Title

One fly—one genome: chromosome-scale genome assembly of a single outbred Drosophila melanogaster

Journal

Nucleic Acids Research, 48(13)

ISSN

0305-1048

Authors

Adams, Matthew
McBroome, Jakob
Maurer, Nicholas
et al.

Publication Date

2020-07-27

DOI

10.1093/nar/gkaa450

Peer reviewed

Published online 3 June 2020

Nucleic Acids Research, 2020, Vol. 48, No. 13 ***e75***
doi: 10.1093/nar/gkaa450

One fly–one genome: chromosome-scale genome assembly of a single outbred *Drosophila melanogaster*

Matthew Adams [1,*], Jakob McBroome [2], Nicholas Maurer[2], Evan Pepper-Tunick[2], Nedda F. Saremi[2], Richard E. Green[2,3,4], Christopher Vollmers[2,3,*] and Russell B. Corbett-Detig[2,3,*]

[1]Department of Molecular, Cellular, and Developmental Biology, University of California Santa Cruz, Santa Cruz, CA 95064, USA, [2]Department of Biomolecular Engineering, University of California Santa Cruz, Santa Cruz, CA 95064, USA, [3]UCSC Genomics Institute, University of California Santa Cruz, Santa Cruz, CA 95064, USA and [4]Dovetail Genomics, Scotts Valley, CA 95066, USA

Received January 17, 2020; Revised April 16, 2020; Editorial Decision May 13, 2020; Accepted May 18, 2020

ABSTRACT

A high quality genome assembly is a vital first step for the study of an organism. Recent advances in technology have made the creation of high quality chromosome scale assemblies feasible and low cost. However, the amount of input DNA needed for an assembly project can be a limiting factor for small organisms or precious samples. Here we demonstrate the feasibility of creating a chromosome scale assembly using a hybrid method for a low input sample, a single outbred *Drosophila melanogaster*. Our approach combines an Illumina shotgun library, Oxford nanopore long reads, and chromosome conformation capture for long range scaffolding. This single fly genome assembly has a N50 of 26 Mb, a length that encompasses entire chromosome arms, contains 95% of expected single copy orthologs, and a nearly complete assembly of this individual's *Wolbachia* endosymbiont. The methods described here enable the accurate and complete assembly of genomes from small, field collected organisms as well as precious clinical samples.

INTRODUCTION

The creation of high quality genome assemblies is a key step for the study of organisms on both the level of individuals and populations (1). Conventional genome sequencing projects rely on whole-genome shotgun sequencing approaches that generate huge numbers of short sequence reads at low cost. While short reads can be reassembled into larger contiguous genome segments by identifying overlapping reads, they often fail to generate chromosome length assemblies due to the challenge of assembling repetitive DNA sequences. Consequently, many published genomes are highly fragmented (2). Fragmented genomes can be valuable for gene-level studies but many genomic analyses such as understanding chromosome-scale evolution, resolving full-length haplotypes, association studies, and quantitative trait locus mapping require high-quality chromosome-scale assemblies. New hybrid genome assembly approaches can produce highly contiguous assemblies that represent true chromosome length genomes (3).

Two recent advances in genomic technologies have dramatically raised the quality of genome assemblies (4). First, third generation long-read sequencing technologies are capable of sequencing entire long repetitive sequences, but they suffer from higher error rates and lower throughput (2). Second, proximity-ligation sequencing, or Hi-C, produces short-read pairs representing sequences that are close together in three-dimensional space (5). This allows high throughput 'scaffolding' of challenging genomic regions (6). However, these impressive gains in genome assembly quality have not been realized across all species due to important biological constraints.

Genome projects can be complicated by the small size of many organisms, which yield corresponding low amounts of DNA from a single individual. Consequently it is not always feasible to obtain sufficient input material for the genomic approaches described above without pooling individuals (7). Nonetheless, developing applications for single individual genome assemblies offers several key advantages. First, it may not be possible to obtain more than a single individual for some species. Second, even if many could be

*To whom correspondence should be addressed. Email: msadams@ucsc.edu
Correspondence may also be addressed to Russell B. Corbett-Detig. Email: rucorbet@ucsc.edu
Correspondence may also be addressed to Christopher Vollmers. Email: vollmers@ucsc.edu

found, pooling several individuals increases the genetic diversity in the DNA input, imposing challenges for accurate genome assembly. For wild caught samples, the possibility of combining cryptic species has the potential to impact assembly quality and introduce spurious biological conclusions. Finally, low input sequencing methods could be used to assemble genomes from precious clinical samples. There is therefore a clear need for new methods that can assemble highly contiguous genomes from a single isolate with limited available DNA.

Recently, Kingan *et al*. released a whole-genome assembly obtained from a single mosquito, *Anopheles coluzzii,* sequenced using three PacBio SMRT Cells (8). Although the assembly has high contiguity (contig N50 3.5 Mb), the authors were unable to obtain chromosome-scale contigs or scaffolds and the resulting assembly does not include biologically important regions of the genome that contain chromosomal inversion breakpoints (8,9). Additionally, the input material used, approximately 100 ng of high quality DNA, may still be challenging to obtain from a single field-collected individual in many species. Nonetheless, this pioneering work suggests a powerful solution in developing low-input protocols for simultaneously obtaining Hi-C and long-read data from single individuals.

Here, we present a chromosome scale hybrid genome assembly of a single *Drosophila melanogaster* female. From this single individual, we produce long reads, short reads and proximity ligation sequencing data. Our assembly approach leverages the unique value added by each data type to produce a chromosome-scale and accurate genome assembly. This approach is applicable for millions of small species and for irreplaceable clinical samples.

MATERIALS AND METHODS

DNA extraction

High molecular weight DNA was extracted from one half of a single *D. melanogaster* female using a Qiagen MagAttract HMW DNA kit. One half of a single fly was placed in a 1.5 ml tube with lysis buffer and proteinase k then crushed with a pestle using an up and down motion as to not shear DNA. The lysis and proteinase k digestion was incubated overnight at 37°C. The rest of the purification was performed according to the manufacturer's protocol. The total amount of DNA recovered was 104.4 ng measured with a Thermo Fisher Qubit fluorometer and Qubit dsDNA HS assay kit. This sample was subsequently used for the Tn5 and nanopore library prep.

