Illumina TruSeq synthetic long-reads empower *de novo* assembly and resolve complex, highly repetitive transposable elements

Rajiv C. McCoy[1], Ryan W. Taylor[1], Timothy A. Blauwkamp[2], Joanna L. Kelley[3], Michael Kertesz[4], Dmitry Pushkarev[5], Dmitri A. Petrov*[1] and Anna-Sophie Fiston-Lavier*[1,6]

[1]Department of Biology, Stanford University, Stanford, California 94305, USA

[2]Illumina Inc., San Diego, California 92122, USA

[3]School of Biological Sciences, Washington State University, Pullman, Washington 99164, USA

[4]Department of Bioengineering, Stanford University, Stanford, California 94035, USA

[5]Department of Physics, Stanford University, Stanford, California 94035, USA

[6]Institut des Sciences de l'Evolution-Montpellier, Montpellier, Cedex 5, France

Corresponding authors: Rajiv C. McCoy rmccoy@stanford.edu
Dmitri Petrov dpetrov@stanford.edu, and Anna-Sophie Fiston-Lavier asfiston@univ-montp2.fr

*DAP and ASFL are joint senior authors on this work.

Running title: Long read assembly of *D. melanogaster* genome

Keywords: genome assembly, Moleculo, LR-seq, repeats, *Drosophila*

Abstract

High-throughput DNA sequencing technologies have revolutionized genomic analysis, including the *de novo* assembly of whole genomes. Nevertheless, assembly of complex genomes remains challenging, in part due to the presence of dispersed repeats which introduce ambiguity during genome reconstruction. Transposable elements (TEs) can be particularly problematic, especially for TE families exhibiting high sequence identity, high copy number, or present in complex genomic arrangements. While TEs strongly affect genome function and evolution, most current *de novo* assembly approaches cannot resolve long, identical, and abundant families of TEs. Here, we applied a novel Illumina technology called TruSeq synthetic long-reads, which are generated through highly parallel library preparation and local assembly of short read data and achieve lengths of 1.5-18.5 Kbp with an extremely low error rate (∼0.03% per base). To test the utility of this technology, we sequenced and assembled the genome of the model organism *Drosophila melanogaster* (reference genome strain *y;cn,bw,sp*) achieving an N50 contig size of 69.7 Kbp and covering 96.9% of the euchromatic chromosome arms of the current reference genome. TruSeq synthetic long-read technology enables placement of individual TE copies in their proper genomic locations as well as accurate reconstruction of TE sequences. We entirely recovered and accurately placed 4,229 (77.8%) of the 5,434 of annotated transposable elements with perfect identity to the current reference genome. As TEs are ubiquitous features of genomes of many species, TruSeq synthetic long-reads, and likely other methods that generate long reads, offer a powerful approach to improve *de novo* assemblies of whole genomes.

Introduction

Tremendous advances in DNA sequencing technology, computing power, and assembly approaches, have enabled the assembly of genomes of thousands of species from the sequences of DNA fragments, but several challenges still remain. All assembly approaches are based on the assumption that similar sequence reads originate from the same genomic region, thereby allowing the reads to be overlapped and merged to reconstruct the underlying genome sequence (Nagarajan and Pop, 2013). Deviations from this assumption, including those arising due to polymorphism and repeats, complicate assembly and may induce assembly failure. When possible, performing multiple rounds of inbreeding, using input DNA from a single individual, or even sequencing mutant haploid embryos (Langley et al., 2011) can limit heterozygosity and improve assembly results.

By spanning regions of high diversity and regions of high identity, the use of longer input sequences can also help overcome problems posed by both polymorphism and repeats. The recent application of Pacific Biosciences (PacBio) long read technology to resolve complex segmental duplications (Huddleston et al., 2014) is a case in point. Illumina recently introduced TruSeq™ synthetic long-read technology, which builds upon underlying short read data to generate accurate synthetic reads up to 18.5 Kbp in length. The technology was already used for the *de novo* assembly of the genome of the colonial tunicate, *Botryllus schlosseri* (Voskoboynik et al., 2013). However, because no high quality reference genome was previously available for that species, advantages, limitations, and general utility of the technology for genome assembly were difficult to assess. By performing assembly of the *Drosophila melanogaster* genome, our study uses comparison to a high quality reference to evaluate the application of synthetic long-read technology for *de novo* assembly. While future work will be required to investigate the use of the technology for resolving polymorphism in outbred species, our work specifically focuses on the accuracy of assembly of repetitive DNA sequences.

In some species, repetitive DNA accounts for a large proportion of the total genome size, for example comprising more than half of the human genome (Lander et al., 2001; de Koning et al., 2011) and 80% of some plant genomes (Feschotte et al., 2002). Here, we focus on one class of dynamic repeats, called transposable elements (TEs), which are a common feature of almost all eukaryotic genomes sequenced to date. Some families of TEs are represented in hundreds or even thousands of nearly-identical copies, and some copies span up to tens of kilobases. Consequently, TEs dramatically affect genome size and structure, as well as genome function; transposition has the potential to induce complex genomic rearrangements that detrimentally affect the host, but can also provide the raw material for adaptive evolution (González et al., 2008; González and Petrov, 2009; Casacuberta and González, 2013), for example, by creating new transcription factor binding sites (Rebollo et al., 2012) or otherwise affecting expression of nearby genes (González et al., 2009).

Despite their biological importance, knowledge of TE dynamics is hindered by technical limitations resulting in the absence of certain TE families from genome assemblies. Many software packages for whole genome assembly use coverage-based heuristics, distinguishing putative unique regions from putative repetitive regions based on deviation from average coverage (e.g., Celera (Myers et al., 2000), Velvet (Zerbino and Birney, 2008)). While TE families with sufficient divergence among copies may be properly assembled, recently diverged families are often present in sets of disjointed reads or small contigs that cannot be placed with respect to the rest of the assembly. For example, the *Drosophila* 12 Genomes Consortium (Clark et al., 2007) did not even attempt to evaluate accuracy or completeness of TE assembly. Instead, they used four separate programs to estimate abundance of TEs and other repeats within each assembled genome, but the resulting upper and lower bounds commonly differed by more than three fold. The recent improvement to the draft assembly of *Drosophila simulans* reported that the majority of TE sequences (identified by homology to *D. melanogaster* TEs) were contained in fragmented contigs less than 500 bp in length (Hu et al., 2013).

