Highly contiguous assemblies of 101 drosophilid genomes

Bernard Y. Kim[1,*,†], Jeremy R. Wang[2,†], Danny E. Miller[3], Olga Barmina[4], Emily Delaney[4], Ammon Thompson[4], Aaron A. Comeault[5], David Peede[6], Emmanuel R. R. D'Agostino[6], Julianne Pelaez[7], Jessica M. Aguilar[7], Diler Haji[7], Teruyuki Matsunaga[7], Ellie E. Armstrong[1], Molly Zych[8], Yoshitaka Ogawa[9], Marina Stamenković-Radak[10], Mihailo Jelić[10], Marija Savić Veselinović[10], Marija Tanasković[11], Pavle Erić[11], Jian-jun Gao[12], Takehiro K. Katoh[12], Masanori J. Toda[13], Hideaki Watabe[14], Masayoshi Watada[15], Jeremy S. Davis[16], Leonie C. Moyle[17], Giulia Manoli[18], Enrico Bertolini[18], Vladimír Košťál[19], R. Scott Hawley[20], Aya Takahashi[9], Corbin D. Jones[6], Donald K. Price[21], Noah Whiteman[7], Artyom Kopp[4], Daniel R. Matute[6,*,‡], Dmitri A. Petrov[1,*,‡]

1. Department of Biology, Stanford University, Stanford, CA. USA.
2. Department of Genetics, University of North Carolina, Chapel Hill, North Carolina. USA.
3. Department of Pediatrics, Division of Genetic Medicine, University of Washington and Seattle Children's Hospital, Seattle, WA. USA.
4. Department of Evolution and Ecology, University of California Davis, Davis, CA. USA.
5. School of Natural Sciences, Bangor University, Bangor, Gwynedd, LL57 2DGA, UK.
6. Biology Department , University of North Carolina, Chapel Hill, North Carolina. USA.
7. Department of Integrative Biology, University of California, Berkeley, CA. USA.
8. Molecular and Cellular Biology Program, University of Washington, Seattle, WA. USA.
9. Department of Biological Sciences, Tokyo Metropolitan University, Hachioji. Japan.
10. Faculty of Biology, University of Belgrade, Belgrade, Serbia.
11. University of Belgrade, Institute for Biological Research "Siniša Stanković", National Institute of Republic of Serbia, Belgrade, Serbia.
12. School of Ecology and Environmental Science, Yunnan University, Kunming, Yunnan 650500, China.
13. Hokkaido University Museum, Hokkaido University, Sapporo, Japan.
14. Biological Laboratory, Sapporo College, Hokkaido University of Education, Sapporo, Japan.
15. Graduate School of Science and Engineering, Ehime University, Matsuyama, Ehime, Japan.
16. Department of Biology, University of Kentucky, Lexington, KY. USA.
17. Department of Biology, Indiana University, Bloomington, IN. USA.
18. Neurobiology and Genetics, Theodor Boveri Institute, Biocentre, University of Würzburg, Würzburg. Germany.
19. Institute of Entomology, Biology Centre, Academy of Sciences of the Czech Republic. Czech Republic.
20. Department of Molecular and Integrative Physiology, University of Kansas Medical Center, Kansas City, Kansas 66160 and Stowers Institute for Medical Research, Kansas City, MO. USA.
21. School of Life Science, University of Nevada, Las Vegas. USA.

* Correspondence: bernard.kim@stanford.edu, dmatute@email.unc.edu, dpetrov@stanford.edu

†,‡ These authors contributed equally to this work.

ABSTRACT (150 word limit)

Over 100 years of studies in *Drosophila melanogaster* and related species in the genus *Drosophila* have facilitated key discoveries in genetics, genomics, and evolution. While high-quality genome assemblies exist for several species in this group, they only encompass a small fraction of the genus. Recent advances in long read sequencing allow high quality genome assemblies for tens or even hundreds of species to be generated. Here, we utilize Oxford Nanopore sequencing to build an open community resource of high-quality assemblies for 101 lines of 95 drosophilid species encompassing 14 species groups and 35 sub-groups with an average contig N50 of 10.5 Mb and greater than 97% BUSCO completeness in 97/101 assemblies. These assemblies, along with detailed wet lab protocol and assembly pipelines, are released as a public resource and will serve as a starting point for addressing broad questions of genetics, ecology, and evolution within this key group.

Introduction

The biological and genetic tractability of fruit flies (*Drosophila* and related genera) has led to their status as a premier model system for biological research, particularly in the genomic era (Clark et al., 2007; Hales et al., 2015). Current publicly available genome assemblies number in the tens of species, some with accompanying gene expression and regulation databases (Chen et al., 2014; modENCODE Consortium et al., 2010), comparative genomics tools (Stark et al., 2007), or population genomic data (Guirao-Rico & González, 2019; Lack et al., 2016; Signor et al., 2018). Unfortunately, these genomic resources are far from comprehensive for this remarkably biodiverse group, which encompasses over 1,600 described species (O'Grady & DeSalle, 2018). Expanding the phylogenetic scope of these resources will enable further study of the ecological and evolutionary forces that shape this large and diverse clade.

Recent developments in long-read sequencing make it increasingly feasible to quickly generate high-quality genomes at the level of whole clades. Long-read sequencing simplifies many genome assembly challenges by fully spanning complex regions, such as repetitive elements (generally <10kb in length), and allows generation of chromosome-level assemblies at a reasonable cost. A number of recent studies have used long-read technology to assemble high-quality *Drosophila* genomes for several species groups (Bracewell et al., 2019; Chakraborty et al., 2019; Hill et al., 2020; Mai et al., 2020; Miller et al., 2018). Notably, Miller et al. (2018) estimated the cost of a high-quality long-read assembly at US $1,000, a significant milestone in the democratization of large genome assembly projects.

Here we improve upon that benchmark to present a community resource of 101 *de novo* genome assemblies, from 95 drosophilid species contributed by *Drosophila* researchers from across the world, representing a diversity of ecologies and geographical distributions. We used a hybrid assembly approach with Oxford Nanopore (ONT) long-read sequencing to construct the draft genome and Illumina short reads for polishing. The quality of these genomes is assessed. We propose that under ideal conditions, at least two samples of a typical *Drosophila* genome can be sequenced per ONT release 9.4.1 (rev D) flow cell for as little as $350 (USD) per final high-quality genome. In conjunction with this manuscript and data, we provide a wet lab protocol on Protocols.io specifically optimized for *Drosophila* genome assembly, along with containerized computational pipelines on GitHub. These genome assemblies and technical resources should facilitate the process of conducting large-scale genome projects in this key model clade and beyond.

Results & Discussion

Taxon sampling

Briefly, our selection of species and strains for sequencing (**Table 1**) improves the geographic, ecological, and phylogenetic diversity of *Drosophila* genomes available to the public. The lines sequenced here represent 13 species groups (Toda, 2020) in both major subgenera (*Drosophila* and *Sophophora*); originate from mainland and island locations in North America, Europe, Africa, and Asia; are found from northern (e.g., *D. tristis, D.littoralis*) to equatorial (e.g., *D. bocqueti*) latitudes; represent notable independent transitions to herbivory (*Scaptomyza* and *Lordiphosa*); and include samples of a pest (*Zaprionus indianus*) taken from its native and invasive range. For some species, for instance some *Lordiphosa* spp., only wild-caught flies were sequenced. We sequenced lines in use for active research projects so additional resources like gene expression or population data are expected in the near future. Despite our efforts to improve species diversity, we acknowledge that this initial sampling is heavily biased towards taxa that can be maintained in the lab. Additional details of sample collection are provided in the **Methods**. Future work to improve biological and taxonomic diversity, particularly for species difficult to culture, should employ single fly sequencing and assembly workflows (Adams et al., 2020).

Near chromosome-scale assembly with ultra-long reads

We sequenced 101 fly strains using a modified ONT 1D ligation kit approach, optimized for DNA extractions from 15–30 whole flies and to reduce library prep cost while balancing read lengths and overall throughput. Sequencing runs varied with sample quality and type, and in general read lengths increased over the course of this work. Under optimal conditions, libraries prepared with the supplied protocol should yield 12–15 Gb of data per R9.4.1 flow cell with a read N50 greater than 20kb, and about 30% of data in reads longer than 50kb. We generated paired-end, 150bp Illumina reads for most strains unless public datasets were available.