Illumina short-insert Tn5 sequencing

From the HMW DNA sample, 10 ng of gDNA was tagmented with Tn5 transposase for 8 minutes at 55°C. The reaction was halted by adding 0.2% SDS and incubated at room temperature for 7 min. Four separate PCR reactions were set up using the KAPA Biosystems HiFi Polymerase Kit and amplified for 16 cycles using uniquely indexed i5 and i7 primers. The amplified libraries were pooled and purified using the $\geq$ 300 bp cutoff on the ZYMO Select-a-Size DNA Clean and Concentrator Kit. 500 ng of the purified library pool was run on a Thermo Fisher 2% E-Gel EX Agarose Gel and cut between 550 and 800 bp. The gel cut was purified with the NEB Monarch DNA Gel Extraction Kit and quantified using the Qubit dsDNA HS Assay Kit and the Agilent TapeStation.

Nanopore sequencing

From the HMW DNA sample, 78.3 ng was used as input. The sample was first sheared using a Covaris g-TUBE centrifuged for 30 s at 8600 RCF. The sheared DNA was size selected using Solid Phase Reversible Immobilization (SPRI) beads at 0.7 beads:1 sample ratio and eluted in 25 ul ultrapure water.

End repair and A-tailing was performed using NEBNext Ultra II End Repair/dA-Tailing Module followed by ligation of Nextera adapters using NEB Blunt/TA Ligase Master Mix following the manufacturer's protocol. The adaptor ligated sample was purified by SPRI beads at a 1:1 ratio and eluted in 50 ul of ultrapure water. The sample was divided into six, 25 ul PCR reactions with Nextera primers and KAPA HiFi Readymix 2× (95 C for 30 s, followed by 12 cycles of 98°C for 10 s, 63°C for 30 s 72°C for 6 min, with a final extension at 72°C for 8 min then hold at 4°C). The PCR reactions were pooled and purified by SPRI beads at a 1:1 ratio and eluted in 60 ul of ultrapure water. Concentration was measured to be 110 ng/ul using the Qubit dsDNA HS assay. The entire sample was size selected by gel electrophoresis using a 1% low melting agarose gel. An area from 6–10 kb was cut out and digested using NEB Beta Agarase I following the manufacturer's protocol then purified using SPRI beads at a 1:1 ratio.

One hundred nanograms of size selected DNA was mixed with 50 ng of a DNA splint and circularized by Gibson assembly using 2× NEBuilder HiFi DNA Assembly Master Mix incubated for 60 min at 50°C. Non circularized DNA was digested overnight at 37°C using Exonuclease I, Exonuclease III and Lambda Exonuclease (all NEB). Circularized DNA was purified by SPRI beads at a 0.8:1 ratio and eluted in 40 ul of ultrapure water.

The circularized DNA was split into 8 50 ul rolling circle amplification (RCA) reactions (5 ul 10× Phi29 buffer (NEB), 2.5 ul 10 mM dNTPs (NEB), 2.5 ul 10 uM exonuclease resistant random hexamer primers (Thermo), 5 ul DNA, 1 ul Phi29 polymerase (NEB), 34 ul ultrapure water). Reactions were incubated overnight at 30°C. All reactions were pooled and debranched using T7 Endonuclease (NEB) for 2 h at 37°C. To shear ultra-long RCA products the sample was run through a Zymo Research DNA Clean and Concentrator-5 column and eluted in 40 ul ultrapure water. A final size selection was performed by gel electrophoresis using a 1% low melting agarose gel. An area at approximately 10 kb was cut out and digested using NEB Beta Agarase I following the manufacturer's protocol then purified using SPRI beads at a 1:1 ratio.

The cleaned and size selected RCA product was sequenced using the ONT 1D Genomic DNA by Ligation sample prep kit (SQK-LSK109) and a single MinION flow cell following the manufacturer's protocol. The raw data was basecalled using the Guppy basecaller. Consensus reads were generated by Concatemeric Consensus Caller with Partial Order alignments (C3POa).

HiC library

The anterior half of the fly was placed into a 1.5 ml tube with 1 ml of cold 1× PBS. 31.25 ul of 32% paraformaldehyde was added. The sample was briefly vortexed and incubated for 30 min at room temperature with rotation. After incubation the supernatant was removed and washed twice with 1 ml of cold 1× PBS. 50 ul of lysate wash buffer was added before grinding with pestle. 5 ul of 20% SDS was added then vortexed for 30 s and incubated at 37°C for 15 min with shaking. 100 ul of SPRI beads were added to bind chromatin. Bound sample was washed 3 times with SPRI wash buffer.

Beads were resuspended in 50 ul of Dpn II digestion mix (42.5 ul water, 5 ul 10× DpnII buffer, 0.5 ul 100 mM DTT, 2 ul DpnII) and digested for 1 h at 37°C with shaking. Beads were washed twice with SPRI wash buffer and resuspended in 50 ul of end fill-in mix (37 ul water, 5 ul 10× NEB Buffer 2, 4 ul 1 mM biotin-dCTP, 1.5 ul 10 mM dATP dTTP dGTP, 0.5 ul 100 mM DTT, 2 ul Klenow fragment) then incubated for 30 min at room temperature while shaking. Beads were washed twice with SPRI wash buffer and resuspended in 200 ul of intra-aggragete mix (171 ul water, 1 ul 100 mM ATP, 20 ul 10× NEB T4 DNA Ligase Buffer, 1 ul 20 mg/ml BSA, 5 ul 10% Triton X-100, 2 ul T4 DNA ligase) then incubated at 16°C overnight while shaking. Beads were placed on a magnet to remove supernatant then resuspended in 50 ul of crosslink reversal buffer (48.5 ul crosslink reversal mix, 1.5 ul proteinase K) then incubated for 15 min at 55°C, followed by 45 min at 68°C while shaking. Beads were placed on a magnet and the supernatant was transferred to a clean 1.5 ml tube. 100 ul of SPRI beads were added to the supernatant and allowed to bind before washing twice with 80% ethanol and eluting sample with 50 ul of 1× TE buffer.

