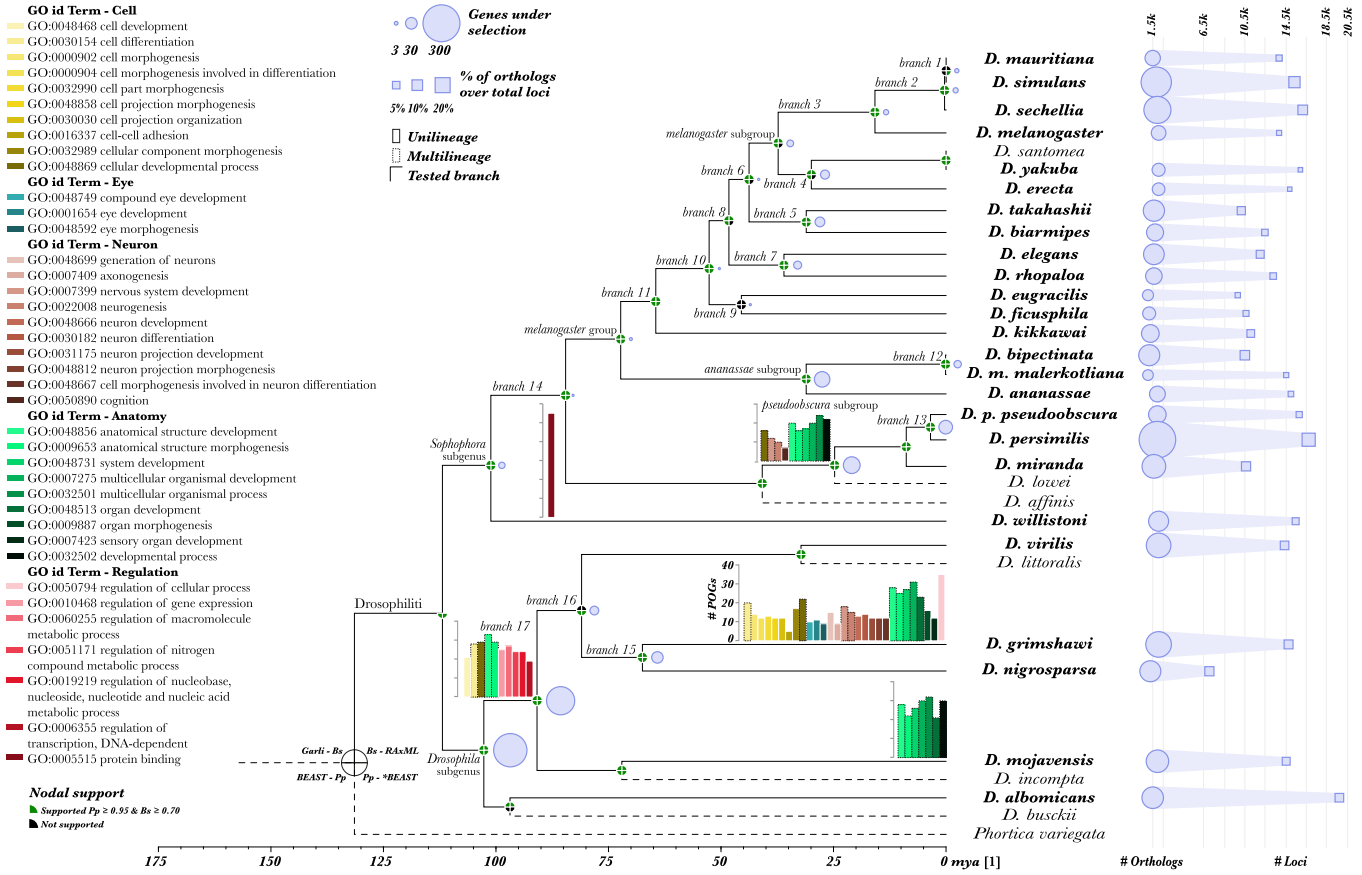

Fig. 1. Phylogenetic relationships of 31 *Drosophila* species based on 109 complete mtDNA sequences. Branches tested for positive diversifying selection are depicted with a continuous line. Histograms on branches show the number of genes under positive diversifying selection for the specific biological processes (BP) (colours). BP for which bars have dotted outlines appeared multiply in the phylogeny (multilineage). The plot on the right side describes the number of annotated loci and orthologs for each species; circle and square size describe the number of genes under selection for a branch or species and the percentage of orthologs over the total number of loci, respectively. The root was scaled according to Wheat and Wahlberg 2013a, 2013b and shown as scale bar in millions of years ago (mya).

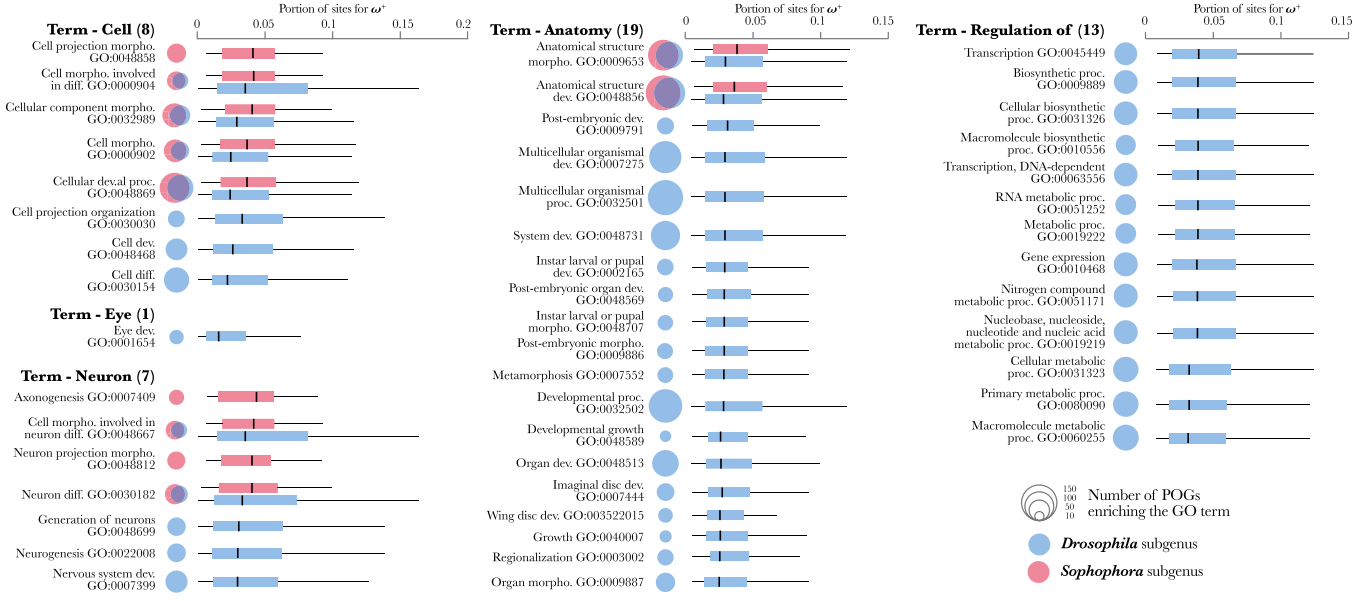

Fig. 2. Venn diagrams of the numbers of genes involved in each GO term under diversifying selection (d_N-d_S) exclusive to and shared by the two subgenera; boxplots of the portion of sites under selection for genes in each GO term.

shared, *Sophophora* had a higher number of genes involved (Fig. 2): The *Sophophora* lineages had 1049 pOGs, whereas the *Drosophila* lineages had just 700, with 475 pOGs common to both. In contrast, *Sophophora* had in general significantly fewer pOGs per branch than *Drosophila* (37 vs. 111; one-tailed Wilcoxon rank-sum test, p-value = 0.003) (Fig. S9).

3.5. Enrichment analysis: Drosophila lineages-specific

Grouping lineage-specific pOGs gave 25 groups for terminal branches (one per species) and 21 groups for internal ones, out of the 23 possible groups. Five branches (three terminal, two internal) showed a significant enrichment of 39 GO term categories (Fig. 1, Table S4). The number of GO terms per lineage ranged from one in *D. willistoni* to 29 in *D. grimshawi*. These GO terms included numerous biological processes related to cell development and differentiation, sensorial development, nervous system, anatomy, regulation of genes, and cellular and metabolic processes. The magnitude of this enrichment ranged from 1.39- to 8.25-fold, and from five to 55 genes per category (Fig. 1, Table S4). Of these 39 GO terms, 28 appeared once in the phylogeny (henceforth unilineage terms), whereas 11 were multilineage terms (Fig. 1).

