Comparative gene annotation of 304 species of Drosophilidae

Pankaj Dhakad[1]*, Bernard Kim[2], Dmitri Petrov[3,4], Darren J. Obbard[1]

[1]Institute of Ecology and Evolution, University of Edinburgh, Edinburgh, United Kingdom

[2]Department of Ecology and Evolutionary Biology, Princeton University, Princeton, United States of America

[3]Department of Biology, Stanford University, Stanford, California, United States of America

[4]CZ Biohub, Investigator, San Francisco, California, United States of America

* Corresponding author

Email: Pankaj.Dhakad@ed.ac.uk

Abstract

High-quality genome annotations are essential if we are to address central questions in comparative genomics, such as the origin of new genes, the drivers of genome size variation, and the evolutionary forces shaping gene content and structure. Here, we present protein-coding gene annotations for 304 species of the family Drosophilidae, generated using the Comparative Annotation Toolkit (CAT) and BRAKER3, and incorporating available RNA-seq and protein evidence. We take a comparative phylogenetic approach to annotation, with the aim of improving consistency and accuracy, and to generate a robust set of gene annotations and orthology assignments. We analyze our annotations using a phylogenetic mixed-model approach and find that gene number and CDS length exhibit moderate phylogenetic heritability (43.3% and 12.3%, respectively). This suggests that while evolutionary history contributes to variation in these traits, species-specific factors, including assembly error, play a substantial role in shaping observed differences. To illustrate the utility of our annotations for comparative analyses, we investigate codon usage bias and amino acid composition across Drosophilidae. We find that codon usage is correlated with overall GC content and evolves slowly, but that it is also strongly shaped by selection, such that, in general, species with the strongest selection on synonymous codon usage show the lowest GC bias in third codon positions. This comparative annotation dataset forms part of an on-going collaborative project to sequence and annotate all species of Drosophilidae, with data and annotations being made rapidly and freely available on an on-going basis. We hope that this effort will serve as a foundation for studies in evolutionary and functional genomics and comparative biology across Drosophilidae.

Introduction

Fundamental questions in comparative genomics include the origin of new genes, the causes of genome size variation, and the factors that shape rates of genome evolution. However, addressing these questions requires not just the genome sequence, but also accurate identification and characterization of genomic features such as coding DNA sequence. High-quality genome annotations are thus essential for the identification of loci that underpin phenomena such as adaptation and speciation [1]. However, the annotation of genomes remains challenging, due to complex gene structures, long non-coding regions, and species-specific features [1-3]. Automated annotation methods, while efficient, often produce artefacts that can mislead interpretations of gene function and evolutionary relationships [4-6]. For example, an early annotation of the *Daphnia pulex* genome may have over-estimated the number of genes and over-predicted paralogs of genes involved in environmental responsiveness, potentially leading to initial misinterpretations of the basis of its adaptive capabilities [7, 8].

Over the past decade, advances in long-read sequencing technologies and scaffolding methods, together with the declining cost of sequencing, have led to a dramatic increase in the number and quality of genome assemblies across the tree of life [9, 10]. This surge is exemplified by large-scale initiatives such as the Darwin Tree of Life project [11], which aims to sequence around 70,000 species in the UK and Ireland, the Vertebrate Genome Project (VGP)

[12], which has the goal of sequencing all extant vertebrate species, and the African BioGenome Project [13], which seeks to sequence more than 105 thousand species in Africa. However, despite advances in sequencing and assembly, genome annotation remains a major bottleneck [14]. Most genomes from large-scale sequencing projects are annotated independently using automated pipelines such as BRAKER, Augustus, and MAKER [15-17]. While these annotations can incorporate evidence from reference protein databases (e.g., OrthoDB), they typically don't take advantage of closely related species or conserved gene order to improve accuracy and consistency across genomes [18, 19]. As a result, different pipelines can identify slightly different sets of genes due to variation in prediction algorithms, parameter settings, and assumptions about gene structure [20]. This inconsistency means that genes missed by one pipeline might be detected by another, while certain gene models may only partially align between annotations [14]. Consequently, such independent annotations make it difficult to achieve a standardized gene set for comparative analyses [7, 8, 20]. Comparative annotation, on the other hand, aims to address these issues by using alignment with well-annotated reference genomes to help guide predictions, reducing discrepancies and aligning gene models more consistently across closely related species [21, 22]. This approach can not only increase the accuracy of gene predictions, but can also ensure a more robust, comparable gene set for evolutionary studies. The Comparative Annotation Toolkit (CAT) is one such method, designed to annotate genomes by projecting known gene models from a reference genome onto target genomes within a phylogenetic framework [22]. CAT integrates evidence from such a 'lift-over' with short and long read RNA-seq data, Iso-seq data, and protein alignments to refine gene model predictions, weighting features that are shared among close relatives more heavily.

For over a century, *Drosophila melanogaster* and its relatives have been at the forefront of genetics, genomics, and evolutionary research, leading to influential discoveries that have shaped these fields [23]. High-quality genome assemblies and annotations have been developed for several key species, beginning with the pioneering sequencing of the *Drosophila melanogaster* genome, which served as a reference for subsequent genomic studies [24, 25]. The *Drosophila* 12 Genomes Project expanded this foundation, offering comparative insights across multiple species, while initiatives such as the ModENCODE project further enriched our knowledge with detailed transcriptomic and epigenomic data [26, 27]. Individual research groups have continued to sequence target species [28-30, 31, such as Miranda, Guanche etc], making possible resources such as DrosOMA (https://drosoma.dcsr.unil.ch/)—providing genus-wide orthology information for 36 Drosophila species [32]. This progress has now culminated in the on-going community effort to achieve a comprehensive genomic study of the entire family Drosophilidae [33, 34]. This effort includes the *de novo* sequencing of new species, scaffolding and improvement of existing genomes, and the generation of new transcriptomic data [33]. As of April 2024, around 360 different drosophilid species had been sequenced to varying levels of completeness—some fragmentary, from short-read data alone [e.g. Drosophila setifemur and Drosophila ironensis; 31], but many to chromosome-level assemblies, using long-read data and/or scaffolding information from HiC [e.g. Chmomyza fuscimana; 35].

Here we contribute to this continuing community effort by providing a comparative coding-sequence annotation for 304 of the highest quality drosophilid genomes. We do this using a combination of CAT and BRAKER3, combining publicly available RNAseq data and previous RefSeq reference genomes [15, 22]. We use phylogenetic linear mixed models to assess and compare the annotations, and as an example of the utility of our comprehensive annotation we analyze codon and amino-acid usage bias across the family. To facilitate its use in both single-gene and genome-wide studies, the annotation is made freely available in the form of genome annotation files and also as aligned (and optionally masked) orthology groups, with annotations linked to gene orthology. In the future, we plan to continue updating this resource with regular new releases as new genomes become available. We hope that this will be a key resource, enabling gene-based analyses of evolution within this important model system.

Methods

Genome Assemblies

We selected an initial candidate set of genome assemblies by supplementing those of Kim et al [33] with all other publicly available drosophilid genomes available as of February 2024. This included genomes available in the RefSeq database (Release 222) [36], those generated by the Darwin Tree of Life project [11], and many assemblies generated by individual labs [37-40]. For each species, we short-listed genome assemblies that had an N50 greater than 50 Kbp and a BUSCO (Benchmarking Universal Single-Copy Orthologs) completeness score of over 90% [41], selecting the assembly with the highest N50 where multiple assemblies were available. To identify and mask repetitive elements, we used RepeatMasker v4.1.2 [42] and Dfam release 3.7 repeat library [43]. These soft-masked genomes were used for all subsequent analyses.

RNA-seq and Protein Data

To annotate the genomes we used the Comparative Annotation Toolkit [CAT; 22], a pipeline that leverages external evidence ('hints') combining data such as RNA-seq, Iso-seq, proteins, and attempted lift-over from aligned references. For each species, we first gathered available RNA-seq data to provide transcript evidence that can help resolve ambiguities in gene. We identified suitable RNA-seq datasets using the ENA Portal API (https://www.ebi.ac.uk/ena/portal/api), selecting up to 10 paired-end RNA-seq datasets and prioritizing those with Poly-A selection to enrich mRNA [44]. Where available, we included data from up to 10 tissue types, including whole body, carcass, thorax, brain, testes and ovaries, as well as different developmental stages and both sexes (selected SRA numbers for each species are in S1 Table). The chosen RNA-seq reads were then down-sampled and normalized to 100x coverage using BBNorm of BBMap v38.95 [45]. The normalized RNA-seq reads were aligned to their respective species genomes using the STAR v2.7.9a with default parameters [46].

To provide protein ‘hints’, we extracted predicted protein sequences from Arthropoda using the OrthoDB v10 protein database [https://orthodb.org; 47]. Such protein sequences provide evolutionary conserved evidence that complements RNA-seq data, particularly for genes that may be underrepresented or absent in the RNA-seq datasets. We aligned these protein sequences to the genomes using miniprot v0.12 [48] with parameters “-ut8 --gtf genome_file”, which are optimized for mapping proteins to genomic sequences. The alignment files generated by miniprot were then converted into hints files using the aln2hints.pl script from the GALBA toolkit [49].

Reference species & Cactus alignment

In addition to RNAseq and protein hints, the CAT pipeline attempts a lift-over of annotations [22]. This uses genome-scale alignments in the Hierarchical Alignment Format [50], each comprising a single reference species and several target species. To define reference clades for annotation, we first generated a preliminary species tree of the 304 drosophilid species (plus seven outgroup species) using 1824 single-copy BUSCO loci [41]. Nucleotide sequences from each locus were aligned separately using MAFFT v7.520 [51] and used to infer a maximum likelihood (ML) gene tree using IQ-TREE v2.2.6 [52] under a GTR+I+G4 substitution model. These ML gene trees were then combined to infer a species tree using ASTRAL-III v5.15.5 [53], which aims to resolve gene-tree species tree incongruence under a model of incomplete lineage sorting.

We selected 37 ‘reference’ species for lift-over annotations based on the completeness and quality of their genomes, as indicated by RefSeq annotations [36]. Using the ETE 3 python package [54], we applied a preorder tree traversal strategy to identify subclades that contained at least one reference species and included the most distantly related leaf within a predefined phylogenetic distance (measured as the expected number of substitutions per site). We varied the phylogenetic distance threshold between 0.005 and 0.35 to ensure each subclade included 3 to 15 species, with lower thresholds for densely populated regions of Drosophilidae tree and higher thresholds to include more distantly related species. From these, we then chose the reference species with gene counts most similar to those of the best-annotated drosophilid model, *Drosophila melanogaster,* as its annotation is likely to be the most complete and accurate within the family, having benefited from extensive manual curation and detailed transcriptomic and functional data [55]. Each selected reference was then used to define a clade for subsequent genome alignment and lift-over processes. For species located on very long branches (i.e. divergence >0.35) and for subclades lacking any reference-species annotations, we attempted a lift-over directly from *Drosophila melanogaster*.

We then used these ‘lift-over subclades’ as guide trees to generate multiple whole-genome alignments with ProgressiveCactus [56]. This approach ensured that the alignments were computationally feasible, and that closely-related genomes aligned together. Finally, we employed the Comparative Annotation Toolkit [22] to annotate multiple target genomes simultaneously, using a lift-over of the selected reference annotation for each subclade.

Running CAT

To perform the genome annotations, we first prepared the reference annotations and extrinsic 'hints' for use in CAT [22]. RefSeq annotation files were converted using the "convert_ncbi_gff3" script provided by CAT, and the resulting GFF3 files were validated with the "validate_gff3" script to ensure compatibility. We then employed three modes of AUGUSTUS [16] in CAT: two based on transMap projections (AugustusTM/R) that project annotations from reference genomes onto target genomes, and one using ab-initio and comparative gene predictions (AugustusCGP) guided by extrinsic hints [57]. We used Comparative Gene Prediction (CGP) parameters trained on 12 well annotated Drosophila species from the Drosophila 12 Genomes Project, based on exon and intron scoring [https://bioinf.uni-greifswald.de/augustus/datasets/; 57].

Complementation with BRAKER3

To complement the comparative annotations generated by CAT, and to reduce potential reference bias, we additionally incorporated *de novo* non-comparative CDS predictions made by BRAKER3 [15]. BRAKER3 is an automated gene prediction pipeline that integrates RNA-seq data with gene prediction algorithms to generate gene models without reference to a reference annotation, improving the identification of novel genes and gene duplicates that may be absent from the reference. However, among the annotations generated by BRAKER3 we identified a number of transposable elements (TEs) which are not the focus of our study. Therefore, to simplify the annotation, we removed these using EarlGrey v4.1.0 [58], applying TEstrainer (https://github.com/jamesdgalbraith/TEstrainer) to ensure that recently duplicated non-TE genes were retained. Finally, to combine the BRAKER3 annotations with those from CAT we compared the coding sequences (CDS) of overlapping genes between the two annotation sets. For one-to-one overlapping genes, we selected the annotation with the longest CDS. In cases of one-to-many or many-to-one overlaps, we preferred the CDS annotations from CAT. Additionally, we retained all non-overlapping genes with a CDS length greater than 150 nucleotides.