TEs, as with other classes of repeats, may also induce mis-assembly. For example, TEs that lie in tandem may be erroneously collapsed, and unique interspersed sequences may be left out or appear as isolated contigs. Several studies have assessed the impact of repeat elements on *de novo* genome assembly. For example, Alkan et al. (2010) showed that the human assemblies are on average 16.2% shorter than expected, mainly due to failure to assemble repeats, especially TEs and segmental duplications. A similar observation was made for the chicken genome, despite the fact that repeat density in this genome is lower than humans (Ye et al., 2011). In addition to coverage, current approaches to deal with repeats such as TEs generally rely on paired-end data (Alkan et al., 2010; Miller et al., 2010; Li et al., 2010). Paired-end reads can help resolve the orientation and distance between assembled flanking sequences, but do not resolve the repeat sequence itself. Likewise, if read pairs do not completely span an identical repeat and are not anchored in unique sequence, alternative possibilities for contig extension cannot be ruled out. Long inserts, commonly referred to as mate-pair libraries, are therefore useful to bridge across long TEs to link and orient contigs, but produce stretches of unknown sequence.

A superior way to resolve TEs is to generate reads that exceed TE length, obviating assembly and allowing TEs to be unambiguously placed based on unique flanking sequence. Pacific Biosciences (PacBio) represents the only high throughput long read (up to ∼15 Kbp) technology available to date, though Oxford Nanopore (Clarke et al., 2009) platforms may soon be available. While single-pass PacBio sequencing has a high error rate of 15-18%, multiple-pass circular consensus sequencing (Jiao et al., 2013) and hybrid or self error correction (Koren et al., 2012) improve read accuracy to greater than 99.9%. Meanwhile, other established sequencing technologies, such as Illumina, 454 (Roche), and Ion Torrent (Life Technologies), offer high throughput and low error rates of 0.1-1%, but much shorter read lengths (Glenn, 2011). Illumina TruSeq synthetic long-reads, which are assembled from underlying Illumina short read data, achieve lengths and error rates comparable to PacBio corrected sequences, but their utility for *de novo* assembly has yet to be demonstrated in cases where a high quality reference genome is available for comparison.

Using a pipeline of standard existing tools, we demonstrate the ability of TruSeq synthetic long-reads to facilitate *de novo* assembly and resolve TE sequences in the genome of the fruit fly *Drosophila melanogaster*, a key model organism in both classical genetics and molecular biology. We further investigate how coverage of synthetic long-reads affects assembly results, an important practical consideration for experimental design. While the *D. melanogaster* genome is moderately large (∼180 Mbp) and complex, it has already been assembled to unprecedented accuracy. Through a massive collaborative effort, the initial genome project (Adams et al., 2000) recovered nearly all of the 120 Mbp euchromatic sequence using a whole-genome shotgun approach that involved painstaking molecular cloning and the generation of a bacterial artificial chromosome physical map. Since that publication, the reference genome has been extensively annotated and improved using several resequencing, gap-filling, and mapping strategies, and currently represents a gold standard for the genomics community (Osoegawa et al., 2007; Celniker et al., 2002; Hoskins et al., 2007). By performing the assembly in this model system with a high quality reference genome, our study is the first to systematically document the advantages and limitations posed by this synthetic long-read technology. *D. melanogaster* harbors a large number (∼100) of families of active TEs, some of which contain many long and virtually identical copies distributed across the genome, thereby making their assembly a particular challenge. This is distinct from other species, including humans, which have TE copies that are shorter and more diverged from each other, and therefore easier to assemble. Our demonstration of accurate TE assembly in *D. melanogaster* should therefore translate favorably to many other systems.

Results

TruSeq synthetic long-reads

Library preparation

This study used Illumina TruSeq synthetic long-read technology generated with a novel highly-parallel next-generation library preparation method (Figure S1). The basic protocol was previously presented by Voskoboynik et al. (2013) (who referred to it as LR-seq) and was patented by Stanford University and licensed to Moleculo, which was later acquired by Illumina. The protocol (see Methods) involves initial mechanical fragmentation of gDNA into ∼10 Kbp fragments. These fragments then undergo end-repair and ligation of amplification adapters, before being diluted onto 384-well plates so that each well contains DNA representing approximately 1-2% of the genome (∼200 molecules, in the case of *Drosophila melanogaster*). Polymerase chain reaction (PCR) is used to amplify molecules within wells, followed by highly parallel Nextera-based fragmentation and barcoding of individual wells. DNA from all wells is then pooled and sequenced on the Illumina HiSeq 2000 platform. Data from individual wells are demultiplexed *in silico* according to the barcode sequences. Synthetic long-reads are then assembled from the short reads using an assembly pipeline that accounts for properties of the molecular biology steps used in the library preparation (see Supplemental Materials). Because each well represents DNA from only ∼200 molecules, even identical repeats can be resolved into synthetic reads as long as they are not so abundant in the genome as to be represented multiple times within a single well.

We applied TruSeq synthetic long-read technology to the fruit fly *D. melanogaster*, a model organism with a high quality reference genome, including extensive repeat annotation (Fiston-Lavier et al., 2007; Quesneville et al., 2003, 2005). The version of the reference genome assembly upon which our analysis is based (Release 5.56; `ftp://ftp.flybase.net/genomes/Drosophila_melanogaster/dmel_r5.56_FB2014_02/fasta/dmel-all-chromosome-r5.56.fasta.gz`) contains a total of 168.7 Mbp of sequence. For simplicity, our study uses the same naming conventions as the reference genome sequence, where the sequences of chromosome arms X, 2L, 2R, 3L, 3R, and 4 contain all of the euchromatin and part of the centric heterochromatin. The sequences labelled XHet, 2LHet, 2RHet, 3LHet, 3RHet, and YHet represent scaffolds from heterochromatic regions that have been localized to chromosomes, but have not been joined to the rest of the assembly. Some of these sequences are ordered, while others are not, and separate scaffolds are separated by stretches of N's with an arbitrary length of 100 bp. Meanwhile, the genome release also includes 10.0 Mbp of additional heterochromatic scaffolds (U) which could not be mapped to chromosomes, as well as 29.0

Mbp of additional small scaffolds that could not be joined to the rest of the assembly (Uextra). Because the Uextra sequences are generally lower quality and partially redundant with respect to the other sequences, we have excluded them from all of our analyses of assembly quality. Assembly assessment based on comparison to the Het and U sequences should also be interpreted with caution, as alignment breaks and detected misassemblies will partially reflect the incomplete nature of these portions of the reference sequence. Finally, we extracted the mitochondrial genome of the sequenced strain from positions 5,288,527-5,305,749 of reference sequence U using BEDTools (version 2.19.1), replacing the mitochondrial reference sequence included with Release 5.56, which represents a different strain (see `http://bergmanlab.smith.man.ac.uk/?p=2033`).