Deep (average 52×) sequencing coverage with a substantial fraction of ultra-long (50-100 kb+) ONT reads (**Table S1**) resulted in highly contiguous genome assemblies (**Figure S1**) comparable to existing reference genomes in contiguity and completeness (**Figure 1**, **Table S2**). We used Flye (Kolmogorov et al., 2019) based on superior assembly contiguity and favorable runtimes relative to Miniasm (Li, 2016) and Canu (Koren et al., 2017) (**Figure S2**, **Methods**). Of 101 total assemblies, 94 contain over 98% of the assembly in contigs larger than 10kb, and contig N50s exceed 1 Mb for all but 7. In cases where DNA was extracted from pools of wild-caught flies or a single fly (*Leucophenga varia*) resulting in sub-optimal read lengths and output, the assembly was comparable to existing short read assemblies (**Figure 1A & 1B**). High contiguity resulted in BUSCO completeness in the range of 97–99+% for all but the 4 most fragmented genomes (**Figure 1C**).

While we estimate the sequencing cost of a single genome assembly, under the typical conditions presented here, to be $350 (USD), there are opportunities for further optimization in future work. Currently, sequencing runs optimized for ultra-long reads suffer from low throughput due to pore clogging during sequencing. We find that near chromosome level contiguity can be achieved even with minimal (~10×) coverage of reads longer than 25kb (**Figure S3**). Additional read depth will improve consensus sequence accuracy, important for downstream tasks like annotation. However this can be obtained from shorter Nanopore and Illumina reads, which are both easier and cheaper to generate.

A comparative genomics resource

To demonstrate the potential this dataset holds for the study of genome evolution and chromosome organization, we revisit a classic result with our highly contiguous assemblies. Although the ordering of genes in drosophilid chromosomal (Muller) elements has been extensively shuffled throughout ~53 million years of evolution (Suvorov et al., *in prep*), the gene

content of each element remains largely conserved (Bracewell et al., 2019; Ranz et al., 2001; Sturtevant & Novitski, 1941). To examine synteny in our assemblies, many of which contain several contigs tens of megabases in length, we constructed an undirected graph using single-copy orthologous markers (i.e., BUSCOs). The number of times two markers were connected by assemblies determined the weight of the graph's edges. When a graph layout method was applied to visualize these relationships (**Methods**), we found that orthologs clustered by the *D. melanogaster* chromosome on which they are found, consistent with the expected conservation of gene content in Muller elements across drosophilids. Furthermore, the lack of clear order within groups is consistent with extensive shuffling within Muller elements. This demonstrates that our dataset can be used for studies of genome evolution. New reference-free, whole-genome alignment methods (Armstrong et al., 2020) should substantially facilitate these kinds of comparative analyses.

Repeat content

A large number of genome assemblies enables comparative analysis of repeat variation against a wide range of genome sizes (140–450Mb), for example the independent expansions of satellite repeats in *D. grimshawi* or retroelements in *D. paulistorum*, *D. bipectinata*, or *D. subpulchrella* (**Figure 3**). Within our dataset alone, RepeatMasker annotations show large variation in repeat content among drosophilids (**Figure 3**). No correlation exists between assembly contiguity and repeat content (**Figure S4**), suggesting long-read sequencing overcomes many of the challenges to drosophilid genome assembly posed by repetitive sequences. Additionally, we observe a positive relationship between the size of repetitive sequences and non-repetitive sequences, suggesting that genome size is influenced by expansions and contractions of both portions of the genome (**Figure S5**). The high continuity of these assemblies should allow the identification of complete transposable elements in the genomes and allow for the analyses of transposable element evolution at the level of individual transposable elements or transposable element families in a way that is not feasible with more fragmented genome assemblies (Clark et al., 2007).

Reproducibility

Detailed laboratory protocols, computational pipelines, and computational container recipes are provided as a reference and to maximize reproducibility. The protocol is publicly available at Protocols.io (dx.doi.org/10.17504/protocols.io.bdfqi3mw) and pipeline scripts along

with associated compute containers are provided in a public GitHub repository (https://github.com/flyseq/drosophila_assembly_pipelines). See **Methods** for additional details on compute containers.

Future directions

We have described an open community resource of 101 nearly chromosome-level drosophilid genome assemblies, adding to or improving upon many assemblies already available for this group (Suvorov et al., n.d.) as well as providing detailed protocols for adding additional genomes economically and easily. We envision the provided dataset being used to address a large number of outstanding questions entailing large comparative analyses among species and to improve the genome annotation process for these species (Armstrong et al., 2020; Fiddes et al., 2018; Shumate & Salzberg, 2020). Finally, these data will allow population genomic data to be compared for a large number of species, providing unprecedented resolution to investigate fundamental questions about the evolutionary process.

Materials and Methods

Taxon Sampling and Sample Collection

The selection of species used for this study was driven by several key objectives. First, we aimed to provide data for ongoing research projects. Second, we aimed to supplement existing genomic data, both as a benchmarking resource against well-studied references (e.g. *D. melanogaster*) and to provide a technological update to some older assemblies (modENCODE Consortium et al., 2010). Third, we aimed to increase the phylogenetic and ecological diversity of publically available *Drosophila* genome assemblies.

In most cases, genomic DNA was collected from lab-raised flies, which were either derived from lines maintained at public *Drosophila* stock centers and individual labs or, in a few cases, from F1 or F2 progeny of flies recently collected in the wild. We collected specimens from the wild with standard fruit or mushroom-baited traps, sweep netting, and aspiration. We established isofemale lines from individual females collected using these baits unless otherwise specified (**Table S1**). For species difficult to culture in the lab (all *Lordiphosa* spp. except *Lo. clarofinis*, *D. sproati, D. murphyi*, *Le. varia, S. graminum*), either wild-caught flies or flies from a transient lab culture were used. In accordance with domestic and international shipping laws, these flies were either fixed in ethanol before transport (*Lordiphosa* spp., *D. subobscura, D.*

obscura, C. costata, D. littoralis, D. tristis, D. ambigua) or transported with permits (P526P-15-02964 to D. Matute, P526P-20-02787 and P526P-19-01521 to A. Kopp, and Hawaii State permit I1302 to D. Price).

Of 101 total assemblies, we include 13 genomes assembled with re-analyzed sequences from Miller *et al*. (Miller et al., 2018); 60 genomes from stock center lines or established lab cultures; 22 genomes from lab-raised flies derived from recent wild collections; and 6 genomes from wild-caught flies. Of note, 6 *Zaprionus* lines used in this study (*Z. africanus, Z. indianus, Z. tsacasi, Z. nigranus, Z. taronus*) were assembled by Comeault et al. (Comeault et al., 2020), but updated higher contiguity assemblies are provided with this manuscript with the exception of *Z. indianus* line 16GNV01 (see "Alternative hybrid assembly process" section below). Details on each sample including (if available) line designations and collection information, are provided in **Table S3**.

DNA extraction and Nanopore sequencing

A high molecular weight (HMW) genomic DNA (gDNA) extraction and Nanopore library prep was performed for each sample, with slight variation in the protocol through time and to deal with differences in sample quality or preservation. Here, we briefly describe a recommended general protocol for HMW gDNA extraction and library prep from 15–30 flies. This protocol is sufficient to reproduce all results from this manuscript at the same or higher levels of data quality. Detailed step-by-step instructions are provided at Protocols.io (dx.doi.org/10.17504/protocols.io.bdfqi3mw). We note one exception made necessary by sample availability and shipping laws. *Scaptomyza graminum* gDNA was extracted by using the Qiagen Blood & Cell Culture DNA Mini Kit from 30 unfrozen flies and prepared with the ONT LSK109 kit without any modifications to the manufacturer's instructions.

Genomic DNA was prepared from about 30 flash frozen or ethanol fixed adult flies. For non-inbred samples, we tried to use 15 flies or less to minimize the genetic diversity of the sample. In the absence of amplification, about 1.5–3 µg of input DNA is needed to prepare 3–4 library loads with the ONT LSK109 kit. Sufficient input DNA is particularly important when selecting for longer reads. Ethanol preserved samples were soaked in a rehydration buffer (400 mM NaCl, 20 mM Tris-HCl pH 8.0, 30 mM EDTA) for 30 minutes at room temperature (~23ºC), dabbed dry with a Kimwipe, then frozen for 1 hour at -80ºC before extraction. Frozen flies were ground in 1.5 mL of homogenization buffer (0.1M NaCl, 30mM Tris HCl pH 8.0, 10 mM EDTA, 0.5% Triton X-100) with a 2 mL Kontes Dounce homogenizer. The homogenate was centrifuged

for 5 minutes at 2,000 ×*g*, the supernatant discarded by decanting, and the pellet resuspended in 100 µL of fresh homogenization buffer. This mixture was then added to a tube with 380 µL extraction buffer (0.1M Tris-HCl pH 8.0, 0.1M NaCl, 20mM EDTA) along with 10 µL Proteinase K (20 mg/mL), 10 µL SDS (10% w/v), and 2 µL RNAse A (10 mg/mL). This tube was incubated at 50°C for 4 hours, with mixing at 30–60 minute intervals by gentle inversion.