Annotation quality assessment

To assess the annotation quality, we used OMArk and BUSCO for assessing completeness based on evolutionary informed expectations of gene content [41, 59]. BUSCO was run in protein mode using the Diptera OrthoDB dataset to estimate the proportion of conserved single-copy orthologs recovered in each annotation [60]. For OMArk, we first generated omamer search databases for each species using the LUCA.h5 orthology database (https://omabrowser.org/oma/current/). We then ran OMArk with taxon ID 7214 (Drosophilidae), assigning proteins to orthologous groups based on phylogenetically informed gene family classifications.

Orthogroup assignment and CDS alignment

We identified coding sequence homology across the 304 *Drosophila* species and one outgroup species (*Musca domestica*) using OrthoFinder v2.5.5 [61]. OrthoFinder first identifies homology using an all-vs-all blast similarity search and then clusters sequences using a Markov clustering algorithm (MCL inflation parameter of 1.5). Subsequently, it can then identify "Hierarchical Orthologous Groups" (HOGs) that comprise the genes descended from a common ancestral gene at a specific taxonomic level—with HOGs defined for different clades nested within each other along the species phylogeny. We extracted HOGs at the level of Drosophilidae (i.e. sets of homologous sequences that have their MRCA at the base of Drosophilidae, or later) and retained for further analysis those HOGs that included at least two species and contained more than three sequences.

The sequences for each of the chosen HOGs were aligned using the MACSE v2 [62] and MAFFT v7.520 [51] aligners, an approach intended to minimize the impact of any frameshifts and in-frame stop codon errors that might arise from sequencing, assembly, or annotation problems, while maximizing codon-aligned sequence length. MACSE v2 incorporates several steps to improve alignment quality, including a prefilter to remove long non-homologous insertions that may result from incorrect annotations, such as intron inclusions or alternative splicing. It then uses HMMCleaner to mask residues that appear misaligned and applies post-processing filters to mask isolated codons and patchy sequences, removing sequences if more than 80% of the residues are masked [63]. Finally, the alignments were trimmed at both ends until a nucleotide position represented by at least 70% of the sequences was reached, ensuring a high-quality alignment for downstream analyses of protein coding sequence. From the 304 annotated drosophilid genomes, we generated 35,836 HOGs and 23,150 high-quality alignments. These HOGs form the basis of subsequent analyses of evolutionary relationships and functional conservation of genes across family Drosophilidae, and all masked and unmasked aligned sequences are made available at https://doi.org/10.5281/zenodo.15016917.

Phylogenetic Generalized Linear Mixed Model analyses

We used Generalized Linear Mixed Model (GLMM) analyses of gene number and CDS length to identify clade-to-clade variation and outlier genomes across the 304 species. Such variation may result from true evolutionary divergence, or from reference bias and the availability (or otherwise) of RNA-seq data, or from systematic errors in assembly or annotation quality. The relative impact of such factors is naturally addressed in a linear mixed model framework, treating species as a random effect and phylogenetic distance from reference species, genome size, assembly contiguity (N50), and the availability of RNAseq as fixed-effect predictors. Because related species exhibit correlated traits [leading to pseudo-replication, if not accounted for; 64], and because the phylogenetic correlation among related species (e.g. 'phylogenetic inertia' or 'phylogenetic heritability') may be of direct interest—reflecting clade-to-clade variation in gene content or the efficacy of selection—we employed a Phylogenetic mixed model approach [65], implemented in the R package MCMCglmm [66]. This incorporates phylogenetic

relationships to model the covariance among species, while evaluating the influence of fixed predictors such as phylogenetic distance from the lift-over reference and RNA-seq availability.

To do this, we first generated a revised species-tree topology of the 304 drosophilid species using 251 single-copy HOGs (those HOGs containing ≥ 300 species) employing IQTREE2 and ASTRAL-III, as for BUSCO genes above. To infer relative branch lengths in approximate time, we randomly selected 10,000 amino acid sites from the HOGs, and (conditioning on the IQTREE/ASTRAL topology) we used BEAST [67] to re-infer branch lengths on the fixed ASTRAL tree topology under a LG+G+I model with 7 gamma categories and an uncorrelated relaxed log-normal clock [68]. For the tree prior, we used birth-death process model [69], setting the fully-informative prior for the MRCA of the subgenera *Drosophila* and *Sophophora* to 47 mya [95% prior density 42-52 mya; 34], a uniform step prior between 0 and 1 on the birth-death growth rate, and the remaining priors to their default values. We ran the MCMC chain for 10^9 generations sampling every 20,000 steps, and stationarity and mixing were assessed from visual inspection of the MCMC chain and Effective Sample Size (ESS) in Tracer [70]. After discarding 20% of the sampled estimates as burn-in, we report divergence times as median node height for each of the clades in the summary tree.

To assess the factors influencing gene number and CDS length across drosophilid species, we fitted a multivariate phylogenetic mixed model using MCMCglmm [65]. Our model included gene number and mean CDS length as response variables, allowing us to analyze their (co-)variation with respect to predictors such as distance from reference, RNAseq availability, status as a liftover reference, assembled genome size, and assembly scaffold N50. This phylogenetic approach can help to disentangle the effects of biological and technical factors on genome annotation metrics, while accounting for phylogenetic relatedness among species. We inferred statistical ‘significance’ on the basis of 95% highest posterior density (HPD) credibility intervals. The model was specified as follows (MCMCglmm syntax):

```
prior <- list(B = list(mu = rep(0, 12), V = diag(12) * 1e+10),
    G = list(G1 = list(V=diag(2), nu=2, alpha.mu=rep(0,2),
    alpha.V=diag(2)*1000)),
    R = list(V=diag(2), nu=2.002))

model <- MCMCglmm(cbind(mean_len, genes) ~ trait - 1 +
    trait:distances + trait:RNA_seq + trait:ref +
    trait:genome_size + trait:N50, random = ~us(trait):Phylo,
    rcov = ~us(trait):units, family = rep("gaussian", 2),
    ginverse = list(Phylo = InverseTree),
    prior = prior, data = annot_stat,
    nitt = 1000000, burnin = 100000,
    thin = 1000, pr = TRUE)
```

Briefly, `trait` is a reserved variable that indexes columns of the response matrix in multi-response models, with -1 removing the global intercept so that each trait has its own baseline estimate. Fixed effects included phylogenetic distance from the reference species (`trait:distances`), availability of RNA-seq data (`trait:RNA_seq`), whether the species

was itself a lift-over reference (`trait:ref`), assembled genome size (`trait:genome_size`), assembly contiguity (`trait:N50`). The random effects term `us(trait):Phylo` describes the phylogenetic (co)variance matrix between gene length and gene number, and the residual variance term `us(trait):units` describes the residual covariance matrix. The argument `ginverse` fits a covariance structure among species to model non-independence due to common ancestry [65]. Variance in CDS length and gene number were treated as Gaussian.

Evolution of GC, codon, and amino acid composition across Drosophilidae

Genome-wide GC content, codon usage, and amino-acid composition are shaped by a combination of mutational bias and natural selection [71, 72]. To illustrate the utility of our coding sequence annotations and alignments for large-scale evolutionary sequence analyses, we examined the evolution of these coding-sequence traits across the 304 drosophilid species. We used the R package 'cubar' [73] to calculate the overall GC content of coding sequences ('GC'), and GC content at third codon positions (GC3). Additionally, we estimated the whole-genome GC content and the GC content of non-coding sequences using 'geecee' [74]. We used 'cusp' tool from EMBOSS [75] to calculate amino acid frequencies in each species, and the nitrogen to carbon ratio (N/C) for protein sequences was calculated as the weighted average of the N/C ratios of individual amino acids, with weights corresponding to the proportion of each amino acid in the sequence. We used a PCA analysis of amino acids frequencies to reduce the dimensionality.

We estimated the strength of selection on codon usage bias (*S*) using the approach of Dos Reis and Wernisch [76], which compares codon frequencies in highly expressed versus reference gene sets. We ranked *Drosophila melanogaster* genes according to their overall expression level (S2 Table; expression data obtained from FlyBase: https://flybase.org/) and analyzed the HOGs that contained these genes, assuming that the globally most highly-expressed genes in *Drosophila melanogaster* are also highly expressed in other species. As expected, these genes were dominated by those encoding ribosomal proteins, yolk proteins, salivary gland secretions and elongation factors whose high expression is likely to be conserved. HOGs were then ranked in order of *Drosophila melanogaster* expression level and binned in 20 expression categories of *ca.* 600 genes in each category (S3 Table). *S* was then estimated as the log-odds ratio of optimal to non-optimal codon frequencies for two-fold degenerate codons, where the preferred codon was identified from the most highly expressed gene category [76, 77]. Note that, to distinguish selection from mutational bias, this method assumes the reference and highly expressed gene sets have similar mutational patterns. Finally, we obtained bootstrap confidence intervals for *S* by resampling genes within expression categories.

As above, we used a multivariate Phylogenetic Generalized Linear Mixed Model (PGLMM) implemented in MCMCglmm to assess the among-species phylogenetic (co)variance in GC3,

non-coding GC, estimated strength of selection on codon usage bias (*S*), observed frequencies of amino acids and N/C ratio, while accounting for phylogenetic effects.

Data availability

All code used for data processing, generating figures, and statistical analyses are available through GitHub (https://github.com/DhakadPankaj/Fly-annotation). Drosophila annotation data is publicly available at link. Supplementary material is available at https://doi.org/10.5281/zenodo.15016917.

Results & Discussion

Gene annotation of 304 species

We selected 17 NCBI 'RefSeq' annotations for use as potential lift-over references and combined this information with RNAseq data from 91 species, protein hints (from mapping of OrthoDB proteins) from all species, and *de novo* prediction to annotate the remaining genomes using CAT and Braker3. On average, we identified 14,543 genes in each genome, with a mean CDS length of 1.6 Kbp. This is very similar to the gold-standard reference, *Drosophila melanogaster*, which currently has 13,904 protein-coding genes of mean CDS length 1.54 Kbp (Genome assembly Release 6.53; GCF_000001215.4). However, there was substantial variation, reflecting both evolutionary variation among species and potential variation in genome assembly quality and RNAseq availability (Fig 1).

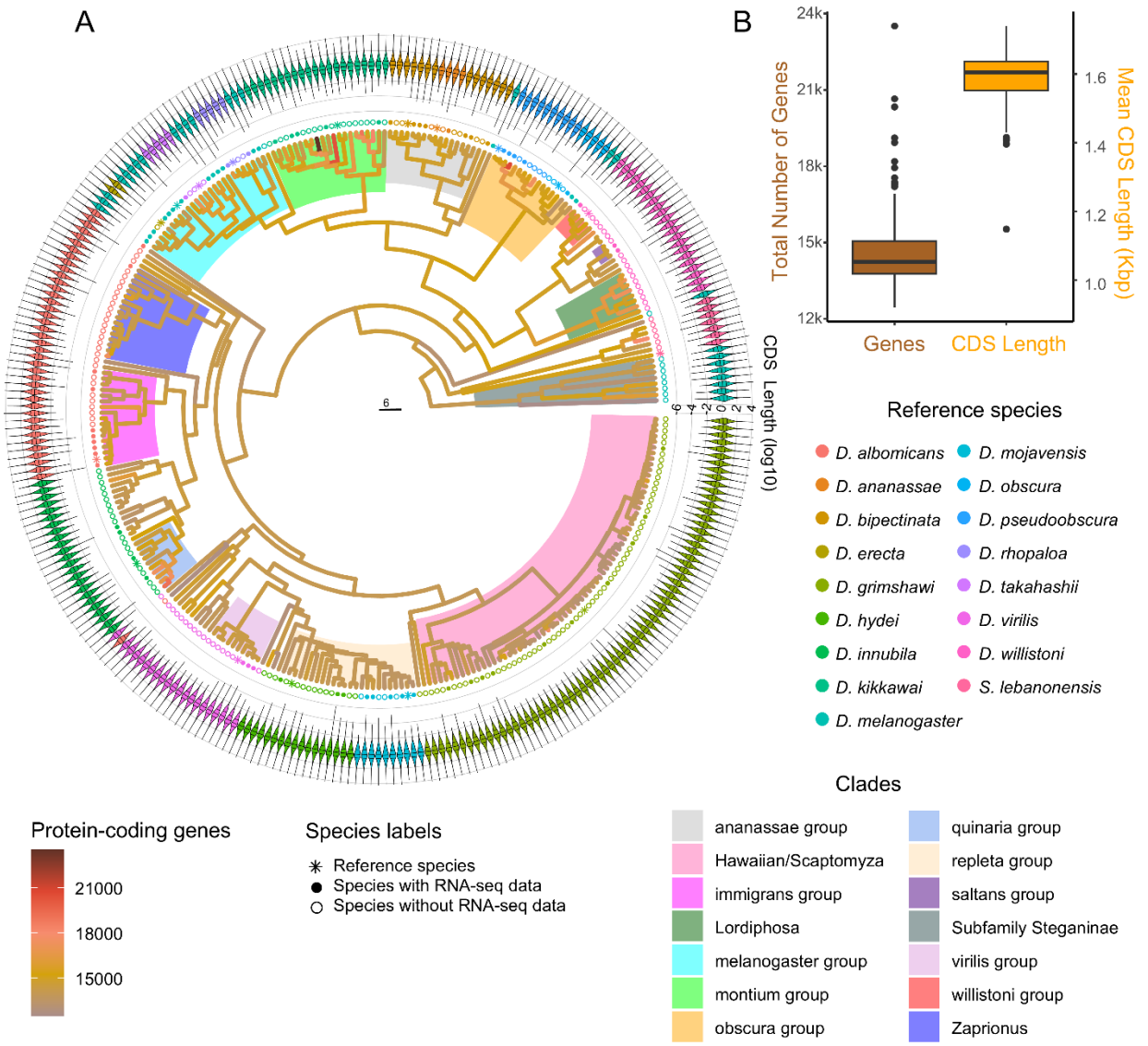


Fig 1. Overview of 304 Drosophila genome annotations.
(A) Phylogenetic tree with ancestral reconstruction of the number of protein-coding genes mapped onto branches. Tip labels are colored according to the reference species for the clade, stars indicate the reference species, and filled versus open circles indicate the availability of RNAseq data for that species. Violin plots (outermost layer) represent the distribution of coding sequence (CDS) lengths on a logarithmic scale. (B) The box plot represents the range in gene number and mean CDS length across family Drosophilidae. An alternative version of the tree, with taxon labels, is provided in S1 File.