The sample was then fragmented by sonication. Fragmented sample was end repaired and adapter ligated using the NEBNext Ultra II kit following the manufacturer's protocol. The sample was purified from ligation reaction by SPRI beads, washed twice with 80% ethanol, and eluted in 30 ul of 1× TE. Biotin tagged fragments were enriched using streptavidin C1 Dynabeads. Enriched fragments were indexed by PCR (23 ul water, 25 ul 2× Kapa mix, 1 ul 10 uM i7 index primer, 1 ul 10 uM i5 index primer) and amplified for 11 cycles. Reaction was purified by SPRI beads and quantified using the Qubit dsDNA HS Assay Kit and the Agilent TapeStation.

Assembly

We produced short-read assemblies using the variation-aware *de Bruijn* graph algorithm, Meraculous (10). Long-read data was assembled using Wtdbg2 (11) using the following options 'wtdbg2 -x ont -g 120m -p 0 -k 15 -S 1 -l 512 -L 1024 –edge-min 2 –rescue-low-cov-edges' followed by the wtdbg2 consensus caller wtpoa-cns (11). The two primary long and short-read assemblies were combined using quickmerge default merge_wrapper.py command.

Scaffolding

We polished the hybrid shotgun and long-read assembly using the Illumina shotgun dataset using the bwa mem algorithm (version 0.7.17) (12) to map the Illumina reads back to the genome and samtools (version 1.7) to sort the reads. We input the sorted alignment to the consensus for wtdbg (wtpoa-cns) (version 2.5) using the command '-x sam-sr' to polish the contigs of the hybrid assembly. We scaffolded the polished assembly using the scaffolding tool HiRise (version 2.1.1) run in Hi-C mode using the default parameters with the Hi-C library as input. After the first round of scaffolding, we sought to remove putative misjoins in our assembly. To do this, we computed the insulator score across the genome using a 1Mb window on either side of a focal test point. We obtained the expected insulation score for a misjoin between two unlinked contigs by computing the same metric for artificial false-joins between random pairs of unlinked contigs. We then broke the assembly at one aberrantly low insulation score site—indicating little Hi-C support for a specific join consistent with our between contig comparisons.

Polishing

The draft assembly went through a total of four iterative rounds of polishing using the automated software tool Pilon using default settings. For each round the short and long-read data was mapped to the draft assembly using minimap2. After each round, the assembly was evaluated for misassemblies, indels, mismatches, N50, and assembly size using QUAST (13) to determine if further polishing would increase the assembly correctness.

Evaluation

To evaluate the completeness of the H3 assembly we searched for conserved genes using Benchmarking Universal Single-Copy Orthologs v3, (BUSCO) with the metazoa odb9 lineage gene set (14). To compare to the current reference genome we used the genome quality assessment tool QUAST using the '–large –k-mer-stats' options (13). Misassemblies are defined by the following criteria, a position in the assembled contigs where (i) the left flanking sequence aligns over 1 kbp away from the right flanking sequence on the reference, (ii) flanking sequences overlap on >1 kb, (iii) flanking sequences align to different strands or different chromosomes. Local misassemblies are defined by the following criteria (a) the gap or overlap between left and right flanking sequences is less than 1 kbp, and larger than the maximum indel length (85 bp), (b) the left and right flanking sequences both are on the same strand of the same chromosome of the reference genome.

Repetitive and genic region coverage analysis

We aligned three separate versions of H3 assembly with zero, one, and two rounds of polishing with Pilon to the *Drosophila melanogaster* reference using Minimap2 with default parameters and sam output (15,16). We then applied samtools compression and sorting to produce sorted bam files (17,18), to which we applied bedtools genomecov with options -ibam and -bga to produce a file of region coordinates and coverage values of 0 or more for each region across the genome (17,18). We combined this information with the annotation gff3 file with a custom script that assigned coverage values to all annotated spans base by base

(18). The average coverage per base was calculated for each annotated span, then the average and mean value of coverages for all spans for each annotation type was calculated. As a control for comparison we performed this procedure on a complete non-reference *melanogaster* assembly and calculated similar values to elucidate any particular weakness our assembly exhibits.

Phasing

To phase the genome, we realigned all short-read data to our final genome assembly using BWA mem (19). We then called all heterozygous variants using GATK (20) on the four largest scaffolds in our assembly, and we filtered this set to exclude SNPs and indels in the bottom 10% or top 10% of observed sequencing depths. As the H3 genome is a mosaic of I38 and dm6 alleles, we 'polarized' each heterozygous variant by realigning the dm6 genome using minimap2 (16) to determine whether H3 contained the dm6 allele. We then aligned all Hi-C data using BWA mem (19) and the ONT data using minimap2 (16) and attempted to phase the genome using varying combinations of these data using hapcut2 (21). We quantified mismatch and switch errors as described in (21).

RESULTS

Sample selection

Although numerous studies have assembled genomes from completely (22) or partially (8) inbred arthropods, the genomes of a field collected samples will likely be highly heterozygous outbred individuals. To make our assembly task conservatively challenging yet straightforward to evaluate, we generated an outbred fly by crossing females of the *D. melanogaster* reference strain *y; cn, bw; sp* or ISO1 (22), to males of another inbred and genetically distinct strain, I38 (23). Importantly, I38's genome is collinear with the reference on broad scales, although smaller rearrangements, such as small-scale indels and copy number variants, are almost certainly present in the genome (23,24). We can therefore use progeny from this cross to demonstrate the applicability of our method for assembling genomes of outbred field-collected arthropod individuals and we can easily verify the accuracy of the assembly by comparison to the ISO1 reference genome. To facilitate the use of several sequencing methods, the single outbred fly chosen for sequencing (referred to as H3) was first laterally dissected (Figure 1).

Primary sequencing datasets

From a single outbred adult female fly, we produced short-read shotgun, long-read shotgun and Hi-C libraries (Figure 1). From the posterior half, we extracted high molecular weight (HMW) DNA and we obtained approximately 104 ng in total. We used 78 ng to produce an Oxford Nanopore Technology (ONT) sequencing library following the R2C2 protocol (25) with slight modification for genomic DNA (see Methods). The R2C2 protocol generated ONT raw reads that contain tandem repeats of *Drosophila* genomic DNA sequence separated by splint sequences. The R2C2 post-processing pipeline (C3POa) processes these raw reads and generates two types of output reads: (i) consensus reads are generated if an ONT raw read is long enough to cover an insert sequence more than once which is evaluated by detecting a splint sequence in the raw read and (ii) regular '1D' reads for which no splint could be detected in the raw read. In total, 277 305 consensus reads and 1 769 380 '1D' reads were generated from a single ONT MinION flow cell. Both read types were included in the assembly. We additionally produced an Illumina sequencing library using a standard Tn5-based protocol (Materials and Methods) and from this we obtained 133 135 777 total paired-end reads (Table 1).