High molecular weight gDNA was purified with a standard phenol-chloroform extraction. The lysate was extracted twice with an equal volume of phenol chloroform isoamyl alcohol (25:24:1 v/v) in a 2 mL light phase lock gel tube. Next, the aqueous layer was decanted into a fresh 2mL phase lock gel tube then extracted once with an equal volume of chloroform. The use of the phase lock gel tube reduces DNA shearing at this stage by minimizing pipette handling. HMW DNA was precipitated by adding 0.1 volume of 3M sodium acetate and 2.0–2.4 volumes of cold absolute ethanol. Gentle mixing resulted in the precipitation of a white, stringy clump of DNA, which was then transferred to a DNA LoBind tube and washed twice with 70% ethanol. After washing, the DNA was pelleted by centrifugation and all excess liquid removed from the tube. The pellet was allowed to air dry until the moment it became translucent, resuspended in 65 µL of 1× Tris-EDTA buffer on a heat block at 50°C for 60 minutes, then incubated for at least 48 hours at 4°C. After 48 hours, the viscous DNA solution was mixed by gentle pipetting with a P1000 tip. This controlled shearing step encourages resuspension of HMW DNA and improves library prep yield. DNA was quantified with Qubit and Nanodrop absorption ratios were checked to ensure 260/280 was greater than 1.8 and 260/230 was greater than 2.0.

The sequencing library was prepared following the ONT Ligation Sequencing Kit (SQK-LSK109) protocol, with two important modifications. First, we started with approximately 3 µg of input DNA, three times the amount recommended by the manufacturer. Second, we utilized size-selective polymer precipitation (Paithankar & Prasad, 1991) with the Circulomics Short Read Eliminator (SRE) buffer plus centrifugation to isolate DNA instead of magnetic beads. We found this to be necessary because magnetic beads irreversibly clumped with viscous HMW gDNA, decreasing library yield and limiting read lengths. The manner in which this was performed was specific to the cleanup step. After the end-prep/repair step, the SRE buffer was used according to the manufacturer's instructions. After adapter ligation, DNA was pelleted by centrifuging the sample at 10,000×*g* for 30 minutes without the addition of any reagents, since DNA readily precipitated upon addition of the ligation buffer. Ethanol washes were avoided past this step since ethanol will denature motor proteins in the prepared library. Instead, the DNA pellet was washed with 100 µL SFB or LFB (interchangeably) from the ligation

sequencing kit instead of 70% ethanol. If library yield was sufficient (>50 ng/μL), the Circulomics SRE buffer was used for a final round of size selection, replacing the ethanol wash with LFB/SFB as described above. Of note, a cheaper and open-source alternative made with polyethylene glycol MW 8000 (PEG 8000), although less effective at size selection, to the SRE buffer is described by Tyson (Tyson, 2020) (dx.doi.org/10.17504/protocols.io.7euhjew). A 1:1 dilution of the PEG 8000 solution described in that protocol can be substituted for SFB or LFB in the washing steps described above.

The typical yield of a library prepared in this manner is in the range of 1–1.5 μg. Approximately 350 ng of the prepared library was loaded for each sequencing run. To maintain flow cell throughput and read length, flow cells were flushed every 8–16 hours with the ONT Flow Cell Wash Kit (EXP-WSH003) and reloaded with a fresh library.

Short read data for polishing

We performed 2×150 bp Illumina sequencing for most of the strains that did not have publicly available short read data available. Illumina libraries were prepared from the same gDNA extractions as the Nanopore library for most samples, with some exceptions as described in **Table S2**. The libraries were prepared in either of two manners. For the majority of samples, sequencing libraries were prepared with a modified version of the Nextera DNA Library Kit protocol (Baym et al., 2015) and sequencing was performed by Admera Health on NextSeq 4000 or HiSeq 4000 machines. Alternatively, Illumina libraries were prepared with the KAPA Hyper DNA kit according to the manufacturer's protocol and sequenced at the UNC sequencing core on a HiSeq 4000 machine. In either case, all samples on a lane were uniquely dual indexed. Illumina sequencing was not performed for *D. equinoxialis*, *D. funebris*, *D. subpulchrella*, *D. tropicalis*, *Le. varia*, *Z. lachaisei*, *Z. taronus*, and the unidentified São Tomé mushroom feeder due to material unavailability (line extinction/culling). Details for each sample, including accession numbers for any public data used in this work, are provided in **Table S2**.

Choice of long read assembly program

Flye v2.6 (Kolmogorov et al., 2019) was used due to its quick CPU runtime, low memory requirements, excellent assembly contiguity, and its consistent performance on benchmarking datasets (Wick & Holt, 2020). We additionally validated the performance of Flye for *Drosophila* genomes using Nanopore data previously generated by Miller *et al*. (2018) and 60× depth of new Nanopore sequencing of the Berkeley Drosophila Genome Project ISO-1 strain of *D.*

melanogaster. We assembled genomes with Flye v2.6 and Canu v1.8 (Koren et al., 2017) to evaluate simple benchmarks of assembly contiguity and run time and to provide a comparison to the Miniasm (Li, 2016) assemblies from Miller *et al*. (Miller et al., 2018) Canu produced relatively contiguous assemblies, but a single assembly took several days on a 92-core cloud server and even longer when a large number of extra-long (>50kb) reads were present in the data. This was determined to be too costly when scaled to >100 species. In addition to a much shorter (8–12 hours wall-clock time) runtime, Flye also produced significantly more contiguous assemblies than those reported by Miller *et al*. (**Figure S2**). Note, several new long read assemblers have been released and these assembly programs have been significantly updated since this work was performed. Assembler performance should be evaluated with up-to-date versions in any future work.

Assembly and long read polishing

After Nanopore sequencing was performed, raw Nanopore data were basecalled with Guppy v3.2.4, using the high-accuracy caller (option: -c dna_r0.4.1_450bps_hac.cfg). Raw Nanopore data previously generated by Miller *et al*. (Miller et al., 2018) were processed in the same manner.

Next, basecalled reads were assembled using Flye v2.6 with default settings. Genome size estimates (option: --genomeSize) were obtained through a web search or taken from a closely related species. If no such information was available, an initial estimate of 200 Mb was used. The specific genome size estimate is provided in **Table S2**.

After generating a draft assembly, we performed long read polishing using Medaka (https://nanoporetech.github.io/medaka/draft_origin.html). Reads were aligned to the draft genome with Minimap2 v2.17 (Li, 2016) before each round of polishing (option: -ax ont). The draft was polished with two rounds of Racon v1.4.3 (Vaser et al., 2017) (options: -m8 -x 6 -g 8 -w 500) and then a single round of Medaka v0.9.1.

Haplotig identification and removal

Next, we assessed each assembly for the presence of multiple haplotypes (haplotigs) using BUSCO v3.0.2 (Simão et al., 2015; Waterhouse et al., 2018) on the Medaka-polished sequences. If the BUSCO duplication rate exceeded 1%, haplotig identification and removal was performed, but on the draft assembly produced by Flye rather than the polished assembly. Purge_haplotigs v1.1.1 (Roach et al., 2018) was run on these sequences following the

guidelines provided by the developer (https://bitbucket.org/mroachawri/purge_haplotigs). Illumina reads were mapped to the draft assembly with Minimap2 (option: -ax sr) to obtain read depth information. The optional clipping step was performed to remove overlapping (duplicate) contig ends. Finally, remaining contigs were re-scaffolded with Nanopore reads using npScarf v1.9-2b (Cao et al., 2017), with support from at least 4 long reads required to link two contigs (option: --support=4). These sequences were polished with Racon and Medaka as described above.

Final polishing and decontamination

The Medaka-polished assembly was further polished with Illumina data and any contigs identified as microbial sequences were removed. For each polishing round, Illumina reads were mapped to the draft assembly with Minimap2 (option: -ax sr) and polished with Pilon v1.23 (Walker et al., 2014). If a genome did not have an accompanying short read dataset but Illumina reads were available from a different strain of the same species (**Table S1**), Pilon was run without correcting SNVs (option: --fix indels,gaps,local). After Pilon polishing, assembly completeness was assessed again with BUSCO v3.0.2. We used BLAST (version 2.10.0) (Altschul et al., 1990) to remove any contigs not associated with at least one BUSCO that were also of bacterial, protozoan, or fungal origin. Finally, any sequences flagged by the NCBI Contamination Screen were excluded or trimmed.