Most species (291 of 304) were annotated with between 12,500 and 17,000 protein-coding genes, but with notable outliers. These included *Drosophila vulcana* [28], *Drosophila miranda* [78] and *Drosophila punjabiensis* [28], which each appeared to possess more than 20,000 genes (Fig 1 and S4 Table). Previous studies have suggested that gene gain on *Drosophila miranda*'s neo-Y chromosome may account for the elevated gene count in this species [78], and similar major structural changes could extend to other species with unexpectedly high gene

numbers[78]. However, it could also be that the miss-assembly or annotation errors led to unusually high numbers of genes [79, 80].

The mean CDS length for most species (296 of the 304 species) ranged between 1.42 and 1.74 Kbp. The outliers included *Drosophila pseudoobscura* ssp. *pseudoobscura* (GenBank accession: GCA_000001765.3), which exhibits unusually short CDSs (mean length of 1.15 Kbp) but a total gene count of 14,546 but close to the family median of 14,249 genes (Fig 1 and S4 Table). This observation raises the possibility that many gene models in such assemblies may be fragmented, incomplete, or represent short repetitive elements, possibly reflecting issues with assembly quality or annotation [79, 80]. Another possibility is that these observed variations in annotated gene number and CDS length might arise due to duplication/loss of gene families in the ancestors of group of species or at tips.

Establishing whether this variation reflects differences in annotation or assembly quality, or true evolutionary processes, is necessarily challenging in the absence of ground truth data for comparison. We first used BUSCO to assess the genome completeness at both the genome and annotated-protein levels. A close agreement between them suggests that the new annotations do not lack significant portion of BUSCO genes detectable at the genome level (S1 Fig). Only *Drosophila pseudoobscura* ssp. *pseudoobscura* and *Drosophila americana* had difference of more than 10% between genome and protein levels BUSCO completeness. We also used OMArk to further assess and compare the quality of protein-coding genes. OMArk estimates the completeness, consistency, fragmentation and contamination of gene-repertoire in a species by comparison with conserved orthologous groups (HOGs). We observed high levels of HOG completeness across most species. However, *Drosophila recens* and *Drosophila miranda* showed a high number of duplications (S5 Table). For *Drosophila miranda* it has previously been shown that there are occurrences of gene gain on its neo-Y chromosome, but the source of duplicates in *Drosophila recens* remains unclear and not previously reported. Interestingly, we also found a high number of duplications in *Drosophila recens*' close relatives *Drosophila subquinaria* and *Drosophila suboccidentalis* (S5 Table). Additionally, OMArk identified a significant level of contamination in the genome of *Drosophila vulcana* (5,235 genes identified as contaminant), a genome that is also flagged as 'contaminated' in NCBI database [28]. Other genomes with apparently high levels of contamination include *Drosophila punjabiensis* (3,131 genes), *Drosophila prolaticilia* (3,811 genes) and *Drosophila nannoptera* (2,006 genes).

Orthology inference

We used OrthoFinder [61] to infer CDS orthology across 304 species of Drosophilidae, using *Musca domestica* as an outgroup. OrthoFinder assigned 98.9% of predicted proteins to orthogroups (OGs), with 95.5% further classified into Hierarchical Orthologous Groups (HOGs). We identified a total of approximately 35 thousand HOGs across the 304 drosophila species genomes (Table 1). More than 90% of genes in each species were assigned to HOGs, although some species—such as *Drosophila pseudoobscura* ssp. *pseudoobscura*, *Leucophenga varia*,

Drosophila vulcana, *Drosophila quasianomalipes*, *Drosophila americana*, and *Drosophila differens*—had lower assignment rates (S2 Fig).

Table 1. Summary statistics of HOGs.

Feature	Count
Number of species	304
Total number of proteins	4438506
Number of Orthogroups (OGs)	38692
Number of proteins in OGs	4391361(98.9%)
Number of root-level HOGs	35836
Number of proteins in HOGs	4240044 (95.5%)
Number of universal single-copy HOGs (≥300 species)	251
Number of HOGs with all species present	775
Number of HOGs with ~99% of species present (≥301)	4994
Number of HOGs containing Dmel genes	12151
Number of ancient HOGs (mrca ≥50 mya)	15974

To interpret patterns of gene conservation and turnover, we classified HOGs into two broad categories: widely conserved HOGs, which include genes shared across most Drosophilidae species, and species- or clade-specific HOGs, which appear to be restricted in their distribution. The widely conserved HOGs were further divided into ‘universal’ HOGs, present in nearly all species (≥99%), and ancient HOGs, shared by species with MRCA of at least 50 mya. Universal HOGs likely represent genes experiencing strong evolutionary constraint, encoding core biological functions necessary across all species. Ancient HOGs, while also conserved, may include genes that have been differentially retained or lost in some major lineages. In contrast, species- and clade-specific HOGs could represent recent gene family expansions or lineage-specific adaptations. Where restricted HOGs encode functional proteins, they may contribute to species-specific traits; however, they also likely arise from methodological artifacts, such as genome annotation errors or orthology misassignment.

Over half of all the predicted protein-coding genes fell into universal HOGs, reinforcing the idea that most genes are highly conserved. However, a substantial fraction of the HOGs (~20,000) contained genes from a small number of species (<30), suggesting either recent evolutionary gains or problems with orthology inference. Interestingly, some HOGs classified as ancient were present in only a few species, and it seems probable that many of these ‘sparse’ HOGs reflect

fragmented assemblies or annotation inconsistencies, rather than true biological patterns. This highlights the importance of caution when interpreting or analyzing low-representation HOGs in comparative studies.

By classifying HOGs based on their evolutionary conservation and phylogenetic distribution, we provide a framework for future studies of gene conservation and turnover in Drosophilidae. Universal and ancient HOGs likely represent functionally essential genes, while restricted HOGs may point to lineage-specific innovations or methodological challenges. This classification allows us to distinguish between broad evolutionary patterns and potential technical noise, improving the reliability of our comparative analyses.

To obtain a well-supported set of orthogroups, and to mitigate potential errors from sparse HOGs, we examined HOGs that contain at least one *Drosophila melanogaster* gene in more detail. Given that *Drosophila melanogaster* has one of the most complete and thoroughly verified gene sets among multicellular eukaryotes, its representation in a HOG provides additional confidence that the group represents a biologically meaningful gene family rather than an artifact. We identified 12,151 such HOGs, which were broadly shared across the majority of Drosophilidae species (S3 Fig). This approach provides greater confidence when analyzing conserved gene families and allows us to assess how widely well-characterized genes are distributed across phylogeny. Our analysis substantially expands the most recently-available gene-orthology set for Drosophilidae from 36 species [32] to 304 species, offering a more comprehensive understanding of gene families across the entire clade, and we hope that this dataset will be valuable for future studies focusing on comparative genomics, evolutionary biology, and functional genomics within Drosophilidae.

Phylogenetic inference using BUSCO and HOG genes

Comparative analyses require an ultrametric phylogenetic tree describing relationships, and thus the expected covariance in traits, among species [65]. The most comprehensive molecular phylogeny of Drosophilidae to date encompasses 704 species, but is based on only 17 reference genes and thus has many deep branches that are not resolved with high confidence [81]. More recently, genome-sequencing of 360 species has enabled a BUSCO-gene tree based on one thousand loci—greatly improving confidence in deeper relationships [33]. However, branch lengths were inferred using 4-fold degenerate sites, which will tend to underestimate deep branch lengths due to substitution saturation. While both BUSCO genes and single-copy HOGs are conserved across Drosophilidae species and thus likely to experience greater purifying selection, BUSCO genes are identified based on their universal presence across a broader evolutionary scale (i.e., Diptera) whereas HOGs are defined within Drosophilidae. Here we infer species trees using two alternative gene-sets, one using 251 single-copy HOGs, and one using 1,824 BUSCO genes. We found that the HOG tree and the BUSCO tree were highly concordant and showed highly supported relationships for all but five species (S4 Fig). For example, in the HOG tree *Drosophila fuyamai* was positioned as a sister to *Drosophila carrolli*, *Drosophila rhopaloa*, and *Drosophila prolongata*, whereas the BUSCO tree also included *Drosophila kurseongensis* in this clade. Other conflicting relationships can be

found in the S4 Fig. Most internal branches were well supported in both trees, but in some places, the HOG tree exhibited slightly lower local posterior probabilities, particularly for short branches (S1 File). These discrepancies likely reflect increased discordance due to incomplete lineage sorting [34], as resolving such branches requires a larger number of gene trees.

Factors affecting annotated gene number and CDS Length

To better understand apparent variation in gene number and CDS length across Drosophilidae, we fitted a phylogenetic generalized linear mixed model [66] to assess the impact of ‘reference’ genome status, phylogenetic distance from the reference, genome size, assembly quality (N50), and the availability of RNA-seq data on these two traits.

Our analysis found no significant differences in gene number or CDS length between our annotations and the established reference annotations, indicating that our annotations are of comparable quality (gene number: p=0.7, 95% HPD CI [-565, 402]; CDS length: p=0.6, 95% HPD CI [-0.04, 0.005]). However, species lost an average of *ca.*14 genes for each extra million years of divergence from their liftover reference (p<0.001, CI [-22, -6.2]), while on average CDS length increased by just one nucleotide per million years divergence from reference (p=0.002; CI [0.03, 1.45]) (Fig 2). This reflects the increased challenges associated with lift-over between more divergent genomes, but—likely because lift-over was only one source of data among many—the effect is relatively small. Increased assembly contiguity (higher N50) in these generally highly contiguous genomes (minimum N50 for inclusion 50 Kbp) was unexpectedly associated with an average of *ca.* 13 *fewer* predicted genes. This may be consistent with a reduction in fragmented gene models (p=0.004; CI [-22, -3.5]) (Fig 2). The inclusion of RNA sequencing data also unexpectedly *reduced* gene number (by *ca.* 400 genes; p=0.002; CI [-692, -143])—without affecting CDS length (p=0.2, CI [-42, 8.26]). This may reflect a reduction in the overall false-positive rate, or a reduction in annotated pseudogenes, or the joining of disjunct exons. We found that *ca.* 13 genes were on average gained with each additional 1Mbp of genome assembly size (p<0.001; CI [10, 16.5]), but mean CDS length decreased by less than 1 bp (p<0.001; CI [-0.9, -0.4]) (Fig 2)—that is, larger genome assemblies contained more genes, but these genes were shorter on average. This pattern could reflect biological processes, such as gene duplications or the expansion of non-coding regions. Alternatively, it might result from the annotation of transposable elements (TEs) as genes, or more fragmented assemblies (perhaps as a result of more repetitive sequence) and thus more fragmented gene models [82].