Because both R2C2 and our Tn5 protocol are optimized for low DNA inputs, they require some amplification to produce suitable quantities of libraries for high throughput sequencing. Likely as a consequence, the variance in sequencing depth exceeds the theoretically expected variance if reads were sampled uniformly at random from the genome. Indeed, for libraries with mean depths 236× and 39.7× we obtained depth variances of 8382 and 1038 for Tn5 and ONT respectively. Nonetheless, we show below that moderately long contigs can still be generated from these data (Supplementary Figure S1).

We also produced a Hi-C library to enable long-range scaffolding across the genome. We optimized a chromatin conformation capture sequencing method (5,26) for application to samples with minimal input materials (see Materials and Methods). Using this approach and just the anterior half of the fly, we were able to produce 68 400 787 reads in total from a Hi-C library (Table 1). This represents an average of approximately 93 991 clone coverage across the genome. Furthermore, despite low-input, the PCR duplication rate is quite modest (12%). These data therefore indicate that our single-fly Hi-C approach can produce high complexity libraries suitable for scaffolding high quality genomes.

Primary assemblies

To accommodate the unique features of each input data type we produced two primary assemblies. First, we assembled the short-read shotgun dataset using the heterozygosity aware *de Bruijn* graph-based algorithm Meraculous (10). As we are interested in assembling a single haploid genome sequence, we collapsed the program's resulting diplotigs into a single haploid assembly (i.e. 'diploid mode 1'). Second, we assembled the processed ONT reads using wtdbg2 (11) (Table 2). As expected given the substantially larger input read lengths, we obtained a much larger contig N50 using this program, than in our short-read based primary assembly (Table 2).

Merging primary assemblies

To combine the short and long-read primary assemblies we used the meta-assembler quickmerge. Quickmerge combines two input assemblies to produce an assembly with higher contiguity. Since the input assemblies come from the same individual, gaps in one assembly can be bridged by the other using the alignment of contigs from each input (27). The resulting merged assembly had a contig N50 of 274.6 kb (Table 2)

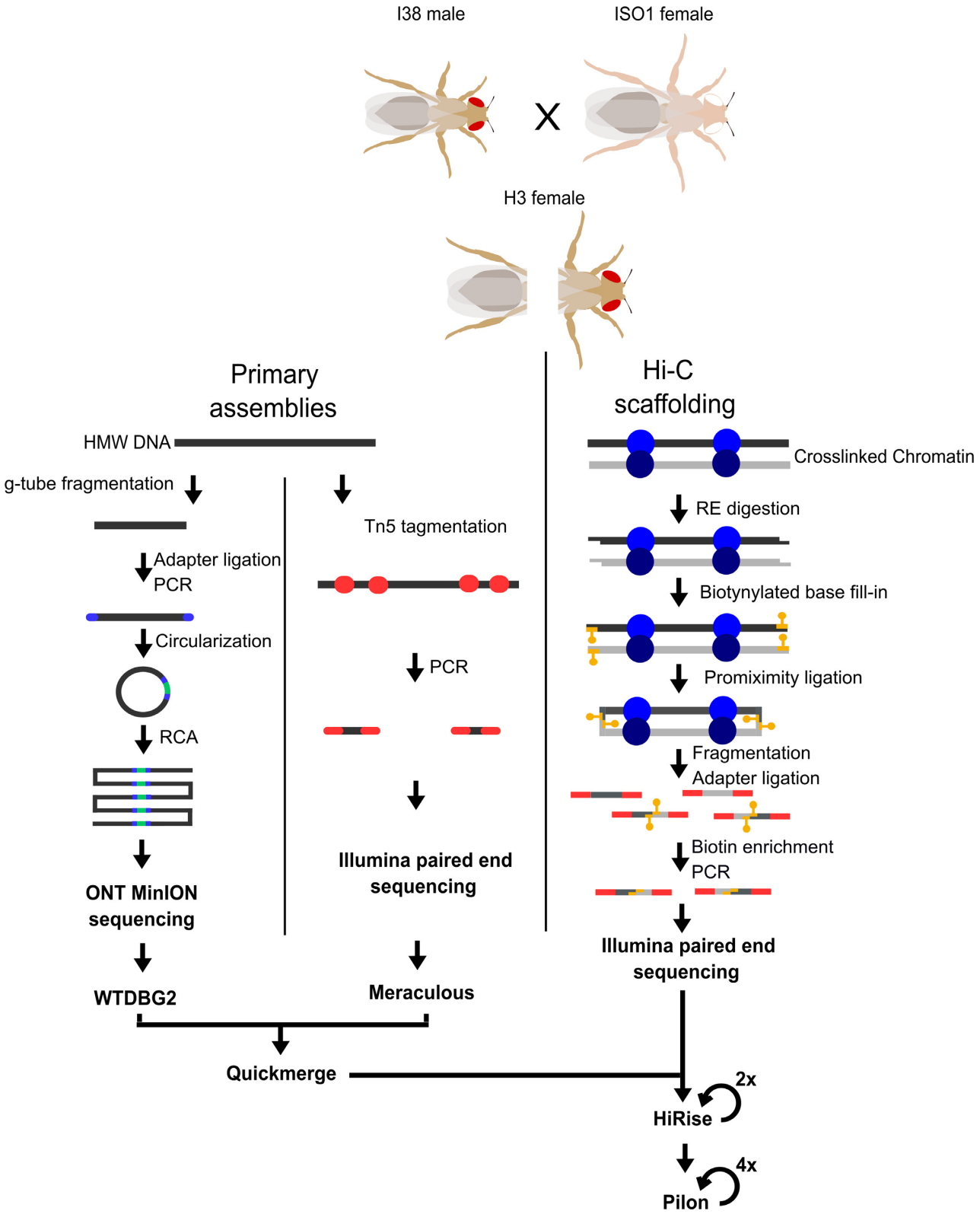

Figure 1. Experimental flow chart. A heterozygous fly (H3) was produced by crossing ISO1 and I38 strains. A single female offspring was laterally dissected. From the posterior half, HMW DNA was extracted and used to prepare the two primary assemblies, a R2C2 genomic library for nanopore sequencing, and a Tn5 tagmentation library for paired end Illumina sequencing. The anterior portion was used to isolate intact chromatin to generate a Hi-C paired end Illumina library. The two primary assemblies were merged into one then arranged into chromosome length scaffolds using the Hi-C contact frequency data.