Approximately 50 adult individuals from the *y;cn,bw,sp* strain of *D. melanogaster* were pooled for the isolation of high molecular weight DNA, which was used to generate TruSeq synthetic long-read libraries using the aforementioned protocol (Figure S1). The strain *y;cn,bw,sp* is the same strain which was used to generate the *D. melanogaster* reference genome (Adams et al., 2000). The fact that the strain is isogenic not only facilitates genome assembly in general, but also ensures that our analysis of TE assembly is not confounded by TE polymorphism. A total of 955,836 synthetic long-reads exceeding 1.5 Kbp (an arbitrary length cutoff) were generated with six libraries (Table S1), comprising a total of 4.20 Gbp. Synthetic long-reads averaged 4,394 bp in length, but have a local maximum near 8.5 Kbp, slightly smaller than the ~10 Kbp DNA fragments used as input for the protocol (Figure 1A).

Error rates

In order to evaluate the accuracy of TruSeq synthetic long-reads, we mapped sequences to the reference genome of *D. melanogaster*, identifying differences between the mapped synthetic reads and the reference sequence. Of 955,836 input synthetic long-reads, 99.84% (954,276 synthetic reads) were successfully mapped to the reference genome, with 90.88% (868,685 synthetic reads) mapping uniquely and 96.36% (921,090 synthetic reads) having at least one alignment with a MAPQ score $\geq$20. TruSeq synthetic long-reads had very few mismatches to the reference at 0.0509% per base (0.0448% for synthetic reads with MAPQ $\geq$20) as well as a very low insertion rate of 0.0166% per base (0.0144% for synthetic reads with MAPQ $\geq$20) and a deletion rate of 0.0290% per base (0.0259% for synthetic reads with MAPQ $\geq$20). Error rates estimated with this mapping approach are conservative, as residual heterozygosity in the sequenced line mimics errors. We therefore used the number of mismatches overlapping known SNPs to calculate a corrected error rate of 0.0286% per base (see Methods). Along with this estimate, we also estimated that the sequenced strain still retains 0.0550% residual heterozygosity relative to the time that the line was established. We note that

TruSeq synthetic long-reads achieve such low error rates due to the fact that they are built as a consensuses of underlying Illumina short reads, which have an approximately ten times higher error rate. We further observed that mismatches are more frequent near the beginning of synthetic long-reads, while error profiles of insertions and deletions are relatively uniform (Figures 1B, 1C, & 1D). Minor imprecision in the trimming of adapter sequence and the error distribution along the lengths of the underlying short reads are likely responsible for this distinct error profile. Based on the observation of low error rates, no pre-processing steps were necessary in preparation for assembly, though overlap-based trimming and detection of chimeric and spurious reads are performed by default by the Celera Assembler.

Analysis of coverage

We quantified the average depth of coverage of the mapped synthetic long-reads for each reference chromosome arm. We observed 33.3-35.2× coverage averages of the euchromatic chromosome arms of each major autosome (2L, 2R, 3L, 3R; Figure 2). Coverage of the heterochromatic scaffolds of the major autosomes (2LHet, 2RHet, 3LHet, 3RHet) was generally lower (24.8-30.6×), and also showed greater coverage heterogeneity than the euchromatic reference sequences. This is explained by the fact that heterochromatin has high repeat content relative to euchromatin, making it more difficult to assemble into synthetic long reads. Nevertheless, the fourth chromosome had an average coverage of 34.4×, despite the enrichment of heterochromatic islands on this chromosome (Haynes et al., 2006). Depth of coverage on sex chromosomes was expected to be lower: 75% relative to the autosomes for the X and 25% relative to the autosomes for the Y, assuming equal numbers of males and females in the pool. Observed synthetic long-read depth was lower still for the X chromosome (21.2×) and extremely low for the Y chromosome (3.84×), which is entirely heterochromatic. Synthetic long-read depth for the mitochondrial genome was also relatively low (19.1×) in contrast to high mtDNA representation in short read genomic libraries, which we suspect to be a consequence of the fragmentation and size selection steps of the library preparation protocol.

Assessment of assembly content and accuracy

Assembly length and genome coverage metrics

To perform *de novo* assembly, we used the Celera Assembler (version 8.1)(Myers et al., 2000), an overlap-layout-consensus assembler developed and used to reconstruct the first genome sequence of a multicellular organism, *D. melanogaster* (Adams et al., 2000), as well as one of the first diploid human genome sequences (Levy et al., 2007). Our Celera-generated assembly contained 6,617 contigs of lengths ranging from 1,506 bp to 567.5 Kbp, with an N50 contig length of 64.1 Kbp. Note that because the TruSeq synthetic long-read data are effectively single end reads, only contig rather than scaffold metrics are reported. The total length of the assembly (i.e. the sum of all contig lengths) was 152.2 Mbp, with a GC content of 42.18% (compared to 41.74% GC content in the reference genome). Upon aligning contigs to the reference genome with NUCmer (Delcher et al., 2002; Kurtz et al., 2004), we observed that the ends of several contigs overlapped with long stretches (>1 Kbp) of perfect sequence identity. We therefore used the assembly program Minimus2 (Sommer et al., 2007) to merge across these regions to generate supercontigs. All statistics in the following sections are based on this two-step assembly procedure combining Celera and Minimus2. The merging step resulted in the additional merging of 1,652 input contigs into 633 supercontigs, resulting in an improved assembly with a total of 5,598 contigs spanning a total of 147.4 Mbp and an N50 contig length of 69.7 Kbp (Table 1).

We used the program QUAST (Gurevich et al., 2013) to evaluate the quality of our assembly based on alignment to the high quality reference genome. This program analyzes the NUCmer (Delcher et al., 2002; Kurtz et al., 2004) alignment to generate a reproducible summary report that quantifies alignment length and accuracy, as well as cataloging mis-assembly events for further investigation. Key results from the QUAST analysis are reported in Table 1, while the mis-assembly event list is included as supplemental material. The NA50 (60.1 Kbp; 63.0 Kbp upon including heterochromatic reference scaffolds) is a key metric from this report that is analogous to N50, but considers lengths of alignments to the reference genome rather than the lengths of the contigs. Contigs are effectively broken at the locations of putative mis-assembly events, including translocations and relocations. As with the synthetic long-reads, the QUAST analysis revealed that indels and mismatches in the assembly are rare, each occurring fewer than an average of 10 times per 100 Kbp (Table 1).