For all lineages where GO term enrichment was detected, at least one unilineage GO term was found, with the exception of the *D. mojavensis* lineage. Two lineages had several unilineage GO terms, such as *branch 17* with six, and *D. grimshawi* with 20 terms. *Branch 17* was enriched in genes related to cell development, gene expression regulation, and metabolic processes. Of the 39 genes involved in these biological processes, six were shared among all processes: *Ets at 21C* (*Ets21C*), a transcription factor that regulates the wound-dependent expression of epidermal wound response and immune system genes (Patterson et al., 2013); *engrailed* (*en*), which controls neuron/glia fate decisions, neuronal identity, and axon pathfinding and alters specificity of synaptic connections between auditory neurons and the giant fiber in *Drosophila* (Pezier et al., 2014); *germ cell-less* (*gcl*), required for the specification of pole cells and germ cell formation (Leatherman et al., 2002); the two transcription factors *glial cells missing* (*gcm* and *gcm2*), both required for the proliferation of plasmatocyte precursors, the expression of Croquemort protein, and the ability of plasmatocytes to convert into macrophages (Kammerer and Giangrande 2001; Alfonso and Jones 2002); and *hedgehog* (*hh*), a signaling protein involved in many functions (for a review, see Ingham et al., 2011) (Table 1).

In the *D. grimshawi* lineage, enrichment in genes involved in cell morphogenesis, sensory organs, and eye and nervous system development was observed. Among these 43 genes, four were shared among at least 18 of the 20 GO terms: *Anaplastic lymphoma kinase* (*Alk*), a receptor which belongs to the tyrosine kinase superfamily, well known for its role in the development of the visceral mesoderm and motor and visual circuitry (Sopko and Perrimon 2013); *runt* (*run*), a transcription factor important in regulating the expression of other *pair-rule* genes such as *en* (also positively selected in this species) (Wheeler et al., 2002); *frizzled* (*fz*) and *starry night* (*stan*), both central to the planar cell polarity and to shaping the morphology of the insect exoskeleton, such as trichomes (cuticular hairs) that cover much of the exoskeleton, sensory bristles, and ommatidia. All of the protein products of these genes accumulate asymmetrically in wing cells, and there is good evidence that this involves local intercellular signaling between protein complexes on the distal edge of one cell and the juxtaposed proximal edge of its neighbour (Chae et al., 1999; Usui et al., 1999; Seifert and Mlodzik 2007; Adler 2012) (Table 1).

Eleven multilineage GO terms occurred in four of the five lineages (Fig. 1, Table S4). They primarily belonged to biological processes related to anatomy and only secondarily to cell differentiation and development and nervous system development. Thus, possible functional convergence was evaluated across branches. On average, only 13% of the genes (3) were shared across those GO terms. Two biological processes (anatomical structure development, morphogenesis) had not a single gene in common; two other categories (system development, multicellular organismal development) shared three genes. Of all 148 genes, only 14 occurred in all categories (Fig. 1).

Enrichment of these functional categories in unrelated lineages can be explained by convergence of functions or by evolutionary pressure towards diversification of biological functions. As very dissimilar sets of genes with only limited overlapping occurred in different branches (Fig. 1), we assume, firstly, if genes enriching the same category are diversifying their functions, they should interact with different genes and address dissimilar functions. More similar functions should occur in closely related lineages, while dissimilarity of functions should increase with divergence. Secondly, if diversification among lineages occurs, positive diversifying selection should target different protein coding regions, with only few sites in common. To test these hypotheses, gene functions were calculated from each biological function and compared across lineages (see Methods). Each of the 28 gene sets gave a unique set of gene function. In three of those gene sets, no statistically significant function was found, while for the remaining 25 sets of functions, between three and 201 significant single functions were inferred (p-adjusted < 0.05). Overlapping functions between sets ranged from none to more than 60 (Fig. 4; Supplementary file 1). The degrees of overlap and phylogenetic distance (ML branch length) were significantly negatively correlated (Spearman correlation: ρ = -0.74, p-value = 0.0007, R^2 = 0.484) (Fig. S10). To test if positive diversifying selection is occurring in the same position, a branch-site test was applied (MEME) (Murrell et al., 2012) to all 14 genes shared across multilineage GO terms. On average, 30 ± 9 (mean ± standard deviation) sites of each locus were found to be under positive diversifying selection, and 4 ± 2 were shared across at least two lineages. One gene showed no overlap at all, and 13 of the 23 pairwise and triplet permutation tests significantly differed from a random distribution (10^7 permutations, p < 0.01) (Table S5). This result could be seen as an indication of an evolutionary trajectory of adaptive convergence rather than divergence. However, this result is not conclusive, and more detailed analyses should be done (Parker et al., 2013).

3.6. Selective pressure on genes related to temperature stress and lifespan

While most enriched categories fell in functions related to communication, behaviour, and/or morphological adaptations, no enrichment for genes related to thermal stress and aging/lifespan was found. To counter GO term annotation bias, we collected information for candidate genes from the literature and created two gene sets. Thermal adaptation was evaluated searching for positive diversifying selection in 38 loci associated with thermal stress response, including members of the *heat shock protein* (*Hsp*) and *Turandot* (*Tot*) gene families (Table S6). Probably due to a lack of sequencing (see above), just five of these genes were included in the OGs but excluded from the branch-site test due to paralogs within lineages and therefore not tested for positive selection.

Adaptation of genes related to lifespan and/or aging was tested using 71 genes (see Section 2). Of these, 18 (Table S7) were recovered as scOGs and tested for enrichment. Positive diversifying selection was detected in 16 scOGs in at least one branch of the phylogeny. Twenty-two branches showed at least one pOG. Enrichment was tested in all branches; nine showed significant values (Fisher's exact test, p < 0.04) (Fig. 4): five lineages of the *melanogaster* group (*branch 1*, *D. simulans*, *D. melanogaster*, *D. rhopaloa*,

Table 1
Description of biological function of some genes under diversifying positive selection.