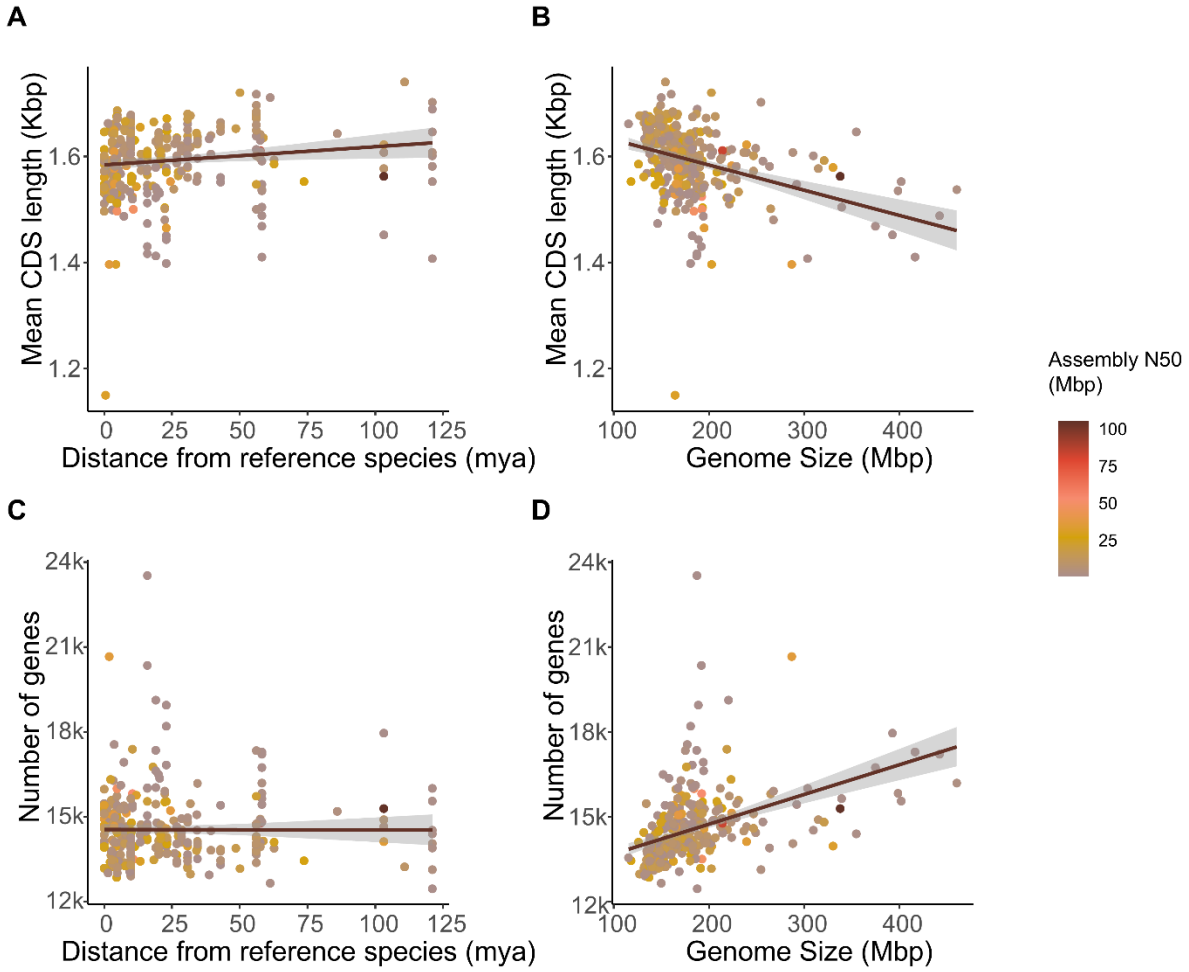


Fig 2. Variation in gene number and CDS length across Drosophilidae.
(A) Coding sequence (CDS) length increases with phylogenetic distance from the reference species used for lift-over annotation. (B) CDS length decreases as genome size increases. (C) Gene number remains largely unaffected by phylogenetic distance from the reference. (D) Gene number increases with genome size, suggesting a relationship between genome expansion and gene content. In all plots the points are colored by genome contiguity (N50). Fitted lines and 95% confidence windows are derived from a non-phylogenetic linear model and are for illustration only; see main text for phylogenetic mixed model analyses.

After accounting for these fixed effects, we found little evidence for major differences in gene number or CDS length among *Drosophila* clades. Posterior estimates of differences among internal nodes generally had confidence intervals overlapping zero, suggesting that number of genes and CDS length have remained relatively stable across lineages (S5 Fig). Correspondingly, phylogenetic heritability took intermediate values for both gene number (43.3%; CI [23.8, 64.3]) and mean CDS length (12.3%; CI [4.2, 23.4]), indicating that while evolutionary history plays a role, species-specific factors contribute substantially to variation in these traits. The negative covariance between gene number and CDS length (-0.42; CI [-0.77, -0.12]) suggests species with more genes tend to have shorter CDS lengths, possibly reflecting

differences in annotation stringency or genome assembly quality. This suggests that large-scale shifts in these genomic traits are rare, and variation in gene number and length are more likely driven by annotation artifacts or species-specific factors rather than broad evolutionary trends. Nevertheless, the montium group had a generally higher inferred gene number and shorter CDS lengths compared to other clades (S5 Fig), which may reflect lineage-specific effects—although this group includes *Drosophila vulcana* and *Drosophila punjabiensis*, which both showed high bacterial contamination in the OMArk analysis. Taken together, these results indicate that while gene number and CDS length variation occur at the species level, they do not seem to have strong phylogenetic structuring at deeper evolutionary timescales, perhaps predominantly reflecting differences in assembly and annotation rather than long-term evolutionary trends.

GC composition and Codon Usage Bias in Drosophilidae

To illustrate the potential utility of our new annotations, we analyzed variation in GC content and its relationship with codon usage bias (CUB) across Drosophilidae [72, 83]. Genomic GC content ranged from 21% in *Drosophila neohypocausta* to 49% in *Drosophila nannoptera* (S6 Table), with coding regions, as expected, showing higher GC content (range: 41–57% GC) than the genome-wide average. GC content at third codon positions (GC3) is a widely used proxy for codon bias, and reflects a balance between mutational pressure and selection acting on synonymous mutations [72, 83]. In our analysis, GC3 was highly correlated between related species, with an estimated phylogenetic heritability of 1 (i.e. no residual variance that is not captured by the phylogenetic effect), indicating strong conservation within clades and little variation among closely related species. GC3 was also positively correlated with non-coding GC content (Fig 3 and Fig 4; phylogenetic correlation from the PGLMM 0.52; $p<0.001$; CI [0.41, 0.59]), suggesting that genome-wide mutational biases contribute to both coding and non-coding base composition. Nevertheless, we observed substantial clade-specific deviations; notably the willistoni and saltans groups, along with subfamily Steganinae, had much lower GC3, whereas the genus *Zaprionus* and the melanogaster, montium, obscura, ananassae, repleta, and virilis species groups exhibited elevated GC3 (S6 Fig). These differences mirror a recent analysis of 29 *Drosophila* species, in which subgenera *Sophophora* and *Drosophila* exhibited distinct codon preferences [72]. Such lineage-specific differences could reflect factors beyond mutation bias or overall GC composition, potentially including selection for translational efficiency or efficacy [84].

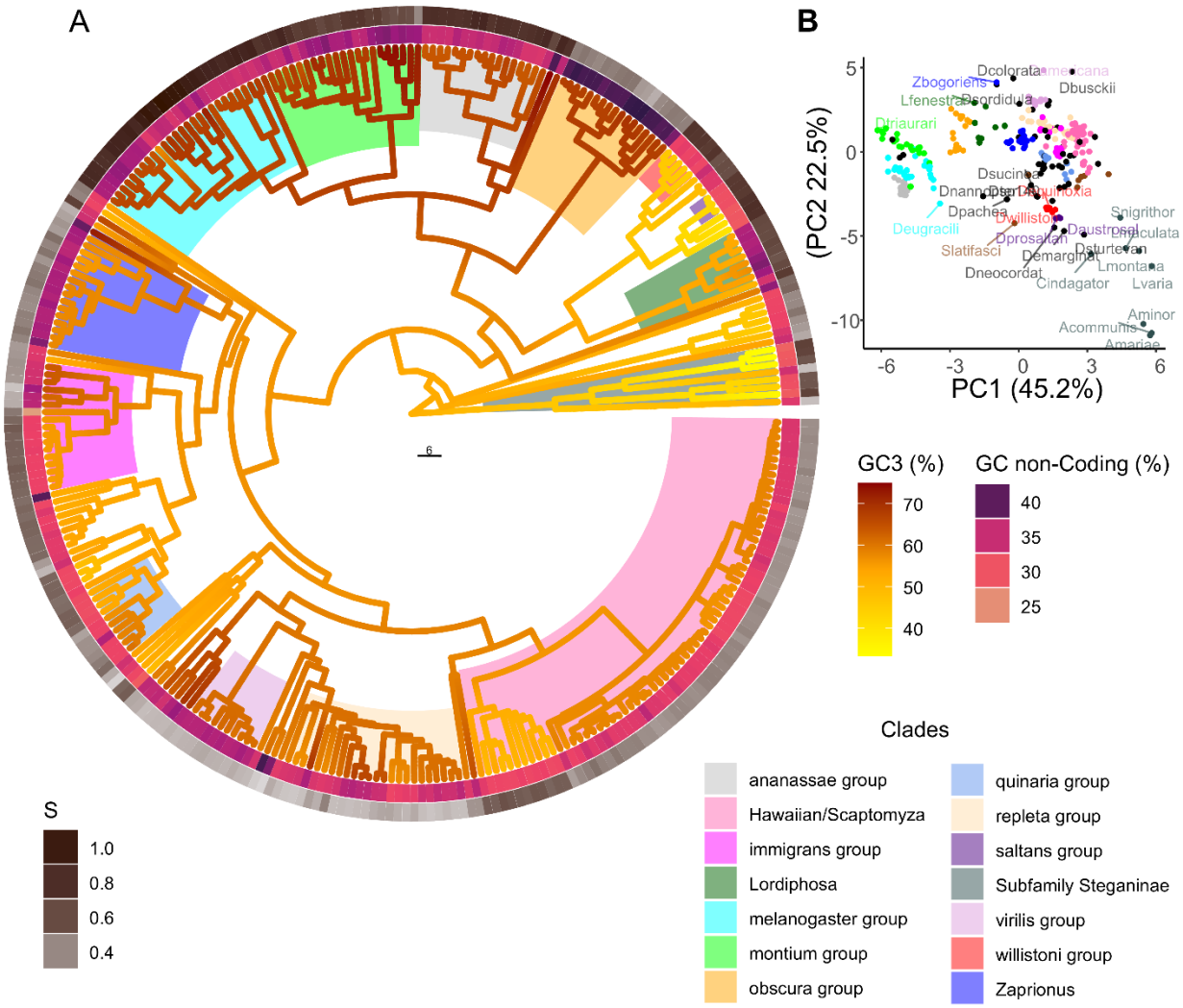


Fig 3. Codon usage, amino acid composition and selection on codon usage across Drosophilidae.

(A) Phylogenetic distribution of GC content at third codon positions (GC3), non-coding GC content, and strength of selection on codon usage bias (*S*) across Drosophilidae. The tree is color-coded by clades, and branches are colored according to GC3. Inner ring shows GC content in non-coding regions and outer ring shows strength of selection on codon usage. (B) Principal component analysis (PCA) of amino acid usage across species, showing that closely related species exhibit similar amino acid composition patterns.

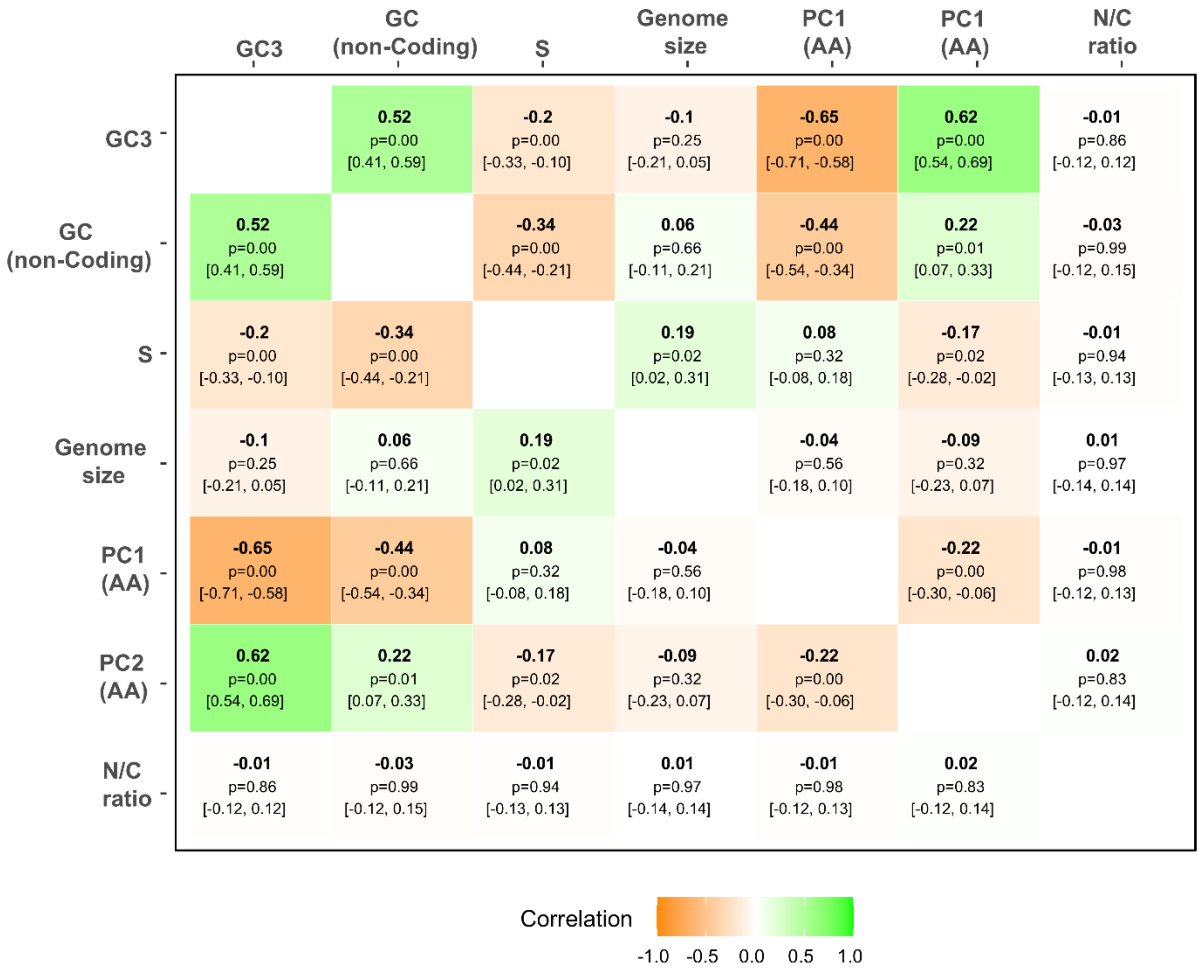


Fig 4. Correlation matrix of codon usage, genome features, and amino acid composition. Pairwise phylogenetic correlations between GC content at third codon positions (GC3), GC content in non-coding regions, strength of selection on codon usage bias (*S*), genome size, principal components PC1 and PC2 of amino acid usage, and nitrogen-to-carbon (N/C) ratio of amino acids. Values represent correlation coefficients with 95% confidence intervals and MCMC p-values. Strong correlations (positive or negative) are highlighted in green and orange, respectively, indicating relationships between nucleotide composition, codon usage, and amino acid preferences.