To gain more insight about the alignment on a per-chromosome basis, we further investigated the NUCmer alignment of the 5,598 assembled contigs to the reference genome. Upon requiring high stringency alignment (>99% sequence identity and >1 Kbp aligned), there were 3,717 alignments of our contigs to the euchromatic portions of chromosomes X, 2, 3, and 4, covering a total of 116.2 Mbp (96.6%) of the euchromatin (Table 2). For the heterochromatic sequence (XHet, 2Het, 3Het, and YHet), there were 817 alignments at this same threshold, covering 8.2 Mbp (79.9%) of the reference. QUAST also identified 179 fully unaligned contigs ranging in size from 1,951 to 26,663 bp, which we investigated further by searching the NCBI nucleotide database with BLASTN (Altschul et al., 1997). Of these contigs, 151 had top hits to bacterial species also identified in the underlying long-read data (Supplemental Materials; Table S2), 113 of which correspond to acetic acid bacteria that are known *Drosophila* symbionts. The remaining 27 contigs with no significant

BLAST hit will require further investigation to determine whether they represent novel fly-derived sequences (Table S6).

Assessment of gene sequence assembly

In order to further assess the presence or absence as well as the accuracy of the assembly of various genomic features, we developed a simple pipeline that reads in coordinates of generic annotations and compares the reference and assembly for these sequences (see Methods). As a first step in the pipeline, we again used the filtered NUCmer (Delcher et al., 2002; Kurtz et al., 2004) alignment, which consists of the best placement of each draft sequence on the high quality reference genome. We then tested whether both boundaries of a given genomic feature were present within the same aligned contig. For features that met this criterion, we performed local alignment of the reference sequence to the corresponding contig using BLASTN (Altschul et al., 1997), evaluating the results to calculate the proportion of the sequence aligned as well as the percent identity of the alignment. We determined that 15,684 of 17,294 (90.7%) FlyBase-annotated genes have start and stop boundaries contained in a single aligned contig within our assembly. A total of 14,558 genes (84.2%) have their entire sequence reconstructed with perfect identity to the reference sequence, while 15,306 genes have the entire length aligned with >99% sequence identity. The presence of duplicated and repetitive sequences in introns complicates gene assembly and annotation, potentially causing genes to be fragmented. For the remaining 1,610 genes whose boundaries were not contained in a single contig, we found that 1,235 were partially reconstructed as part of one or more contigs.

Assessment of assembly gaps

Coverage gaps and variability place an upper bound on the contiguity of genome assemblies. Regions of low coverage in synthetic long-reads may arise from biases in library preparation, sequencing, or the computational processing and assembly from underlying short read data. We therefore performed a simulation of the assembly based solely on breaks introduced by coverage gaps in the synthetic long read alignment to the reference genome (see Supplemental Materials). We required at least one overlap of at least 800 bp in order to merge across synthetic long-reads (thereby simulating a key assembly parameter) and excluded any contiguous covered region (analogous to a contig) of less than 1,000 bp. The resulting pseudo-assembly was comprised of 3,678 contigs spanning a total of 130.4 Mbp with an N50 of 80.5 Kbp. Because this expectation is based on mapping to the current reference genome, the total assembly length cannot be greater than the length of the reference sequence. Together, this simulation suggested that regions of low coverage in synthetic long-reads were primarily responsible for observed cases of assembly failure.

We next analyzed the content of the 3,524 gaps in the NUCmer alignment, which together represent failures of sequencing, library preparation, and genome assembly. We observed that 3,265 of these gaps in the whole genome assembly corresponded to previously-identified reductions in synthetic long-read coverage. Motivated by the observation that 93% of assembly gaps are explained by a deficiency in synthetic long-read coverage, we performed joint analysis of synthetic long-read and underlying short read data to help distinguish library preparation and sequencing biases from biases arising during the computational steps used to assemble the synthetic reads. We used BWA (Li and Durbin, 2009) to map underlying Illumina paired-end short read data to the reference genome, quantifying depth of coverage in the intervals of assembly gaps. While average short read coverage of the genome exceeded 1,500× (for MAPQ>0), mean short read coverage in the assembly gap regions was substantially lower at 263× (for MAPQ>0). However, when coverage was quantified for all mapped reads (including multi-mapped reads with MAPQ=0), average coverage was 1,153×, suggesting an abundance of genomic repeats in these intervals. We also observed a strong reduction in the GC content of the gaps (29.7%; 1,192 intervals with <30% GC content) compared to the overall GC content of the assembly (42.26%). This observation is therefore consistent with a known bias of PCR against high AT (but also high GC) fragments (Benjamini and Speed, 2012). However, low GC content is also a feature of gene-poor and TE-rich regions (Duret and Hurst, 2001), confounding this simple interpretation.

In order to gain further insight about the content of alignment gaps, we applied RepeatMasker (Smit, Hubley, & Green. RepeatMasker Open-4.0.5 1996-2010. <http://www.repeatmasker.org>) to these intervals, revealing that 35.19% of the gap sequence is comprised of TEs, 11.68% of satellites, 2.45% of simple repeats, and 0.21% of other low complexity sequence. These proportions of gap sequences composed of TEs and satellites exceed the overall genomic proportions (the fraction of the reference chromosome arms, excluding scaffolds in Uextra) of 15.07% and 1.12%, respectively, while the proportions composed of simple repeats and low complexity sequences are comparable to the overall genome proportions of 2.44% and 0.34%. Motivated by the overrepresentation of TEs in the gap intervals, we investigated which TE families were most responsible for these assembly failures. A total of 385 of the 3,524 assembly gaps overlapped the coordinates of annotated TEs, with young TE families being highly represented (Table S3). For example, LTR elements from the *roo* family were the most common, with 117 copies (of only 136 copies in the genome) overlapping gap coordinates. TEs from the *roo* family are long (canonical length of 9,092 bp) and recently diverged (mean of 0.0086 substitutions per base), and are therefore difficult to assemble (FlyTE database, http://petrov.stanford.edu/cgi-bin/Tlex_databases/flyTE_home.cgi). Conversely, elements of the high-copy number (2,235 copies) INE-1 family were underrepresented among gaps in the alignment, with only 84 copies overlapping gap coordinates. INE-1 elements tend to be short (611 bp canonical length) and represent older transposition with greater divergence among copies.