Gene name	Symbol	FlyBase id	Biological function	Pathway	Ref.
Anaplastic lymphoma kinase	*Alk*	FBgn0040505	Receptor belonging to the tyrosine kinase superfamily, known for its role in the development of motor and visual circuitry. In the cascade of axon-derived signaling it coordinate neural circuit assembly across different regions of the nervous system	ALK/RTK signaling	Sopko and Perrimon (2013) and Pecot et al. (2014)
Runt	*run*	FBgn0003300	Transcription factor, key to regulating other *pair-rule* gene expressions, such as engrailed (also positively selected in these species)	TGF-β signaling	Wheeler et al. (2002)
Frizzled	*fz*	FBgn0001085	*Frizzled* (*fz*) and *starry night* (*stan*) are central to the planar cell polarity shaping exoskeleton morphology. All protein products of these genes accumulate asymmetrically in wing cells, and there is good evidence that this involves local intercellular signaling between protein complexes on the distal edge of one cell and the juxtaposed proximal edge of its neighbour	Wnt signaling, planar cell polarity	Chae et al. (1999), Usui et al. (1999), Seifert and Mlodzik (2007) and Adler (2012)
Starry night	*stan*	FBgn0024836	"	"	Chae et al. (1999) and Usui et al. (1999), Seifert and Mlodzik (2007) and Adler (2012)
Ets at 21C	*Ets21C*	FBgn0005660	Transcription factor involved in wound and immune response	–	Patterson et al. (2013)
Engrailed	*en*	FBgn0000577	The protein controls the development of neurons including auditory neurons. It is also invovled in the regulation of *yellow* to control wing spot formation	Hedgehog signaling	Wheeler et al. (2002), Gompel et al. (2005) and Pezier et al. (2014)
Germ cell-less	*gcl*	FBgn0005695	Required for the specification of pole cells and germ cell formation	–	Leatherman et al. (2002)
Glial cells missing	*gcm*	FBgn0014179	The transcription factors glial cells missing (gcm and gcm2) are related to immune system through the proliferation of plasmatocyte precursors, the expression of Croquemort protein, and their ability to convert into macrophages	–	Kammerer and Giangrande (2001) and Alfonso and Jones (2002)
Gcm2	*gcm2*	FBgn0019809	"	–	Kammerer and Giangrande (2001) and Alfonso and Jones (2002)
Hedgehog	*hh*	FBgn0004644	Controls *Drosophila* embryonic cuticle patterns and adult appendages and is also vital for numerous aspects related to development and cell fate regulation. It is essential for stem cell maintenance, and its malfunction contributes to a number of disorders including birth defects and cancers. A hallmark of Hh signaling is its ability to act over a long range and control distinct cell fates as a function of Hh concentration	Hedgehog signaling	Ingham et al. (2011) and Sagner et al. (2012)
Forkhead box sub-group O	*foxo*	FBgn0038197	Transcription factor involved in insulin regulation and activated in response to nutrient deprivation.	RTK signaling	Figueroa-Clarevega and Bilder (2015)
Phosphatase and tensin homolog	*Pten*	FBgn0026379	Encodes an enzyme part of a signal pathway regulating cell proliferation, dividing and triggering cells to apoptosis. This enzyme also helps control cell movement and adhesion of cells to surrounding tissues. Additionally, it likely plays a role in maintaining the stability of a cell's genetic information, preventing uncontrolled cell growth	Insulin/PI3 K signaling	Song et al. (2012)
Four wheel drive	*fwd*	FBgn0004373	Transcription factor regulating actin organization. Its overexpression is associated with an increased lifespan, probably affecting an insulin-like pathway	–	Landis et al. (2003)
Stathmin	*stai*	FBgn0266521	Cytosolic phosphoprotein regulating microtubule dynamics; involved in the maintainance of axonal microtubule integrity, neuronal development, plasticity, and regeneration	–	Chauvin and Sobel (2015)
ATP-dependent chromatin assembly factor large subunit	*Acf*	FBgn0027620	Involved in various functions such as the activation of adult stem cell and differentiation	Wnt signaling	Clevers and Nusse (2012)
Ephrin	*Ephrin*	FBgn0040324	Receptor interacting protein regulating topographic mapping along the dorsal-ventral axis of the retinotectal system	EPH-Ephrin signaling	Poliakov et al. (2004)
Regulatory factor X	*Rfx*	FBgn0020379	Transcription factor required for all ciliated neurons; cooperating with cell-type-specific transcription factors to regulate genes required for cilia sensory diversification and specialization	–	Newton et al. (2012)
Dystroglycan	*Dg*	FBgn0034072	Cell surface receptor; its deletion in the brain leads to defects in morphogenesis, including malformation of cortical lamina, improper photoreceptor axon projections, disrupted array of ommatidia on its surface, and eye neuron elongation defects during development	Integrin cell surface interactions	Marrone et al. (2011)
Optix	*Optix*	FBgn0025360	Optix is a homeobox gene expressed in the eye, wing and haltere imaginal discs; involved in the development of eye, clypeolabrum, and several head sensory organs. It was also suggested to drive convergent evolution of butterfly wing pattern mimicry	–	Reed et al. (2011)

D. ananassae), one species of the *pseudoobscura* subgroup (*D. persimilis*), *branch 17*, *D. virilis*, and *D. nigrosparsa* (Fig. 3). Fourteen scOGs related to lifespan were found to be under selection in terminal branches, four of which occurred in at least three lineages: *foxo* (Figueroa-Clarevega and Bilder 2015), *Pten* (Song et al., 2012), *fwd* (Landis et al., 2003), and *stai* (Chauvin and Sobel 2015) (Fig. 3; Table 1).

To evaluate the potential molecular impact of those sites, the 3D structure of two representative proteins, Pten and Sirt1, was reconstructed. In the *D. nigrosparsa* amino-acid sequence of Pten,

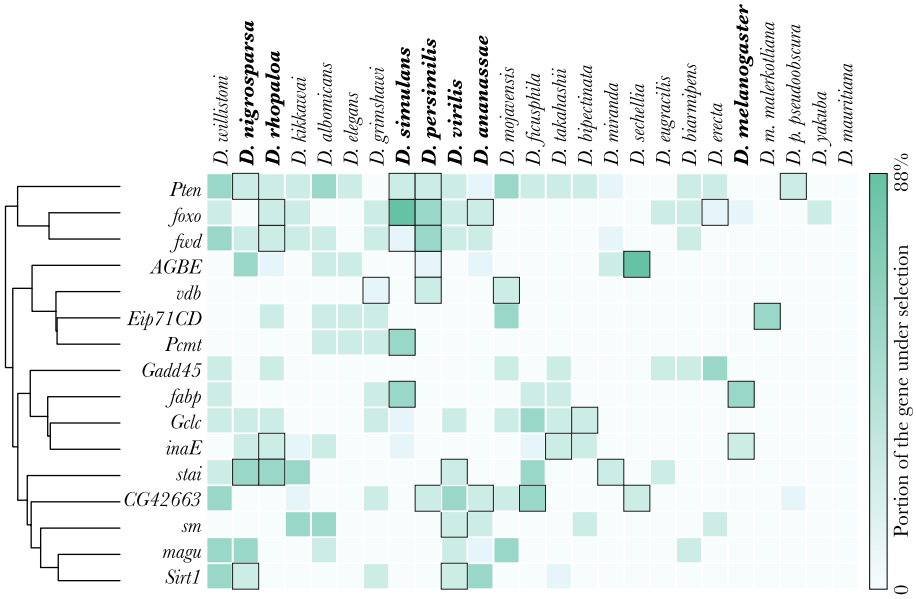

Fig. 3. Heat map of the portion of gene under selection (ω^*) for the 16 lifespan related genes under selection. Species names are in bold when the lifespan gene set was significantly enriched. Black squares show in which species genes were under selection.

multiple mutations occurred in the C2 domain, responsible for the localisation and activation of the protein (Fig. 5), while in Sirt1, two of the four sites under selection were located in the region binding to activating compounds, drastically changing the electrostatic potential of the region (Fig. 6).

4. Discussion

4.1. Pervasive positive diversifying selection in the Drosophila radiation

The search for positive diversifying selection is not an easy task. It has been an area of active research and unrelenting debate (Suzuki and Nei 2004; Wong et al., 2004; Nozawa et al., 2009), and no method is capable of avoiding false positives. Nevertheless, the adaptive branch-site random effects likelihood (aBSREL) method, adopted in this study, proved to deal with some types of false positives better than any method before (Kosakovsky Pond et al., 2011; Smith et al., 2015), due to model flexibility, especially compared with other methods in PAML branch site tools (see Kosakovsky Pond et al., 2011; Smith et al., 2015 for more detailed explanations). A strong bias introduced in our analysis is unlikely to have occurred, as shown by the extended analyses for possible correlations of the rate of positive diversifying selection signal as function of codon usage, branch length, sequence length, and dataset size, without any positive result. It is also noteworthy that in this study we exclusively used single-copy OGs, just like others did (e.g. Roux et al., 2014), that is, we deliberately avoided the analysis of multi-copy OGs. We did so because the presence of several gene copies usually reduces negative selection pressure, introducing a possible source of bias, especially for recently split copies. This is because the presence of several gene copies usually reduces negative selection pressure, introducing a possible source of bias, especially for recently split copies. This is because harmful mutations on one locus become less harmful due to the presence of the other copies (Panchin et al., 2010). Possibly, however, we thus underestimated the overall effect of positive diversifying selection, given that multi-copy OGs are more prone to evolve faster than single-copy OGs, for example in chemosensory receptors (Cicconardi et al., 2017).

The scan of 25 species for positive diversifying selection returned that 66% (1342) of the scOGs are under positive diversifying selection in at least one branch of the phylogeny. In terms of relative numbers of tested pOGs, we found twice as many pOGs compared with a previous study on positive diversifying selection on 12 *Drosophila* species (1175 under selection on 3173 scOGs) (Roux et al., 2014). Unfortunately, the evaluation of different false discovery rates across different studies is not always an easy task and would probably be a separate study; we therefore did not perform an evaluation of this aspect here. Almost all internal (21) and terminal (25) branches (96%) showed at least one pOG. These results suggest that augmenting the number of species might increase instances of genes under positive diversifying selection. This is also likely to reduce the effect of d_S saturation in deep branches, thus improving statistical power. Based on these results, we infer that diversifying selection is likely common and pervasive in Drosophilidae and possibly very significant to their evolution.