Alternative hybrid assembly process

Zaprionus indianus line 16GNV01 had insufficient Nanopore data for a Flye assembly. For this line only and to consolidate all assemblies as a single resource, the same genome assembly from Comeault *et al*. (Comeault et al., 2020) is both reported here and associated with the NCBI BioProject associated with this work. An alternative assembly strategy was taken for this line (Comeault et al., 2020). Briefly, short-read sequence data was assembled first using SPAdes v3.11.1 (Bankevich et al., 2012) using default parameters. Nanopore reads were corrected with Illumina data using FMLRC v.1.0.0 (Wang et al., 2018) and subsequently used to scaffold the SPAdes assembly using LINKS v.1.8.7 (Warren et al., 2015) using the recommended iterative approach of 33 iterations with incrementally increasing *k*-mer distance threshold. The resulting scaffolds were polished with four rounds of Racon followed by four rounds of Pilon (but without Medaka) as described above.

Repeat annotation and masking

Each draft assembly was soft repeat masked with RepeatMasker v4.1.0 (Smit et al., 2013) at medium sensitivity, with both Dfam 3.1 (Hubley et al., 2016) and RepBase RepeatMasker edition (Bao et al., 2015) repeat libraries installed (options: --species drosophila --xsmall). RepeatMasker was initialized with cross_match v1.090518 (Green, 2009) as the sequence search engine and Tandem Repeat Finder v4.0.9 (Benson, 1999).

Assessing assembly contiguity and completeness

Assembly contiguity statistics were computed using a series of custom shell and R scripts. Fasta files were parsed with Bioawk v1.0 and summary statistics were computed in the standard manner with custom scripts. Of particular note, we preferentially presented the *auN* statistic over contig *N50* to summarize assembly contiguity, although both are provided. The *auN* statistic is computed as the area under an *Nx* curve:

$$auN = \sum_i (L_i \frac{L_i}{\sum_j L_j})$$

The advantage of *auN* for comparing assemblies is its sensitivity to assembly breaks: a break may not always affect *N50* but will always affect *auN* (Li, 2020).

Assembly completeness was assessed with BUSCO v4.0.6 (Seppey et al., 2019; Simão et al., 2015), in genome mode with the diptera_odb10 database (options: --m geno -l diptera_odb10 --augustus_species fly). Note that this BUSCO version is different (v4 vs v3). BUSCO v4 was released while the assemblies were in progress, and we wished to evaluate final completeness with the most up-to-date tools while retaining consistency across the assembly pipeline. BUSCO v4 (tested with v4.0.6-v4.1.3) runs with the *D. equinoxialis* genome were unsuccessful. For *D. equinoxialis* only, BUSCO v3.0.4 was run with the diptera_odb9 database (options: --m geno -l diptera_odb9).

Species tree inference from BUSCO orthologs

We inferred species relationships using complete and single-copy orthologs identified by the BUSCO analysis. Amino acid sequences were used instead of nucleotide sequences to achieve better alignments in the face of high sequence divergence (Bininda-Emonds, 2005). Out of 990 single-copy orthologs present in all assemblies, we randomly selected 250 to construct gene trees. The predicted protein sequence of each ortholog was aligned separately

with MAFFT v7.453 (Katoh & Standley, 2013), using the E-INS-i algorithm (options: --ep 0 --genafpair --maxiterate 1000). Gene trees were inferred with RAxML-NG v0.9.0 (Kozlov et al., 2019), using the Le and Gascuel (2008) amino acid substitution model (options: --msa-format FASTA --data-type AA --model LG). The summary method ASTRAL-MP v.5.14.7 (Yin et al., 2019) was run with default settings to reconstruct the species tree. We note that this is not intended to be a definitive phylogenetic reconstruction of species relationships; see Suvorov et al. (Suvorov et al., *in prep*) for a time-calibrated phylogeny utilizing 158 drosophilid whole genomes.

Analysis of chromosome organization

Syntenic comparisons were performed by representing the genome assemblies as paths through an undirected graph. The path each genome traverses can be considered a series of connections between single copy orthologous markers (i.e., BUSCOs). Using BUSCO v4 annotations for each final genome, we constructed a 3,285 by 3,285 symmetric adjacency matrix, with row and column headers (nodes) corresponding to 3,285 possible BUSCOs from the diptera_odb10 database. Off-diagonal entries in each matrix (edges) were the number of times two single-copy BUSCOs were found as connected and immediate neighbors in the assemblies. Sequences of three or more BUSCOs were not considered. The graph was then visualized in two dimensions using the ForceAtlas2 graph layout algorithm (Jacomy et al., 2014) as implemented in the ForceAtlas2 R package (https://github.com/analyxcompany/ForceAtlas2). While this method is primarily designed for flexible, user-friendly tuning of graph visualization, it is similar in effect to other nonlinear dimensionality reduction techniques (Böhm et al., 2020). ForceAtlas2 was run with the settings: tolerance=1, gravity=1, iterations=3000. *D. equinoxialis* was omitted from this analysis due to the BUSCO v4 issues mentioned previously.

Repeat content and genome size analysis

The contribution of repeat content to genome size variation in *Drosophila* was examined by comparing the number of bases in each genome annotated as a type of repeat (previously described) to the number of bases not annotated as repetitive sequence. Phylogenetic independent contrasts (Felsenstein, 1985) were computed for the counts of bases in both categories using the R package ape v5.4.1 (Paradis & Schliep, 2019) using the species tree described above with the root age set to 53 million years following the estimate in Suvorov et al. (Suvorov et al., n.d.).

Compute containers

While the overall computational demands of this work were high, the unique computational challenge we faced was the variety of computational resources used for various stages of the assembly process. Assemblies took place across local servers, institutional clusters, and cloud computing resources. A key factor in ensuring reproducibility across computing environments was the use of computing containers, which is like a lightweight virtual machine that can be customized such that sets of programs and their dependencies are packaged together. Specifically, we used the programs Docker and Singularity to manage containers. These programs allow containers to be built and packaged as an image file which is transferred to another computer. A Dockerfile, a text file containing instructions to set up an image, is used to select the Linux operating system and the suite of programs to be installed within a Docker container. Singularity is used to package the Docker container as an image file that can be transferred to and used in a cluster or cloud environment without the need for administrative permissions. Standard commands are then run inside the container environment. The files and instructions necessary to build these containers, which will allow for the exact reproduction of the computing environment in which this work was performed, are provided at: https://github.com/flyseq/drosophila_assembly_pipelines. We hope these files will facilitate the work of researchers new to Nanopore sequencing or the genome assembly process.

Data availability

Supplementary File 1: Figure S1 Flow chart depiction of the assembly pipeline.

Supplementary File 2: Figure S2 Large improvements in assembly contiguity from an updated assembly workflow.

Supplementary File 3: Figure S3 Highly contiguous assemblies can be obtained with lower coverage of ultra-long reads.

Supplementary File 4: Assembly contiguity is not determined by repeat content.

Supplementary File 5: The non-repetitive and repetitive portions of the genome both contribute to genome size differences in Drosophila.

Supplementary File 6: Table S1 Description of data used for this project, including accession numbers for public data.

Supplementary File 7: Table S2 Assembly summary statistics.

Supplementary File 8: Table S3 Detailed sample information.

Supplementary Files 1-8 are available at: https://doi.org/10.6084/m9.figshare.13377179

Whole genome sequencing data generated by this work are available at NCBI BioProject PRJNA675888. Preliminary access to genome assemblies is provided at https://web.stanford.edu/~bkim331/files/genomes/. Raw Nanopore data are available by request.

Details are also provided at: http://flyseq.org/

Genome assembly pipeline and code: https://github.com/flyseq/drosophila_assembly_pipelines

Full laboratory protocol: dx.doi.org/10.17504/protocols.io.bdfqi3mw

Funding

BYK was supported by the NIH under award no. F32GM135998 and Google Cloud Platform Research Credits. DAP, BYK, and EEA were supported by the NIH under award no. R35GM118165. JRW was supported by the NIH under award no. K01DK119582, Amazon Web Services Cloud Credits for Research, and Google Cloud Platform Research Credits. DRM was supported by the NIH under award nos. R01GM121750 and R01GM125715. AK, OB, ED, and A. Thompson were supported by the NIH under award no. R35GM122592. NW, JP, JMA, DH, and TM were supported by the NIH under award no. NIGMS GM119816. JP was supported by the NSF GRFP and the Mentored Research Award from UC Berkeley. TM was supported by the Uehara Memorial Foundation under award no. 201931028. MS-R, MJ, and MSV were supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia under grant no. 451-03-68/2020-14/200178. MT and PE were supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia under grant no. 451-03-68/2020-14/200007. JJG was supported by the NFSC under award no. 32060112. MW was supported by the JSPS KAKENHI under award no. JP18K06383. GM and EB were supported by the European Union's Horizon 2020 research and innovation program under award no. 765937-CINCHRON. VK was supported by the Czech Science Foundation under award no. 19-13381S. RSH is an American Cancer Society Research Professor. AT was supported by the JSPS KAKENHI under award no. JP19H03276. DKP was supported by the

NSF under award no. 1345247. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgments

We thank Brandon Cooper, Antonio Serrato-Capuchina, and David Turissini for help with collections and field logistics; Sarah C.R. Elgin, Wilson Leung, Elena Gracheva, and Sophia Bieser for help with modENCODE fly lines; Jonathan Chang for helpful discussions about phylogenetic methods; Charlotte Helfrich-Förster for providing lab resources for G. Manoli and E. Bertolini; and John Tyson plus the staff at Circulomics, in particular Kelvin Liu and Michelle Kim, for many helpful discussions about long read library prep and sequencing.