Manual curation of the alignment also revealed that assembly is particularly poor in regions of tandem arrangement of TE copies from the same family, a result that is expected because repeats will be present within individual wells during library preparation (Figure S4A). In contrast, assembly can be successful in regions with high-repeat density, provided that the TEs are sufficiently divergent or from different families (Figure S4B). Together, these observations about the assembly of particular TE families motivated formal investigation of the characteristics of individual TE copies and TE families that affect their assembly, as we describe in the following section.

Assessment of TE sequence assembly

Repeats can induce three common classes of mis-assembly. First, tandem repeats may be erroneously collapsed into a single copy. While the accuracy of TruSeq synthetic long-reads are advantageous in this case, such elements may still complicate assembly because they are likely to be present within a single molecule (and therefore a single well) during library preparation. Second, large repeats may fail to be assembled because reads do not span the repeat anchored in unique sequence, a situation where TruSeq synthetic long-reads are clearly beneficial. Finally, highly identical repeat copies introduce ambiguity into the assembly graph, which can result in breaks or repeat copies placed in the wrong location in the assembly. As TEs are diverse in their organization, length, copy number, GC content, and divergence, we decided to assess the accuracy of TE assembly with respect to each of these factors. We therefore compared reference TE sequences to the corresponding sequences in our assembly. Because a naive mapping approach could result in multiple reference TE copies mapping to the same location in the assembly, our approach was specifically designed to restrict the search space within the assembly based on the NUCmer global alignment (see Methods). Of the 5,434 TE copies annotated in the *D. melanogaster* reference genome, 4,565 (84.0%%) had both boundaries contained in a single contig of our assembly aligned to the reference genome, with 4,229 (77.8%) perfectly reconstructed based on length and sequence identity.

In order to test which properties of TE copies affected faithful reconstruction, we fit a generalized linear mixed model (GLMM) with a binary response variable indicating whether or not each TE copy was perfectly assembled. We included TE length as a fixed effect because we expected assembly to be less likely in cases where individual synthetic reads do not span the length of the entire TE copy. We also included GC content of the interval, including each TE copy and 1 Kbp flanking sequence on each side, as a fixed effect to capture library preparation biases as well as correlated aspects of genomic context (e.g. gene rich vs. gene poor). TE divergence estimates (FlyTE database, `http://petrov.stanford.edu/cgi-bin/Tlex_databases/flyTE_home.cgi`) were included as a predictor because low divergence (corresponding to high sequence identity) can cause TEs to be misplaced or mis-assembled. We also hypothesized that copy number (TE copies per family), could be important, because high copy number represents more opportunities for false joins which can break the assembly or generate chimeric contigs. Finally, we included a random effect of TE family, which accounts for various family-specific factors not represented by the fixed effects, such as sequence complexity. This grouping factor also accounts for pseudo-replication arising due to multiple copies of TEs within families (Hurlbert, 1984). We found that length ($b = -1.633$, $Z = -20.766$, $P < 2 \times 10^{-16}$), divergence ($b = 0.692$, $Z = 7.501$, $P = 6.35 \times 10^{-14}$), and GC content ($b = 0.186$, $Z = 3.171$, $P = 0.00152$) were significant predictors of accurate TE assembly (Figure 3; Table S5). Longer and less divergent TE copies, as well as those in regions of low GC content, resulted in a lower probability of accurate assembly (Figure 3). We found that overall copy number was not a significant predictor of accurate assembly ($b = 0.095$, $Z = 0.162$, $P = 0.871$). However, upon restricting the test to consider only high identity copies (<0.01 substitutions per base compared to the canonical sequence), we observed an expected reduction in the probability of accurate assembly with increasing copy number ($b = -0.529$, $Z = -2.936$, $P = 0.00333$). Plotting initial results also suggested a possible interaction between divergence and the number of high identity copies. Our model therefore additionally includes this significant interaction term, which demonstrates that low divergence of an individual TE copy is more problematic in the presence of many high identity copies from the same family (Figure 3).

In spite of the limitations revealed by our analysis, we observed several remarkable cases where accurate assembly was achieved, distinguishing the sequences of TEs from a single family with few substitutions among the set. For example, elements in the *Tc1* family have an average of 0.039 substitutions per base with respect to the 947 bp canonical sequence, yet 25 of 26 annotated copies were assembled with 100% accuracy (Table S4). The assembled elements from this family range from 131 bp to 1,662 bp, with a median length of 1,023 bp.

Impact of the coverage on assembly results

The relationship between coverage and assembly quality is complex, as we expect a plateau in assembly quality at the point where the assembly is no longer limited by data quantity. To evaluate the impact of depth of synthetic long-read coverage on the quality of the resulting assembly, we randomly down-sampled the full ∼34× dataset to 20×, 10×, 5×, and 2.5×. We then performed separate *de novo* assemblies for each of these down-sampled datasets, evaluating and comparing assemblies using the same size and correctness metrics previously reported for the full-coverage assembly. We observed an expected nonlinear pattern for several important assembly metrics, which begin to plateau as data quantity increases. NG50 contig length (analogous to N50, but normalized to the genome size of 180 Mb to facilitate comparison among assemblies) increases rapidly with coverage up to approximately 10×, increasing only marginally at higher synthetic long-read coverage (Figure 4A). We do not expect the monotonic increase to continue indefinitely, as very high coverage can overwhelm OLC assemblers such as Celera (see documentation, which advises against high coverage such as 80×). Gene content of the assembly also increases only marginally as synthetic long-read coverage increases above approximately 10×, but TE content does not saturate as rapidly (Figure 4B). Our results likewise suggest that even very low synthetic long-read coverage assemblies (5×) can accurately recover approximately half of all genes and TEs.

Discussion

Rapid technological advances and plummeting costs of DNA sequencing technologies have allowed biologists to explore the genomes of species across the tree of life. However, translating the massive amounts of sequence data into a high quality reference genome ripe for biological insight is a substantial technical hurdle. Many assemblers use coverage-based heuristics to classify problematic repeats and either break the assembly at ambiguous repeat regions or place consensus repeat sequences in the assembly. This approach balances the tradeoff between assembly contiguity and the rate of mis-assembly, but the resulting biased representation of certain classes of repeats limits understanding of repeat evolution. Understanding the dynamics of repeats such as TEs is fundamental to the study genome evolution, as repeats affect genome size structure as well as genome function (González and Petrov, 2009; Feschotte et al., 2002; Kidwell and Lisch, 2001; Cordaux and Batzer, 2009; Nekrutenko and Li, 2001). Several tools (e.g. T-lex2 (Fiston-Lavier et al., 2011), RetroSeq (Keane et al., 2013), Tea (Lee et al., 2012), ngs_te_mapper (Linheiro and Bergman, 2012), RelocaTE (Robb et al., 2013), RetroSeq (Keane et al., 2013), PoPoolation TE (Kofler et al., 2012), TE-locate (Platzer et al., 2012)) are currently available for discovery and annotation of TE sequences in high-throughput sequencing data. However, because these tools depend on the quality of the assembly to which they are applied, annotation is generally limited to TE families containing predominantly short and divergent TE copies, biasing our current view of TE organization. Accurate assembly and annotation of TEs and other repeats will dramatically enrich our understanding of the complex interactions between TEs and host genomes as well as genome evolution in general.