4.2. More biological processes are under positive diversifying selection in the subgenus Drosophila than in Sophophora

The large diversity of life forms we see today is the result of various evolutionary processes, one of which is adaptive radiation, the diversification of species enabling them to occupy different ecological niches. Studying diversification across phylogenetic lineages is a useful approach to identify adaptive radiation and gain information about the processes underlying the origin of biodiversity (Magnacca and Price, 2015). Functional analysis approaches greatly improved the biological interpretation of large gene lists, ranging in size from hundreds to thousands of genes, switching from a gene-centric analysis to a more biological module-centric analysis (Huang et al., 2009).

Comparing the two subgenera, similarities and differences in positive diversifying selection targets were identified. In both groups, a substantial portion of scOGs under positive diversifying selection was found, 52% and 34% in *Sophophora* and *Drosophila*, respectively, with a relevant proportion of pOGs in common, 42% and 68%, respectively, and numerous biological processes affected. The analytical approach used in this study allowed to identify more

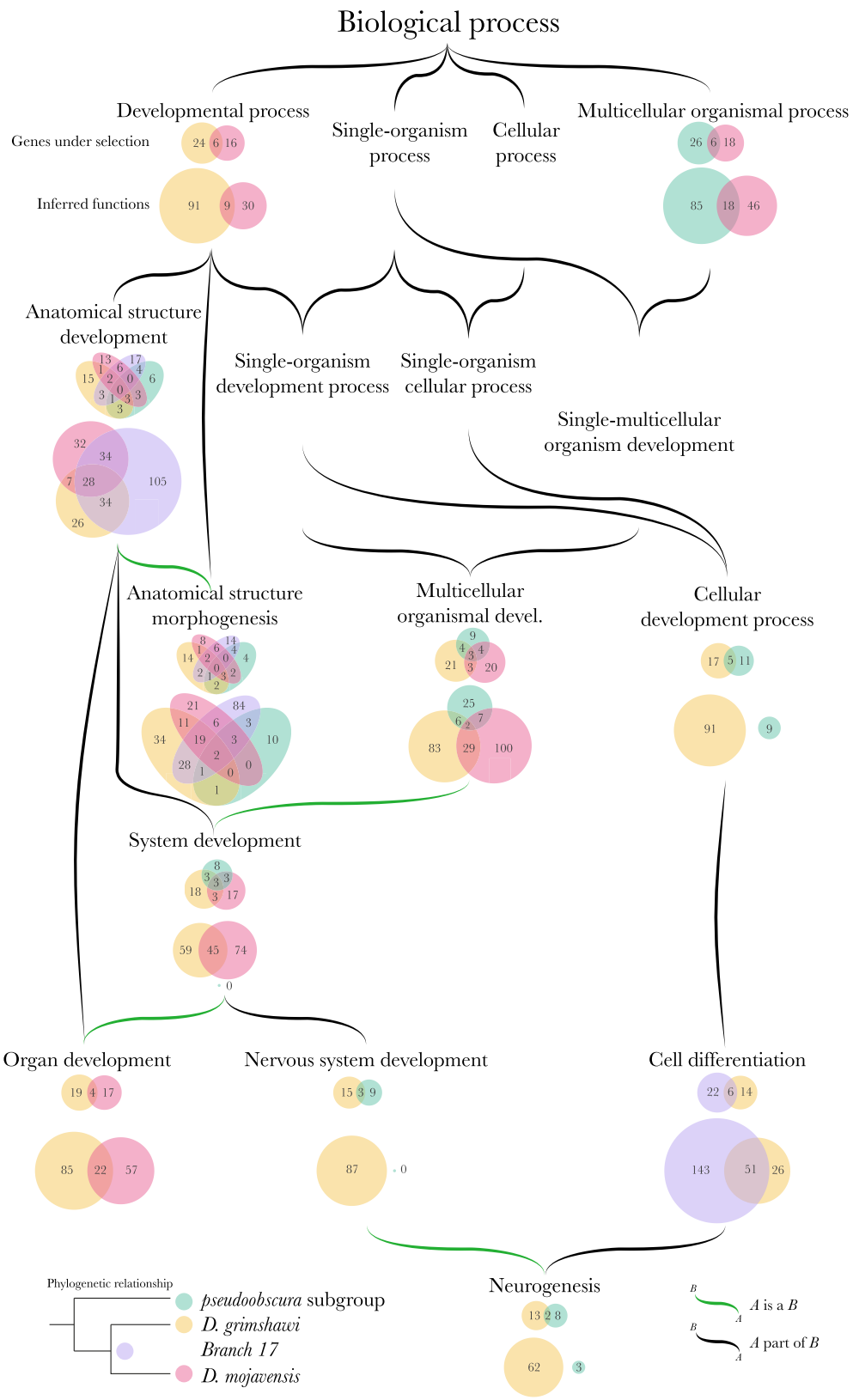

Fig. 4. Gene ontology hierarchy for biological processes enriched in multiple lineages (multilineage terms) according to QuickGO (EMBL-EBI). More general terms are in the upper part, more specific ones in the lower part. For each enriched GO term, two comparisons (Venn diagrams) are depicted: one for genes, the other for the inferred functions generated from the lineage-specific gene set. The amount of inferred function shared was negatively correlated with the phylogenetic relationship of the lineages (bottom-left tree).

instances of positive selection, more and new GO terms compared with previous studies (Clark et al., 2007; Roux et al., 2014), and significant differences between the two subgenera. Cumulatively, *Sophophora* has more pOGs but also more branches than *Drosophila*. In contrast, the number of pOGs per branch is significantly higher in *Drosophila*. We thus propose a scenario of a more focused

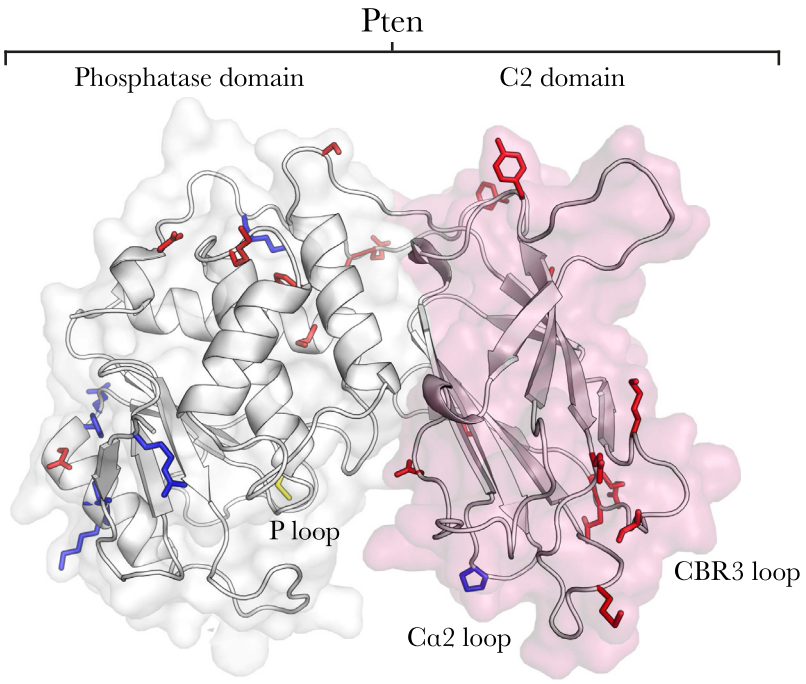

Fig. 5. Pten is a 415 amino acids protein (*D. melanogaster*) composed of five functional domains: a phosphatidylinositol-4,5-bisphosphate (PtdIns(4,5)P_2)-binding domain, a phosphatase domain, a C2 domain, a carboxy-terminal tail, and a PDZ-binding domain. The Pten phosphatase domain (gray) consists of a central five-stranded β-sheet packing with two α-helices on one side and four on the other. This domain contains the active pocket, which in its bottom part has the phosphate-binding loop (P loop) with the catalytic cysteine (yellow). The C-terminal of Pten folds into a β-sandwich, containing two antiparallel β-sheets with two short α-helices intervening between the strands (red). The C2 domain is a membrane-binding regulatory domain and consists of two regions, Cα2 and CBR3. These provide a binding site for the C-terminal tail, which masks the membrane-binding site located in the same interface, and regulate the plasma membrane localisation of Pten. Sites under positive diversifying selection are mapped on the Pten structure (pdb code: 3N0A). Residues in red and blue are under selection in *D. nigrosparsa* and in other species, respectively. The image has been generated with PyMOL.

selection in *Drosophila* than in *Sophophora*. We favour this idea due to the higher number of GO terms enriched in *Drosophila* and therefore interpret the higher absolute number of pOGs in *Sophophora* as correlated with the higher number of branches tested, whereas the higher number of pOGs per branch in *Drosophila* may be explained by a stronger and more canalised selection leading to stronger adaptive radiation.

