

## Single-fly assemblies fill major phylogenomic gaps across the Drosophilidae Tree of Life

Bernard Y. Kim<sup>1</sup>, Hannah R. Gellert<sup>1</sup>, Samuel H. Church<sup>2</sup>, Anton Suvorov<sup>3</sup>, Sean S. Anderson<sup>4\*</sup>, Olga Barmina<sup>5\*</sup>, Sofia G. Beskid<sup>1\*</sup>, Aaron A. Comeault<sup>6\*</sup>, K. Nicole Crown<sup>7\*</sup>, Sarah E. Diamond<sup>7\*</sup>, Steve Dorus<sup>8\*</sup>, Takako Fujichika<sup>9\*</sup>, James A. Hemker<sup>10\*</sup>, Jan Hrcek<sup>11\*</sup>, Maaria Kankare<sup>12\*</sup>, Toru Katoh<sup>13\*</sup>, Karl N. Magnacca<sup>14\*</sup>, Ryan A. Martin<sup>7\*</sup>, Teruyuki Matsunaga<sup>15\*</sup>, Matthew J. Medeiros<sup>16\*</sup>, Danny E. Miller<sup>17\*</sup>, Scott Pitnick<sup>8\*</sup>, Sara Simoni<sup>1\*</sup>, Tessa E. Steenwinkel<sup>18\*</sup>, Michele Schiffer<sup>19\*</sup>, Zeeshan A. Syed<sup>8\*</sup>, Aya Takahashi<sup>9\*</sup>, Kevin H-C. Wei<sup>20</sup>, Tsuya Yokoyama<sup>1\*</sup>, Michael B. Eisen<sup>21,22†</sup>, Artyom Kopp<sup>5†</sup>, Daniel Matute<sup>4†</sup>, Darren J. Obbard<sup>23†</sup>, Patrick M. O'Grady<sup>24†</sup>, Donald K. Price<sup>25†</sup>, Masanori J. Toda<sup>26†</sup>, Thomas Werner<sup>27†</sup>, Dmitri A. Petrov<sup>1,28</sup>

Correspondence: [bernardkim@stanford.edu](mailto:bernardkim@stanford.edu) (B.Y. Kim); [dpetrov@stanford.edu](mailto:dpetrov@stanford.edu) (D.A. Petrov)

\*Authors listed in alphabetical order

†Authors listed in alphabetical order

<sup>1</sup>Department of Biology, Stanford University, USA

<sup>2</sup>Department of Ecology and Evolutionary Biology, Yale University, USA

<sup>3</sup>Department of Biological Sciences, Virginia Tech, USA

<sup>4</sup>Department of Biology, University of North Carolina Chapel Hill, USA

<sup>5</sup>Department of Evolution and Ecology, University of California Davis, USA

<sup>6</sup>School of Environmental and Natural Sciences, Bangor University, UK

<sup>7</sup>Department of Biology, Case Western Reserve University, USA

<sup>8</sup>Center for Reproductive Evolution, Department of Biology, Syracuse University, USA;

<sup>9</sup>Department of Biological Sciences, Tokyo Metropolitan University, Japan

<sup>10</sup>Department of Developmental Biology, Stanford University, USA

<sup>11</sup>Institute of Entomology, Biology Centre, Czech Academy of Sciences, Czechia

<sup>12</sup>Department of Biological and Environmental Science, University of Jyväskylä, Finland

<sup>13</sup>Department of Biological Sciences, Hokkaido University, Japan

<sup>14</sup>Hawaii Invertebrate Program, Division of Forestry & Wildlife, State of Hawaii, USA

<sup>15</sup>Department of Complexity Science and Engineering, The University of Tokyo, Japan

<sup>16</sup>Pacific Biosciences Research Center, University of Hawai'i, Mānoa, USA

<sup>17</sup>Division of Genetic Medicine, Department of Pediatrics; Department of Laboratory Medicine and Pathology, University of Washington, USA

<sup>18</sup>Baylor College of Medicine, USA

<sup>19</sup>Daintree Rainforest Observatory, James Cook University, Australia

<sup>20</sup>Department of Zoology, The University of British Columbia

<sup>21</sup>Department of Cell and Molecular Biology, University of California Berkeley, United States

<sup>22</sup>Howard Hughes Medical Institute, University of California Berkeley, United States

<sup>23</sup>Institute of Ecology and Evolution, University of Edinburgh, UK

<sup>24</sup>Department of Entomology, Cornell University, USA

<sup>25</sup>School of Life Sciences, University of Nevada Las Vegas, USA

<sup>26</sup>Hokkaido University Museum, Hokkaido University, Japan

<sup>27</sup>Department of Biological Sciences, Michigan Technological University, USA

<sup>28</sup>CZ Biohub, Investigator

## Abstract

Long-read sequencing is driving rapid progress in genome assembly across all major groups of life, including species of the family Drosophilidae, a longtime model system for genetics, genomics, and evolution. We previously developed a cost-effective hybrid Oxford Nanopore (ONT) long-read and Illumina short-read sequencing approach and used it to assemble 101 drosophilid genomes from laboratory cultures, greatly increasing the number of genome assemblies for this taxonomic group. The next major challenge is to address the laboratory culture bias in taxon sampling by sequencing genomes of species that cannot easily be reared in the lab. Here, we build upon our previous methods to perform amplification-free ONT sequencing of single wild flies obtained either directly from the field or from ethanol-preserved specimens in museum collections, greatly improving the representation of lesser studied drosophilid taxa in whole-genome data. Using Illumina Novaseq X Plus and ONT P2 sequencers with R10.4.1 chemistry, we set a new benchmark for inexpensive hybrid genome assembly at US \$150 per genome while assembling genomes from as little as 35 ng of genomic DNA from a single fly. We present 183 new genome assemblies for 179 species as a resource for drosophilid systematics, phylogenetics, and comparative genomics. Of these genomes, 62 are from pooled lab strains and 121 from single adult flies. Despite the sample limitations of working with small insects, most single-fly diploid assemblies are comparable in contiguity (>1Mb contig N50), completeness (>98% complete dipteran BUSCOs), and accuracy (>QV40 genome-wide with ONT R10.4.1) to assemblies from inbred lines. We present a well-resolved multi-locus phylogeny for 360 drosophilid and 4 outgroup species encompassing all publicly available (as of August 2023) genomes for this group. Finally, we present a Progressive Cactus whole-genome, reference-free alignment built from a subset of 298 suitably high-quality drosophilid genomes. The new assemblies and alignment, along with updated laboratory protocols and computational pipelines, are released as an open resource and as a tool for studying evolution at the scale of an entire insect family.

## Introduction

Species in the model system Drosophilidae (vinegar or fruit flies) have long served to showcase the power of genomics as a tool for understanding evolutionary pattern and process. Drosophilid species were represented among the very first metazoan genomes (Adams et al., 2000; Richards et al., 2005), comparative genomic datasets (Clark et al., 2007; modENCODE Consortium et al., 2010); population genomic datasets (Mackay et al., 2012); and single-cell atlases (Li et al., 2022). These community-built resources have served as a foundation for subsequent discoveries across many fields of scientific inquiry. A logical next step would be the exhaustive sequencing of the >4,400 species (Finet et al., 2021) in this biologically diverse family, with the ultimate aim of creating a framework for connecting micro- to macro-evolutionary processes through the powerful lens of the *Drosophila* system. This seemingly daunting goal is made entirely feasible by cost effective long-read sequencing approaches (Kim et al., 2021; Miller et al., 2018; Solares et al., 2018), which greatly simplify the process of genome assembly. Further, the genomic tractability of drosophilids (genome sizes ~140-500Mbp), their worldwide abundance, and scientific importance of this model system, allow us to rapidly sequence new species while building upon the extensive scaffold of existing scientific resources and knowledge (O'Grady & DeSalle, 2018).

We previously assembled 101 genomes of 93 different drosophilid species in a major step towards the comprehensive genomic study of Drosophilidae (Kim et al., 2021). A hybrid Oxford Nanopore (ONT) long-read and Illumina short-read assembly proved to be an inexpensive and efficient approach to this task, at the time enabling us to build genomes in under a week at the low cost of US \$350 per genome. Since then, the number of drosophilid genomes has increased significantly: as of writing (August 2023), 468 genome assemblies for 167 drosophilid species are available to the public through NCBI (representative genomes are listed in **Table S1**). While impressive, the genomic resources for the *Drosophila* system have major gaps: one of the most obvious is the sampling bias towards species predisposed to culture in a laboratory environment (Finet et al., 2021; O’Grady & DeSalle, 2018). Large sections of interesting drosophilid biodiversity, such as the Hawaiian *Drosophila* and *Scaptomyza* which may be one of the best examples of an adaptive radiation in nature and possibly a fifth of the species in the family (Church & Extavour, 2022; Magnacca & Price, 2012; O’Grady et al., 2010), along with many other genera interspersed throughout the family, are almost entirely unstudied with modern genomic tools.

Correcting the taxon sampling bias in genome assemblies is an important step on the path towards a comprehensive genomic study of family Drosophilidae (Finet et al., 2021; O’Grady & DeSalle, 2018), but there are technical challenges to address. Establishment of a laboratory culture is a typical first step of generating a drosophilid genome due to the comparatively demanding sample input requirements for long-read sequencing. Recommended genomic DNA (gDNA) inputs range from a few hundred nanograms to several micrograms of high-quality, high molecular weight (HMW) gDNA, often exceeding the amount that can be extracted from an average-sized or smaller drosophilid. Pooling multiple wild individuals increases gDNA yield but increases the number of haplotypes in the data, leading to inflated consensus error rates and lower assembly contiguities (Aury & Istace, 2021; Guan et al., 2020; Kim et al., 2021). This is further exacerbated by the high genetic diversity of many insect populations, as well as issues arising from pools containing mis-identified individuals of similar species.

While single-specimen insect genome assembly methods (Adams et al., 2020; Kingan et al., 2019) circumvent the laboratory culture step and can address the taxonomic sequencing bias, the utility of Nanopore sequencing approaches for this purpose have not been thoroughly explored. It is indeed possible to obtain high-quality genome assemblies from single small organisms like drosophilids (e.g. Adams et al., 2020; Obbard et al., 2023); however, the aforementioned low-input assembly protocols typically increase the material for library construction by employing some form of amplification, and offset contiguity decreases from shorter reads by scaffolding contigs with Hi-C. These methods, while effective on a genome-by-genome basis, can introduce amplification bias but more importantly increase costs and add complexity to the process. Applying these methods at the scale of hundreds or thousands of genomes is not currently feasible without the resources of a large consortium. Furthermore, these methods still require fairly fresh (ideally, flash-frozen) samples. Collections held by field biologists, systematists, and museums are a largely untapped resource for addressing taxonomic bias in available genomes, but the efficacy of long-read sequencing for these types of specimens is similarly untested. We therefore sought out both freshly-collected

and ethanol-preserved (up to two decades old) specimens to test the limits of amplification-free Nanopore sequencing approaches for single flies.