Competing Interests

The authors declare that there is no conflict of interest.

References

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

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

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), 246–251. https://doi.org/10.1038/s41586-020-2871-y

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

Bao, W., Kojima, K. K., & Kohany, O. (2015). Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA*, *6*(1), 11. https://doi.org/10.1186/s13100-015-0041-9

Baym, M., Kryazhimskiy, S., Lieberman, T. D., Chung, H., Desai, M. M., & Kishony, R. (2015). Inexpensive Multiplexed Library Preparation for Megabase-Sized Genomes. *PLOS ONE*, *10*(5), e0128036. https://doi.org/10.1371/journal.pone.0128036

Benson, G. (1999). Tandem repeats finder: A program to analyze DNA sequences. *Nucleic Acids Research*, *27*(2), 573–580. https://doi.org/10.1093/nar/27.2.573

Bininda-Emonds, O. R. (2005). transAlign: Using amino acids to facilitate the multiple alignment of protein-coding DNA sequences. *BMC Bioinformatics*, *6*(1), 156. https://doi.org/10.1186/1471-2105-6-156

Böhm, J. N., Berens, P., & Kobak, D. (2020). A Unifying Perspective on Neighbor Embeddings along the Attraction-Repulsion Spectrum. *ArXiv:2007.08902 [Cs, Stat]*. http://arxiv.org/abs/2007.08902

Bracewell, R., Chatla, K., Nalley, M. J., & Bachtrog, D. (2019). Dynamic turnover of centromeres drives karyotype evolution in *Drosophila*. *ELife*, *8*, e49002. https://doi.org/10.7554/eLife.49002

Bronski, M. J., Martinez, C. C., Weld, H. A., & Eisen, M. B. (2020). Whole Genome Sequences of 23 Species from the *Drosophila montium* Species Group (Diptera: Drosophilidae): A

Resource for Testing Evolutionary Hypotheses. *G3: Genes, Genomes, Genetics*. https://doi.org/10.1534/g3.119.400959

Cao, M. D., Nguyen, S. H., Ganesamoorthy, D., Elliott, A. G., Cooper, M. A., & Coin, L. J. M. (2017). Scaffolding and completing genome assemblies in real-time with nanopore sequencing. *Nature Communications*, *8*(1), 14515. https://doi.org/10.1038/ncomms14515

Chakraborty, M., Emerson, J. J., Macdonald, S. J., & Long, A. D. (2019). Structural variants exhibit widespread allelic heterogeneity and shape variation in complex traits. *Nature Communications*, *10*(1), 4872. https://doi.org/10.1038/s41467-019-12884-1

Chen, Z.-X., Sturgill, D., Qu, J., Jiang, H., Park, S., Boley, N., Suzuki, A. M., Fletcher, A. R., Plachetzki, D. C., FitzGerald, P. C., Artieri, C. G., Atallah, J., Barmina, O., Brown, J. B., Blankenburg, K. P., Clough, E., Dasgupta, A., Gubbala, S., Han, Y., … Richards, S. (2014). Comparative validation of the *D. melanogaster* modENCODE transcriptome annotation. *Genome Research*, *24*(7), 1209–1223. https://doi.org/10.1101/gr.159384.113

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., Larracuente, 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), 203–218. https://doi.org/10.1038/nature06341

Comeault, A. A., Wang, J., Tittes, S., Isbell, K., Ingley, S., Hurlbert, A. H., & Matute, D. R. (2020). Genetic Diversity and Thermal Performance in Invasive and Native Populations of African Fig Flies. *Molecular Biology and Evolution*, *37*(7), 1893–1906. https://doi.org/10.1093/molbev/msaa050

Felsenstein, J. (1985). Phylogenies and the Comparative Method. *The American Naturalist*,

125(1), 1–15.

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

Green, P. (2009). *Phrap, version 1.090518*. Phrap, Version 1.090518. http://phrap.org

Guirao-Rico, S., & González, J. (2019). Evolutionary insights from large scale resequencing datasets in *Drosophila melanogaster*. *Current Opinion in Insect Science*, *31*, 70–76. https://doi.org/10.1016/j.cois.2018.11.002

Hales, K. G., Korey, C. A., Larracuente, A. M., & Roberts, D. M. (2015). Genetics on the Fly: A Primer on the Drosophila Model System. *Genetics*, *201*(3), 815–842. https://doi.org/10.1534/genetics.115.183392

Hill, T., Rosales-Stephens, H.-L., & Unckless, R. L. (2020). Rapid divergence of the copulation proteins in the *Drosophila dunni* group is associated with hybrid post-mating-prezygotic incompatibilities. *BioRxiv*, 2020.05.20.106724. https://doi.org/10.1101/2020.05.20.106724

Hubley, R., Finn, R. D., Clements, J., Eddy, S. R., Jones, T. A., Bao, W., Smit, A. F. A., & Wheeler, T. J. (2016). The Dfam database of repetitive DNA families. *Nucleic Acids Research*, *44*(D1), D81–D89. https://doi.org/10.1093/nar/gkv1272

Jacomy, M., Venturini, T., Heymann, S., & Bastian, M. (2014). ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software. *PLOS ONE*, *9*(6), e98679. https://doi.org/10.1371/journal.pone.0098679

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

Kolmogorov, M., Yuan, J., Lin, Y., & Pevzner, P. A. (2019). Assembly of long, error-prone reads using repeat graphs. *Nature Biotechnology*, *37*(5), 540–546. https://doi.org/10.1038/s41587-019-0072-8

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., & Phillippy, A. M. (2017). Canu: Scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Research*, *27*(5), 722–736. https://doi.org/10.1101/gr.215087.116

Kozlov, A. M., Darriba, D., Flouri, T., Morel, B., & Stamatakis, A. (2019). RAxML-NG: A fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics*, *35*(21), 4453–4455. https://doi.org/10.1093/bioinformatics/btz305

Lack, J. B., Lange, J. D., Tang, A. D., Corbett-Detig, R. B., & Pool, J. E. (2016). A Thousand Fly Genomes: An Expanded Drosophila Genome Nexus. *Molecular Biology and Evolution*, *33*(12), 3308–3313. https://doi.org/10.1093/molbev/msw195

Le, S. Q., & Gascuel, O. (2008). An Improved General Amino Acid Replacement Matrix. *Molecular Biology and Evolution*, *25*(7), 1307–1320. https://doi.org/10.1093/molbev/msn067

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. (2020, April 8). *AuN: a new metric to measure assembly contiguity*. https://lh3.github.io/2020/04/08/a-new-metric-on-assembly-contiguity

Mai, D., Nalley, M. J., & Bachtrog, D. (2020). Patterns of Genomic Differentiation in the

Drosophila nasuta Species Complex. *Molecular Biology and Evolution*, *37*(1), 208–220. https://doi.org/10.1093/molbev/msz215

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

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., Berezikov, 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

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

Paithankar, K. R., & Prasad, K. S. N. (1991). Precipitation of DNA by polyethylene glycol and ethanol. *Nucleic Acids Research*, *19*(6), 1346–1346. https://doi.org/10.1093/nar/19.6.1346

Paradis, E., & Schliep, K. (2019). ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, *35*(3), 526–528. https://doi.org/10.1093/bioinformatics/bty633

Ranz, J. M., Casals, F., & Ruiz, A. (2001). How Malleable is the Eukaryotic Genome? Extreme Rate of Chromosomal Rearrangement in the Genus *Drosophila*. *Genome Research*, *11*(2), 230–239. https://doi.org/10.1101/gr.162901

Roach, M. J., Schmidt, S. A., & Borneman, A. R. (2018). Purge Haplotigs: Allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinformatics*, *19*(1), 460.

https://doi.org/10.1186/s12859-018-2485-7

Seppey, M., Manni, M., & Zdobnov, E. M. (2019). BUSCO: Assessing Genome Assembly and Annotation Completeness. In M. Kollmar (Ed.), *Gene Prediction: Methods and Protocols* (pp. 227–245). Springer. https://doi.org/10.1007/978-1-4939-9173-0_14

Shumate, A., & Salzberg, S. L. (2020). Liftoff: An accurate gene annotation mapping tool. *BioRxiv*, 2020.06.24.169680. https://doi.org/10.1101/2020.06.24.169680

Signor, S. A., New, F. N., & Nuzhdin, S. (2018). A Large Panel of *Drosophila simulans* Reveals an Abundance of Common Variants. *Genome Biology and Evolution*, *10*(1), 189–206. https://doi.org/10.1093/gbe/evx262

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, *31*(19), 3210–3212. https://doi.org/10.1093/bioinformatics/btv351

Smit, A. F. A., Hubley, R., & Green, P. (2013, 2015). *RepeatMasker Open-4.0*. RepeatMasker Open-4.0.