4.3. Genes in Drosophila lineages leading to D. grimshawi are strongly selected in terms related to cell, morphological, neuronal, and sensorial development and function

In this study, we also wanted to track putative adaptive changes along the *Drosophila* radiation trying to gain insight into the possible biological meaning of branch-specific positive diversifying selection, grouping pOGs into lineage-specific groups. One very interesting example of this approach was the result of pOGs of *D. grimshawi*. This species belongs to Hawaiian Drosophilidae, an outstanding example of explosive adaptive radiation, with more than 1000 endemic species (Edwards et al., 2007), of which 120, including *D. grimshawi*, belong to the picture-wing *Drosophila*. Despite their distinct morphology, pigmentation, and behaviour, these species are separated only by a few hundred thousand years, with relatively few genomic DNA differences (Edwards et al., 2007; Magnacca and Price 2015). Many of these species have small populations with limited distribution and often possess elaborate species-specific wing spots and unusual modifications of mouthparts and legs in males. These species are also known for their complex courtship behaviour, differing even among closely related species (Magnacca and Price 2015), and their extreme sexual antennal lobe dimorphism. Sexual selection, geographic subdivision, host plant specialisation, morphological innovation, or a combination of these could explain the high species diversity (Edwards et al., 2007).

With our hypothesis-free (unsupervised) approach, we found that the branches leading to *D. grimshawi* (*branch 17* and *D. grimshawi* branch) have an exceptional number of genes under selection (in 36 GO terms), all related to cell, neuron, and anatomical differentiation, morphogenesis, and development of sensory organ and of eye. In *D. grimshawi*, *fz*, *hh*, *shf*, and *stan*, among others, are detected with a signal of positive diversifying selection. These genes are involved in two very important signaling pathways, the Hedgehog (HH) and *frizzled*/*stan* pathways. The HH pathway mediates fundamental processes during embryo development and induces tissue morphogenesis and homeostasis (Ingham et al., 2011). Its role during imaginal disc development is related to cuticle and wing pigmentation (Sagner et al., 2012). The *frizzled*/*stan* pathway controls planar cell polarity, which influences morphological and sensorial elements such as sensory bristles, the eye, and epidermal hairs on the wing (Adler 2012). Over 30 genes are related to eye and nervous system development (Table S4). Most of them, such as *Acf* (Clevers and Nusse 2012), *Alk* (Pecot et al., 2014), *Ephrin* (Poliakov et al., 2004), *Rfx* (Newton et al., 2012), *Dg* (Marrone et al., 2011), *Optix* (Reed et al., 2011), and *en* (Gompel et al., 2005; Pezier et al., 2014), perform key tasks. We do not have evidence that sexual selection is directly involved, but we know that in closely related species, some positive diversifying selection signals were found in genes related to sensory detection and mating (Kang et al., 2016). Because the main evolutionary force in picture-winged Hawaiian species seems to be related to sexual selection (Kaneshiro 1988; Carson and Carson 1997; Price et al., 2014), it would be promising to experimentally evaluate the role of these genes under positive diversifying selection on sexual behaviour.

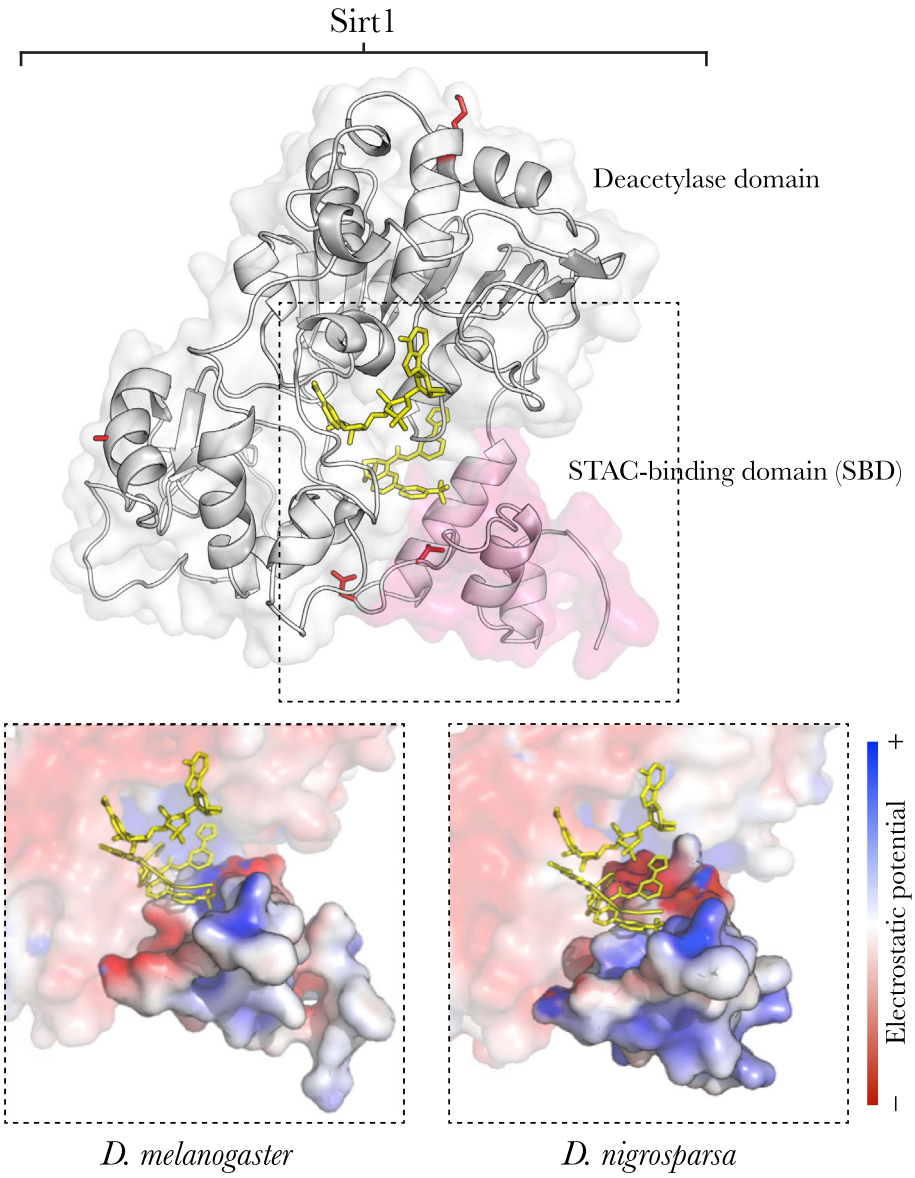

Fig. 6. In *D. melanogaster*, Sirt1 it is composed of 823 amino acids. Nad+ and small molecule sirtuin-activating compounds (STAC), in yellow, were grafted from human SIRT2 (pdb code: 4ZZI) by superimposing the corresponding SIRT1 binding region. Here, we show the structure of the deacetylase domain (gray) and the STAC-binding domain in red. The electrostatic potentials of Sirt1 for *D. melanogaster* and *D. nigrosparsa* are displayed in the two panels below. The colours, ranging from red to white to blue, represent regions with a negative, neutral, and positive electrostatic potential, respectively. The image has been generated with PyMOL.

4.4. Positive diversifying selection acts on the same biological processes in unrelated Drosophila lineages, promoting adaptive diversification of functions

Genes under positive diversifying selection can either become subject to adaptive convergence, when selection enforces identical function, or can differentiate towards adaptive divergence, which may lead to new adaptive functions. Almost one third of all enriched biological processes analysed here (11/39) does not occur in single lineages but is enriched in multiple, mostly unrelated ones. Our data provide evidence for positive diversifying selection acting on the same biological processes via different routes, promoting diversity of functions and adaptive divergence. Three lines of evidence support this hypothesis. (i) The low number of common genes among the same categories in different lineages (on average 3/23 genes); (ii) the minor overlap of inferred functions for genes belonging to the same category in different lineages; (iii) function similarity among lineages correlating negatively with phylogenetic distances; more functions are shared by closely related lineages compared with less related lineages. We speculate that these shared genes represent key elements for the diversification of functions. In other words, these genes may play a significant role in adaptation, as they are relevant in various biological processes and most amino-acid modifications occur in non-random positions.