Table 1. Summary of sequencing data used for assembly and scaffolding

Library	Total number of reads	Read length	Predicted coverage
Illumina Tn5	133 135 777	151 bp (paired end)	333×
ONT R2C2	2 046 685	3541 bp (median length)	60×
Illumina HiC	68 400 787	151 bp (paired end)	171×

Table 2. Summary of primary and scaffold assembly statistics.*Final assembly size of the H3 fly after removal of the endosymbiont *Wolbachia* genome (see section Genomic Bycatch)

	Contig N50 (kb)	Scaffold N50 (kb)	Assembly Size (Mb)
Meraculous	51	N/A	112.1
wtdbg2	97.7	N/A	112.3
Quickmerge	274.6	N/A	111.2
Hi-Rise	N/A	26 182	111.36
Pilon-Polishing	N/A	26 279	112.22
H3 Genome*	N/A	26 279	110.96

Scaffolding

Although the final merged primary assembly is reasonably contiguous, we observed by far the greatest gains in scaffold size after using our Hi-C data. We ran HiRise to scaffold the merged primary assembly and a single punitive misjoin was removed before rerunning HIRise a second time (see methods) from which we obtained a scaffold N50 of 26 Mb. Our final scaffolded assembly contains all the major chromosome arms in the *D. melanogaster* genome represented as single scaffolds, and correctly joins arms 2L and 2R across their heterochromatin-rich centromeric region (Figure 2). It therefore appears that the ability to produce high quality Hi-C libraries from extremely limited input material is the most essential component of our method for making contiguous genome assemblies for single individuals in small species.

Polishing and gap filling

Because we combined diverse data types, and in particular because our primary assembly relies on error-prone long reads, we sought to polish the contigs and fill gaps in the final highly contiguous assembly. In total we performed four rounds of iterative polishing with Pilon ((15), see Materials and Methods), until we did not observe significant additional improvements (Supplementary Table S2). The final assembly produced by this step, which we use for all validation below, is the largest of all of our assemblies at 112.2 Mb (110.96 Mb after removing *wolbachia* contigs), which presumably reflects the success in our polishing and gap filling by incorporating additional sequences.

Quality of the final assembly

We assessed our final assembly quality using several metrics. First, we applied the Benchmarking Universal Single-Copy Orthologs, BUSCO, algorithm (14). Briefly, the program provides an assessment of assembly quality specifically with respect to genic sequences by searching for a set of nearly-universal and single copy genes. In applying this quality metric we obtained a BUSCO score of 95.2% completeness for our final assembly. This is slightly lower than the current *D. melanogaster* ISO1 reference BUSCO score of 98.9%, but it is not dramatically different. We therefore conclude that the majority of the expected genic sequences are complete in our assembly.

Second, to compare the assembly of our H3 fly to the dm6 reference and quantify misassemblies we used the genome quality assessment tool QUAST (13). In addition, we used QUAST to compare another high quality assembly of a different *D. melanogaster* strain, A4 (28), to the dm6 reference to set a benchmark for the expected differences between genetically diverse strains (Table 3). Because A4 was completely inbred and independently isolated from ISO1, whereas our H3 sample is heterozygous for the ISO1 genome, our assembly should more closely match the reference genome. The reason is that we would expect the reference allele to be selected 50% of the time at non-reference sites, and we should therefore observe approximately half as many apparent differences in our final assembly as for A4 relative to the ISO1 reference genome. As expected, our assembly had substantially fewer misassemblies, mismatches and indels than the A4 strain when compared to the dm6 reference, likely because of the relatedness between ISO1 and our assembled individual.

Although our bioinformatic approach has produced a highly contiguous and accurate genome assembly, we acknowledge that alternative approaches might improve on our results. It is typically not possible to extensively optimize a bioinformatic pipeline including all possible variations. We therefore caution that this method should be considered guidelines for processing these types of data, but that researchers should evaluate them carefully for a given assembly task to ensure optimal results can be obtained.

Repeat content

Despite similar BUSCO scores and the modest rate of misassemblies that we observe, our genome assembly is ∼20% smaller than the canonical *D. melanogaster* reference genome. We suspected that much of the difference occurs because our assembly relies on relatively short reads and therefore collapsed repetitive regions. To evaluate this, we used the dm6 annotation data to evaluate coverage across different types of genomic features for both our single-fly assembly and a separate comparison of the A4 assembly. We found that while unique sequence including genes and especially exon sequences were captured in their entirety the majority of the time, highly duplicated elements such as transposons and tRNAs were much less likely to be covered by the H3 assembly (Table 4). This is a general weakness of short-read assemblies (29) and should be acknowledged by any forthcoming analysis applying this method of assembly.