Here, we present the outcome of that work and another major step towards unbiased comprehensive genomic study of an entire insect family: 183 new genome assemblies of 179 drosophilid species representing many of the various genera, species groups, and subgroups across Drosophilidae. With this study, genome sequences of 360 drosophilid species (plus 1 new outgroup genome from Family Diastatidae) are available for public download. Of just the new genomes, 121 are from single adult flies and 62 are from laboratory strains. Genomes were predominantly sequenced with a hybrid ONT R9.4.1 (87 genomes) or R10.4.1 (80 genomes) and Illumina approach, except for a small fraction of samples where sample quality issues or low gDNA yield limited our sequencing. The new data include several R10.4.1 runs for community testing and benchmarking, including 378× depth of coverage of the *D. melanogaster* reference (Adams et al., 2000; dos Santos et al., 2015) strain. We present an updated wet lab protocol and genome assembly pipeline that work for sequencing even the most miniscule drosophilids (e.g., some *Scaptomyza* spp. individuals that we estimate to be less than 1/2 the length of adult *D. melanogaster* and that yield 35 ng gDNA from one adult) at a per-sample cost of US \$150 including short reads. Even at this lower limit, our single-fly sequencing approach regularly produces contiguous (>1 Mb contig N50), complete (>98% BUSCO), and accurate (>QV40 genome-wide with ONT R10.4.1) genomes. Finally, we present a reference-free, whole-genome alignment of 298 drosophilid species to facilitate comparative genomic studies of this model system.

## Results and Discussion

### *Taxon sampling*

We selected additional species for sequencing with the primary objective of improving the taxonomic diversity of genomes of species across the family Drosophilidae (**Figure 1**). Following the TaxoDros database (Bächli, 2023), family Drosophilidae is split into the lesser studied subfamily Steganinae, for which we sequenced 9 species from 5 genera (*Stegana*, *Leucophenga*, *Phortica*, *Cacoxenus*, *Amiota*), and the more extensively studied subfamily Drosophilinae. Within the subfamily Drosophilinae, we sequenced 8 species from 4 genera (*Colocasiomyia*, *Chymomyza*, *Scaptodrosophila*, *Lissocephala*) that were clearly sister to the large, paraphyletic genus *Drosophila*; 22 species from 14 genera that render the genus *Drosophila* paraphyletic (*Collessia*, *Dettopsomyia*, *Dichaetophora*, *Hirtodrosophila*, *Hypselothyrea*, *Liodrosophila*, *Lordiphosa*, *Microdrosophila*, *Mulgravea*, *Mycodrosophila*, *Phorticella*, *Sphaerogastrella*, *Zygothrica*, *Zaprionus*); and improved sampling across the *testacea*, *quinaria*, *robusta*, *melanica*, *repleta*, and Hawaiian *Drosophila* species groups. For the *Scaptomyza*-Hawaiian *Drosophila* radiation specifically, we present 63 new genomes from most major groups, including multiple subgenera of the genus *Scaptomyza* (*Exalloscaphomyza*, *Engiscaphomyza*, *Elmomyza*) and representatives of the *picture-wing*, *haleakale*, *antopocerus*, *modified tarsus*, *ciliated tarsus*, and *modified mouthpart* species groups and their sub-groups. We also included two species, *D. maculinotata*, *D. flavopinicola*, considered to be close relatives of the *Scaptomyza*+Hawaiian *Drosophila* lineage (Finet et al., 2021; Spieth & Heed, 1975).

### *Highly accurate genomes with Nanopore R10.4.1 sequencing*

Oxford Nanopore sequencing hardware and chemistry have seen major upgrades in the shift to R10.4.1 and are now able to read DNA fragments at >99% single-read accuracy (the “Q20 chemistry”). To assess consensus sequence accuracy of drosophilid assemblies using these reads, we extracted HMW gDNA from male adult flies of the *D. melanogaster* Berkeley *Drosophila* Genome Project iso-1 strain and generated 54.4 Gbp (378× depth) of ONT R10.4.1 data, with a read N50 of 28,261 bp and containing 11.2 Gbp (79× depth) of reads 50 kb in length or longer. About 40× depth of publicly available short read data (Solares et al., 2018) were downloaded from NCBI. Using the full dataset, we assembled (Methods) 142.1 Mbp of genomic sequence in 93 contigs, with a contig N50 of 21.7 Mbp and a genome-wide consensus accuracy (QV46.0) comparable to the dm6 reference genome (QV43.4). Of all the 17,867 features annotated in the dm6 reference genome, we were able to transfer with LiftOff (Shumate & Salzberg, 2021) 17,627 (98.7%) to the ONT assembly, and of the subset of 13,967 protein-coding genes, 13,870 (99.3%) were lifted over to the ONT assembly.

Next, we assessed the performance of R10.4.1 for more practical levels of sequencing coverage. We downsampled the original reads to 10 replicate datasets for each of 20×, 25×, 30×, 40×, 50×, and 60× depths of coverage and ran our genome assembly pipeline both with and without additional Illumina polishing (using all 40× of the short-read data each time). Gene annotations were again lifted from the dm6 reference genome to each new assembly, and consensus quality scores were evaluated for the entire genome, then separately for each of the major chromosome scaffolds, coding sequences, introns, and intergenic regions (**Figure 2**, **Tables S2 and S3**).

Even for lower (<30×) coverage ONT-only datasets, genome-wide consensus accuracy (about QV41 at 20× depth) exceeds QV40 (**Figure 2**), the standard recommended by the Vertebrate Genomes Project (Rhie et al., 2021), and the genome-wide consensus accuracy we previously reported (QV37.15) for a *D. melanogaster* R9.4.1 and Illumina hybrid assembly (Kim et al., 2021). Consensus accuracy is fairly stable past 30× depth, but continues to increase by about QV1 from 30× to 60× depth of coverage (**Figure 2**). As we previously observed, consensus accuracy varies across both chromosomes and genomic elements (**Tables S2 and S3**), where autosomes are more accurate than sex chromosomes and coding sequences (>QV55 at ≥30× depth) are significantly more accurate than the rest of the genome. Of the ~23 Mbp of coding regions in the *D. melanogaster* genome we assembled with 60× ONT data, we detect an average of 47 errors (or QV56.9) for R10.4.1-only assemblies and 32 errors (or QV58.6) for Illumina-polished assemblies. For most genomic elements, additional polishing with Illumina data slightly improved consensus accuracy by ~QV1-2 except for contigs mapping to the dm6 Y chromosome, which were not improved by short read polishing.

While we will not delve further into assembly of the *D. melanogaster* reference genome here, our benchmarking clearly shows the R10.4.1 ONT and Illumina hybrid approach is a cost-effective way to generate high-quality assemblies that meets current standards for reference genomes. We also note a significant increase in R10.4.1 chemistry sensitivity over R9.4.1, in other words, that about 1/5 to 1/10 (by mass) of loaded library is needed to achieve similar pore occupancy to R9.4.1. This indicates great potential for low input ONT library preps and is the critical factor that makes it possible to sequence the more challenging samples presented shortly.

Despite the relatively small improvements to sequence accuracy provided by short-read polishing, we still recommend Illumina sequencing for Nanopore-based genome assembly projects of drosophilid species. The input requirements (1-10 ng gDNA) and the per-genome costs of Illumina sequencing (US \$35 for 30× coverage of a 200 Mbp fly genome) are minimal. Short-read data provide an avenue for assessing assembly accuracy in a less biased reference-free manner (Cheng et al., 2021; Rhie et al., 2020), and are easier to integrate into downstream population genomic analyses within a single variant calling pipeline for genomes from wild-collected individuals.

### *Haplotype phasing improves the accuracy of single-fly assemblies*

One of the recent major developments in genome assembly is the class of methods that consider linkage information contained in long reads from a diploid sample to construct a phased diploid assembly, rather than a single haploid assembly. Diploid assembly methods can incorporate phasing during assembly (Cheng et al., 2021), phase while polishing a haploid draft genome (Aury & Istace, 2021; Kolmogorov et al., 2023), or correct long reads prior to genome assembly in a haplotype-aware manner (Holley et al., 2021). These methods all aim to minimize issues arising from random haplotype switches in a haploid consensus of a diploid organism, which may have downstream effects on read mapping, variant calling, and out-of-frame errors in coding regions (Aury & Istace, 2021). While these genomes are not completely phased with respect to the parents and may still randomly switch parental phase within contigs, we find that the diploid assembly approach (Methods) improves assembly consensus accuracy by up to an order of magnitude (from QV29.7-QV37.3 to QV40.3-QV58.2) in a head-to-head comparison with a subset of our single-fly genome samples (**Figure 3**). These results also imply that diploid assembly of single flies is a superior strategy to assembly from pools of wild individuals. Phased diploid assembly of single flies will therefore be the primary strategy for genome assembly with wild-collected flies in this manuscript and in future work.

### *183 New drosophilid whole-genome sequences*

Here we present 183 new genomes for 179 drosophilid species. Of this total, 163 genomes were assembled with a hybrid ONT and Illumina approach, 4 were assembled with only ONT R10.4.1, and 16 were assembled only from Illumina paired-end reads. Sixty-two genomes were assembled with material from a laboratory stock, and the other 121 genomes were assembled from material extracted from a single adult fly. As described, these data improve the depth of sampling for key taxa, such as the Hawaiian *Drosophila*, multiple mycophagous taxa, and the *repleta* group, but also capture at least one species from many, but not all, drosophilid clades without a sequenced representative (**Figure 1**).

High-quality genome assemblies were generated from these data following our pipeline for haploid assembly from laboratory lines (Kim et al., 2021), or a diploid assembly workflow for single, wild-caught flies (see Methods for details). Most of the long-read assemblies are highly contiguous, complete, and accurate (**Figure 4**), in most cases exceeding the QV40 and 1Mb contig N50 minimum standards proposed by the Vertebrate Genomes Project (Rhie et al., 2021). We note that genome-wide sequence quality for R9.4.1 drosophilid hybrid assemblies usually does not exceed QV40 even though we have previously demonstrated coding sequences in R9.4.1 hybrid assemblies to be highly accurate (Kim et al. 2021). The full details

on the samples underlying these assemblies and the corresponding genome QC metrics are provided in **Table S4**.

The major factors limiting assembly quality were contamination and sample quality. Freshly collected samples had minimal issues during the library prep steps, but microbial (e.g., bacterial, nematode) sequences were abundant in wild-collected specimens (particularly in *quinaria* group), reducing on-target read coverage of both Illumina and Nanopore reads and thus genome contiguity and accuracy. For some of these samples, we made several attempts at a genome assembly and presented the best one here. Older ethanol-preserved specimens were also challenging to prepare and assemble. Heavily fragmented degraded gDNA (e.g., *Zaprionus obscuricornis* and *Phortica magna*) limited our ability to remove shorter gDNA fragments with size selection buffers during the ONT library prep. The co-purification of unknown contaminants (frequently present in some of the older Hawaiian picture-wing flies) required us to perform additional sample purification steps that led to significant sample loss. All these factors negatively impacted read throughput and consensus quality.