One of the simplest ways to accurately resolve repeat sequences is to acquire reads longer than the lengths of the repeats themselves. Here, we evaluated a novel library preparation approach that allows the generation of highly accurate synthetic reads up to 18.5 Kbp in length. We tested the utility of this approach for assembling and placing highly repetitive, complex TEs with high accuracy. As a first step in our analysis, we analyzed the content of the synthetic long-read data, evaluating synthetic long-read accuracy as well as coverage of the *D. melanogaster* reference genome. We found that the synthetic long-reads were highly accurate, with error rates comparable to consensus sequences produced using third generation long-read sequencing technologies. We also observed relatively uniform coverage across both the euchromatic and heterochromatic portions of the autosomes, with an expected reduced coverage of the heterochromatin. This observation is explained by the fact that heterochromatin is more difficult to sequence as well as the fact that it is generally more repetitive and therefore more difficult to assemble into long reads from underlying short read data.

Despite general uniformity in synthetic long-read coverage, we identified important biases resulting in coverage gaps and reductions in repeat-dense regions with relatively low average GC content. While GC biases of PCR are well documented, GC content is also correlated with repeat density, thereby confounding this interpretation (Duret and Hurst, 2001). Other biases introduced in the molecular biology, sequencing, and/or computational steps of the data preparation (e.g. the fragility of certain DNA sequences during the random shearing step) are also possible, but cannot be disentangled using this data set and will require further investigation. Enhancements to the protocol enabled by a better understanding of these biases could substantially improve the utility of the technology, as reductions in synthetic long-read coverage explained the vast majority of gaps in the genome assembly. Upper bound expectations of assembly contiguity based solely on synthetic long-read coverage were roughly consistent with actual assembly results.

Our assembly achieved an N50 contig length of 69.7 Kbp, covering 96.9% of the euchromatic scaffolds (and centric heterochromatin) of the reference genome, and containing 84.2% of annotated genes with perfect sequence identity. Using standard assembly size (number of contigs, contig length, etc.) and correctness metrics based on alignment to the reference genome, we demonstrated that our assembly is comparable to other *de novo* assemblies of large and complex genomes (e.g., see Salzberg et al., 2012; Bradnam et al., 2013). Nevertheless, we expect that future methodological advances will unlock the full utility of TruSeq synthetic long-read technology. We used a simple pipeline of existing tools to investigate the advantages and limitations of TruSeq synthetic long-reads, but new algorithms and assembly software will be tailored specifically for this platform in the near future (J. Simpson, pers. comm.).

An important caveat in the interpretation of our results is the fact that the assembly was performed on a highly inbred strain of *D. melanogaster*. This was beneficial to our study because it allowed us to attribute TE sequence differences to divergence among TE copies. For an outbred species, distinguishing between divergence among TE copies and polymorphism within TE copies complicates this analysis. For the same reasons, polymorphism in general is a key feature limiting non-haploid genome assemblies, as algorithms must strike a balance between merging polymorphic haplotypes and splitting slightly diverged repeat copies to produce a haploid representation of the genome sequence. Forthcoming assemblies of other *Drosophila* species demonstrate the importance of polymorphism, achieving N50 contig lengths up to 436 Kbp for inbred species, but only 19 Kbp for the species that could not be inbred (Chen et al., in press). The first application of the synthetic long-read technology presented here was to assemble the genome of the colonial tunicate *Botryllus schlosseri*, but assessment of assembly quality was difficult as no high quality reference genome exists for comparison. Likewise, recent work demonstrated the utility of the same technology for assigning polymorphisms to individual haplotypes (Kuleshov et al., 2014), but this problem is somewhat distinct from the *de novo* resolution of polymorphism in the absence of a reference genome. Future work will be required to systematically evaluate the ability of synthetic long-read data to help resolve polymorphism in outbred species.

Our study demonstrates that TruSeq synthetic long-reads enable accurate assembly of complex, highly repetitive TE sequences. Previous approaches to *de novo* assembly generally fail to assemble and place long, abundant, and identical TE copies with respect to the rest of the assembly. For example, the majority of TE-containing contigs in the improved draft assembly of *Drosophila simulans* (which combined Illumina short read and Sanger data) were smaller than 500 bp (Hu et al., 2013). Likewise, short read assemblies from the *Drosophila* 12 Genomes Consortium (Clark et al., 2007) estimated TE copy number, but did not even attempt to place TE sequences with respect to the rest of the assemblies. Our assembly contains 77.8% of annotated TEs perfectly identical in sequence to the current reference genome. Despite the high quality of the current reference, errors undoubtedly exist in the current TE annotations, and it is likely that there is some divergence between the sequenced strain and the reference strain from which it was derived, making our estimate of the quality of TE assembly conservative. Likewise, we used a generalized linear modeling approach to demonstrate that TE length is the main feature limiting the assembly of individual TE copies, a limitation that could be partially overcome by future improvements to the library preparation technology to achieve even longer synthetic reads. This analysis also revealed a significant interaction between divergence and the number of high identity copies within TE families. Low divergence among copies is problematic for families with a large number of high identity copies, but is less important for families with overall copies. Further dilution during library preparation may therefore enhance assembly of dispersed TE families. By performing this assessment in *D. melanogaster*, a species with particularly active, abundant, and identical TEs, our results suggest that synthetic long-read technology can empower studies of TE dynamics for many non-model species.