Stark, A., Lin, M. F., Kheradpour, P., Pedersen, J. S., Parts, L., Carlson, J. W., Crosby, M. A., Rasmussen, M. D., Roy, S., Deoras, A. N., Ruby, J. G., Brennecke, J., Hodges, E., Hinrichs, A. S., Caspi, A., Paten, B., Park, S.-W., Han, M. V., Maeder, M. L., … Kellis, M. (2007). Discovery of functional elements in 12 *Drosophila* genomes using evolutionary signatures. *Nature*, *450*(7167), 219–232. https://doi.org/10.1038/nature06340

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. R., Armstrong, E. E., Peede, D., D'Agostino, E. R. R., Price, D. K., Lang, M., Courtier-Orgogozo, V., David, J. R., Petrov, D. A., Matute, D. R., Schrider,

D. R., & Comeault, A. A. (2020). Widespread introgression across a phylogeny of 155 *Drosophila* genomes. *In Prep*.

Toda, M. J. (2020). *bioinfo—Taxonomic information*. https://bioinfo.museum.hokudai.ac.jp/db/index.php

Tyson, J. (2020, January 29). *Bead-free long fragment LSK109 library preparation*. Bead-Free Long Fragment LSK109 Library Preparation. https://dx.doi.org/10.17504/protocols.io.7euhjew

Vaser, R., Sović, I., Nagarajan, N., & Šikić, M. (2017). Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Research*, *27*(5), 737–746. https://doi.org/10.1101/gr.214270.116

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

Wang, J. R., Holt, J., McMillan, L., & Jones, C. D. (2018). FMLRC: Hybrid long read error correction using an FM-index. *BMC Bioinformatics*, *19*(1), 50. https://doi.org/10.1186/s12859-018-2051-3

Warren, R. L., Yang, C., Vandervalk, B. P., Behsaz, B., Lagman, A., Jones, S. J. M., & Birol, I. (2015). LINKS: Scalable, alignment-free scaffolding of draft genomes with long reads. *GigaScience*, *4*(1), 35. https://doi.org/10.1186/s13742-015-0076-3

Waterhouse, R. M., Seppey, M., Simão, F. A., Manni, M., Ioannidis, P., Klioutchnikov, G., Kriventseva, E. V., & Zdobnov, E. M. (2018). BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. *Molecular Biology and Evolution*, *35*(3), 543–548. https://doi.org/10.1093/molbev/msx319

Wick, R., & Holt, K. E. (2020). Benchmarking of long-read assemblers for prokaryote whole genome sequencing. *F1000 Research*, *8*(2138), 1–22. https://doi.org/10.12688/f1000research.21782.3

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

Figures

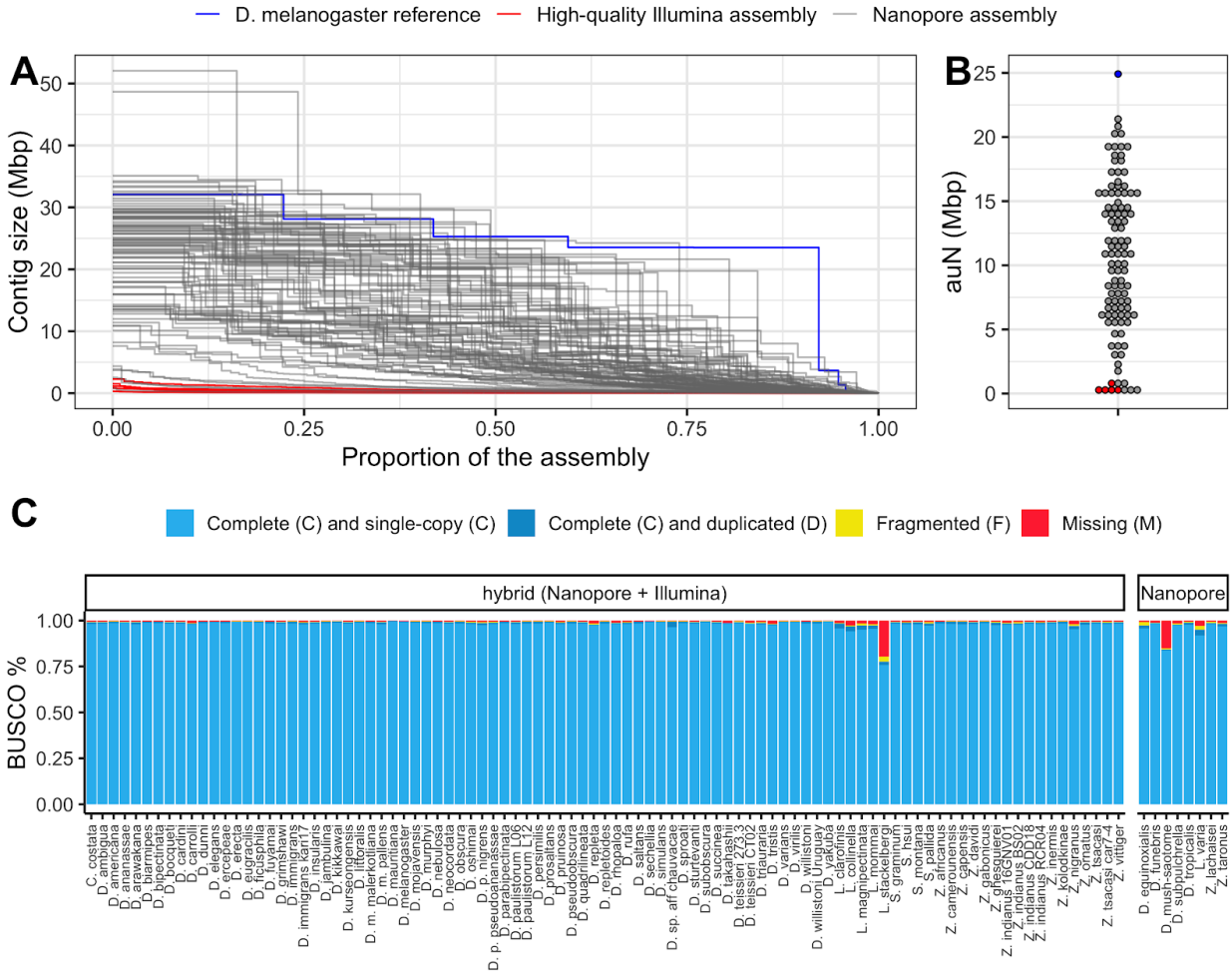

Figure 1. Nanopore-based assemblies are highly contiguous and complete. (**A,B**) Assembly contiguity is compared to the *D. melanogaster* v6.22 reference genome (blue) as well as 5 recently published, highly contiguous Illumina assemblies (red lines, *D. birchii, D. bocki, D. bunnanda, D. kanapiae, D. truncata*; Bronski et al., 2020). (**A**) *Nx* curves, or the (y-axis) size of each contig when contigs are sorted in descending size order, in relation to the (x-axis) cumulative proportion of the genome assembly that is covered. (**B**) The distribution of the *auN* statistic, a measure of contiguity obtained by calculating the area under the *Nx* curve. A contig break will always result in lower *auN* but not necessarily a lower *N50*. (**C**) Assembly completeness assessed by BUSCO v4.0.6 (Simão et al., 2015). Note, *D. equinoxialis* was evaluated with BUSCO v3.0.2 due to an unfixable bug. Individual assembly summary statistics are provided in **Table S2**.