In the worst-case scenario, where sample limitations allowed us to only generate Illumina data, a simple draft assembly was generated from paired-end reads with only a contaminant removal post-processing step. While the utility of the latter type of assembly for population or comparative genomics is limited, the sequences are useful for phylogenetic inference (Dylus et al., 2023; F. Zhang et al., 2019).

#### *Comparative resources based on whole-genome data*

We present a number of additional resources to assist users of these data with downstream comparative genomic analyses. First, we provide a curated (as of August 2023) list of representative whole genome datasets of 360 drosophilid species and 4 outgroups (1 agromyzid, 2 ephydrids, and 1 diastatid) in **Table S1**. This table provides a convenient reference for readers to keep track of the genomes incorporated into the resources we present. This can be especially challenging in such a rapidly changing field. Updated versions of this list will be presented in future work.

We inferred species relationships of these 364 genomes using 1,000 dipteran BUSCO genes (Kriventseva et al., 2019; Manni et al., 2021) identified as complete and single copy across the most genomes (**Figure S1**). Our whole-genome approach significantly improves the resolution of the phylogeny (**Figure S1**), especially for deep-branching nodes that could not be resolved with smaller datasets (Finet et al., 2021). Interestingly, we still observe uncertainty at some nodes in our phylogeny, possibly due to incomplete lineage sorting or introgression (Suvorov et al., 2022). A more detailed discussion of the taxonomic implications of whole-genome sequencing will be reserved for a forthcoming study. Lastly, the phylogeny was scaled by the substitution rate at 4-fold degenerate sites in the BUSCO genes to provide a guide tree for whole-genome alignment.

A Progressive Cactus (Armstrong et al., 2020) reference-free whole-genome alignment was computed for a subset of 298 genomes (**Table S5**) out of the 364 species listed in **Table S1**. The Cactus alignments and HAL file format utilities (Hickey et al., 2013) are part of a large suite of comparative genomics tools that allow users to quickly perform liftover of genomic features, comparative annotation (Fiddes et al., 2018), compute evolutionary rate scores (Pollard et al., 2010), and more. To minimize issues from low-quality assemblies, genomes for

the alignment were selected based on minimum contiguity (N50 > 20kb) and minimum completeness (BUSCO >95%) filters. For species with multiple assemblies present we deferred to the NCBI representative genome unless our new genome was a major improvement. Alignment coverage of *D. melanogaster* genomic elements in other species is depicted in **Figure 5**, showing that most of the protein-coding genome aligns across Drosophilidae. The inferred ancestral drosophilid genome is 33.5 Mbp in size, about 10 Mbp larger than the sum of *D. melanogaster* coding sequences, and contains 97.3% complete BUSCO genes. Moreover, the estimated average substitution rate at 4-fold degenerate sites is 44.8 substitutions/site (**Figure S1**), suggesting a complete saturation of substitutions at every neutrally evolving site in the alignable genome and base-level resolution of comparative genomics approaches for evolutionary rate estimation (e.g., Christmas et al., 2023). Together, these results demonstrate the immense potential for comparative genomics provided by this dataset even across deep evolutionary timescales. We note that fitting branch lengths with 4-fold sites alone may mis-estimate deeper branch lengths due to substitution saturation, and will reexamine this problem with more complex codon substitution models in forthcoming work on the Drosophilidae Tree of Life.

### *Remaining challenges*

As another major step towards clade-scale genomics of the family Drosophilidae, we have demonstrated that assembling genomes from individual flies, even those preserved in ethanol for up to two decades, is now feasible. We are still far from sequencing every, or even most, species in the family. We are faced with several immediate challenges as this resource grows, and it is important for users to be aware of them.

There is significant variation in genome quality and completeness despite the generally optimistic genome quality metrics we have reported. Read lengths for single-fly library preps are usually around an order of magnitude shorter than those from inbred lines due to sample material limitations. This difference is particularly egregious for the most fragmented gDNA extractions from the oldest ethanol samples, which are close to the lower limits of what is acceptable for ONT sequencing and assembly (1,000 bp read N50). The reconstruction of repetitive sequences and other complex genomic regions will be severely limited from such short sequences, although these data are still superior to paired-end short reads for genome assembly. We urge that comparative genomic analyses, particularly those of structural variation, carefully consider these possibilities and do not readily consider absence of sequence as evidence of absence.

As this resource grows beyond the most commonly studied drosophilids, ambiguity in species identification and taxonomy may become significant issues. All wild-collected samples are, to the best of our ability, identified through key morphological characters prior to sequencing (e.g., Werner et al., 2018, 2020). It is nevertheless inevitable that we will misidentify or sequence ambiguous specimens, especially ones that belong to rare and/or cryptic species (e.g., *D. colorata*) or those with historically inconsistent taxonomic placement (e.g., *D. flavopinicola*). In several cases, we have discovered through sequencing that even laboratory/stock center lines were misidentified or contaminated. We try to minimize these issues by checking genome assemblies against known marker sequences on the Barcode of Life (Ratnasingham & Hebert, 2007) and NCBI Nucleotide databases (Methods), but even then



there are many species, some not formally described, for which basic genetic markers are not readily available. It is possible for these markers to come from incorrectly identified samples too. Whole-genome sequencing approaches applied alongside comprehensive field collections will help address the majority of these taxonomic and data quality problems. We will continue to update and correct records as new datasets are generated or if errors in identification are detected.

### *Next steps*

In creating these resources, we aim to build a powerful open resource for drosophilid evolutionary genomics and ultimately a framework for connecting micro- to macro-evolutionary processes in Drosophilidae and beyond. This study is just one part of an ongoing set of community-level projects developing this genomic resource. We will briefly describe ongoing efforts to inform readers about additional new resources that will be available in the near future.

In addition to continuously assembling new genomes, we are working on improvements to existing genomes. We are generating new R10.4.1 data with the specific intent of improving consensus accuracy for existing R9.4.1 assemblies along with new transcriptomic data for gene annotation. For species with lines readily available from the National *Drosophila* Species Stock Center center, heads and bodies from pools of adult males and females (4 libraries per species) will be sequenced. For wild-collected specimens, libraries will be prepared from single whole adults. Full-length transcripts from select species will be sequenced with long-read approaches. Importantly, we will submit sequences for annotation by NCBI RefSeq to maintain consistency in gene annotation pipelines (Weisman et al., 2022).

A well-resolved phylogeny is crucial for comparative analyses, but there are many open phylogenetic and taxonomic issues to be resolved in Drosophilidae. While it is clear that whole-genome sequencing is the best way to infer a reliable phylogeny, there are legacy Sanger datasets for hundreds more species from many past molecular phylogenetic studies of this group (Finet et al., 2021; O'Grady & DeSalle, 2018) that remain useful for inference of the Drosophilidae Tree of Life. The whole-genome data we are generating will provide both a comprehensive topology for a broader phylogenetic analysis that includes classic marker data, and a way to connect disparate markers from past studies using complete genomes.

Until now, our sampling efforts have been focused on readily available specimens and easily accessible populations of drosophilids. There are still large geographical gaps in the data, particularly in Africa, South America, and Australia – all regions that host significant fractions of global drosophilid biodiversity. Efforts to collect in these regions will be especially fruitful if in partnership with local field biologists – collaborations that we are making efforts to establish now. Even among new groups sampled here, such as the Hawaiian drosophilids, there are major lineages (e.g., the *Scaptomyza* subgenus *Titanochaeta*, and representatives of the *nudidrosophila*, *modified mouthpart*, and *rustica* species groups) that remain to be sampled. Museum specimens, like the ethanol ones presented here, but also dry-pinned specimens (Shpak et al., 2023), are proving to be viable material for drosophilid genomics, potentially even providing a valuable look into the past genomic diversity of many species.

Whereas drosophilid species, especially *D. melanogaster*, have played an important role in the development of population genetics, they have also served as early subjects of comparative population genetic studies (e.g. Langley et al., 2012; Machado et al., 2016; Ohta,

1993; Zhao & Begun, 2017), that is, the study of population genetic processes across species. The family Drosophilidae is uniquely well positioned as a system to accelerate the development of this growing field, and the tools provided in this manuscript provide a coherent framework for comparative population genomics at the scale of entire large clades. We are working on polymorphism data from wild populations for many (100s) of the genomes featured here, that will be presented in a forthcoming study of population genomic variation across the entire family. The relationship of population genomic variation with broad-scale macro-evolutionary patterns such as species diversity is still largely unknown (Rolland et al., 2023) and this will be a major step towards addressing this fundamental question in evolutionary biology.

#### *Reproducible workflows with Protocols.IO, Snakemake, and compute containers*

Reproducibility and transparency are top priorities for this open-science community dataset. Full, detailed protocols are provided through Protocols.io. Genome assembly from lab stocks (DOI: [dx.doi.org/10.17504/protocols.io.81wgbxbz1lpk/v1](https://doi.org/10.17504/protocols.io.81wgbxbz1lpk/v1)); genome assembly from single flies (DOI: [dx.doi.org/10.17504/protocols.io.ewov1q967gr2/v1](https://doi.org/10.17504/protocols.io.ewov1q967gr2/v1)). A list of laboratory reagents is provided in **Table S6**. Genome assembly pipelines are available from GitHub ([https://github.com/flyseq/2023\\_drosophila\\_assembly](https://github.com/flyseq/2023_drosophila_assembly)). This repository contains instructions for building compute containers that include all programs used for this manuscript as well as Snakemake (Köster & Rahmann, 2012) based workflows for running full computational pipelines.

#### *Data availability*

All new data generated for this work are deposited at NCBI SRA and GenBank under BioProject PRJNA1020440. The whole-genome alignment is archived at Dryad (DOI: [dx.doi.org/10.5061/dryad.x0k6djhrd](https://doi.org/10.5061/dryad.x0k6djhrd)). NCBI accessions and citations for public data used for this study are provided in **Table S1**. Raw Nanopore data (fast5, pod5) will be provided upon email request due to their large file sizes.

#### *CRedit author contributions*

Conceptualization: BYK, DAP

Methodology: BYK, HRG

Software: BYK

Validation: BYK, SHC, HRG

Formal analysis: BYK, HRG, SHC, AS

Investigation: BYK, HRG, SHC, AS, MBE, AK, DM, DJO, PMO, DKP, MJT, TW, DAP

Resources: All authors.

Data curation: BYK, HG, SHC, JH, MK, KNM, ZS, AK, DJO, PMO, DKP, MJT, TW

Writing - Original draft: BYK, HRG, SHC, AS, DAP.

Writing - Review & Editing: All authors.