To quantify the role of selection in determining codon usage, we estimated the strength of selection on two-fold codons [quantified by the 'S' statistic of 76]. Estimates of *S* ranged from 0.24 in *Drosophila pachea* (95% bootstrap interval across genes [0.22, 0.29]) up to 1.08 [0.96, 1.20] in *Drosophila takahashii* (Fig 3). As expected, the melanogaster, montium, and ananassae groups showed elevated GC3 and *S*, confirming stronger selection in favour of GC-ending codons in these groups [72, 84, 85]. However, the willistoni and saltans groups—which display low GC3—also showed relatively high *S*, confirming that the AT-bias seen in *Drosophila willistoni* is (at least in part) a result of selection [84, 86, 87]. A similar pattern was also seen in the subfamily Steganinae and overall GC3 and *S* are negatively correlated across Drosophilidae—indicating that species with higher GC3 are, on average, experiencing weaker selection on codon usage (Fig 4; phylogenetic correlation from the PGLMM: -0.2; p<0.001; CI [-

0.33, -0.10]). Interestingly, a weak *positive* correlation between *S* and genome size (Fig 4; phylogenetic correlation from the PGLMM 0.15; p=0.02; CI: [0.03, 0.32]) suggests that species with larger genomes tend to experience slightly stronger selection on codon usage—in contrast to what might be expected under relaxed constraint in species with small effective population size [88].

Amino Acid Composition

To investigate whether the variation in codon usage is associated with variation in amino acid composition, we analyzed the relative proportions of all 20 amino acids across the annotated proteins of Drosophilidae. In general, it is thought that amino acid usage is influenced by a combination of mutational biases, translational selection, and functional constraints—but genome-wide nucleotide composition has been shown to play a significant role in shaping amino acid frequencies [83, 89]. Our principal component analysis (PCA) revealed that more closely related species share more similar amino acid usage patterns (Fig 3). PC1 primarily separated species based on GC content on the codons, with high-GC3 genomes enriched GC-rich amino acids (Pro, Gly, Ala, Arg) and low-GC3 genomes enriched AT-rich amino acids (Asn, Tyr, Ile), see Fig 5. To assess whether the patterns were linked to biochemical properties of the amino-acids, such as nitrogen-to-carbon (N/C) ratio or amino acid essentiality [measured in Drosophila melanogaster; 90, 91], we examined the remaining PCA loadings. However, we found no clear patterns, suggesting that other factors, such as protein structure or functional constraints, may play a small role in shaping variation in amino acid use among species of Drosophilidae (Fig 5).

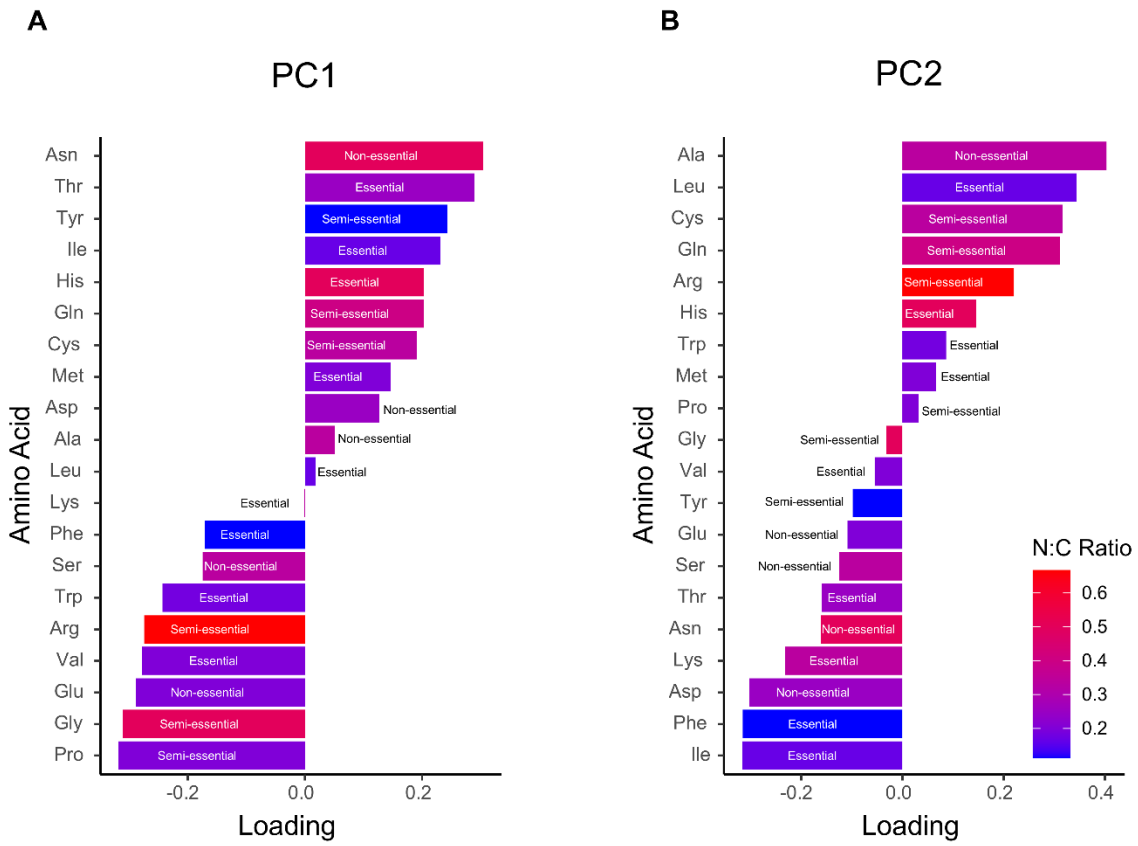


Fig 5. Principal component analysis (PCA) loadings of amino acid usage. Bar plots show the loadings of individual amino acids on the first two principal components: (A) PC1 and (B) PC2. Amino acids are coloured according to their nitrogen to carbon (N:C) ratio, with higher ratios in red and lower ratios in blue. Essentiality categories (essential, semi-essential, non-essential) are indicated alongside each bar. Positive and negative loadings reflect the relative contribution of each amino acid to the corresponding principal component.

Conclusions

This work, to generate standardized, simultaneous multi-species coding DNA sequence annotations across 304 species of Drosophilidae, forms part of an ongoing community effort working toward a comprehensive genomic study of the entire family [33]. We envisage that these new annotations, orthology assignments, and multiple sequence alignments will provide a valuable resource for both single-gene and genome-wide evolutionary studies. And, along with future updates as new genomes are sequenced, this resource will support future research in studies of adaptation and functional genomics within this key model clade.

Supporting information

S1 Fig. BUSCO completeness scores for genome assemblies and annotated protein sets.

S2 Fig. Overview of orthology assignments across 304 Drosophilidae species.
S3 Fig. Classification of Hierarchical Orthologous Groups (HOGs).
S4 Fig. Posterior estimates of gene number and mean CDS length reconstructed at internal nodes of the species phylogeny.
S5 Fig. Posterior estimates of GC3 content and strength of selection (S) on codon usage bias at internal nodes of the species phylogeny.
S6 Fig. Posterior estimates of GC3 and S at internal nodes.
S1 Table. SRA accession numbers for RNA-seq datasets used in genome annotation.
S2 Table. *Drosophila melanogaster* genes ranked by overall expression level (FPKM).
S3 Table. Expression category assignments for HOGs based on *D. melanogaster* gene expression ranks.
S4 Table. Summary statistics for genome annotations and orthology assignments across species.
S5 Table. BUSCO and OMArk completeness assessments for genome annotations.
S6 Table. Estimates of codon usage bias metrics (GC3, GC content in non-coding regions and selection strength S) across species.
S1 File. Species trees inferred from BUSCO and HOG gene sets.

Funding

This work was supported by a Darwin Trust PhD studentship to PD, Biotechnology and Biological Sciences Research Council grant BB/T007516/1 to DJO, and National Institutes of Health (NIH) National Institute of General Medical Sciences (NIGMS) grant R35 GM118165 to DAP. DAP is a CZ Biohub Investigator. For the purpose of open access, the author has applied a Creative Commons Attribution (CC BY) license to any Author Accepted Manuscript version arising from this submission.

Acknowledgements

We wish to thank members of the University of Edinburgh Institute for Ecology and Evolution for the collaborative provision of shared computational resources, James Galbraith for help with EarlGrey and TEstrainer, and Eric Lai and Garima Setia for feedback on missing gene models. We would also like to acknowledge the contributions of the broader Drosophila research community, whose collaborative efforts in genome sequencing have made this work possible.

References

1. Ejigu GF, Jung J. Review on the Computational Genome Annotation of Sequences Obtained by Next-Generation Sequencing. Biology (Basel). 2020;9(9):295. Epub 20200918. doi: 10.3390/biology9090295.
2. Kwon T, Hanschen ER, Hovde BT. Addressing the pervasive scarcity of structural annotation in eukaryotic algae. Sci Rep. 2023;13(1):1687. Epub 20230130. doi: 10.1038/s41598-023-27881-0.

3. Vuruputoor VS, Monyak D, Fetter KC, Webster C, Bhattarai A, Shrestha B, et al. Welcome to the big leaves: Best practices for improving genome annotation in non-model plant genomes. Appl Plant Sci. 2023;11(4):e11533. Epub 20230808. doi: 10.1002/aps3.11533.
4. Mathe C, Dunand C. Automatic Prediction and Annotation: There Are Strong Biases for Multigenic Families. Front Genet. 2021;12:697477. Epub 20210916. doi: 10.3389/fgene.2021.697477.
5. Promponas VJ, Iliopoulos I, Ouzounis CA. Annotation inconsistencies beyond sequence similarity-based function prediction - phylogeny and genome structure. Stand Genomic Sci. 2015;10(1):108. Epub 20151119. doi: 10.1186/s40793-015-0101-2.
6. Scalzitti N, Jeannin-Girardon A, Collet P, Poch O, Thompson JD. A benchmark study of ab initio gene prediction methods in diverse eukaryotic organisms. BMC Genomics. 2020;21(1):293. Epub 20200409. doi: 10.1186/s12864-020-6707-9.
7. Colbourne JK, Pfrender ME, Gilbert D, Thomas WK, Tucker A, Oakley TH, et al. The ecoresponsive genome of Daphnia pulex. Science. 2011;331(6017):555-61. doi: 10.1126/science.1197761.
8. Ye Z, Xu S, Spitze K, Asselman J, Jiang X, Ackerman MS, et al. A New Reference Genome Assembly for the Microcrustacean Daphnia pulex. G3 (Bethesda). 2017;7(5):1405-16. Epub 20170505. doi: 10.1534/g3.116.038638.
9. van Dijk EL, Naquin D, Gorrichon K, Jaszczyszyn Y, Ouazahrou R, Thermes C, et al. Genomics in the long-read sequencing era. Trends Genet. 2023;39(9):649-71. Epub 20230523. doi: 10.1016/j.tig.2023.04.006.
10. Espinosa E, Bautista R, Larrosa R, Plata O. Advancements in long-read genome sequencing technologies and algorithms. Genomics. 2024;116(3):110842. Epub 20240411. doi: 10.1016/j.ygeno.2024.110842.
11. Darwin Tree of Life Project C. Sequence locally, think globally: The Darwin Tree of Life Project. Proc Natl Acad Sci U S A. 2022;119(4):e2115642118. doi: 10.1073/pnas.2115642118.
12. Rhie A, McCarthy SA, Fedrigo O, Damas J, Formenti G, Koren S, et al. Towards complete and error-free genome assemblies of all vertebrate species. Nature. 2021;592(7856):737-46. Epub 20210428. doi: 10.1038/s41586-021-03451-0.
13. Sharaf A, Ndiribe CC, Omotoriogun TC, Abueg L, Badaoui B, Badiane Markey FJ, et al. Bridging the gap in African biodiversity genomics and bioinformatics. Nat Biotechnol. 2023;41(9):1348-54. doi: 10.1038/s41587-023-01933-2.
14. Freedman AH, Sackton TB. Building better genome annotations across the tree of life. 2024.
15. Gabriel L, Bruna T, Hoff KJ, Ebel M, Lomsadze A, Borodovsky M, et al. BRAKER3: Fully automated genome annotation using RNA-seq and protein evidence with GeneMark-ETP, AUGUSTUS, and TSEBRA. Genome Res. 2024;34(5):769-77. Epub 20240625. doi: 10.1101/gr.278090.123.
16. Stanke M, Diekhans M, Baertsch R, Haussler D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics. 2008;24(5):637-44. Epub 20080124. doi: 10.1093/bioinformatics/btn013.
17. Cantarel BL, Korf I, Robb SM, Parra G, Ross E, Moore B, et al. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res. 2008;18(1):188-96. Epub 20071119. doi: 10.1101/gr.6743907.
18. Venkatraman M, Fleischer RC, Tsuchiya MTN. Comparative Analysis of Annotation Pipelines Using the First Japanese White-Eye (Zosterops japonicus) Genome. Genome Biol Evol. 2021;13(5). doi: 10.1093/gbe/evab063.
19. Nachtweide S, Romoth L, Stanke M. Comparative Genome Annotation. Methods in Molecular Biology: Springer US; 2024. p. 165-87.