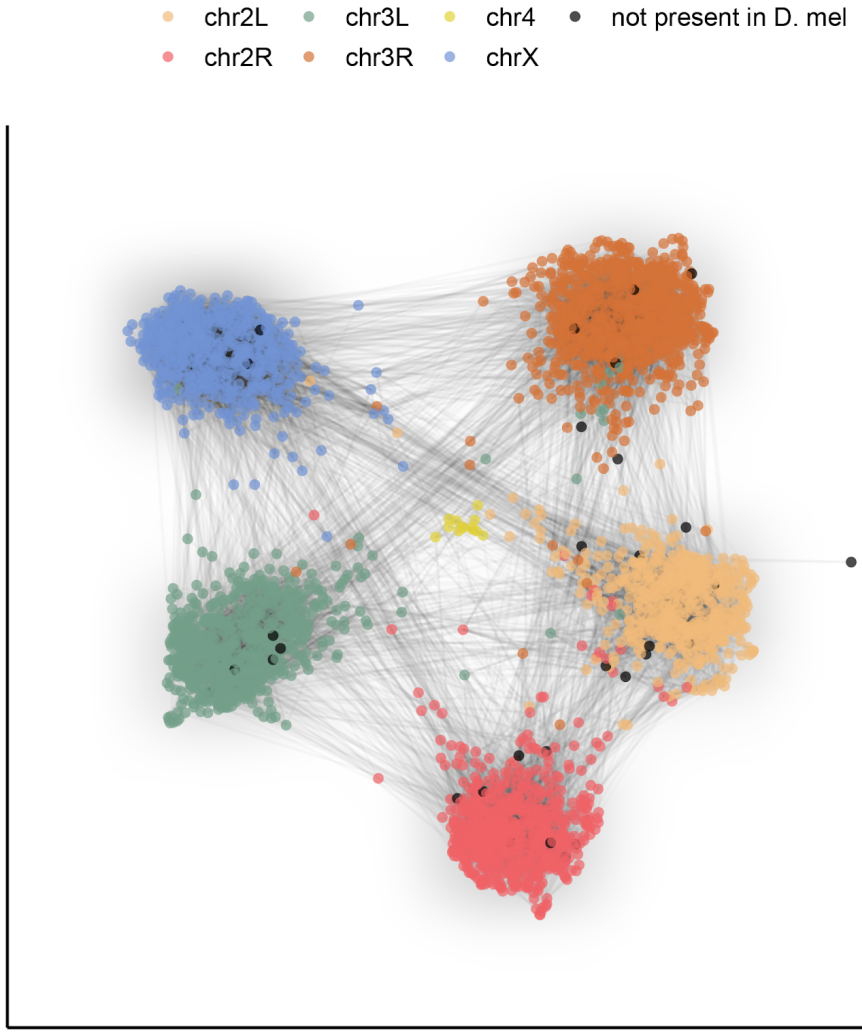

Figure 2. Gene content of Muller elements is conserved across drosophilids while gene order changes. Each node in this graph represents an orthologous marker corresponding to single-copy orthologs annotated by BUSCOv4 (Seppey et al., 2019; Simão et al., 2015). An edge between two nodes represents the number of times that BUSCO pair is directly connected within an assembly. Each BUSCO is colored by the chromosome arm in *D. melanogaster* that it is found on. The ForceAtlas2 (Jacomy et al., 2014) graph layout algorithm was used for visualization.

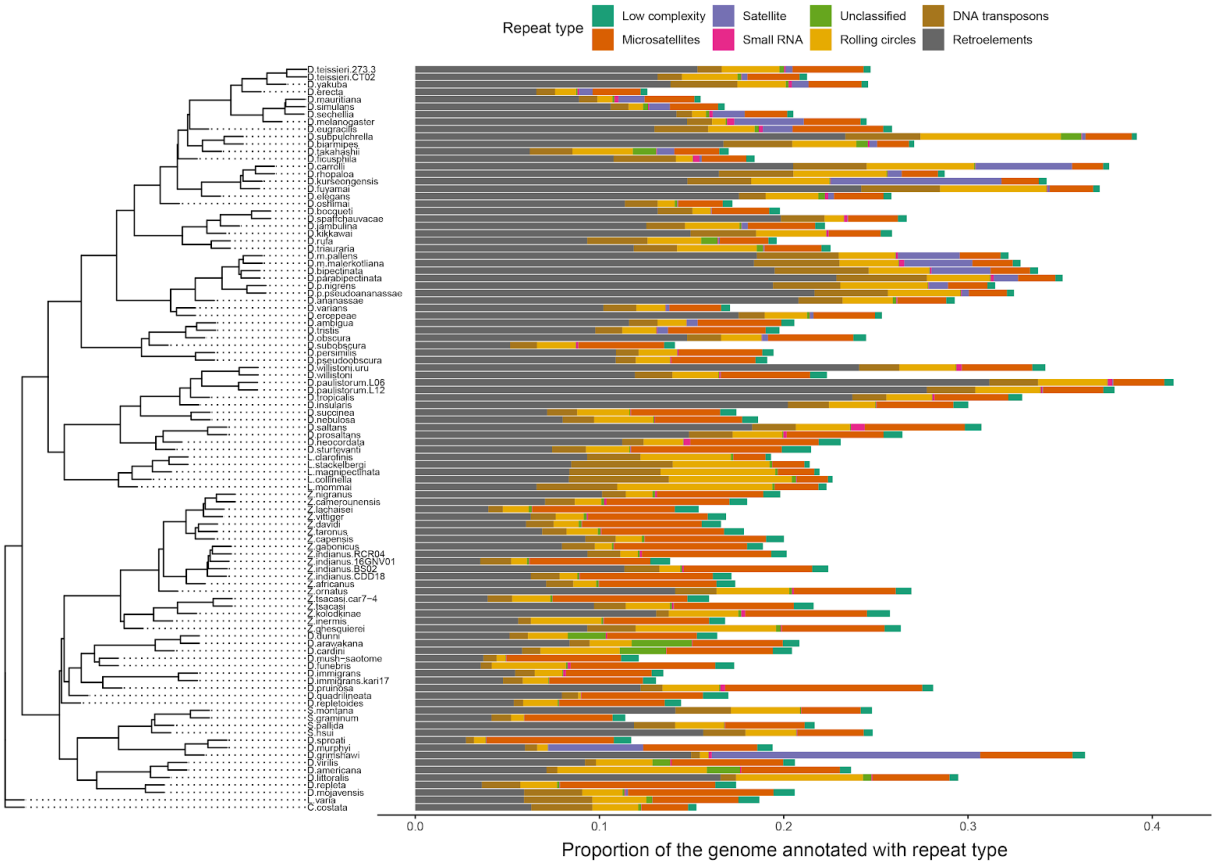

Figure 3. Repeat content varies greatly between drosophilid groups. For each species, the proportion of each genome annotated with a particular repeat type is depicted. Species relationships were inferred by randomly selecting 250 of the set of BUSCOs (Simão et al., 2015) that were complete and single-copy in all assemblies. RAxML-NG (Kozlov et al., 2019) was used to build gene trees for each BUSCO then ASTRAL-MP (Yin et al., 2019) to infer a species tree. Repeat annotation was performed with RepeatMasker (Smit et al., 2013) using the Dfam 3.1 (Hubley et al., 2016) and RepBase RepeatMasker edition (Bao et al., 2015) databases.

Tables

Table 1. Species and strain information for all samples assembled for this work. Note: Species group and subgroup information is taken from the NCBI Taxonomy Browser with slight modifications following O'Grady and DeSalle (2018). Strain names along with corresponding NDSSC and Kyoto DGRC stock center numbers are provided to the best of our knowledge. See **Tables S1 and S3** for detailed information on samples and data.