Visualization: BYK, HRG

Supervision: BYK, DAP

Project administration: BYK, DAP

Funding acquisition: BYK, DAP, MBE, DM

### *Funding*

SGB was supported by an NSF GRFP. SHC was supported by NSF PRFB DBI 2109502. KNC was supported by NIGMS R35 GM137834. SED and RM were supported by The Expanding Horizons Initiative, Case Western Reserve University. SD, SP, and ZS were supported by NSF DEB-1811805. MBE was supported by the Howard Hughes Medical Institute. TF was supported by the Grant-in-Aid for JSPS Research Fellow (22J11897). JAH was supported by NHGRI 5T32HG000044-27. JH was supported by the Czech Ministry of Education, Youth and Sports grant ERC CZ LL2001. MK was supported by the Academy of Finland 322980. BYK was supported by NIGMS F32 GM135998. AK was supported by NIGMS R35 GM122592. TM was supported by the Grant-in-Aid for Research Activity Start-up 22K20565. DM was supported by NIGMS R35 GM148244. PMO was supported by NSF DEB2030129, DEB1839598, and DEB1241253. DJO was supported by BBSRC BB/T007516/1. DAP, BYK, HRG, SGB, JAH, and TY were supported by NIGMS R35 GM118165 and the Chan-Zuckerberg Biohub Investigator Program to DAP. AT was supported by JSPS KAKENHI JP23H02530. KHW was supported by NIGMS K99GM137041. TW was supported by NSF DOB/DEB-1737877 and Huron Mountain Wildlife Foundation (Michigan Tech Agreement #1802025).

### *Acknowledgements*

We thank Anna Mácová for assistance with shipping fly samples. Some of the computing for this project was performed on the Sherlock cluster. We would like to thank Stanford University and the Stanford Research Computing Center for providing computational resources and support that contributed to these research results.

## **Materials and Methods**

### *Fly collection*

Adult flies were collected for sequencing from laboratory cultures (63 species) or from field collections (101 species). Laboratory strains were grown on species-appropriate media. Mature adults were collected, sexed, and starved for 1 day before dry collection at -80 °C or into ethanol. Wild individuals were obtained from older ethanol collections or fresh through collection in the field. Field samples were obtained from banana, mushroom, and watermelon baits and by sweep netting, then stored in ethanol, RNA later, or RNA/DNA Shield (single samples) depending on sample purpose and transport method for the samples. Once in the lab, all samples were promptly stored at -20/-80 °C until sequencing. Collection details are listed in **Tables S1 and S2**. Older ethanol-fixed specimens were obtained from collections held by M.J. Toda and D.K. Price.

Specimens were collected and/or maintained under the following permits. To S.H. Church: (1) National Parks HAVO-SHC-2022-SCI-0022; (2) DOFAW Native Invertebrate Research Permit with NARS I5081; (3) Kaua'i Forest Reserves KPI-2022-289; (4) Kaua'i State Parks DSP-KPI-2022-288; (5) Hawai'i Forest Reserves Access Permit (no number issued). To J. Hrcek: (1) WITK16977516 and (2) PTU20-002501 from Queensland Government. To A. Kopp: (1) USDA APHIS P526P-22-06553 and (2) P526P-20-02787. To D. Matute: Permit granted by

the São Toméan and Zambian authorities. To M.J. Medeiros: (1) Hawai'i State Department of Land and Natural Resources I5297 and (2) Kaua'i Access Permit KPI-2022-298. To D. Obbard: land-owner permissions granted by email from Keith Obbard, Hugh Gibson, Effie Gibson, Sandy Bayne, and City of Edinburgh Council. To M.J. Toda: (1) Economic Planning Unit of Malaysian Government (research permissions UPE: 40/200/19 SJ.732 and UPE: 40/200/19 SJ.1194 and 1195) and (2) Ministry of Research and Technology of Indonesia (research permissions: 5816/SU/KS/2004, 6967/SU/KS/2004 and 416/SIP/FRP/SM/XI/2013). To T. Werner: (1) USDA Forest Service, Custer Gallatin National Forest (file code 2720); (2) State of Idaho IDFG Wildlife Bureau #764-22-000052; (3) United States Department of the Interior, National Park Service, Olympic #OLYM-2022-SCI-0033; (4) United States Department of the Interior, National Park Service, Olympic #OLYM-2023-SCI-0023; (5) USDA Forest Service, Umpqua National Forest (file code 2720). To T. Werner and B.Y. Kim: State of California, Natural Resources Agency, #23-635-014.

### *Genomic DNA extraction and library prep*

The genomic DNA extraction and library prep methods employed here follow the protocols described in Kim et al. 2021, with slight simplifications to streamline the process. Ethanol-fixed flies were rehydrated by 30 min of incubation in Buffer STE (400 mM NaCl 20 mM; Tris-HCl pH 8.0; 30 mM EDTA). Rehydrated or frozen flies were crushed with a plastic pestle and immediately mixed with warmed 200-500  $\mu$ L lysis buffer (0.1 M Tris-HCl pH 8.0; 0.1 M NaCl; 20 mM EDTA; 250  $\mu$ g/mL proteinase K; 0.6% SDS) and incubated at 50-60  $^{\circ}$ C for up to 4 hours. RNase A was added to a final concentration of  $\sim$ 190  $\mu$ g/mL (about 1  $\mu$ L of 20  $\mu$ g/mL stock per 100  $\mu$ L lysis buffer) 30 min before phenol-chloroform extraction. The lysate was transferred to a phase lock gel tube and subjected to two phenol-chloroform and one chloroform extraction. The aqueous layer was decanted to a new 1.5-mL tube and DNA precipitated by the addition of 10% 3 M NaOAc and 2-2.5 volumes of cold absolute ethanol. The DNA was pelleted by centrifugation, resuspended in 26  $\mu$ L 10 mM Tris pH 8.0 at 4  $^{\circ}$ C overnight or until homogeneously resuspended, then quantified by Nanodrop and Qubit. One to three aliquots of purified gDNA (a minimum of 20 ng unless sample limited) were reserved for Illumina library prep and the rest (range 30-2000 ng of gDNA) taken for Nanopore sequencing.

Nanopore libraries were prepared with the ONT ligation kit workflow (SQK-LSK110 for R9.4.1 flow cells and SQK-LSK114 for R10.4.1 flow cells), reducing reactions to half volumes from the ONT recommendations at all steps to reduce library prep cost. We do not observe reduced library yield or lower sequencing throughput from this reduction for drosophilid library preps. If starting with gDNA concentrations of  $>$ 50 ng/ $\mu$ L, we performed DNA fragment size selection with the PacBio SRE kit prior to starting the prep. The DNA repair and end-prep incubation steps were increased to 60 min at 25  $^{\circ}$ C and 30 min at 65  $^{\circ}$ C from the standard 5 min for each step. Elution steps during bead cleanups were extended to 30 min or longer (4-6 hours) on the heat block at 37  $^{\circ}$ C or until beads were homogeneously resuspended. The adapter ligation step was also increased to 30-60 min from the standard 5 min. Prepared library was sequenced on an R9.4.1 or R10.4.1 flow cell according to the manufacturer's instructions with fast live basecalling on.

For Illumina library preps, 5-10 ng DNA in 6  $\mu$ L were prepared with the Illumina DNA library preparation kit. We followed the standard workflow but reduced reaction volumes by 1/5

as a cost-saving measure. Libraries were amplified with 6 PCR cycles, except for the most sample-limited specimens, which were amplified with 7 PCR cycles (with more reads to compensate). Prepared libraries were individually cleaned up with beads and groups of samples with comparable genome sizes pooled by ng/ $\mu$ L concentration. Fragment size distributions were checked with BioAnalyzer, and the molar concentration of library pools were quantified by qPCR in-house, by the Stanford PAN facility, or by Admera Health. Sequencing was performed on NovaSeq S4 and NovaSeq X Plus machines at the Chan-Zuckerberg Biohub (San Francisco, CA, USA) and Admera Health (South Plainfield, NJ, USA).

A detailed step-by-step sequencing protocol is available at [Protocols.io](https://www.protocols.io) and as a spreadsheet in the Supplementary Materials.

### *Genome assembly pipeline*

ONT raw sequencing data were processed with standard tools. After initial sequencing, reads were basecalled with Guppy 6 and the appropriate (for R9.4.1 and R10.4.1) super-accuracy basecalling model with the adapter trimming and read splitting options on. Adapter trimming was performed on Illumina data with BBtools (Bushnell, 2022).

Two separate genome assembly pipelines were executed for inbred lines and wild-collected flies. Inbred lines were assembled with a haploid assembly pipeline. ONT reads were assembled with Flye (Kolmogorov et al., 2019) and haplotigs were identified and removed from the draft assembly with one round of `purge_dups` (Guan et al., 2020). The assembly was then polished using Nanopore data with one round of Medaka and further polished using Illumina data with one round of Pilon (Walker et al., 2014) with only base-level corrections enabled for the latter. Finally, contaminant sequences were flagged and removed with NCBI Foreign Contamination Screen (FCS, Astashyn et al., 2023). Single wild-caught flies were assembled with a diploid assembly pipeline. Haplotype-aware correction was applied to ONT reads using Illumina data with Ratanosk (Holley et al., 2021), corrected reads were assembled with Flye, contaminant sequences flagged and removed with NCBI FCS, haplotigs identified and removed by one round of `purge_dups`, and a phased dual assembly was generated with Hapdup (Kolmogorov et al., 2023). Repetitive sequences were identified in both haploid and the primary diploid assemblies with RepeatModeler2 (Flynn et al., 2020). Sequences were soft-masked with RepeatMasker (Smit et al., 2013), using the genome-specific repeat library. The *D. miranda* genome was scaffolded into chromosomes with RagTag (Alonge et al., 2022), using the current NCBI RefSeq genome (Mahajan et al., 2018) as a reference. In this specific case, we ignored the NCBI RefSeq genome due to extensive BUSCO duplication in the assembly.

Genome quality metrics were assessed with a number of tools. Genome contiguity statistics were computed with GenomeTools (Gremme et al., 2013). Completeness of protein-coding genes was assessed with BUSCO v5 (Manni et al., 2021) and the OrthoDB diptera\_odb10 set (Kriventseva et al., 2019). If Illumina reads were available, genome accuracy (QV) was assessed with the reference-free, k-mer based approach implemented in Yak (Cheng et al., 2021). If only Nanopore reads were available, consensus QV was assessed with a mapping and variant calling approach (Kim et al., 2021; Solares et al., 2018). Nanopore reads were mapped to the draft assembly with minimap2 (Li, 2016) and variant calling was performed with PEPPER-Margin-DeepVariant (Shafin et al., 2021). Homozygous non-reference genotypes

were considered to be errors and QV was computed as  $QV = -10 \cdot \log_{10}(\text{number of homozygous derived} / \text{number of callable sites})$ .

### *Short-read assembly*

Genomes were assembled from short-read only datasets with minimal additional processing. Paired-end reads were assembled with SPAdes (Bankevich et al., 2012). Similar to long-read assemblies, contaminant sequences were identified and removed with NCBI FCS.

### *Additional quality control*

We performed further quality control steps in addition to the standard assembly workflow to minimize issues related to incorrectly sexed flies and species misidentification.

To verify fly sex, we leveraged the observation that drosophilid genes rarely translocate across Muller elements (Kim et al., 2021; Sturtevant & Novitski, 1941) and reasoned that single copy BUSCO genes on the X chromosome (typically Muller A in drosophilids and in *D. melanogaster*) should have half read coverage of autosomes in sequencing data of males and equal coverage in female reads. We note that sex chromosome transitions are fairly common in dipterans (Vicoso & Bachtrog, 2015), but these transitions typically involve chromosome fusions in drosophilids and are unlikely to completely mislead coverage-based sex verification as a heuristic approach. We created a list of the X-linked BUSCO orthologs on the *D. melanogaster* X chromosome, then for each test genome, examined the depth of ONT read coverage over the same orthologs identified as single-copy and complete in our other assemblies.