20. Weisman CM, Murray AW, Eddy SR. Mixing genome annotation methods in a comparative analysis inflates the apparent number of lineage-specific genes. Curr Biol. 2022;32(12):2632-9 e2. Epub 20220518. doi: 10.1016/j.cub.2022.04.085.
21. König S, Romoth L, Stanke M. Comparative Genome Annotation. Methods in Molecular Biology: Springer New York; 2018. p. 189-212.
22. Fiddes IT, Armstrong J, Diekhans M, Nachtweide S, Kronenberg ZN, Underwood JG, et al. Comparative Annotation Toolkit (CAT)-simultaneous clade and personal genome annotation. Genome Res. 2018;28(7):1029-38. Epub 20180608. doi: 10.1101/gr.233460.117.
23. Beller M, Oliver B. One hundred years of high-throughput Drosophila research. Chromosome Res. 2006;14(4):349-62. doi: 10.1007/s10577-006-1065-2.
24. Richards S, Liu Y, Bettencourt BR, Hradecky P, Letovsky S, Nielsen R, et al. Comparative genome sequencing of Drosophila pseudoobscura: chromosomal, gene, and cis-element evolution. Genome Res. 2005;15(1):1-18. doi: 10.1101/gr.3059305.
25. Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, et al. The genome sequence of Drosophila melanogaster. Science. 2000;287(5461):2185-95. doi: 10.1126/science.287.5461.2185.
26. Drosophila 12 Genomes C, Clark AG, Eisen MB, Smith DR, Bergman CM, Oliver B, et al. Evolution of genes and genomes on the Drosophila phylogeny. Nature. 2007;450(7167):203-18. doi: 10.1038/nature06341.
27. mod EC, Roy S, Ernst J, Kharchenko PV, Kheradpour P, Negre N, et al. Identification of functional elements and regulatory circuits by Drosophila modENCODE. Science. 2010;330(6012):1787-97. Epub 20101222. doi: 10.1126/science.1198374.
28. Bronski MJ, Martinez CC, Weld HA, Eisen MB. Whole Genome Sequences of 23 Species from the Drosophila montium Species Group (Diptera: Drosophilidae): A Resource for Testing Evolutionary Hypotheses. G3 (Bethesda). 2020;10(5):1443-55. Epub 20200504. doi: 10.1534/g3.119.400959.
29. Puerma E, Orengo DJ, Cruz F, Gómez-Garrido J, Librado P, Salguero D, et al. The High-Quality Genome Sequence of the Oceanic Island Endemic Species Drosophila guanche Reveals Signals of Adaptive Evolution in Genes Related to Flight and Genome Stability. Genome Biology and Evolution. 2018;10(8):1956-69. doi: 10.1093/gbe/evy135.
30. Mahajan S, Wei KH-C, Nalley MJ, Gibilisco L, Bachtrog D. De novo assembly of a young Drosophila Y chromosome using single-molecule sequencing and chromatin conformation capture. PLOS Biology. 2018;16(7):e2006348. doi: 10.1371/journal.pbio.2006348.
31. Li F, Rane RV, Luria V, Xiong Z, Chen J, Li Z, et al. Phylogenomic analyses of the genusDrosophilareveals genomic signals of climate adaptation. Molecular Ecology Resources. 2022;22(4):1559-81. doi: 10.1111/1755-0998.13561.
32. Thiebaut A, Altenhoff AM, Campli G, Glover N, Dessimoz C, Waterhouse RM. DrosOMA: the Drosophila Orthologous Matrix browser. F1000Res. 2023;12:936. Epub 20240116. doi: 10.12688/f1000research.135250.2.
33. Kim BY, Gellert HR, Church SH, Suvorov A, Anderson SS, Barmina O, et al. Single-fly genome assemblies fill major phylogenomic gaps across the Drosophilidae Tree of Life. PLoS Biol. 2024;22(7):e3002697. Epub 20240718. doi: 10.1371/journal.pbio.3002697.
34. Suvorov A, Kim BY, Wang J, Armstrong EE, Peede D, D'Agostino ERR, et al. Widespread introgression across a phylogeny of 155 Drosophila genomes. Curr Biol. 2022;32(1):111-23 e5. Epub 20211116. doi: 10.1016/j.cub.2021.10.052.
35. Obbard DJ WSIToLp. The genome sequence of a drosophilid fruit fly, Chymomyza fuscimana (Drosophilidae) (Zetterstedt, 1838). Wellcome Open Res. 2023. doi: 10.12688/wellcomeopenres.20122.1.
36. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional

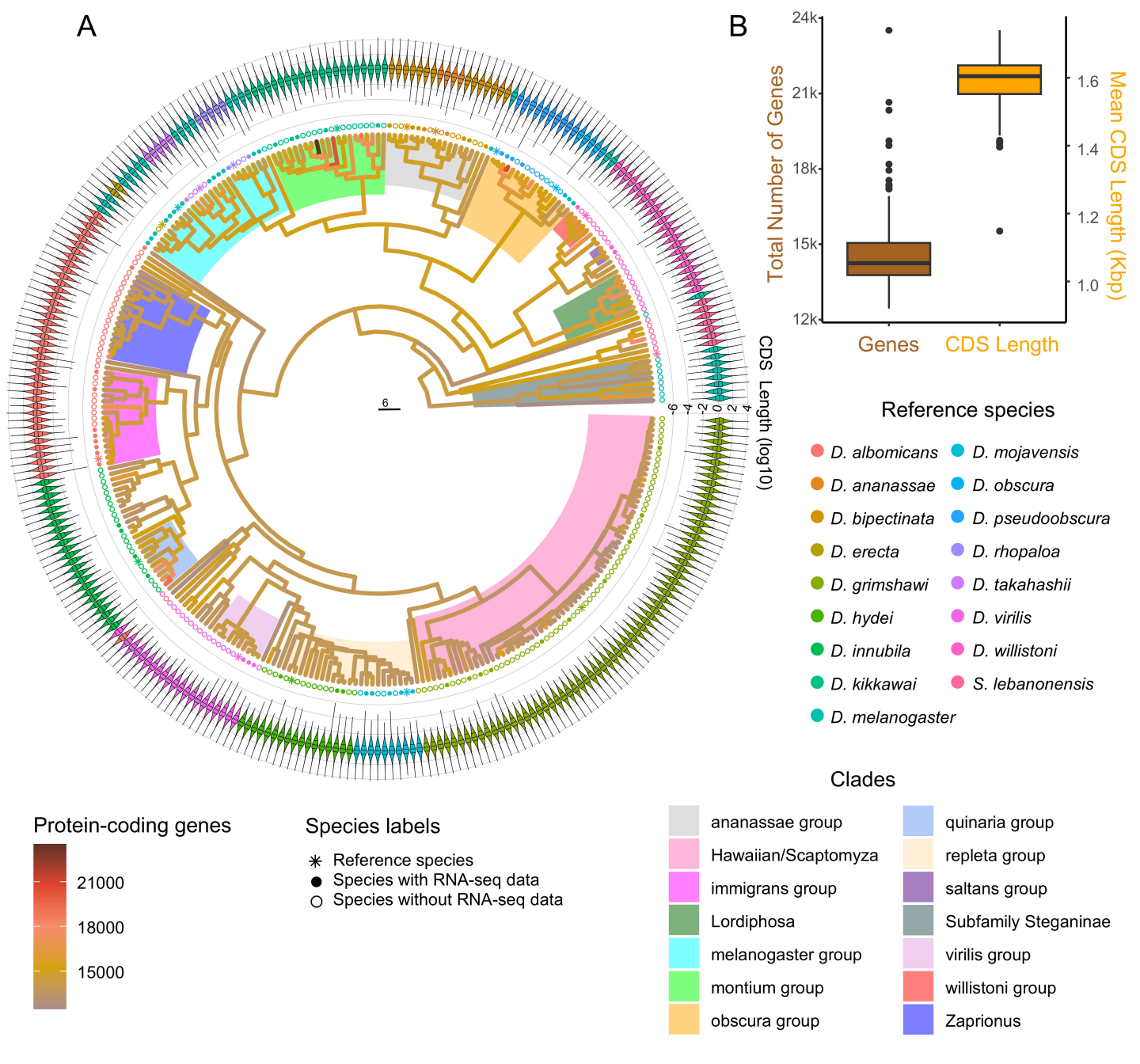
annotation. Nucleic Acids Res. 2016;44(D1):D733-45. Epub 20151108. doi: 10.1093/nar/gkv1189.