Phasing

We next evaluated our prospects for phasing the genome of this outbred individual, i.e. assigning each heterozygous allele to a chromosome. To do this, we realigned our short-read data to our final genome assembly and called all heterozygous variants using GATK (20). We then realigned the

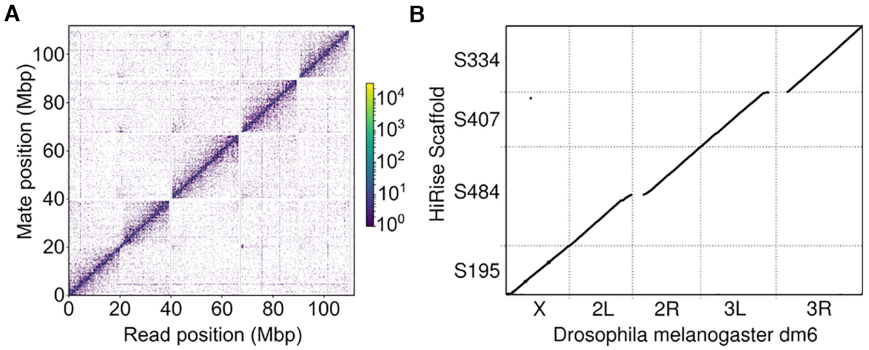

Figure 2. Genome Contiguity. (**A**) The read density map for Hi-C read pairs mapped onto the five largest contigs in our final assembly. (**B**) Dot plot of Hi-C scaffold assembly mapped to the dm6 reference genome. Continuous diagonal lines represent full length scaffolds of all major chromosome arms. For clarity of visualization, we restricted this plot to alignments of 5 kb or more using delta-filter in the mummerplot package.

Table 3. Summary of QUAST output comparing H3 and A4 assemblies to the dm6 reference genome

	H3 against dm6 reference	A4 against dm6 reference
# misassemblies	798	2309
# misassembled contigs	15	145
# local misassemblies	1251	3491
# mismatches per 100 kb	525.36	1136.97
# indels per 100 kb	88.7	118.84

Table 4. Sequence uniqueness strongly impacts assembly coverage

	H3 assembly	A4 assembly control
Coding sequence (CDS)	94.0%	97.9%
Exon	93.9%	99.5%
Long noncoding RNA	90.6%	98.8%
microRNA	93.7%	99.6%
tRNA	76.5%	98.7%
Mobile genetic elements	55.3%	82.0%

The columns are H3 assembly without any polishing and a non-reference control assembly of standard coverage and size. The rows are annotation types. The value corresponds to the percent of aligned annotated elements with at least 90% of their sequence captured in our assembly. The coverage distribution of our assembly is bimodal, with the vast majority of elements being either covered by a single assembled contig or not covered at all. An expanded table including more annotation types and counts, polished versions of the assembly, and overall assembly statistics can be found in the supplement (Supplementary Table S1).

Hi-C and long-read data as well and attempted to infer the phase using combinations of these data and the Hapcut2 algorithm (21). Because our individual is outbred and we know the complete genome sequence of both ancestors, it is straightforward to quantify the phase accuracy.

Using just the short-read data to phase heterozygous SNPs in the H3 individual, we achieve a modest combined mismatch and switch error rate (*sensu* (21)) of 0.00147 errors/site. Briefly, mismatch errors denote sites where single variants are phased incorrectly in an otherwise correct block and switch errors denote a change where at least two subsequent variants are phased incorrectly relative to preceding sites. However the mean phase block length is just 14 heterozygous variants or ∼2 kb. When we incorporated our Hi-C data, the combined error rate increased to 0.0147 error/site, but nearly entire chromosomes' variants were included in a single phase block (i.e. 99.95% of variants/chromosome). The addition of Hi-C increased switch errors in particular by 0.0126 errors/site. This is likely a consequence of somatic chromosome pairing in dipterans (30), which has previously been demonstrated to create an excess of sister chromosome contacts in Hi-C data (9,31). The increased switch error rate suggests that ∼17% of Illumina-phasable blocks that are joined by the addition of Hi-C result in switch errors. Therefore, phase inferred from these data could be useful across relatively short distances (e.g. 5 kb), but should be regarded with caution at larger genomic distances. This might not be suitable for all applications of phasing, but would be sufficient for many population genetic questions that rely on short-distance haplotype and linkage information.

Genomic bycatch

Although not a primary consideration in this work, we found that our assembly captures additional material that is potentially of interest and underscores the power of our approach. First, our selected individual was phenotypically female, nonetheless, we discovered a non-trivial rate of Y-chromosome mapping contigs. Importantly, we found a similar Y-mapping rate in all three raw sequencing datasets (Supplementary Table S3), and the relevant Y:Autosome depth closely resembles that of typical phenotypic males (unpublished data). We therefore believe this is an XXY female. Despite the abundance of Y-derived reads, our Y chromosome assembly is exceedingly fragmented, as most Y chromosome assemblies are, reflecting the challenges of assembling extremely repeat-dense chromosomes (32). Nonetheless, this finding highlights the value of sequencing individuals rather than pools because pooling would likely obscure this relationship of relative chromosome depths.

Second, the reference strain is known to harbor the symbiotic bacteria *Wolbachia*, as we used this as the female parent in the cross *Wolbachia* is present in our sample due to infected embryos. Despite the differences in read-depths relative to the nuclear genome, our assembly includes nearly full coverage of the *Wolbachia* genome with few apparent misassemblies (Figure 3 and Supplementary Figure S2). *Wolbachia* in particular (33), and endosymbionts more generally (34), are frequently present in host somatic tissues, likely

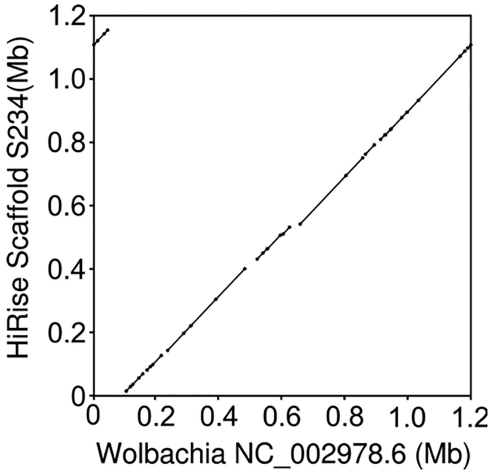

Figure 3. Dot-plot comparison of our nearly-complete *Wolbachia* assembly to the canonical wMel *Wolbachia* genome sequence. Note that the apparent discontinuity in the top right/left, reflects the circular nature of the bacterial genome, and simply indicates that our assembly breaks the circle at a slightly different place.

explaining the similar abundances of *Wolbachia*-derived reads across sequencing libraries prepared from different parts of the fly. This suggests that in addition to nearly complete nuclear genomes, our assembly method might also be a powerful tool for investigating individual's endosymbiont communities – a fundamental consideration in arthropod biology (35). Additionally, the analysis of a single individual obviates important concerns about pooling for interpreting inter-strain endosymbiont diversity (as in, (36)), and again emphasizes the potential impact of this approach. See also, Kingan *et al*. for a related approach assembling complete endosymbiont genomes from the genomic data of a single insect (37).