To verify species identities, we checked each genome assembly against known marker sequences on the NCBI Nucleotide database and COI sequences on the Barcode of Life Database (Ratnasingham & Hebert, 2007). Because `purge_dups` flags contigs in the assembly with abnormal levels of coverage, we performed species verification on the unprocessed Flye draft assemblies. COI and other marker sequences were identified in each genome by a BLAST search with the *D. melanogaster* COI as the query sequence, then extracted sequences were provided to the NCBI BLAST and BOLD web interfaces to search for close matches. We considered markers with 98% sequence identity as positive identifications, but recommend caution here as it is possible that public databases may also contain sequences from misidentified species. Finally, individual species' phylogenetic positions were checked against a large published phylogeny (Finet et al., 2021). We will continue to correct any additional issues we identify in the future.

### *Downloading additional genomes*

A representative genome for every sequenced drosophilid species was obtained from NCBI, giving priority to genomes designated as representative genomes by NCBI. Sample information and NCBI RefSeq/GenBank/SRA accessions are provided in **Table S1**.

### *Species tree inference from BUSCO orthologs*

A species tree was inferred from up to 1,000 single-copy orthologs identified in each genome following our previous workflow (Kim et al., 2021; Suvorov et al., 2022). For each genome, we ran BUSCO v5 (Manni et al., 2021) using the `diptera_odb10` dataset (Kriventseva et al., 2019) with Augustus gene prediction (Stanke et al., 2008) on. Coding and amino acid

sequences for the 1,000 most complete single-copy genes across all genomes were collated into single FASTA files. Coding and amino acid sequences were aligned with MAFFT (Katoch & Standley, 2013) and codon alignments were generated by back-translation with PAL2NAL (Suyama et al., 2006). Individual gene trees were generated from DNA alignments with IQTREE2 (Minh et al., 2020) with automatic model selection. A species tree was inferred from the gene trees with ASTRAL (C. Zhang et al., 2018). Four-fold degenerate codons in the codon alignments were identified using the *D. melanogaster* BUSCO annotations as the reference and 4-d codons and 4-d sites were extracted in sufficient statistics format using the msa\_view tool from the PHAST software package (Hubisz et al., 2011). The 1,000 genes were split into 30 subsets and branch lengths of the ASTRAL species tree were re-estimated with the phyloFit tool for each subset, then average branch lengths across all subsets were computed with the phyloBoot tool. Both tools are in the PHAST package. We note that lengths of deep branches in a phylogeny solely based on 4-fold degenerate sites may be mis-estimated due to substitution saturation, and we will address these issues with more complex substitution models in forthcoming work. Trees were plotted with the Interactive Tree of Life web tool at [itol.embl.de](http://itol.embl.de) (Letunic & Bork, 2021).

### *Progressive Cactus alignment*

A Progressive Cactus (Armstrong et al., 2020) whole-genome, reference-free alignment was built with a subset of 298 out of the 360 whole genomes listed in **Table S1**. The list of genomes used is provided in **Table S4**. Genomes were filtered based on contig N50 (minimum 20 kbp) and BUSCO completeness (minimum 95% complete single copy). One representative genome was used for each species. The species tree scaled by the 4-fold degenerate substitution rate was provided as a guide tree.

Alignment coverage was computed by extracting coding sequences, introns, and intergenic regions from the *D. melanogaster* dm6 genome annotations. Repetitive elements were identified by re-masking the dm6 genome with RepeatMasker set to produce GFF repeat annotation output. Regions were converted into BED file format, merged with BEDtools (Quinlan & Hall, 2010), and lifted to target genomes with the halLiftover tool.

## References

- Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., Scherer, S. E., Li, P. W., Hoskins, R. A., Galle, R. F., George, R. A., Lewis, S. E., Richards, S., Ashburner, M., Henderson, S. N., Sutton, G. G., Wortman, J. R., Yandell, M. D., Zhang, Q., ... Venter, J. C. (2000). The Genome Sequence of *Drosophila melanogaster*. *Science*, *287*(5461), 2185–2195. <https://doi.org/10.1126/science.287.5461.2185>
- Adams, M., McBroome, J., Maurer, N., Pepper-Tunick, E., Saremi, N. F., Green, R. E., Vollmers, C., & Corbett-Detig, R. B. (2020). One fly—one genome: Chromosome-scale genome assembly of a single outbred *Drosophila melanogaster*. *Nucleic Acids Research*, *48*(13), e75–e75. <https://doi.org/10.1093/nar/gkaa450>
- Alonge, M., Lebeigle, L., Kirsche, M., Jenike, K., Ou, S., Aganezov, S., Wang, X., Lippman, Z. B., Schatz, M. C., & Soyk, S. (2022). Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. *Genome Biology*, *23*(1), 258. <https://doi.org/10.1186/s13059-022-02823-7>
- Armstrong, J., Hickey, G., Diekhans, M., Fiddes, I. T., Novak, A. M., Deran, A., Fang, Q., Xie, D., Feng, S., Stiller, J., Genereux, D., Johnson, J., Marinescu, V. D., Alföldi, J., Harris, R. S., Lindblad-Toh, K., Haussler, D., Karlsson, E., Jarvis, E. D., ... Paten, B. (2020). Progressive Cactus is a multiple-genome aligner for the thousand-genome era. *Nature*, *587*(7833), Article 7833. <https://doi.org/10.1038/s41586-020-2871-y>
- Astashyn, A., Tvedte, E. S., Sweeney, D., Sapojnikov, V., Bouk, N., Joukov, V., Mozes, E., Strobe, P. K., Sylla, P. M., Wagner, L., Bidwell, S. L., Clark, K., Davis, E. W., Smith-White, B., Hlavina, W., Pruitt, K. D., Schneider, V. A., & Murphy, T. D. (2023). *Rapid and sensitive detection of genome contamination at scale with FCS-GX* (p. 2023.06.02.543519). bioRxiv. <https://doi.org/10.1101/2023.06.02.543519>
- Aury, J.-M., & Istace, B. (2021). Hapo-G, haplotype-aware polishing of genome assemblies with accurate reads. *NAR Genomics and Bioinformatics*, *3*(2), lqab034. <https://doi.org/10.1093/nargab/lqab034>
- Bächli, G. (2023, July). *TaxoDros*. <https://www.taxodros.uzh.ch/>
- Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. A., & Pevzner, P. A. (2012). SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *Journal of Computational Biology*, *19*(5), 455–477. <https://doi.org/10.1089/cmb.2012.0021>
- Bushnell, B. (2022, October 6). *BBMap*. <https://sourceforge.net/projects/bbmap/>
- Cheng, H., Concepcion, G. T., Feng, X., Zhang, H., & Li, H. (2021). Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature Methods*, *18*(2), Article 2. <https://doi.org/10.1038/s41592-020-01056-5>
- Christmas, M. J., Kaplow, I. M., Genereux, D. P., Dong, M. X., Hughes, G. M., Li, X., Sullivan, P. F., Hindle, A. G., Andrews, G., Armstrong, J. C., Bianchi, M., Breit, A. M., Diekhans, M., Fanter, C., Foley, N. M., Goodman, D. B., Goodman, L., Keough, K. C., Kirilenko, B., ... Karlsson, E. K. (2023). Evolutionary constraint and innovation across hundreds of placental mammals. *Science*, *380*(6643), eabn3943. <https://doi.org/10.1126/science.abn3943>



- Church, S. H., & Extavour, C. G. (2022). Phylotranscriptomics Reveals Discordance in the Phylogeny of Hawaiian *Drosophila* and *Scaptomyza* (Diptera: Drosophilidae). *Molecular Biology and Evolution*, 39(3), msac012. <https://doi.org/10.1093/molbev/msac012>
- Clark, A. G., Eisen, M. B., Smith, D. R., Bergman, C. M., Oliver, B., Markow, T. A., Kaufman, T. C., Kellis, M., Gelbart, W., Iyer, V. N., Pollard, D. A., Sackton, T. B., Larracuent, A. M., Singh, N. D., Abad, J. P., Abt, D. N., Adryan, B., Aguade, M., Akashi, H., ... \*Broad Institute Genome Sequencing Platform. (2007). Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature*, 450(7167), Article 7167. <https://doi.org/10.1038/nature06341>
- dos Santos, G., Schroeder, A. J., Goodman, J. L., Strelets, V. B., Crosby, M. A., Thurmond, J., Emmert, D. B., Gelbart, W. M., & the FlyBase Consortium. (2015). FlyBase: Introduction of the *Drosophila melanogaster* Release 6 reference genome assembly and large-scale migration of genome annotations. *Nucleic Acids Research*, 43(D1), D690–D697. <https://doi.org/10.1093/nar/gku1099>
- Dylus, D., Altenhoff, A., Majidian, S., Sedlazeck, F. J., & Dessimoz, C. (2023). Inference of phylogenetic trees directly from raw sequencing reads using Read2Tree. *Nature Biotechnology*, 1–9. <https://doi.org/10.1038/s41587-023-01753-4>
- Fiddes, I. T., Armstrong, J., Diekhans, M., Nachtweide, S., Kronenberg, Z. N., Underwood, J. G., Gordon, D., Earl, D., Keane, T., Eichler, E. E., Haussler, D., Stanke, M., & Paten, B. (2018). Comparative Annotation Toolkit (CAT)—Simultaneous clade and personal genome annotation. *Genome Research*, 28(7), 1029–1038. <https://doi.org/10.1101/gr.233460.117>
- Finet, C., Kassner, V. A., Carvalho, A. B., Chung, H., Day, J. P., Day, S., Delaney, E. K., De Ré, F. C., Dufour, H. D., Dupim, E., Izumitani, H. F., Gautério, T. B., Justen, J., Katoh, T., Kopp, A., Koshikawa, S., Longdon, B., Loreto, E. L., Nunes, M. D. S., ... Marlétaz, F. (2021). DrosoPhyla: Resources for Drosophilid Phylogeny and Systematics. *Genome Biology and Evolution*, 13(8), evab179. <https://doi.org/10.1093/gbe/evab179>
- Flynn, J. M., Hubley, R., Goubert, C., Rosen, J., Clark, A. G., Feschotte, C., & Smit, A. F. (2020). RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences*, 117(17), 9451–9457. <https://doi.org/10.1073/pnas.1921046117>
- Gremme, G., Steinbiss, S., & Kurtz, S. (2013). GenomeTools: A Comprehensive Software Library for Efficient Processing of Structured Genome Annotations. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 10(03), 645–656. <https://doi.org/10.1109/TCBB.2013.68>
- Guan, D., McCarthy, S. A., Wood, J., Howe, K., Wang, Y., & Durbin, R. (2020). Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics*, 36(9), 2896–2898. <https://doi.org/10.1093/bioinformatics/btaa025>
- Hickey, G., Paten, B., Earl, D., Zerbino, D., & Haussler, D. (2013). HAL: A hierarchical format for storing and analyzing multiple genome alignments. *Bioinformatics*, 29(10), 1341–1342. <https://doi.org/10.1093/bioinformatics/btt128>
- Holley, G., Beyter, D., Ingimundardottir, H., Møller, P. L., Kristmundsdottir, S., Eggertsson, H. P., & Halldorsson, B. V. (2021). Rataosk: Hybrid error correction of long reads enables accurate variant calling and assembly. *Genome Biology*, 22(1), 28. <https://doi.org/10.1186/s13059-020-02244-4>
- Hubisz, M. J., Pollard, K. S., & Siepel, A. (2011). PHAST and RPHAST: Phylogenetic analysis