Subgenus	Group	Subgroup	Species	Sex	Strain name	NDSSC	Kyoto DGRC/ Ehime	Additional notes
Sophophora	*melanogaster*	*melanogaster*	*D. melanogaster*	MF	ISO-1 GENOME	14021-0231.36	NA	BDGP reference strain
			D. mauritiana	F	NA	14021-0241.01	NA	Miller et al. 2018
			D. simulans	F	NA	14021-0251.006	NA	Miller et al. 2018
			D. sechellia	F	NA	14021-0248.01	NA	Miller et al. 2018
			D. teissieri	M	273.3	NA	NA	
			D. teissieri	M	CT02	NA	NA	
			D. yakuba	F	NA	14021-0261.01	NA	Miller et al. 2018
			D. erecta	F	NA	14021-0224.01	NA	Miller et al. 2018
		eugracilis	*D. eugracilis*	F	NA	14026-0451.02	NA	Miller et al. 2018
		suzukii	*D. subpulchrella*	M	L1	NA	NA	
			D. biarmipes	MF	361.0 iso1 l-11 GENOME strain 1	14023-0361.10	NA	modENCODE strain
		takahashii	*D. takahashii*	F	IR98-3 E-12201	NA	E-912201	inbred derivative of Ehime stock IR98-3
		ficusphila	*D. ficusphila*	F	631.0-iso1 l-10 GENOME	14025-0441.05	NA	modENCODE strain
		rhopaloa	*D. carrolli*	MF	KB866	NA	NA	

			D. rhopaloa	MF	BaVi067 GENOME	14029-0021.01	E-24701	modENCODE strain
		rhopaloa	*D. kurseongensis*	F	SaPa58	NA	NA	
			D. fuyamai	F	KB-1217	14029-0011.01	NA	
		elegans	*D. elegans*	F	HK0461.03 GENOME	14027-0461.03	NA	modENCODE strain
		suzukii	*D. oshimai*	M	MT-04	NA	NA	
		montium	*D. bocqueti*	M	YAK3_mont-66	NA	NA	
			D. sp aff chauvacae	M	mont_up-71	NA	NA	
			D. jambulina	MF	st-2	14028-0671.01	NA	
			D. kikkawai	F	561.0-iso4 l-10 GENOME	14028-0561.14	NA	modENCODE strain
Sophophora	*melanogaster*		*D. rufa*	F	EH091 iso-C L_3	NA	914802	inbred derivative of Ehime stock EH091
			D. triauraria	F	NA	14028-0691.9	NA	Miller et al. 2018; previously mis-identified as *D. kikkawai*
		ananassae	*D. malerkotliana pallens*	F	palQ-isoG	NA	NA	
			D. malerkotliana malerkotliana	MF	mal0-isoC	14024-0391.00	NA	inbred derivative of strain 14024-0391.00
			D. bipectinata	MF	4-4-2-3-1-1-1-1-1 BackUp	14024-0381.04	NA	Inbred derivative of NDSSC strain
			D. parabipectinata	MF	par2-isoB	14024-0401.02	NA	inbred derivative of strain 14024-0401.02 (now extinct)
			D. pseudoananassae pseudoananassae	F	Wau 125	NA	NA	
			D. pseudoananassae nigrens	F	VT04-31	NA	NA	
			D. ananassae	F	14024-0371.13	NA	NA	Miller et al. 2018
			D. varians	MF	CKM15-L1	NA	NA	
			D. ercepeace	MF	164-14	14024-0432.00	NA	

Sophophora	*obscura*	*obscura*	*D. ambigua*	M	R42	NA	NA	isofemale strain from the wild
			D. tristis	M	D2	NA	NA	isofemale strain from the wild
			D. obscura	M	BZ-5	NA	NA	isofemale strain from the wild
			D. subobscura	M	Küsnacht	NA	NA	standard laboratory strain
		pseudoobscura	*D. persimilis*	F	NA	14011-0111.01	NA	Miller et al. 2018
			D. pseudoobscura	F	NA	14011-0121.94	NA	Miller et al. 2018
	willistoni	*willistoni*	*D. willistoni* L17	M	L-G3	14030-0811.17	NA	
			D. willistoni	F	NA	14030-0811.00	NA	Miller et al. 2018
			D. paulistorum L06	M	(Heed) H66.1C	14030-0771.06	NA	
			D. paulistorum L12	M	L12	14030-0771.12	NA	
			D. tropicalis	M	(Heed) H65.2	14030-0801.00	NA	
			D. insularis	M	jp01i	NA	NA	isofemale line from J. Powell
		bocainensis	*D. sucinea*	M	49.15	14030-0791.01	NA	
			D. nebulosa	M	H176.10	14030-0761.01	NA	
	saltans	*saltans*	*D. saltans*	M	(Heed) H180.40	14045-0911.00	NA	
			D. prosaltans	M	(Heed) H29.6	14045-0901.02	NA	
		neocordata	*D. neocordata*	M	2536.7	14041-0831.00	NA	
		sturtevanti	*D. sturtevanti*	F	H191.23	14043-0871.01	NA	
	Lordiphosa	*miki*	*L. clarofinis*	MF	Guizhou062018LC	NA	NA	Line inbred for 2 generations in the lab before sequencing
			L. stackelbergi	MF	UCILTSSapporo052019LS	NA	NA	Pool of 50 wild-caught flies
			L. magnipectinata	MF	UCKTSapporo052019LM	NA	NA	Pool of 50 wild-caught flies

Sophophora		*fenestrarum*	*L. collinella*	MF	UCKTSapporo052019LC	NA	NA	Pool of 30 wild-caught flies
			L. mommai	MF	MMSapporo052014LM	NA	NA	
Drosophila	*Zaprionus*	*vittiger*	*Z. nigranus*	M	st01n	NA	NA	line derived from wild collection
			Z. camerounensis	M	jd01cam	NA	NA	isofemale line from J. David
			Z. lachaisei	M	jd01l	NA	NA	line derived from wild collection
			Z. vittiger	M	jd01v	NA	NA	isofemale line from J. David
			Z. davidi	M	jd01d	NA	NA	isofemale line from J. David
			Z. taronus	M	st01t	NA	NA	line derived from wild collection
			Z. capensis	M	jd01cap	NA	NA	isofemale line from J. David
			Z. gabonicus	M	jd01gab	NA	NA	isofemale line from J. David
			Z. indianus RCR04	M	RCR04	NA	NA	
			Z. indianus 16GNV01	M	16GNV01	NA	NA	
			Z. indianus BS02	M	BS02	NA	NA	
			Z. indianus CDD18	M	CDD18	NA	NA	
			Z. africanus	M	BS06	NA	NA	
			Z ornatus	M	jd01o	NA	NA	isofemale line from J. David
		tuberculatus	*Z. tsacasi* car7-4	M	car7-4	NA	NA	
			Z. tsacasi	M	jd01t	NA	NA	isofemale line from J. David
		inermis	*Z. kolodkinae*	M	jd01k	NA	NA	isofemale line from J. David
			Z. inermis	M	18BSZ10	NA	NA	

			Z. ghesquierei	M	jd01ghe	NA	NA	isofemale line from J. David
Drosophila	*cardini*	*dunni*	*D. dunni*	M	H254.21	15182-2291.00	NA	
			D. arawakana	M	MONHI050227(B)-104	15182-2261.03	NA	
		cardini	*D. cardini*	M	NA	15181-2181.03	917701	
	funebris	*funebris*?	*undescribed* (Sao Tome mushroom)	M	st01m	NA	NA	undescribed species collected on mushroom, Sao Tome
		funebris	*D. funebris*	M	fst01	NA	NA	line derived from wild collection
	immigrans	*immigrans*	*D. immigrans*	F	FK05-19	15111.1731.12	NA	
			D. immigrans kari17	M	kari17	NA	NA	
		(incertae sedis)	*D. pruinosa*	M	iso-A1 I-9	NA	NA	
	tumiditarsus	*quadrilineata*	*D. quadrilineata*	M	quad-TMU	NA	914402	
			D. repletoides	M	ISZ-isoB I-10	NA	NA	
	Scaptomyza	*Scaptomyza*	*S. montana*	MF	iso-CA-L1	NA	NA	
			S. graminum	F	TMU-2019	NA	NA	30 wild-caught females
		Parascaptomyza	*S. pallida*	MF	iso-CA-L1	NA	NA	
		Hemiscaptomyza	*S. hsui*	MF	iso-CA-L1	NA	NA	
	Hawaiian *Drosophila*	*orphnopeza*	*D. sproati*	MF	DKPTOMS02	NA	NA	Pool of wild-caught flies
			D. murphyi	MF	DKPHETFM01	NA	NA	Flies from recently established but not inbred lab line
	virilis	*grimshawi*	*D. grimshawi*	F	NA	15287-2541.00	NA	Same line as caf1 genome
		virilis	*D. virilis*	F	NA	15010-1051.87	NA	Miller et al. (2018)

Drosophila	*repleta*	*virilis*	*D. americana*	M	3367.1	15010-0951.00	NA	Also called Anderson strain
			D. littoralis	M	Kilpisjärvi 1	NA	NA	Originally misidentified as *D. ezoana* (Lankinen 1986, J Comp Physiol A 159: 123-142)
		repleta	*D. repleta*	M	kari30	NA	NA	
		mulleri	*D. mojavensis*	F	15081-1352.22	NA	NA	Miller et al. (2018)
genus: *Leucophenga*			*L. varia*	M	nc01v	NA	NA	Sequenced single wild-caught fly, no amplification
genus: *Chymomyza*			*C. costata*	M	Sapporo	NA	NA	