4.5. Positive diversifying selection on lifespan-related genes is pervasive in Drosophila lineages

During aging, functions such as regeneration and reproduction slowly decline. Homeostasis becomes more susceptible to stress, and the loss of functions triggers age-associated diseases and ultimately death. The *antagonistic pleiotropy* theory postulates that genes beneficial to early life can be detrimental in later life, after reproductive success (Moskalev et al., 2014). A special case of this

theory, the *disposable soma* theory (Moskalev et al., 2014), predicts that genes controlling resources redistribute energy from body maintenance to growth and reproduction. Repairing cellular damage requires energy, and competition may rise with the needs of energy for reproduction. In favour of growth and development, longevity-assuring genes reduce or turn off their activity, and aging proceeds. Here, a significant enrichment of genes related to longevity was found in nine *Drosophila* lineages and in both subgenera. Sixteen scOGs, related to longevity, were found to be target of positive diversifying selection and common to many species. According to the *antagonistic pleiotropy* theory, adaptation does not necessarily imply a lifespan extension but more probably a significant contribution to life-cycle tuning to better adapt to the environment. Examples of this pattern are the reproductive behaviour and ecological adaptation of *D. nigrosparsa*, in which genes related to lifespan are enriched. This alpine species has a developmental time longer than a month, a life expectancy of four months, and a fecundity curve (number of laid eggs per day) ten times lower and twice longer than *D. melanogaster* (Kinzner, pers. comm.). Apparently, *D. nigrosparsa* adapts towards body maintenance instead of high reproductive activity. Two targets of positive diversifying selection (Pten and Sirt1) and their structural changes could explain how this species adapted to a harsh alpine environment, where food is scarce and not always present. Pten, a lifespan regulator, promotes longevity by modulating the insulin-like pathway (Song et al., 2012). We show that numerous modifications are affecting Cα2 and CBR3 loops as well as the phosphatidylinositol-4,5-bisphosphate (PtdIns(4,5)P2)-binding domain. It has been shown that modifications of only two residues in Cα2 and CBR3 regions affect the activity of the protein (Nguyen et al., 2014). Because the binding with the plasma membrane promotes the dissociation of the C-terminal tail, which masks the catalytic domain, we can speculate that mutations in those regions could more efficiently release the C-terminal tail and more effectively bind Pten to its substrates, modulating the protein affinity to the cell membrane. This not only could determine the cell localisation of the protein but also promote a different placement of the phosphatase domain with its substrate (Song et al., 2012; Nguyen et al., 2014). Sirt1 belongs to a family of five sirtuins; it is a NAD+-dependent deacetylase involved in the modulation of gene silencing to DNA repair. It is controlled by regulators and mediates lifespan. Sirt1 substrates include key cellular regulators of a wide variety of central signaling pathways. For instance, Sirt1 can act as a deacetylase of Pten, as Pten is hyperacetylated and excluded from the nucleus in Sirt1-deficient cells. It is responsible for switching stress response programs related to, for example, food availability, temperature, and endogenous oxidative stress (Frankel et al., 2011). The N-terminal part of its NAD^+-binding domain, in which we observed two selected sites modulating the overall electrostatic charge of the region, promotes the protein function by binding to small-molecule sirtuin-activating compounds and plays a significant role in the regulation and affinity between Sirt1 and its ligands (Dai et al., 2015).

5. Conclusions

Since the release of the 12 *Drosophila* genomes (Clark et al., 2007), many studies provided more than an initial assessment of a positive diversifying selection landscape of the *Drosophila* radiation. To our knowledge, this is the largest and most detailed genome-wide analysis of positive diversifying selection to date on *Drosophila*, and overall, the second largest, right after a study on 30 primate taxa (Moretti et al., 2014). Including 13 additional *Drosophila* transcriptomes and using a recent, more accurate method for detecting positive diversifying selection (Kosakovsky Pond et al., 2011; Smith et al., 2015), we provide insight into the particular patterns of positive diversifying selection – which sometimes acts on different genes to achieve the same goal – that have helped shape present-day genes. This study may be seen as incremental to previous studies and, more importantly, as the starting point for specific in-depth studies for a multitude of biological adaptations in *Drosophila*.

Competing financial interests

The authors declare no competing financial interests.

Data accessibility

All mitochondrial DNA sequences and annotation (mtDNA.All.scaffolds.tar.gz): Dryad doi: 10.5061/dryad.0961b

Acknowledgements

We thank Michael Fink for help with cluster computing, Mateusz Kaduk for help with Hieranoid2, Sergei Pond for help in implementing aBSREL in HyPhy, Martin-Carl Kinzner for information on *Drosophila nigrosparsa* life history, Anna Tramontano for helpful discussion, and an anonymous reviewer for constructive input. This work was supported by the Austrian Federal Ministry of Science, Research and Economy, BMWFW, as part of the 'Infra strukturprogramm' of the Focal Point Scientific Computing at the University of Innsbruck. F.C. was supported by the Austrian Science Fund (P23949 and P23409).

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ympev.2017.04.023.

References

Abdi, H., 2010. In: Salkind, Neil. (Ed.), Holm's Sequential Bonferroni Procedure. Sage, Thousand Oaks, CA.
Adler, P.N., 2012. The frizzled/stan pathway and planar cell polarity in the *Drosophila* wing. Curr. Top. Dev. Biol. 101, 1–31.
Alfonso, T.B., Jones, B.W., 2002. Gcm2 promotes glial cell differentiation and is required with glial cells missing for macrophage development in *Drosophila*. Dev. Biol. 248, 369–383.
Altenhoff, A.M., Boeckmann, B., Capella-gutierrez, S., et al., 2016. Standardized benchmarking in the quest for orthologs. Nat. Methods 13, 425–430.
Altenhoff, A.M., Dessimoz, C., 2009. Phylogenetic and functional assessment of orthologs inference projects and methods. PLoS Comput. Biol. 5, e1000262.
Bächli, G., Vilela, C.R., Escher, S.A., Saura, A., et al., 2004. The Drosophilidae (Diptera) of Fennoscandia and Denmark. Brill Academic Publishers.
Baker, N.A., 2001. Mathematical and Computational Modeling of Biomolecular Systems. University of California, San Diego.
Bankevich, A., Nurk, S., Antipov, D., et al., 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol. 19, 455–477.
Benson, D.A., Clark, K., Karsch-Mizrachi, I., et al., 2014. GenBank. Nucleic Acids Res. 43, D30–D35.
Bernt, M., Donath, A., Jühling, F., et al., 2013. MITOS: improved de novo metazoan mitochondrial genome annotation. Mol. Phylogenet. Evol. 69, 313–319.
Biegert, A., Mayer, C., Remmert, M., Söding, J., Lupas, A.N., 2006. The MPI bioinformatics toolkit for protein sequence analysis. Nucleic Acids Res. 34, 335–339.
Bouckaert, R., Heled, J., Kühnert, D., et al., 2014. BEAST 2: a software platform for Bayesian evolutionary analysis. PLoS Comput. Biol. 10, e1003537.
Carson, H.L., Carson, H.L., 1997. Sexual selection: A driver of genetic change in Hawaiian *Drosophila*. J. Hered. 88, 343–352.
Castresana, J., 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol. Biol. Evol. 17, 540–552.
Chae, J., Kim, M.J., Goo, J.H., et al., 1999. The Drosophila tissue polarity gene starry night encodes a member of the protocadherin family. Development (Cambridge, England) 126, 5421–5429.