Alongside this synthetic long-read technology, several third-generation sequencing platforms have been developed to sequence long molecules directly. One such technology, Oxford Nanopore (Oxford, UK) sequencing (Clarke et al., 2009), possesses several advantages over existing platforms, including the generation of reads exceeding 5 Kbp at a speed of 1 bp per nanosecond. Pacific Biosciences' (Menlo Park, CA, USA) single-molecule real-time (SMRT) sequencing platform likewise uses direct observation of enzymatic reactions to produce base calls in real time with reads averaging ∼8.5 Kbp in length (for P5-C3 chemistry), and fast sample preparation and sequencing (1-2 days each) (Roberts et al., 2013, `http://investor.pacificbiosciences.`

com/releasedetail.cfm?ReleaseID=794692). Perhaps most importantly, neither Nanopore nor PacBio sequencing requires PCR amplification, thereby reducing biases and errors that place an upper limit on the sequencing quality of most other platforms. By directly sequencing long molecules, these third-generation technologies will likely outperform TruSeq synthetic long-reads in certain capacities, such as assembly contiguity enabled by homogeneous genome coverage. Indeed, preliminary results from the assembly of a different *y;cn,bw,sp* substrain of *D. melanogaster* using corrected PacBio data achieved an N50 contig length of 15.3 Mbp and closed two of the remaining gaps in the euchromatin of the Release 5 reference sequence (Landolin et al., 2014 [http://dx.doi.org/10.6084/m9.figshare.976097]). While not yet systematically assessed, it is likely that PacBio long reads will also help resolve high identity repeats, though current raw error rates may be limiting.

Most current approaches to *de novo* assembly fare poorly on long, abundant, and recently diverged repetitive elements, including some families of TEs. The resulting assemblies offer a biased perspective of evolution of complex genomes. In addition to accurately recovering 96.9% of the euchromatic portion of the high quality reference genome, our assembly using TruSeq synthetic long-reads accurately placed and perfectly reconstructed the sequence of 84.2% of genes and 77.8% of TEs. Improvements to *de novo* assembly, facilitated by TruSeq synthetic long-reads and other long read technologies, will empower comparative analyses that will enlighten the understanding of the dynamics of repeat elements and genome evolution in general.

Methods

Reference genome and annotations

The latest release of the *D. melanogaster* genome sequence at the time of the preparation of this manuscript (Release 5.56) and corresponding TE annotations were downloaded from FlyBase (`http://www.fruitfly.org/`). All TE features come from data stored in the FlyTE database (`http://petrov.stanford.edu/cgi-bin/Tlex_databases/flyTE_home.cgi`), and were detected using the program BLASTER (Quesneville et al., 2003, 2005).

Library preparation

High molecular weight DNA was separately isolated from pooled samples of the *y;cn,bw,sp* strain of *D. melanogaster* using a standard ethanol precipitation-based protocol. Approximately 50 adult individuals, both males and females, were pooled for the extraction to achieve sufficient gDNA quantity for preparation of multiple TruSeq synthetic long-read libraries.

Six libraries were prepared by Illumina's FastTrack Service using the TruSeq synthetic long-read technology, previously known as Moleculo. To produce each library, extracted gDNA is sheared into approximately 10 Kbp fragments, ligated to amplification adapters, and then diluted to the point that each well on a 384-well plate contains approximately 200 molecules, representing approximately 1.5% of the entire genome. These pools of DNA are then amplified by long range PCR. Barcoded libraries are prepared within each well using Nextera-based fragmentation and PCR-mediated barcode and sequencing adapter addition. The libraries undergo additional PCR amplification if necessary, followed by paired-end sequencing on the Illumina HiSeq 2000 platform.

Assembly of synthetic long-reads from short read data

Based on the unique barcodes, assembly is performed among molecules originating from a single well, which means that the likelihood of individual assemblies containing multiple members of gene families (that are difficult to distinguish from one another and from polymorphism within individual genes) is greatly reduced. The assembly process, which is described in detail in the Supplemental Materials, consists of several modules. First, raw short reads are pre-processed to remove low-quality sequence ends. Digital normalization (Brown et al., 2012) is then performed to reduce coverage biases introduced by PCR, such that the corrected short read coverage of the highest covered fragments is $\sim$40×. The next step uses overlap-based error correction to generate higher quality consensus sequences for each short read. The main assembly steps implement the String Graph Assembler (SGA) (Simpson and Durbin, 2012) which generates contigs using an overlap approach, then scaffolds contigs from the same fragment using paired-end information. Gap filling is then conducted to fill in scaffold gaps. The original paired-end reads are then mapped back to the assembled synthetic long-reads and contigs are either corrected or broken based on inconsistencies in the alignment.

Assessment of synthetic long-read quality

To estimate the degree of contamination of the *D. melanogaster* libraries prepared by Illumina, we used BLASTN (version 2.2.28+) (Altschul et al., 1997) to compare the synthetic long-reads against reference sequences from the NCBI nucleotide database (`http://www.ncbi.nlm.nih.gov/nuccore`), selecting the target sequences with the lowest e-value for each query sequence.

The TruSeq synthetic long-reads were then mapped to the *D. melanogaster* reference genome as single-end reads using BWA-MEM (Li and Durbin, 2009). Depth of coverage was estimated by applying the GATK DepthOfCoverage tool to the resulting alignment. To estimate error rates, we then parsed the BAM file to calculate position-dependent mismatch, insertion, and deletion profiles. Because a portion of this effect would result from accurate sequencing of genomes harboring residual heterozygosity, we used data from the *Drosophila* Genetic Reference Panel (DGRP) (Mackay et al., 2012) to estimate both the rate of residual heterozygosity as well as a corrected error rate of the TruSeq synthetic long-reads. We applied the jvarkit utility (<`https://github.com/lindenb/jvarkit/wiki/SAM2Tsv`>) to identify positions in the reference genome where mismatches occurred. We then used the relationship that the total number sites with mismatches to the euchromatic reference chromosome arms $(M) = 1{,}105{,}831 = Lm + pL\theta$, where L is the 120,381,546 bp length of the reference sequence to which we aligned, m is the per base error rate, p is the proportion of heterozygous sites still segregating in the inbred line, and θ is the average proportion of pairwise differences between *D. melanogaster* genome sequences, estimated as 0.141 from DGRP. Meanwhile, the number of mismatches that overlap with SNP sites in DGRP $(M_{SNP}) = 53,515 = Lm\theta_D + pL\theta$, where θ_D is the proportion of sites that are known SNPs within DGRP (0.0404). Note that this formulation makes the simplifying assumption that all segregating SNPs would have been previously observed in DGRP, which makes the correction conservative. Solving for the unknown variables:

$$m = \frac{M - M_{SNP}}{L(1-\theta_D)} \qquad\qquad p = \frac{M_{SNP} - M\theta_D}{L\theta(1-\theta_D)}$$

To convert m to the TruSeq synthetic long-read error rate, we simply divide by the average depth of coverage of the euchromatic sequence (31.81×), estimating a corrected error rate of 0.0286% per base. This estimate is still conservative in that it does not account for mismatches observed multiple times at a single site, which should overwhelmingly represent residual polymorphism.