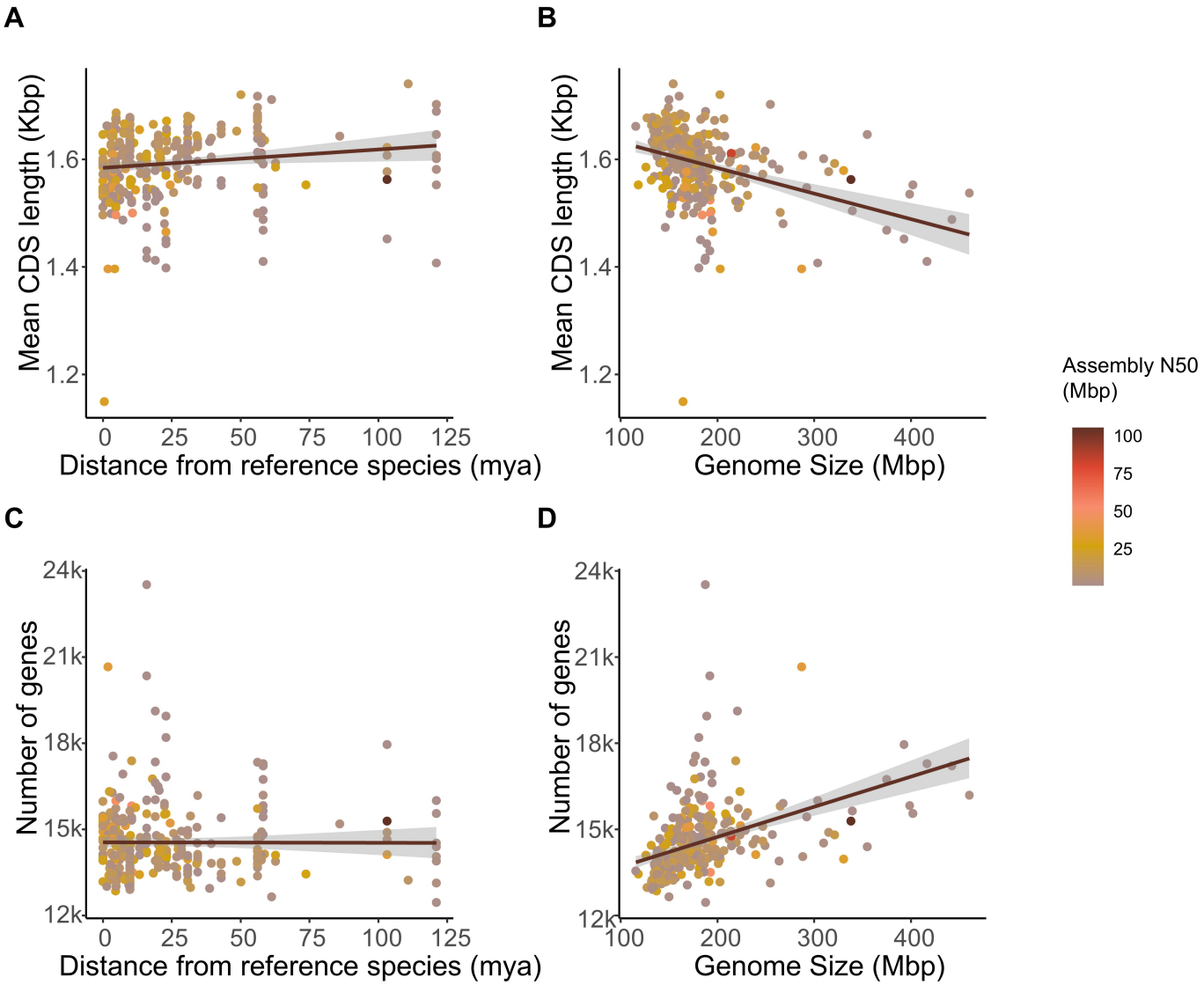
37. Zhou Q, Bachtrog D. Sex-specific adaptation drives early sex chromosome evolution in Drosophila. Science. 2012;337(6092):341-5. doi: 10.1126/science.1225385.
38. Renschler G, Richard G, Valsecchi CIK, Toscano S, Arrigoni L, Ramirez F, et al. Hi-C guided assemblies reveal conserved regulatory topologies on X and autosomes despite extensive genome shuffling. Genes Dev. 2019;33(21-22):1591-612. Epub 20191010. doi: 10.1101/gad.328971.119.
39. Sanchez-Flores A, Penaloza F, Carpinteyro-Ponce J, Nazario-Yepiz N, Abreu-Goodger C, Machado CA, et al. Genome Evolution in Three Species of Cactophilic Drosophila. G3 (Bethesda). 2016;6(10):3097-105. Epub 20161013. doi: 10.1534/g3.116.033779.
40. Wei KH, Mai D, Chatla K, Bachtrog D. Dynamics and Impacts of Transposable Element Proliferation in the Drosophila nasuta Species Group Radiation. Mol Biol Evol. 2022;39(5). doi: 10.1093/molbev/msac080.
41. Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31(19):3210-2. Epub 20150609. doi: 10.1093/bioinformatics/btv351.
42. Smit A, Hubley, R & Green, P. RepeatMasker Open-4.0 2013-2015. Available from: http://www.repeatmasker.org.
43. Storer J, Hubley R, Rosen J, Wheeler TJ, Smit AF. The Dfam community resource of transposable element families, sequence models, and genome annotations. Mob DNA. 2021;12(1):2. Epub 20210112. doi: 10.1186/s13100-020-00230-y.
44. Yuan D, Ahamed A, Burgin J, Cummins C, Devraj R, Gueye K, et al. The European Nucleotide Archive in 2023. Nucleic Acids Res. 2024;52(D1):D92-D7. doi: 10.1093/nar/gkad1067.
45. Bushnell B. BBMap: A Fast, Accurate, Splice-Aware Aligner 2014. Available from: https://sourceforge.net/projects/bbmap/.
46. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1):15-21. Epub 20121025. doi: 10.1093/bioinformatics/bts635.
47. Kriventseva EV, Kuznetsov D, Tegenfeldt F, Manni M, Dias R, Simao FA, et al. OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. Nucleic Acids Res. 2019;47(D1):D807-D11. doi: 10.1093/nar/gky1053.
48. Li H. Protein-to-genome alignment with miniprot. Bioinformatics. 2023;39(1). doi: 10.1093/bioinformatics/btad014.
49. Bruna T, Li H, Guhlin J, Honsel D, Herbold S, Stanke M, et al. Galba: genome annotation with miniprot and AUGUSTUS. BMC Bioinformatics. 2023;24(1):327. Epub 20230831. doi: 10.1186/s12859-023-05449-z.
50. Hickey G, Paten B, Earl D, Zerbino D, Haussler D. HAL: a hierarchical format for storing and analyzing multiple genome alignments. Bioinformatics. 2013;29(10):1341-2. Epub 20130316. doi: 10.1093/bioinformatics/btt128.
51. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30(4):772-80. Epub 20130116. doi: 10.1093/molbev/mst010.
52. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, et al. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Mol Biol Evol. 2020;37(5):1530-4. doi: 10.1093/molbev/msaa015.
53. Zhang C, Rabiee M, Sayyari E, Mirarab S. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. BMC Bioinformatics. 2018;19(Suppl 6):153. Epub 20180508. doi: 10.1186/s12859-018-2129-y.

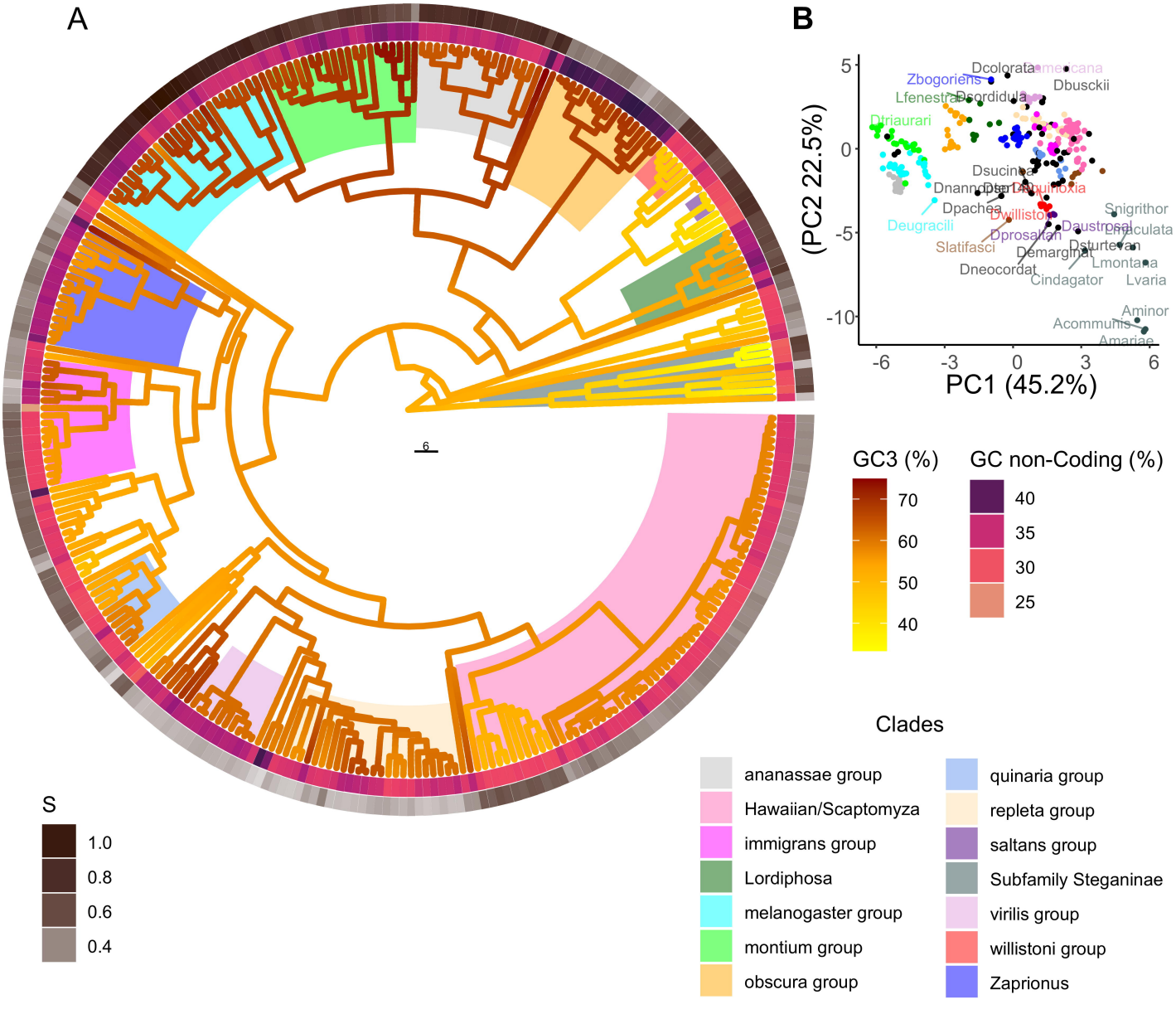
54. Huerta-Cepas J, Serra F, Bork P. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. Mol Biol Evol. 2016;33(6):1635-8. Epub 20160226. doi: 10.1093/molbev/msw046.
55. Matthews BB, Dos Santos G, Crosby MA, Emmert DB, St Pierre SE, Gramates LS, et al. Gene Model Annotations for Drosophila melanogaster: Impact of High-Throughput Data. G3 (Bethesda). 2015;5(8):1721-36. Epub 20150624. doi: 10.1534/g3.115.018929.
56. Armstrong J, Hickey G, Diekhans M, Fiddes IT, Novak AM, Deran A, et al. Progressive Cactus is a multiple-genome aligner for the thousand-genome era. Nature. 2020;587(7833):246-51. Epub 20201111. doi: 10.1038/s41586-020-2871-y.
57. Konig S, Romoth LW, Gerischer L, Stanke M. Simultaneous gene finding in multiple genomes. Bioinformatics. 2016;32(22):3388-95. Epub 20160727. doi: 10.1093/bioinformatics/btw494.
58. Baril T, Galbraith J, Hayward A. Earl Grey: A Fully Automated User-Friendly Transposable Element Annotation and Analysis Pipeline. Mol Biol Evol. 2024;41(4). doi: 10.1093/molbev/msae068.
59. Nevers Y, Warwick Vesztrocy A, Rossier V, Train CM, Altenhoff A, Dessimoz C, et al. Quality assessment of gene repertoire annotations with OMArk. Nat Biotechnol. 2025;43(1):124-33. Epub 20240221. doi: 10.1038/s41587-024-02147-w.
60. Kuznetsov D, Tegenfeldt F, Manni M, Seppey M, Berkeley M, Kriventseva EV, et al. OrthoDB v11: annotation of orthologs in the widest sampling of organismal diversity. Nucleic Acids Res. 2023;51(D1):D445-D51. doi: 10.1093/nar/gkac998.
61. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. 2019;20(1):238. Epub 20191114. doi: 10.1186/s13059-019-1832-y.
62. Ranwez V, Douzery EJP, Cambon C, Chantret N, Delsuc F. MACSE v2: Toolkit for the Alignment of Coding Sequences Accounting for Frameshifts and Stop Codons. Mol Biol Evol. 2018;35(10):2582-4. doi: 10.1093/molbev/msy159.
63. Di Franco A, Poujol R, Baurain D, Philippe H. Evaluating the usefulness of alignment filtering methods to reduce the impact of errors on evolutionary inferences. BMC Evol Biol. 2019;19(1):21. Epub 20190111. doi: 10.1186/s12862-019-1350-2.
64. Freckleton RP. The seven deadly sins of comparative analysis. J Evol Biol. 2009;22(7):1367-75. Epub 20090605. doi: 10.1111/j.1420-9101.2009.01757.x.
65. Hadfield JD, Nakagawa S. General quantitative genetic methods for comparative biology: phylogenies, taxonomies and multi-trait models for continuous and categorical characters. J Evol Biol. 2010;23(3):494-508. Epub 20100107. doi: 10.1111/j.1420-9101.2009.01915.x.
66. Hadfield JD. MCMC Methods for Multi-Response Generalized Linear Mixed Models: The MCMCglmm R Package. Journal of Statistical Software. 2010;33.
67. Suchard MA, Lemey P, Baele G, Ayres DL, Drummond AJ, Rambaut A. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. Virus Evol. 2018;4(1):vey016. Epub 20180608. doi: 10.1093/ve/vey016.
68. Drummond AJ, Ho SY, Phillips MJ, Rambaut A. Relaxed phylogenetics and dating with confidence. PLoS Biol. 2006;4(5):e88. Epub 20060314. doi: 10.1371/journal.pbio.0040088.
69. Gernhard T. The conditioned reconstructed process. J Theor Biol. 2008;253(4):769-78. Epub 20080415. doi: 10.1016/j.jtbi.2008.04.005.
70. Rambaut A, Drummond AJ, Xie D, Baele G, Suchard MA. Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. Syst Biol. 2018;67(5):901-4. doi: 10.1093/sysbio/syy032.
71. Hershberg R, Petrov DA. Selection on codon bias. Annu Rev Genet. 2008;42(Volume 42, 2008):287-99. doi: 10.1146/annurev.genet.42.110807.091442.
72. Kokate PP, Techtmann SM, Werner T. Codon usage bias and dinucleotide preference in 29 Drosophila species. G3 (Bethesda). 2021;11(8). doi: 10.1093/g3journal/jkab191.