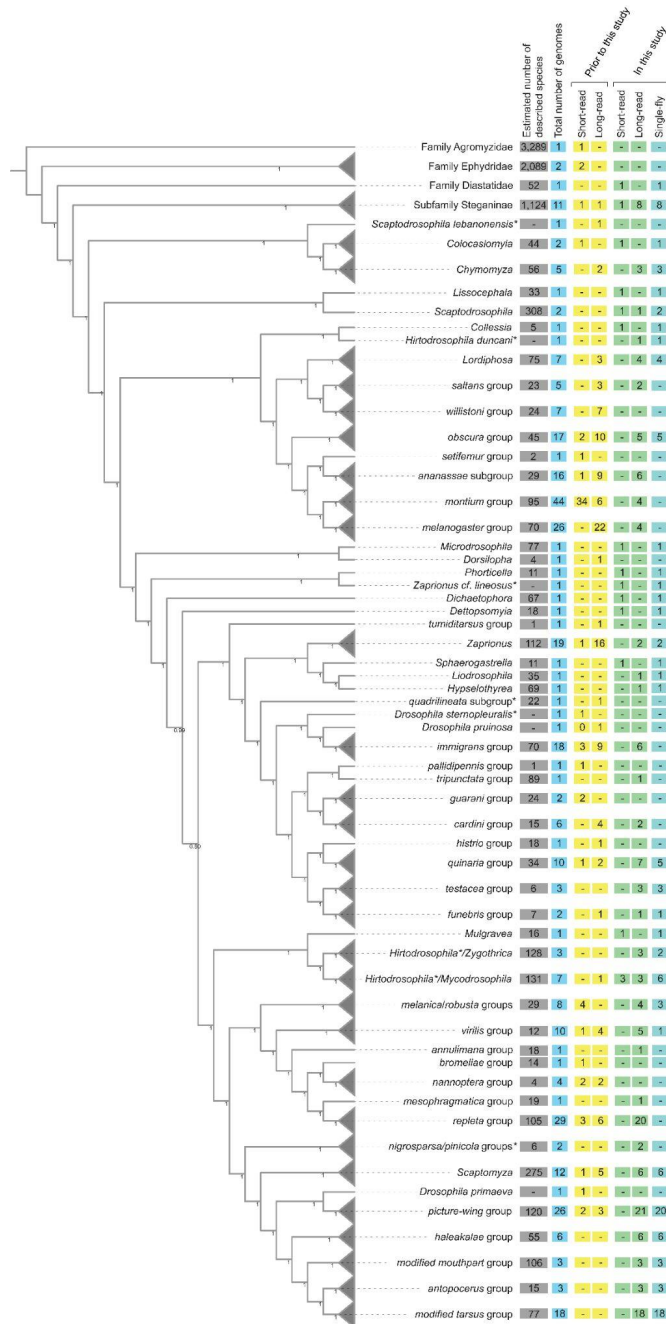
- with space/time models. *Briefings in Bioinformatics*, 12(1), 41–51.  
<https://doi.org/10.1093/bib/bbq072>
- Katoh, K., & Standley, D. M. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution*, 30(4), 772–780. <https://doi.org/10.1093/molbev/mst010>
- Kim, B. Y., Wang, J. R., Miller, D. E., Barmina, O., Delaney, E., Thompson, A., Comeault, A. A., Peede, D., D’Agostino, E. R., Pelaez, J., Aguilar, J. M., Haji, D., Matsunaga, T., Armstrong, E. E., Zych, M., Ogawa, Y., Stamenković-Radak, M., Jelić, M., Veselinović, M. S., ... Petrov, D. A. (2021). Highly contiguous assemblies of 101 drosophilid genomes. *eLife*, 10, e66405. <https://doi.org/10.7554/eLife.66405>
- Kingan, S. B., Heaton, H., Cudini, J., Lambert, C. C., Baybayan, P., Galvin, B. D., Durbin, R., Korlach, J., & Lawnczak, M. K. N. (2019). A High-Quality De novo Genome Assembly from a Single Mosquito Using PacBio Sequencing. *Genes*, 10(1), Article 1.  
<https://doi.org/10.3390/genes10010062>
- Kolmogorov, M., Billingsley, K. J., Mastoras, M., Meredith, M., Monlong, J., Lorig-Roach, R., Asri, M., Jerez, P. A., Malik, L., Dewan, R., Reed, X., Genner, R. M., Daida, K., Behera, S., Shafin, K., Pesout, T., Prabakaran, J., Carnevali, P., Consortium (NABEC), N. A. B. E., ... Paten, B. (2023). Scalable Nanopore sequencing of human genomes provides a comprehensive view of haplotype-resolved variation and methylation (p. 2023.01.12.523790). bioRxiv. <https://doi.org/10.1101/2023.01.12.523790>
- Kolmogorov, M., Yuan, J., Lin, Y., & Pevzner, P. A. (2019). Assembly of long, error-prone reads using repeat graphs. *Nature Biotechnology*, 37(5), Article 5.  
<https://doi.org/10.1038/s41587-019-0072-8>
- Köster, J., & Rahmann, S. (2012). Snakemake—A scalable bioinformatics workflow engine. *Bioinformatics*, 28(19), 2520–2522. <https://doi.org/10.1093/bioinformatics/bts480>
- Kriventseva, E. V., Kuznetsov, D., Tegenfeldt, F., Manni, M., Dias, R., Simão, F. A., & Zdobnov, E. M. (2019). OrthoDB v10: Sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Research*, 47(D1), D807–D811. <https://doi.org/10.1093/nar/gky1053>
- Langley, C. H., Stevens, K., Cardeno, C., Lee, Y. C. G., Schrider, D. R., Pool, J. E., Langley, S. A., Suarez, C., Corbett-Detig, R. B., Kolaczkowski, B., Fang, S., Nista, P. M., Holloway, A. K., Kern, A. D., Dewey, C. N., Song, Y. S., Hahn, M. W., & Begun, D. J. (2012). Genomic Variation in Natural Populations of *Drosophila melanogaster*. *Genetics*, 192(2), 533–598. <https://doi.org/10.1534/genetics.112.142018>
- Letunic, I., & Bork, P. (2021). Interactive Tree Of Life (iTOL) v5: An online tool for phylogenetic tree display and annotation. *Nucleic Acids Research*, 49(W1), W293–W296.  
<https://doi.org/10.1093/nar/gkab301>
- Li, H. (2016). Minimap and miniasm: Fast mapping and de novo assembly for noisy long sequences. *Bioinformatics*, 32(14), 2103–2110.  
<https://doi.org/10.1093/bioinformatics/btw152>
- Li, H., Janssens, J., De Waegeneer, M., Kolluru, S. S., Davie, K., Gardeux, V., Saelens, W., David, F. P. A., Brbić, M., Spanier, K., Leskovec, J., McLaughlin, C. N., Xie, Q., Jones, R. C., Brueckner, K., Shim, J., Tattikota, S. G., Schnorrer, F., Rust, K., ... Aerts, S. (2022). Fly Cell Atlas: A single-nucleus transcriptomic atlas of the adult fruit fly. *Science*, 375(6584), eabk2432. <https://doi.org/10.1126/science.abk2432>

- Machado, H. E., Bergland, A. O., O'Brien, K. R., Behrman, E. L., Schmidt, P. S., & Petrov, D. A. (2016). Comparative population genomics of latitudinal variation in *Drosophila simulans* and *Drosophila melanogaster*. *Molecular Ecology*, 25(3), 723–740. <https://doi.org/10.1111/mec.13446>
- Mackay, T. F. C., Richards, S., Stone, E. A., Barbadilla, A., Ayroles, J. F., Zhu, D., Casillas, S., Han, Y., Magwire, M. M., Cridland, J. M., Richardson, M. F., Anholt, R. R. H., Barrón, M., Bess, C., Blankenburg, K. P., Carbone, M. A., Castellano, D., Chaboub, L., Duncan, L., ... Gibbs, R. A. (2012). The *Drosophila melanogaster* Genetic Reference Panel. *Nature*, 482(7384), Article 7384. <https://doi.org/10.1038/nature10811>
- Magnacca, K., & Price, D. (2012). New species of Hawaiian picture wing *Drosophila* (Diptera: Drosophilidae), with a key to species. *Zootaxa*, 3188, 1–30. <https://doi.org/10.11646/zootaxa.3188.1.1>
- Mahajan, S., Wei, K. H.-C., Nalley, M. J., Gibilisco, L., & Bachtrog, D. (2018). De novo assembly of a young *Drosophila* Y chromosome using single-molecule sequencing and chromatin conformation capture. *PLOS Biology*, 16(7), e2006348. <https://doi.org/10.1371/journal.pbio.2006348>
- Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A., & Zdobnov, E. M. (2021). BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Molecular Biology and Evolution*, 38(10), 4647–4654. <https://doi.org/10.1093/molbev/msab199>
- Miller, D. E., Staber, C., Zeitlinger, J., & Hawley, R. S. (2018). Highly Contiguous Genome Assemblies of 15 *Drosophila* Species Generated Using Nanopore Sequencing. *G3: Genes, Genomes, Genetics*, 8(10), 3131–3141. <https://doi.org/10.1534/g3.118.200160>
- Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A., & Lanfear, R. (2020). IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution*, 37(5), 1530–1534. <https://doi.org/10.1093/molbev/msaa015>
- modENCODE Consortium, T., Roy, S., Ernst, J., Kharchenko, P. V., Kheradpour, P., Negre, N., Eaton, M. L., Landolin, J. M., Bristow, C. A., Ma, L., Lin, M. F., Washietl, S., Arshinoff, B. I., Ay, F., Meyer, P. E., Robine, N., Washington, N. L., Stefano, L. D., Berezhikov, E., ... Kellis, M. (2010). Identification of Functional Elements and Regulatory Circuits by *Drosophila* modENCODE. *Science*, 330(6012), 1787–1797. <https://doi.org/10.1126/science.1198374>
- Obbard, D. J., Wellcome Sanger Institute Tree of Life programme, Wellcome Sanger Institute Scientific Operations: DNA Pipelines collective, Tree of Life Core Informatics collective, & Darwin Tree of Life Consortium. (2023). *The genome sequence of a drosophilid fruit fly, *Hirtodrosophila cameraria* (Haliday, 1833)*. <https://wellcomeopenresearch.org/articles/8-361>
- O'Grady, P. M., & DeSalle, R. (2018). Phylogeny of the Genus *Drosophila*. *Genetics*, 209(1), 1–25. <https://doi.org/10.1534/genetics.117.300583>
- O'Grady, P., Magnacca, K., & Lapoint, R. (2010). Taxonomic relationships within the endemic Hawaiian Drosophilidae (Insecta: Diptera). *Records of the Hawaii Biological Survey*, 108, 1–34.
- Ohta, T. (1993). Amino acid substitution at the Adh locus of *Drosophila* is facilitated by small population size. *Proceedings of the National Academy of Sciences*, 90(10), 4548–4551. <https://doi.org/10.1073/pnas.90.10.4548>