Chauvin, S., Sobel, A., 2015. Neuronal stathmins: a family of phosphoproteins cooperating for neuronal development, plasticity and regeneration. Prog. Neurobiol. 126, 1–18.
Chen, F., Mackey, A.J., Vermunt, J.K., Roos, D.S., 2007. Assessing performance of orthology detection strategies applied to eukaryotic genomes. PLoS One 2, e383.
Chen, Z.X., Sturgill, D., Qu, J., et al., 2014. Comparative validation of the *D. melanogaster* modENCODE transcriptome annotation. Genome Res. 24, 1209–1223.
Cicconardi, F., D., Di Marino, Olimpieri, P.P., Arthofer, W., Schlick-Steiner, B.C., Steiner, F.M., 2017. Chemosensory adaptations of the mountain fly *Drosophila nigrosparsa* (Insecta: Diptera) through genomics' and structural biology's lenses. Sci. Rep. 7, 43770.
Clark, A.G., Eisen, M.B., Smith, D.R., et al., 2007. Evolution of genes and genomes on the *Drosophila* phylogeny. Nature 450, 203–218.
Clevers, H., Nusse, R., 2012. Wnt/β-Catenin signaling and disease. Cell 149, 1192–1205.
Dai, H., Case, A.W., Riera, T.V., et al., 2015. Crystallographic structure of a small molecule SIRT1 activator-enzyme complex. Nature Communications 6, 7645.
Le Duc, D., Renaud, G., Krishnan, A., et al., 2015. Kiwi genome provides insights into evolution of a nocturnal lifestyle. Genome Biol. 16, 147.
Eddy, S.R., 1998. Profile hidden Markov models. Bioinformatics 14, 755–763.
Edgar, R.C., 2010. Supplementary material - search and clustering orders of magnitude faster than BLAST. Bioinformatics 26, 2460–2461.
Edwards, K.a., Doescher, L.T., Kaneshiro, K.Y., Yamamoto, D., 2007. A database of wing diversity in the Hawaiian *Drosophila*. PLoS One 2, e487.
Engel, I., Seumois, G., Chavez, L., et al., 2016. Innate-like functions of natural killer T cell subsets result from highly divergent gene programs. Nat. Immunol. 17, 728–739.
Fay, J.C., Wyckoff, G.J., Wu, C.-I., 2002. Testing the neutral theory of molecular evolution with genomic data from *Drosophila*. Nature 415, 1024–1026.
Feijão, P.C., Neiva, L.S., de Azeredo-Espin, A.M.L., Lessinger, A.C., 2006. AMiGA: the arthropodan mitochondrial genomes accessible database. Bioinformatics 22, 902–903.
Figueroa-Clarevega, A., Bilder, D., 2015. Malignant *Drosophila* tumors interrupt insulin signaling to induce cachexia-like wasting. Dev. Cell 33, 47–55.
Fletcher, W., Yang, Z., 2010. The effect of insertions, deletions, and alignment errors on the branch-site test of positive selection. Mol. Biol. Evol. 27, 2257–2267.
Frankel, S., Ziafazeli, T., Rogina, B., 2011. DSir2 and longevity in *Drosophila*. Exp. Gerontol. 46, 391–396.
Gardiner, A., Barker, D., Butlin, R.K., Jordan, W.C., Ritchie, M.G., 2008. *Drosophila* chemoreceptor gene evolution: Selection, specialization and genome size. Mol. Ecol. 17, 1648–1657.
Gompel, N., Prud'homme, B., Wittkopp, P.J., Kassner, V.A., Carroll, S.B., 2005. Chance caught on the wing: cis-regulatory evolution and the origin of pigment patterns in *Drosophila*. Nature 433, 481–487.
Heger, A., Ponting, C.P., 2007. Evolutionary rate analyses of orthologs and paralogs from 12 *Drosophila* genomes. Genome Res. 17, 1837–1849.
Heled, J., Drummond, A.J., 2010. Bayesian inference of species trees from multilocus data. Mol. Biol. Evol. 27, 570–580.
Hoffmann, A.a., Willi, Y., 2008. Detecting genetic responses to environmental change. Nat. Rev. Genet. 9, 421–432.
Huang, D.W., Sherman, B.T., Lempicki, R.A., 2009. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat. Protoc. 4, 44–57.
Hulsen, T., Huynen, M., de Vlieg, J., Groenen, P., 2006. Benchmarking ortholog identification methods using functional genomics data. Genome Biol. 7, R31.
Ingham, P.W., Nakano, Y., Seger, C., 2011. Mechanisms and functions of Hedgehog signalling across the metazoa. Nat. Rev. Genet. 12, 393–406.
Iyer, M.K., Chinnaiyan, A.M., 2011. RNA-Seq unleashed. Nat. Biotechnol. 29, 599–600.
Jiggins, F.M., Kim, K.W., 2007. A screen for immunity genes evolving under positive selection in *Drosophila*. J. Evol. Biol. 20, 965–970.
Jordan, G., Goldman, N., 2012. The effects of alignment error and alignment filtering on the sitewise detection of positive selection. Mol. Biol. Evol. 29, 1125–1139.
Kammerer, M., Giangrande, A., 2001. Glide2, a second glial promoting factor in *Drosophila melanogaster*. EMBO J. 20, 4664–4673.
Kaneshiro, K.Y., 1988. Speciation in the Hawaiian "*Drosophila*": sexual selection appears to play an important role. Bioscience 38, 258–263.
Kang, L., Settlage, R., McMahon, W., et al., 2016. Genomic signatures of speciation in sympatric and allopatric hawaiian picture-winged drosophila. Geno. Biol. Evolut. 8, evw095.
Kimura, M., 1968. Genetic variability maintained in a finite population due to mutational production of neutral and nearly neutral isoalleles. Genet. Res. 11, 247–269.
King, J.L., Jukes, T.H., 1969. Non-Darwinian evolution. Science 164, 788–798.
Kosakovsky Pond, S.L., Frost, S.D.W., Muse, S.V., 2005. HyPhy: hypothesis testing using phylogenies. Bioinformatics 21, 676–679.
Kosakovsky Pond, S.L., Murrell, B., Fourment, M., et al., 2011. A random effects branch-site model for detecting episodic diversifying selection. Mol. Biol. Evol. 28, 3033–3043.
Landis, G.N., Bhole, D., Tower, J., 2003. A search for doxycycline-dependent mutations that increase *Drosophila melanogaster* life span identifies the VhaSFD, Sugar baby, filamin, fwd and Cctl genes. Genome Biol. 4, 1–14.
Lanfear, R., Calcott, B., Kainer, D., Mayer, C., Stamatakis, A., 2014. Selecting optimal partitioning schemes for phylogenomic datasets. BMC Evol. Biol. 14, 82.
Larkin, M.a., Blackshields, G., Brown, N.P., et al., 2007. Clustal W and Clustal X version 2.0. Bioinformatics 23, 2947–2948.
Larracuente, A.M., Sackton, T.B., Greenberg, A.J., et al., 2008. Evolution of protein-coding genes in *Drosophila*. Trends Genet. 24, 114–123.
Leatherman, J.L., Levin, L., Boero, J., Jongens, T.a., 2002. Germ cell-less acts to repress transcription during the establishment of the *Drosophila* germ cell lineage. Curr. Biol. 12, 1681–1685.
Lechner, M., Hernandez-Rosales, M., Doerr, D., et al., 2014. Orthology detection combining clustering and synteny for very large datasets. PLoS One 9, e105015.
Leinonen, R., Sugawara, H., Shumway, M., 2011. The sequence read archive. Nucleic Acids Res. 39, 2010–2012.
Liu, L., Yu, L., Edwards, S.V., 2010. A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. BMC Evol. Biol. 10, 302.
Magnacca, K.N., Price, D.K., 2015. Rapid adaptive radiation and host plant conservation in the Hawaiian picture wing *Drosophila* (Diptera: Drosophilidae). Mol. Phylogenet. Evol. 92, 226–242.
Markova-Raina, P., Petrov, D., 2011. High sensitivity to aligner and high rate of false positives in the estimates of positive selection in the 12 *Drosophila* genomes. Genome Res. 21, 863–874.
Marrone, A.K., Kucherenko, M.M., Rishko, V.M., Shcherbata, H.R., 2011. New Dystrophin/Dystroglycan interactors control neuron behavior in *Drosophila* eye. BMC Neurosci. 12, 93.
Moretti, S., Laurenczy, B., Gharib, W.H., et al., 2014. Selectome update: quality control and computational improvements to a database of positive selection. Nucleic Acids Res. 42, 917–921.
Morrow, G., Tanguay, R.M., 2015. *Drosophila melanogaster* Hsp22: a mitochondrial small heat shock protein influencing the aging process. Front. Genet. 6, 1–7.
Moskalev, A., Aliper, A., Smit-McBride, Z., Buzdin, A., Zhavoronkov, A., 2014. Genetics and epigenetics of aging and longevity. Cell Cycle 13, 1063–1077.
Murrell, B., Wertheim, J.O., Moola, S., et al., 2012. Detecting individual sites subject to episodic diversifying selection. PLoS Genet. 8, e1002764.
Newton, F.G., zur Lage, P.I., Karak, S., et al., 2012. Forkhead transcription factor Fd3F cooperates with Rfx to regulate a gene expression program for mechanosensory cilia specialization. Dev. Cell 22, 1221–1233.
Nguyen, H.-N., Yang, J.-M., Afkari, Y., et al., 2014. Engineering ePTEN, an enhanced PTEN with increased tumor suppressor activities. Proc. Natl. Acad. Sci. USA 111, 2684–2693.
Nielsen, R., 2005. Molecular signatures of natural selection. Annu. Rev. Genet. 39, 197–218.
Nozawa, M., Suzuki, Y., Nei, M., 2009. Reliabilities of identifying positive selection by the branch-site and the site-prediction methods. Proc. Natl. Acad. Sci. 106, 6700–6705.
O'Grady, P.M., Markow, T.a., 2009. Phylogenetic taxonomy in *Drosophila*. Fly 3, 10–14.
Paaby, A.B., Blacket, M.J., Hoffmann, A.a., Schmidt, P.S., 2010. Identification of a candidate adaptive polymorphism for *Drosophila* life history by parallel independent clines on two continents. Mol. Ecol. 19, 760–774.
Panchin, A.Y., Gelfand, M.S., Ramensky, V.E., Artamonova, I.I., 2010. Asymmetric and non-uniform evolution of recently duplicated human genes. Biol. Direct 5, 54.
Parkash, R., Ramniwas, S., Kajla, B., Aggarwal, D.D., 2012. Divergence of desiccation-related traits in two *Drosophila* species of the *takahashii* subgroup from the western Himalayas. J. Exp. Biol. 215, 2181–2191.
Parker, J., Tsagkogeorga, G., Cotton, J.a., et al., 2013. Genome-wide signatures of convergent evolution in echolocating mammals. Nature 502, 1–9.
Patterson, R.a., Juarez, M.T., Hermann, A., et al., 2013. Serine proteolytic pathway activation reveals an expanded ensemble of wound response genes in *Drosophila*. PLoS One 8, e61773.
Pecot, M.Y., Chen, Y., Akin, O., et al., 2014. Sequential axon-derived signals couple target survival and layer specificity in the *Drosophila* visual system. Neuron 82, 320–333.
Peng, Y., Leung, H.C.M., Yiu, S.M., Chin, F.Y.L., 2012. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinform. (Oxford, England) 28, 1420–1428.
Pezier, a., Jezzini, S.H., Marie, B., Blagburn, J.M., 2014. Engrailed alters the specificity of synaptic connections of *Drosophila* auditory neurons with the giant fiber. J. Neurosci. 34, 11691–11704.
Poliakov, A., Cotrina, M., Wilkinson, D.G., 2004. Diverse roles of eph receptors and ephrins in the regulation of cell migration and tissue assembly. Dev. Cell 7, 465–480.
Price, M.N., Dehal, P.S., Arkin, A.P., 2010. FastTree 2 – approximately maximum-likelihood trees for large alignments. PLoS One 5, e9490.
Price, D.K., Souder, S.K., Russo-Tait, T., 2014. Sexual selection, epistasis and species boundaries in sympatric Hawaiian picture-winged *Drosophila*. J. Insect Behav. 27, 27–40.
Proshkina, E.N., Shaposhnikov, M.V., Sadritdinova, A.F., Kudryavtseva, A.V., Moskalev, A.a., 2015. Basic mechanisms of longevity: a case study of *Drosophila* pro-longevity genes. Ageing Res. Rev. 25, 218–231.
Ranwez, V., Harispe, S., Delsuc, F., Douzery, E.J.P., 2011. MACSE: multiple alignment of coding SEquences accounting for frameshifts and stop codons. PLoS One 6, e22594.
Reed, R.D., Papa, R., Martin, A., et al., 2011. Optix drives the repeated convergent evolution of butterfly wing pattern mimicry. Science 333, 1137–1141.
Remmert, M., Biegert, A., Hauser, A., Söding, J., 2011. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nat. Methods 9, 173–175.