73. Zhang H. cubar: Codon Usage Bias Analysis. R package 2024. Available from: https://CRAN.R-project.org/package=cubar.
74. Blankenberg D, Taylor J, Schenck I, He J, Zhang Y, Ghent M, et al. A framework for collaborative analysis of ENCODE data: making large-scale analyses biologist-friendly. Genome Res. 2007;17(6):960-4. doi: 10.1101/gr.5578007.
75. Rice P, Longden I, Bleasby A. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet. 2000;16(6):276-7. doi: 10.1016/s0168-9525(00)02024-2.
76. dos Reis M, Wernisch L. Estimating translational selection in eukaryotic genomes. Mol Biol Evol. 2009;26(2):451-61. Epub 20081125. doi: 10.1093/molbev/msn272.
77. Eyre-Walker A, Bulmer M. Synonymous substitution rates in enterobacteria. Genetics. 1995;140(4):1407-12. doi: 10.1093/genetics/140.4.1407.
78. Bachtrog D, Mahajan S, Bracewell R. Massive gene amplification on a recently formed Drosophila Y chromosome. Nat Ecol Evol. 2019;3(11):1587-97. Epub 20191028. doi: 10.1038/s41559-019-1009-9.
79. Torresen OK, Star B, Mier P, Andrade-Navarro MA, Bateman A, Jarnot P, et al. Tandem repeats lead to sequence assembly errors and impose multi-level challenges for genome and protein databases. Nucleic Acids Res. 2019;47(21):10994-1006. doi: 10.1093/nar/gkz841.
80. Ko BJ, Lee C, Kim J, Rhie A, Yoo DA, Howe K, et al. Widespread false gene gains caused by duplication errors in genome assemblies. Genome Biol. 2022;23(1):205. Epub 20220927. doi: 10.1186/s13059-022-02764-1.
81. Finet C, Kassner VA, Carvalho AB, Chung H, Day JP, Day S, et al. DrosoPhyla: Resources for Drosophilid Phylogeny and Systematics. Genome Biol Evol. 2021;13(8). doi: 10.1093/gbe/evab179.
82. Konstantinidis KT, Tiedje JM. Trends between gene content and genome size in prokaryotic species with larger genomes. Proc Natl Acad Sci U S A. 2004;101(9):3160-5. Epub 20040218. doi: 10.1073/pnas.0308653100.
83. Behura SK, Severson DW. Comparative analysis of codon usage bias and codon context patterns between dipteran and hymenopteran sequenced genomes. PLoS One. 2012;7(8):e43111. Epub 20120817. doi: 10.1371/journal.pone.0043111.
84. Heger A, Ponting CP. Variable strength of translational selection among 12 Drosophila species. Genetics. 2007;177(3):1337-48. doi: 10.1534/genetics.107.070466.
85. Vicario S, Moriyama EN, Powell JR. Codon usage in twelve species of Drosophila. BMC Evol Biol. 2007;7(1):226. Epub 20071115. doi: 10.1186/1471-2148-7-226.
86. Singh ND, Arndt PF, Petrov DA. Minor shift in background substitutional patterns in the Drosophila saltans and willistoni lineages is insufficient to explain GC content of coding sequences. BMC Biol. 2006;4(1):37. Epub 20061018. doi: 10.1186/1741-7007-4-37.
87. Powell JR, Sezzi E, Moriyama EN, Gleason JM, Caccone A. Analysis of a shift in codon usage in Drosophila. J Mol Evol. 2003;57 Suppl 1(0):S214-25. doi: 10.1007/s00239-003-0030-3.
88. Petit N, Barbadilla A. Selection efficiency and effective population size in Drosophila species. J Evol Biol. 2009;22(3):515-26. Epub 20081218. doi: 10.1111/j.1420-9101.2008.01672.x.
89. Williford A, Demuth JP. Gene expression levels are correlated with synonymous codon usage, amino acid composition, and gene architecture in the red flour beetle, Tribolium castaneum. Mol Biol Evol. 2012;29(12):3755-66. Epub 20120723. doi: 10.1093/molbev/mss184.
90. Croset V, Schleyer M, Arguello JR, Gerber B, Benton R. A molecular and neuronal basis for amino acid sensing in the Drosophila larva. Sci Rep. 2016;6(1):34871. Epub 20161216. doi: 10.1038/srep34871.

91. Park J, Carlson JR. Physiological responses of the Drosophila labellum to amino acids. J Neurogenet. 2018;32(1):27-36. Epub 20171130. doi: 10.1080/01677063.2017.1406934.


A
B
6
CDS Length (log10)
-6
-4
-2
0
2
4
24k
21k
18k
15k
12k
Total Number of Genes
1.6
1.4
1.2
1.0
Mean CDS Length (Kbp)
Genes
CDS Length
Reference species
D. albomicans
D. ananassae
D. bipectinata
D. erecta
D. grimshawi
D. hydei
D. innubila
D. kikkawai
D. melanogaster
D. mojavensis
D. obscura
D. pseudoobscura
D. rhopaloa
D. takahashii
D. virilis
D. willistoni
S. lebanonensis
Protein-coding genes
21000
18000
15000
Species labels
Reference species
Species with RNA-seq data
Species without RNA-seq data
Clades
ananassae group
Hawaiian/Scaptomyza
immigrans group
Lordiphosa
melanogaster group
montium group
obscura group
quinaria group
repleta group
saltans group
Subfamily Steganinae
virilis group
willistoni group
Zaprionus

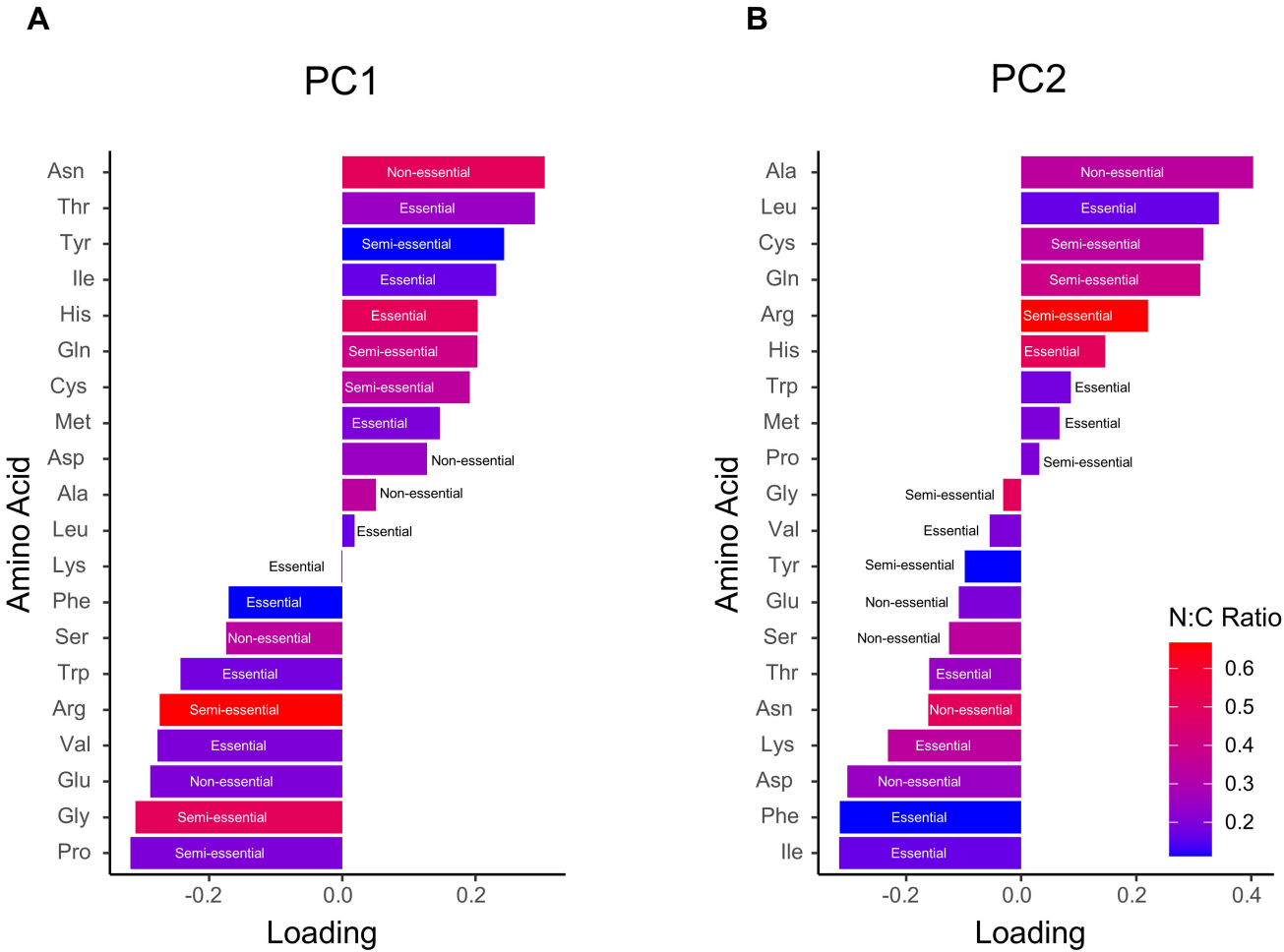

A
B
C
D
Mean CDS length (Kbp)
1.6
1.4
1.2
0
25
50
75
100
125
Distance from reference species (mya)
100
200
300
400
Genome Size (Mbp)
Number of genes
24k
21k
18k
15k
12k
Assembly N50 (Mbp)
100
75
50
25


A
B
6
S
1.0
0.8
0.6
0.4
(PC2 22.5%)
PC1 (45.2%)
5
0
-5
-10
-6
-3
0
3
6
Dcolorata
Dbusckii
Zbogoriens
Lfenestra
Dsordidula
Dtriaurari
Dsucinea
Dnannopter
Dquinoxia
Dpachea
Dwillistoni
Deugracili
Dprosaltan
Slatifasci
Demarginat
Dneocordat
Cindagator
Snigrithor
Daustrosal
Lmaculata
Dsturtevan
Lmontana
Lvaria
Aminor
Acommunis
Amariae
GC3 (%)
70
60
50
40
GC non-Coding (%)
40
35
30
25
Clades
ananassae group
Hawaiian/Scaptomyza
immigrans group
Lordiphosa
melanogaster group
montium group
obscura group
quinaria group
repleta group
saltans group
Subfamily Steganinae
virilis group
willistoni group
Zaprionus

	GC3	GC (non-Coding)	S	Genome size	PC1 (AA)	PC1 (AA)	N/C ratio
GC3		**0.52** p=0.00 [0.41, 0.59]	**-0.2** p=0.00 [-0.33, -0.10]	**-0.1** p=0.25 [-0.21, 0.05]	**-0.65** p=0.00 [-0.71, -0.58]	**0.62** p=0.00 [0.54, 0.69]	**-0.01** p=0.86 [-0.12, 0.12]
GC (non-Coding)	**0.52** p=0.00 [0.41, 0.59]		**-0.34** p=0.00 [-0.44, -0.21]	**0.06** p=0.66 [-0.11, 0.21]	**-0.44** p=0.00 [-0.54, -0.34]	**0.22** p=0.01 [0.07, 0.33]	**-0.03** p=0.99 [-0.12, 0.15]
S	**-0.2** p=0.00 [-0.33, -0.10]	**-0.34** p=0.00 [-0.44, -0.21]		**0.19** p=0.02 [0.02, 0.31]	**0.08** p=0.32 [-0.08, 0.18]	**-0.17** p=0.02 [-0.28, -0.02]	**-0.01** p=0.94 [-0.13, 0.13]
Genome size	**-0.1** p=0.25 [-0.21, 0.05]	**0.06** p=0.66 [-0.11, 0.21]	**0.19** p=0.02 [0.02, 0.31]		**-0.04** p=0.56 [-0.18, 0.10]	**-0.09** p=0.32 [-0.23, 0.07]	**0.01** p=0.97 [-0.14, 0.14]
PC1 (AA)	**-0.65** p=0.00 [-0.71, -0.58]	**-0.44** p=0.00 [-0.54, -0.34]	**0.08** p=0.32 [-0.08, 0.18]	**-0.04** p=0.56 [-0.18, 0.10]		**-0.22** p=0.00 [-0.30, -0.06]	**-0.01** p=0.98 [-0.12, 0.13]
PC2 (AA)	**0.62** p=0.00 [0.54, 0.69]	**0.22** p=0.01 [0.07, 0.33]	**-0.17** p=0.02 [-0.28, -0.02]	**-0.09** p=0.32 [-0.23, 0.07]	**-0.22** p=0.00 [-0.30, -0.06]		**0.02** p=0.83 [-0.12, 0.14]
N/C ratio	**-0.01** p=0.86 [-0.12, 0.12]	**-0.03** p=0.99 [-0.12, 0.15]	**-0.01** p=0.94 [-0.13, 0.13]	**0.01** p=0.97 [-0.14, 0.14]	**-0.01** p=0.98 [-0.12, 0.13]	**0.02** p=0.83 [-0.12, 0.14]	

Correlation

-1.0 -0.5 0.0 0.5 1.0


A
PC1
Amino Acid
Asn
Thr
Tyr
Ile
His
Gln
Cys
Met
Asp
Ala
Leu
Lys
Phe
Ser
Trp
Arg
Val
Glu
Gly
Pro
Non-essential
Essential
Semi-essential
Essential
Essential
Semi-essential
Semi-essential
Essential
Non-essential
Non-essential
Essential
Essential
Essential
Non-essential
Essential
Semi-essential
Essential
Non-essential
Semi-essential
Semi-essential
-0.2
0.0
0.2
Loading
B
PC2
Amino Acid
Ala
Leu
Cys
Gln
Arg
His
Trp
Met
Pro
Gly
Val
Tyr
Glu
Ser
Thr
Asn
Lys
Asp
Phe
Ile
Non-essential
Essential
Semi-essential
Semi-essential
Semi-essential
Essential
Essential
Essential
Semi-essential
Semi-essential
Essential
Semi-essential
Non-essential
Non-essential
Essential
Non-essential
Essential
Non-essential
Essential
Essential
-0.2
0.0
0.2
0.4
Loading
N:C Ratio
0.6
0.5
0.4
0.3
0.2