DISCUSSION

Recent advances in technology have greatly increased the quality of genome assemblies but generally require a relatively large DNA input. This limitation reduces the applicability of these methods for many precious, rare, and/or field collected specimens. Here, from a single fly we were able to construct a chromosome scale genome assembly with an N50 of 26 Mb. The primary assemblies were made with less than 90 ng of total input DNA. Therefore, our approach demonstrates that high quality chromosome-scale assemblies can be obtained from limited sample inputs.

Our method also compares favorably for total cost outlay. The DNA isolation and library preparation involves only basic molecular biology methods and equipment. We produced all necessary sequencing data on approximately one half of a HiSeq 4000 lane and a single MinION flow cell. We can therefore produce a contiguous, high quality genome for approximately $1,200 in total materials and reagent costs. For cost effectiveness, our approach compares quite favorably with available alternatives such as Pacbio SMRT cells at $2,000 each.

There are many genome assembly approaches available, and ours may not be optimal for all applications. When input materials are severely limited, the approach we describe here provides an appealing set of trade-offs and may be the only option to produce highly contiguous genome assemblies. Indeed, we have been able to make R2C2 libraries with as little as 10 ng of input DNA. Nonetheless, if more DNA is available, recent advances in PacBio library preparations (8) might be a more appealing option for the long-read assembly. This method does not require amplification, and results in a less biased coverage. However, without Hi-C data for scaffolding, chromosome-scale assemblies are unlikely to be achievable. We therefore consider the addition of our Hi-C approach a necessary prerequisite for high quality genomes.

Perhaps the most fundamental concern for the suitability of our approach is the researcher's specific questions and motivations for making a genome. Applications that require high contiguity in an assembly would be enhanced significantly using this approach. For example, association studies and quantitative trait locus mapping approaches generally require knowledge of large-scale linkage among sites to be successful (38). Similarly, many population genetic frameworks, e.g. those for local ancestry inference (39,40), and for estimating past effective population sizes (41), are based on the spatial distribution of markers along a reference genome. Finally, comparative studies of large-scale chromosome structure would be significantly enhanced by contiguous genome assemblies (9). However, if the distributions of repetitive elements across the genome are of interest, our specific method is unlikely to perform well. Many studies are concerned primarily with coding regions, and for those our approach presents a reasonably high quality option.

This approach can serve as a guide point for genome projects of small organisms which make a large majority of the diversity of life. Approximately 80% of known species are insects, and ∼5 million total insect species are believed to exist on earth (42). Additionally, any research projects dealing with minimal DNA could achieve chromosome scale genomic information from this approach. This approach is therefore positioned to revolutionize our understanding of genome structure across diverse species.

DATA AVAILABILITY

The sequencing data and final assembly generated in this study has been submitted to the NCBI BioProject database (https://www.ncbi.nlm.nih.gov/bioproject/) under accession number PRJNA591165.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

NIH [R35 GM128932 to R.B.C.-D., R35 GM133569-01 to C.V.]; Alfred P. Sloan Fellowship (to R.B.C.-D.); NIH training grant [T32 HG008345-01 to J.M.]. Funding for open access charge: NIH.

Conflict of interest statement. R.E.G. is co-founder and paid consultant of Dovetail Genomics.