Roux, J., Privman, E., Moretti, S., et al., 2014. Patterns of positive selection in seven ant genomes. Mol. Biol. Evol. 31, 1661–1685.
Sagner, A., Merkel, M., Aigouy, B., et al., 2012. Establishment of global patterns of planar polarity during growth of the *Drosophila* wing epithelium. Curr. Biol. 22, 1296–1301.
Schreiber, F., Sonnhammer, E.L.L., 2013. Hieranoid: hierarchical orthology inference. J. Mol. Biol. 425, 2072–2081.
Seifert, J.R.K., Mlodzik, M., 2007. Frizzled/PCP signalling: a conserved mechanism regulating cell polarity and directed motility. Nat. Rev. Genet. 8, 126–138.
Shaw, T.I., Ruan, Z., Glenn, T.C., Liu, L., 2013. STRAW: species TRee analysis web server. Nucleic Acids Res. 41, 238–241.
Shiao, M.-S., Chang, J.-M., Fan, W.-L., et al., 2015. Expression divergence of chemosensory genes between *Drosophila sechellia* and its sibling species and its implications for host shift. Geno. Biol. Evolut. 7, evv183.
Smith, M.D., Wertheim, J.O., Weaver, S., et al., 2015. Less is more: an adaptive branch-site random effects model for efficient detection of episodic diversifying selection. Mol. Biol. Evol. 32, 1342–1353.
Song, M.S., Salmena, L., Pandolfi, P.P., 2012. The functions and regulation of the PTEN tumour suppressor. Nat. Rev. Mol. Cell Biol. 13, 283–296.
Sonnhammer, E.L.L., Östlund, G., 2015. InParanoid 8: orthology analysis between 273 proteomes, mostly eukaryotic. Nucleic Acids Res. 43, D234–D239.
Sopko, R., Perrimon, N., 2013. Receptor tyrosine kinases in *Drosophila* development. Cold Spring Harbor Perspect. Biol. 5, 1–31.
Studer, R.A., Robinson-Rechavi, M., 2009. How confident can we be that orthologs are similar, but paralogs differ? Trends Genet. 25, 210–216.
Suzuki, Y., Nei, M., 2004. False-positive selection identified by ML-based methods: examples from the Sig1 gene of the diatom *Thalassiosira weissflogii* and the tax gene of a human T-cell lymphotropic virus. Mol. Biol. Evol. 21, 914–921.
Talavera, G., Castresana, J., 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst. Biol. 56, 564–577.
Trachana, K., Forslund, K., Larsson, T., et al., 2014. A phylogeny-based benchmarking test for orthology inference reveals the limitations of function-based validation. PLoS One 9.
Trachana, K., Larsson, T.a., Powell, S., et al., 2011. Orthology prediction methods: a quality assessment using curated protein families. BioEssays: News Rev. Molec. Cell. Develop. Biol. 33, 769–780.
Usui, T., Shima, Y., Shimada, Y., et al., 1999. Flamingo, a seven-pass transmembrane cadherin, regulates planar cell polarity under the control of Frizzled. Cell 98, 585–595.
Wang, G.-D., Zhai, W., Yang, H.-C., et al., 2015. Out of southern East Asia: the natural history of domestic dogs across the world. Cell Res. 26, 1–13.
Webb, B., Sali, A., 2014. Protein structure modeling with MODELLER. Meth. Molec. Biol. (Clifton, N.J.) 1137, 1–15.
Wheat, C.W., Wahlberg, N., 2013a. Critiquing blind dating: the dangers of over-confident date estimates in comparative genomics. Trends Ecol. Evol. 28, 636–642.
Wheat, C.W., Wahlberg, N., 2013b. Phylogenomic insights into the Cambrian explosion, the colonization of land and the evolution of flight in Arthropoda. Syst. Biol. 62, 93–109.
Wheeler, J.C., VanderZwan, C., Xu, X., et al., 2002. Distinct in vivo requirements for establishment versus maintenance of transcriptional repression. Nat. Genet. 32, 206–210.
Wong, W.S.W., Yang, Z., Goldman, N., Nielsen, R., 2004. Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. Genetics 168, 1041–1051.
Yang, Z., Nielsen, R., Goldman, N., Krabbe Pedersen, A.-M., 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics 155, 431–449.
Yang, Z., Dos Reis, M., 2011. Statistical properties of the branch-site test of positive selection. Mol. Biol. Evol. 28, 1217–1228.
Zhang, J., Nielsen, R., Yang, Z., 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol. Biol. Evol. 22, 2472–2479.
Zuberi, K., Franz, M., Rodriguez, H., et al., 2013. GeneMANIA prediction server 2013 update. Nucleic Acids Res. 41, 115–122.