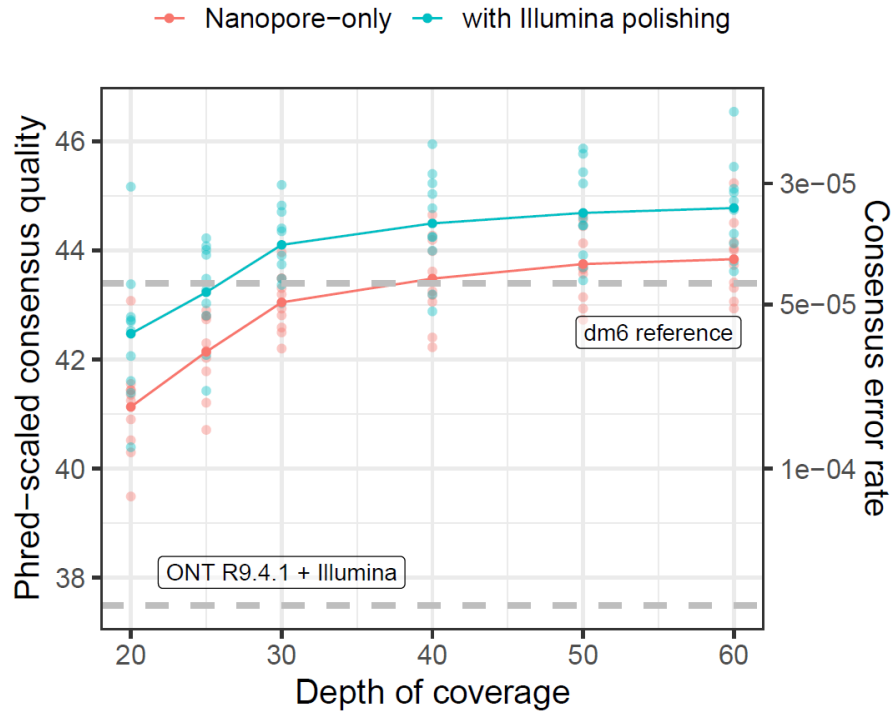
- Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R., & Siepel, A. (2010). Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Research*, *20*(1), 110–121. <https://doi.org/10.1101/gr.097857.109>
- Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*(6), 841–842. <https://doi.org/10.1093/bioinformatics/btq033>
- Ratnasingham, S., & Hebert, P. D. N. (2007). bold: The Barcode of Life Data System (<http://www.barcodinglife.org>). *Molecular Ecology Notes*, *7*(3), 355–364. <https://doi.org/10.1111/j.1471-8286.2007.01678.x>
- Rhie, A., McCarthy, S. A., Fedrigo, O., Damas, J., Formenti, G., Koren, S., Uliano-Silva, M., Chow, W., Functamman, A., Kim, J., Lee, C., Ko, B. J., Chaisson, M., Gedman, G. L., Cantin, L. J., Thibaud-Nissen, F., Haggerty, L., Bista, I., Smith, M., ... Jarvis, E. D. (2021). Towards complete and error-free genome assemblies of all vertebrate species. *Nature*, *592*(7856), Article 7856. <https://doi.org/10.1038/s41586-021-03451-0>
- Rhie, A., Walenz, B. P., Koren, S., & Phillippy, A. M. (2020). Merqury: Reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biology*, *21*(1), 245. <https://doi.org/10.1186/s13059-020-02134-9>
- Richards, S., Liu, Y., Bettencourt, B. R., Hradecky, P., Letovsky, S., Nielsen, R., Thornton, K., Hubisz, M. J., Chen, R., Meisel, R. P., Couronne, O., Hua, S., Smith, M. A., Zhang, P., Liu, J., Bussemaker, H. J., Batenburg, M. F. van, Howells, S. L., Scherer, S. E., ... Gibbs, R. A. (2005). Comparative genome sequencing of *Drosophila pseudoobscura*: Chromosomal, gene, and cis-element evolution. *Genome Research*, *15*(1), 1–18. <https://doi.org/10.1101/gr.3059305>
- Rolland, J., Henao-Diaz, L. F., Doebeli, M., Germain, R., Harmon, L. J., Knowles, L. L., Liow, L. H., Mank, J. E., Machac, A., Otto, S. P., Pennell, M., Salamin, N., Silvestro, D., Sugawara, M., Uyeda, J., Wagner, C. E., & Schluter, D. (2023). Conceptual and empirical bridges between micro- and macroevolution. *Nature Ecology & Evolution*, *7*(8), Article 8. <https://doi.org/10.1038/s41559-023-02116-7>
- Shafin, K., Pesout, T., Chang, P.-C., Nattestad, M., Kolesnikov, A., Goel, S., Baid, G., Kolmogorov, M., Eizenga, J. M., Miga, K. H., Carnevali, P., Jain, M., Carroll, A., & Paten, B. (2021). Haplotype-aware variant calling with PEPPER-Margin-DeepVariant enables high accuracy in nanopore long-reads. *Nature Methods*, *18*(11), Article 11. <https://doi.org/10.1038/s41592-021-01299-w>
- Shpak, M., Ghanavi, H. R., Lange, J. D., Pool, J. E., & Stensmyr, M. C. (2023). *Genomes from 25 historical Drosophila melanogaster specimens illuminate adaptive and demographic changes across more than 200 years of evolution* (p. 2023.04.24.538033). bioRxiv. <https://doi.org/10.1101/2023.04.24.538033>
- Shumate, A., & Salzberg, S. L. (2021). Liftoff: Accurate mapping of gene annotations. *Bioinformatics*, *btaa1016*. <https://doi.org/10.1093/bioinformatics/btaa1016>
- Smit, A. F. A., Hubble, R., & Green, P. (2013, 2015). *RepeatMasker Open-4.0*. RepeatMasker Open-4.0.
- Solares, E. A., Chakraborty, M., Miller, D. E., Kalsow, S., Hall, K., Perera, A. G., Emerson, J. J., & Hawley, R. S. (2018). Rapid Low-Cost Assembly of the *Drosophila melanogaster* Reference Genome Using Low-Coverage, Long-Read Sequencing. *Genes|Genomes|Genetics*, *8*(10), 3143–3154. <https://doi.org/10.1534/g3.118.200162>
- Spieth, H. T., & Heed, W. B. (1975). The *Drosophila pinicola* species group. (Diptera:

- Drosophilidae). *Pan-Pacific Entomologist*, 51(4), 287–295.
- Stanke, M., Diekhans, M., Baertsch, R., & Haussler, D. (2008). Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics*, 24(5), 637–644. <https://doi.org/10.1093/bioinformatics/btn013>
- Sturtevant, A. H., & Novitski, E. (1941). The Homologies of the Chromosome Elements in the Genus *Drosophila*. *Genetics*, 26(5), 517–541.
- Suvorov, A., Kim, B. Y., Wang, J., Armstrong, E. E., Peede, D., D’Agostino, E. R. R., Price, D. K., Waddell, P. J., Lang, M., Courtier-Orgogozo, V., David, J. R., Petrov, D., Matute, D. R., Schrider, D. R., & Comeault, A. A. (2022). Widespread introgression across a phylogeny of 155 *Drosophila* genomes. *Current Biology*, 32(1), 111-123.e5. <https://doi.org/10.1016/j.cub.2021.10.052>
- Suyama, M., Torrents, D., & Bork, P. (2006). PAL2NAL: Robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Research*, 34(suppl\_2), W609–W612. <https://doi.org/10.1093/nar/gkl315>
- Vicoso, B., & Bachtrog, D. (2015). Numerous Transitions of Sex Chromosomes in Diptera. *PLOS Biology*, 13(4), e1002078. <https://doi.org/10.1371/journal.pbio.1002078>
- Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C. A., Zeng, Q., Wortman, J., Young, S. K., & Earl, A. M. (2014). Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. *PLOS ONE*, 9(11), e112963. <https://doi.org/10.1371/journal.pone.0112963>
- Weisman, C. M., Murray, A. W., & Eddy, S. R. (2022). Mixing genome annotation methods in a comparative analysis inflates the apparent number of lineage-specific genes. *Current Biology*, 32(12), 2632-2639.e2. <https://doi.org/10.1016/j.cub.2022.04.085>
- Werner, T., Steenwinkel, T., & Jaenike, J. (2018). The Encyclopedia of North American Drosophilids Volume 1: Drosophilids of the Midwest and Northeast. *Open Access Books*. <https://digitalcommons.mtu.edu/oabooks/1>
- Werner, T., Steenwinkel, T., & Jaenike, J. (2020). The Encyclopedia of North American Drosophilids Volume 2: Drosophilids of the Southeast. *Open Access Books*. <https://digitalcommons.mtu.edu/oabooks/3>
- Yin, J., Zhang, C., & Mirarab, S. (2019). ASTRAL-MP: Scaling ASTRAL to very large datasets using randomization and parallelization. *Bioinformatics*, 35(20), 3961–3969. <https://doi.org/10.1093/bioinformatics/btz211>
- Zhang, C., Rabiee, M., Sayyari, E., & Mirarab, S. (2018). ASTRAL-III: Polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics*, 19(6), 153. <https://doi.org/10.1186/s12859-018-2129-y>
- Zhang, F., Ding, Y., Zhu, C.-D., Zhou, X., Orr, M. C., Scheu, S., & Luan, Y.-X. (2019). Phylogenomics from low-coverage whole-genome sequencing. *Methods in Ecology and Evolution*, 10(4), 507–517. <https://doi.org/10.1111/2041-210X.13145>
- Zhao, L., & Begun, D. J. (2017). Genomics of parallel adaptation at two timescales in *Drosophila*. *PLOS Genetics*, 13(10), e1007016. <https://doi.org/10.1371/journal.pgen.1007016>