REFERENCES

1. Dudchenko,O., Batra,S.S., Omer,A.D., Nyquist,S.K., Hoeger,M., Durand,N.C., Shamim,M.S., Machol,I., Lander,E.S., Aiden,A.P. *et al.* (2017) De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science*, **356**, 92–95.
2. Worley,K.C., Richards,S. and Rogers,J. (2017) The value of new genome references. *Exp. Cell Res.*, **358**, 433–438.
3. Rice,E.S. and Green,R.E. (2019) New approaches for genome assembly and scaffolding. *Annu. Rev. Anim. Biosci.*, **7**, 17–40.
4. Yuan,Y., Bayer,P.E., Batley,J. and Edwards,D. (2017) Improvements in genomic Technologies: Application to crop genomics. *Trends Biotechnol.*, **35**, 547–558.
5. Lieberman-Aiden,E., van Berkum,N.L., Williams,L., Imakaev,M., Ragoczy,T., Telling,A., Amit,I., Lajoie,B.R., Sabo,P.J., Dorschner,M.O. *et al.* (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, **326**, 289–293.
6. Putnam,N.H., O'Connell,B.L., Stites,J.C., Rice,B.J., Blanchette,M., Calef,R., Troll,C.J., Fields,A., Hartley,P.D., Sugnet,C.W. *et al.* (2016) Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. *Genome Res.*, **26**, 342–350.
7. Li,F., Zhao,X., Li,M., He,K., Huang,C., Zhou,Y., Li,Z. and Walters,J.R. (2019) Insect genomes: progress and challenges. *Insect Mol. Biol.*, **28**, 739–758.
8. Kingan,S.B., Heaton,H., Cudini,J., Lambert,C.C., Baybayan,P., Galvin,B.D., Durbin,R., Korlach,J. and Lawniczak,M.K.N. (2019) A High-Quality de novo genome assembly from a single mosquito using PacBio sequencing. *Genes*, **10**, 62.
9. Corbett-Detig,R.B., Said,I., Calzetta,M., Genetti,M., McBroome,J., Maurer,N.W., Petrarca,V., Torre,A.D. and Besansky,N.J. (2019) Fine-Mapping complex inversion breakpoints and investigating somatic pairing in the pecies complex using proximity-ligation sequencing. *Genetics*, **213**, 1495–1511.
10. Chapman,J.A., Ho,I., Sunkara,S., Luo,S., Schroth,G.P. and Rokhsar,D.S. (2011) Meraculous: de novo genome assembly with short paired-end reads. *PLoS One*, **6**, e23501.
11. Ruan,J. and Li,H. (2019) Fast and accurate long-read assembly with wtdbg2. *Nature Methods*, **17**, 155–158.
12. Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.
13. Gurevich,A., Saveliev,V., Vyahhi,N. and Tesler,G. (2013) QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, **29**, 1072–1075.
14. Simão,F.A., Waterhouse,R.M., Ioannidis,P., Kriventseva,E.V. and Zdobnov,E.M. (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, **31**, 3210–3212.
15. Walker,B.J., Abeel,T., Shea,T., Priest,M., Abouelliel,A., Sakthikumar,S., Cuomo,C.A., Zeng,Q., Wortman,J., Young,S.K. *et al.* (2014) Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One*, **9**, e112963.
16. Li,H. (2018) Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, **34**, 3094–3100.
17. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G., Durbin,R. and 1000 Genome Project Data Processing Subgroup (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
18. Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.
19. Li,H. and Durbin,R. (2010) Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics*, **26**, 589–595.
20. McKenna,A., Hanna,M., Banks,E., Sivachenko,A., Cibulskis,K., Kernytsky,A., Garimella,K., Altshuler,D., Gabriel,S., Daly,M. *et al.* (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.*, **20**, 1297–1303.
21. Edge,P., Bafna,V. and Bansal,V. (2017) HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies. *Genome Res.*, **27**, 801–812.
22. Adams,M.D. (2000) The genome sequence of Drosophila melanogaster. *Science*, **287**, 2185–2195.
23. Grenier,J.K., Arguello,J.R., Moreira,M.C., Gottipati,S., Mohammed,J., Hackett,S.R., Boughton,R., Greenberg,A.J. and Clark,A.G. (2015) Global diversity lines - a five-continent reference panel of sequenced Drosophila melanogaster strains. *G3*, **5**, 593–603.
24. Lack,J.B., Cardeno,C.M., Crepeau,M.W., Taylor,W., Corbett-Detig,R.B., Stevens,K.A., Langley,C.H. and Pool,J.E. (2015) The Drosophila genome nexus: a population genomic resource of 623 Drosophila melanogaster genomes, including 197 from a single ancestral range population. *Genetics*, **199**, 1229–1241.
25. Volden,R., Palmer,T., Byrne,A., Cole,C., Schmitz,R.J., Green,R.E. and Vollmers,C. (2018) Improving nanopore read accuracy with the R2C2 method enables the sequencing of highly multiplexed full-length single-cell cDNA. *Proc. Natl. Acad. Sci. U.S.A.*, **115**, 9726–9731.
26. Belton,J.-M., McCord,R.P., Gibcus,J.H., Naumova,N., Zhan,Y. and Dekker,J. (2012) Hi-C: a comprehensive technique to capture the conformation of genomes. *Methods*, **58**, 268–276.
27. Chakraborty,M., Baldwin-Brown,J.G., Long,A.D. and Emerson,J.J. (2016) Contiguous and accurate de novo assembly of metazoan genomes with modest long read coverage. *Nucleic Acids Res.*, **44**, e147.
28. Chakraborty,M., VanKuren,N.W., Zhao,R., Zhang,X., Kalsow,S. and Emerson,J.J. (2018) Hidden genetic variation shapes the structure of functional elements in Drosophila. *Nat. Genet.*, **50**, 20–25.
29. Treangen,T.J. and Salzberg,S.L. (2011) Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat. Rev. Genet.*, **13**, 36–46.
30. Cooper,K.W. (1948) The evidence for long range specific attractive forces during the somatic pairing of dipteran chromosomes. *J. Exp. Zool.*, **108**, 327–335.
31. AlHaj Abed,J., Erceg,J., Goloborodko,A., Nguyen,S.C., McCole,R.B., Saylor,W., Fudenberg,G., Lajoie,B.R., Dekker,J., Mirny,L.A. *et al.* (2019) Highly structured homolog pairing reflects functional organization of the Drosophila genome. *Nat. Commun.*, **10**, 4485.
32. Kuderna,L.F.K., Lizano,E., Julià,E., Gomez-Garrido,J., Serres-Armero,A., Kuhlwilm,M., Alandes,R.A., Alvarez-Estape,M., Juan,D., Simon,H. *et al.* (2019) Selective single molecule sequencing and assembly of a human Y chromosome of African origin. *Nat. Commun.*, **10**, 4.
33. Pietri,J.E., DeBruhl,H. and Sullivan,W. (2016) The rich somatic life of Wolbachia. *Microbiologyopen*, **5**, 923–936.
34. Russell,S.L., Chappell,L. and Sullivan,W. (2019) A symbiont's guide to the germline. *Curr. Top. Dev. Biol.*, **135**, 315–351.
35. Blow,F. and Douglas,A.E. (2019) The hemolymph microbiome of insects. *J. Insect Physiol.*, **115**, 33–39.
36. Medina,P., Russell,S.L. and Corbett-Detig,R. (2019) Deep data mining reveals variable abundance and distribution of microbial reproductive manipulators within and among diverse host species. bioRxiv doi: https://doi.org/10.1101/679837, 23 June 2019, preprint: not peer reviewed.
37. Kingan,S.B., Urban,J., Lambert,C.C., Baybayan,P., Childers,A.K., Coates,B., Scheffler,B., Hackett,K., Korlach,J. and Geib,S.M. (2019) A high-quality genome assembly from a single, field-collected spotted lanternfly (Lycorma delicatula) using the PacBio Sequel II system. *Gigascience*, **8**, doi:10.1093/gigascience/giz122.
38. Ashton,D.T., Ritchie,P.A. and Wellenreuther,M. (2017) Fifteen years of quantitative trait loci studies in fish: challenges and future directions. *Mol. Ecol.*, **26**, 1465–1476.
39. Maples,B.K., Gravel,S., Kenny,E.E. and Bustamante,C.D. (2013) RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. *Am. J. Hum. Genet.*, **93**, 278–288.
40. Corbett-Detig,R. and Nielsen,R. (2017) A hidden markov model approach for simultaneously estimating local ancestry and admixture time using next generation sequence data in samples of arbitrary ploidy. *PLoS Genet.*, **13**, e1006529.
41. Li,H. and Durbin,R. (2011) Inference of human population history from individual whole-genome sequences. *Nature*, **475**, 493–496.
42. Stork,N.E. (2018) How many species of insects and other terrestrial arthropods are there on Earth? *Annu. Rev. Entomol.*, **63**, 31–45.