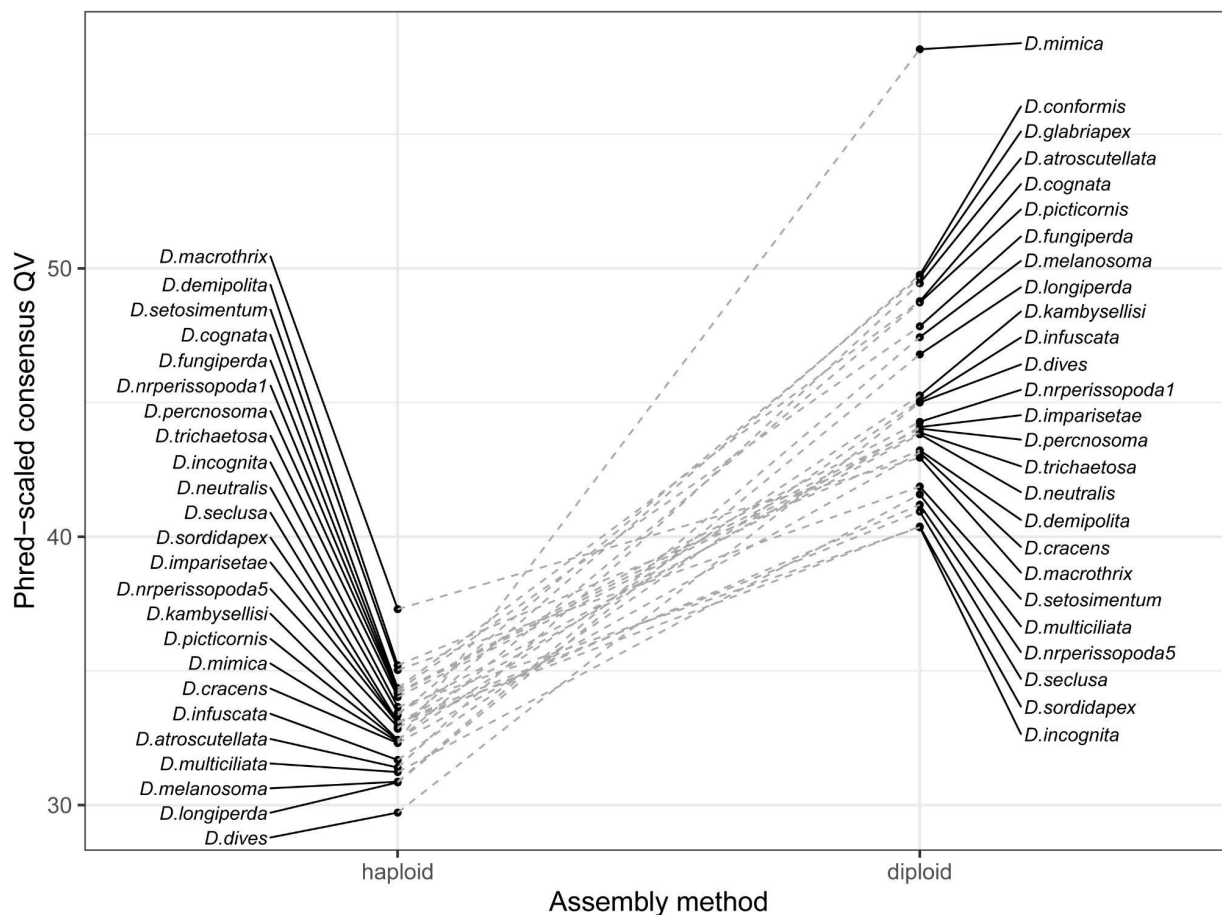
## Figures



**Figure 1.** Cladogram of drosophilid species with whole-genome data, with some groups collapsed (gray triangles). Species relationships were inferred from 1,000 orthologs (see Methods). Node values are the local posterior probabilities reported by ASTRAL-MP (Yin et al., 2019). Values in the boxes indicate, as of August 2023, the number of species with whole-genome sequences for each taxon. The count of short-read and long-read datasets, and data available before this study (including Kim et al., 2021) and new genomes presented here are shown separately. \*Note that *Scaptodrosophila*, *Hirtodrosophila*, *Zaprionus*, *immigrans*, and *histro* groups are potentially rendered polyphyletic by these samples. The positions of *nigrosparsa* and *pinicola* groups are currently considered to be uncertain.

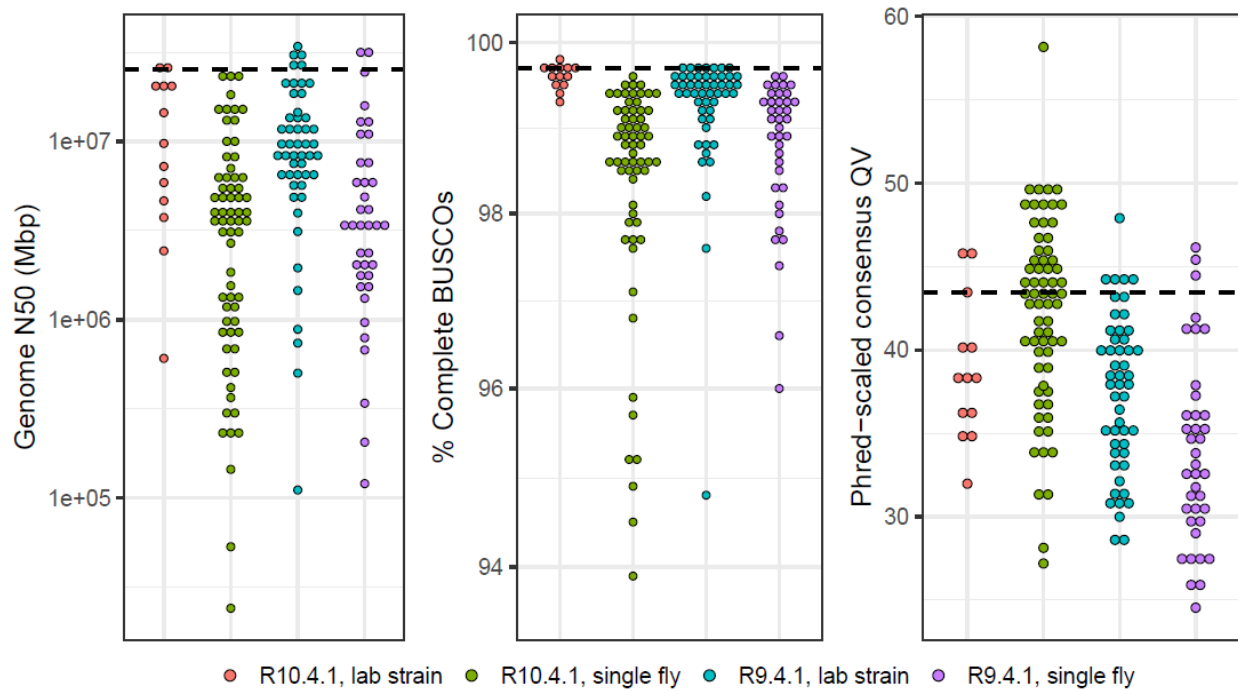


**Figure 2.** High genome-wide consensus accuracy with Nanopore R10.4.1 sequencing. The Phred-scaled consensus accuracy (left axis) and per-base consensus error rate (rate) are shown for genomes built with 20× to 60× coverage of ONT reads. Dashed gray lines show consensus accuracy estimates for R9.4.1 + Illumina (Kim et al. 2021) and the dm6 reference genome (dos Santos et al., 2015).

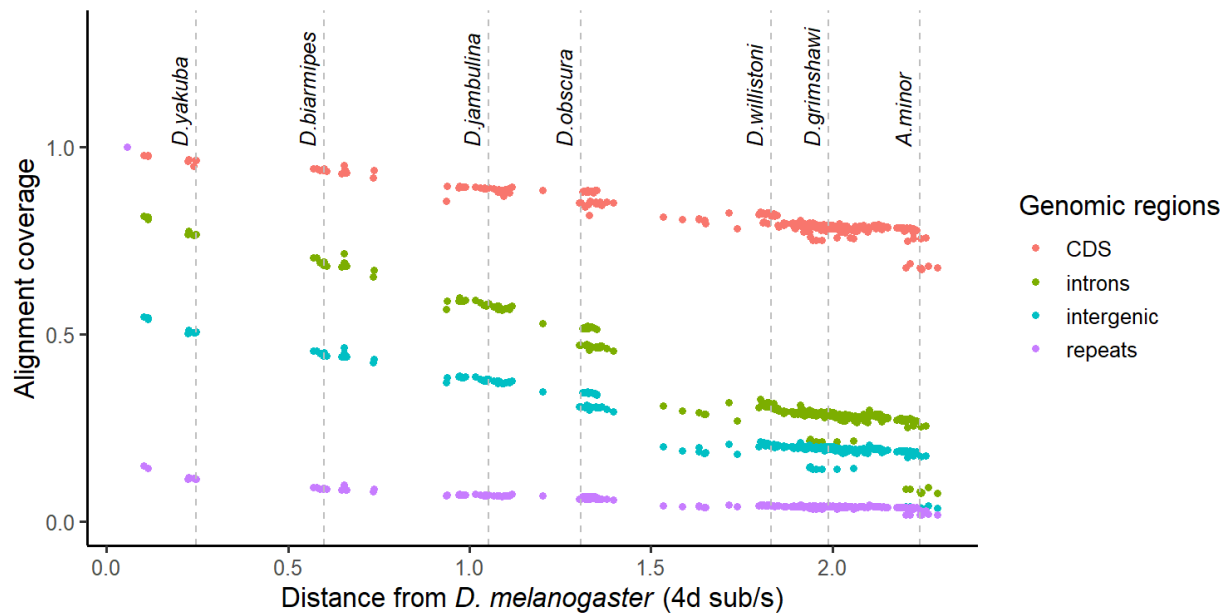


**Figure 3.** Consensus accuracy for single-fly genomes is greatly improved by diploid assembly. Phred-scaled consensus quality (QV) is shown for a subset of 25 R10.4.1 and Illumina hybrid single-fly genomes assembled with haploid (left) and diploid (right) pipelines.





**Figure 4.** The distribution of genome quality metrics for 168 new long-read assemblies. Distributions of genome N50, the percentage of complete dipteran BUSCOs (Manni et al., 2021), and Phred-scaled QV are plotted separately for R10.4.1 and R9.4.1 assemblies from lab strains and from single flies. The black dashed line is the value computed for the *D. melanogaster* dm6 reference genome. The 16 samples that were only sequenced with Illumina are omitted from these plots. Numerical values of statistics and additional sample information is provided in **Table 2**.



**Figure 5.** Proportion of *D. melanogaster* genomic elements aligning to other species as a function of 4-fold divergence from *D. melanogaster*. Each dot represents one species. Alignment coverage is defined as the proportion of genomic elements in *D. melanogaster* that uniquely map to another species.

## Supplementary Tables

Tables can be accessed at: [https://github.com/flyseq/2023\\_drosophila\\_assembly/tables/](https://github.com/flyseq/2023_drosophila_assembly/tables/)

**Table S1.** Representative genome assemblies for 360 drosophilid species and 4 outgroups.

**Table S2.** Consensus accuracy by chromosome for *D. melanogaster* BDGP iso-1 reassembly.

**Table S3.** Consensus accuracy by genomic element type for *D. melanogaster* BDGP iso-1 reassembly.

**Table S4.** Sample and genome quality information on 183 new genomes by this study.

**Table S5.** Information on genomes used for the Progressive Cactus alignment.

**Table S6.** List of laboratory reagents and materials used for this study.

## Supplementary Figures

**Figure S1.** Multi-locus species tree of 360 drosophilid and 4 outgroup taxa estimated from 1,000 single-copy orthologs (Methods). Branch lengths are scaled by the substitution rate at 4-fold degenerate sites. Node values are posterior probabilities computed by ASTRAL.