Genome assembly

Most recent approaches to *de novo* genome assembly are based on the de Bruijn graph paradigm, which offers a substantial computational advantage over overlap-layout-consensus (OLC) approaches when applied to large datasets. Nevertheless, for datasets with moderate sequencing depth (such as TruSeq synthetic long-read libraries), OLC approaches can be computationally tractable and tend to be less affected by both repeats and sequencing errors than de Bruijn graph-based algorithms. Likewise, many modern Bruijn graph-based assemblers simply do not permit reads exceeding arbitrary length cutoffs. We therefore elected to use the Celera Assembler (Myers et al., 2000), an OLC assembler developed and used to generate the first genome sequence of a multicellular organism, *Drosophila melanogaster* (Adams et al., 2000), as well as one of the first diploid human genome sequences (Levy et al., 2007).

As TruSeq synthetic long-reads share some characteristics with consensus-corrected PacBio reads, we applied Celera Assembler parameters recommended for these PacBio data to take advantage of the read length and low error rate (Koren et al., 2012, 2013). In particular, the approach uses a different unitigger algorithm, decreases the unitig error rates (which is made possible by low synthetic long-read error rates and low rates of polymorphism) and increased the k-mer size to increase overlap specificity. Upon observing partially overlapping contigs among the output of the Celera Assembler, we decided to use the program Minimus2 (Sommer et al., 2007) to merge these contigs into supercontigs, reducing redundancy and improving assembly contiguity. Parameters used for both assembly programs are further described in the Supplemental Materials.

For the down-sampled assemblies with lower coverage, we based the expected coverage on the average euchromatic autosomal depth of coverage of 34× for the full dataset. We randomly sampled reads from a concatenated FASTQ of all six libraries until the total length of the resulting dataset was equal to the desired coverage.

Assessment of assembly quality

We aligned the contigs produced by the Celera Assembler to the reference genome sequence using the NUCmer pipeline (version 3.23) (Delcher et al., 2002; Kurtz et al., 2004). From this alignment, we used the delta-filter tool to extract the best mapping of each query draft contig onto the high quality reference sequence (see Supplemental Materials). We then used to coordinates of these alignments to both measure overall assembly quality and investigate assembly of particular genomic features, including genes and TEs. Using this alignment, we identified the locations of reference-annotated gene and TE sequences in our assembly and used local alignment with BLASTN (Altschul et al., 1997) to determine sequence identity and length ratio (assembled length/reference length) for each sequence. To calculate correctness metrics, we used the tool QUAST (version 2.3) which again uses the NUCmer alignment to the reference genome to calculate the prevalence of mismatches, indels, and other mis-assembly events.

The GLMM used to test the characteristics of TEs that affected accurate assembly were built using the *lme4* package (Bates et al., 2013) within the R statistical computing environment (R Core Team, 2013). TE features (predictor variables) were available for all but the Y family of TEs, which was recently annotated (Release 5.56). The response variable was represented by a binary indicator denoting whether or not the entire length of the TE was accurately assembled. This model assumed a binomial error distribution with a logit link function. TE copy length, GC content (including 1 Kbp flanking regions on each side), divergence (number of substitutions per base compared to the canonical sequence of the TE family), number of high identity (<0.01 substitutions per base compared to the canonical sequence) copies per family, and the interaction between high identity copies and divergence were included as fixed effects, while TE family was included as a random effect. All predictor variables were standardized to zero mean and unit variance prior to fitting, in order to compare the magnitude of the effects.

All figures with the exception of those in the supplement were generated using the *ggplot2* package (Wickham, 2009).

Data access

Sequence data can be found under the NCBI BioProject: PRJNA235897, BioSample: SAMN02588592. Experiment SRX447481 references the synthetic long-read data, while experiment SRX503698 references the underlying short read data. The main genome assembly is available from FigShare at `http://dx.doi.org/10.6084/m9.figshare.985645` and the QUAST contig report is available at `http://dx.doi.org/10.6084/m9.figshare.985916`. Scripts written to assess presence or absence of genomic features in the *de novo* assembly can be found in a GitHub repository at <`https://github.com/rmccoy7541/assess-assembly`> while other analysis scripts, including those to reproduce down-sampled assemblies, can be found in a separate GitHub repository at <`https://github.com/rmccoy7541/dmel-longread-assembly`>. The parameter choices for various software packages are described in the Supplemental Materials.

Acknowledgements

Thank you to Alan Bergland for performing the DNA extractions and to Anthony Long for providing the strain. Thanks also to Julie Collens and Courtney McCormick for preparing and delivering the synthetic long-read and underlying short read libraries. And thank you to Jared Simpson, Brian Walenz, and Sergey Koren for advice regarding the genome assembly, as well as the engineers and administrators of the Proclus and SCG computing clusters. This work was supported by NIH grants R01 GM100366, R01 GM097415, and R01 GM089926 to DAP.

Author contributions

RCM, RWT, and ASFL contributed to the data analysis. RCM prepared the manuscript, ASFL also contributed to the writing of the manuscript, and all other authors contributed comments and revisions. JLK provided guidance on analyses throughout, and DP provided input regarding descriptions of the synthetic long-read technology. TAB and MK contributed to the data generation and provided guidance during planning stages of the experiment. DAP and ASFL helped design the experiment and provided guidance on analyses. All authors read and approved the final manuscript.

Disclosure declaration

TAB was Head of Molecular Biology at Moleculo Inc. from January 16, 2012 to December 31, 2012. Upon acquisition of Moleculo Inc. by Illumina Inc. on December 31, 2012, TAB was retained as a Staff Scientist at Illumina Inc. The sequencing libraries presented herein were prepared and sequenced at Illumina Inc. under TAB's supervision as part of a collaboration between Illumina Inc. and the lab of DAP.

Figure legends

Figure 1: Characteristics of TruSeq synthetic long-reads. **A:** Read length distribution. **B, C, & D:** Position-dependent profiles of **B:** mismatches, **C:** insertions, and **D:** deletions compared to the reference genome. Error rates presented in these figures represent all differences with the reference genome, and can be due to errors in the reads, mapping errors, errors in the reference genome, or accurate sequencing of residual polymorphism.

Figure 2: Depth of coverage per chromosome arm. The suffix "Het" indicates the heterochromatic portion of the corresponding chromosome. M refers to the mitochondrial genome of the *y;cn,bw,sp* strain. U and Uextra are additional scaffolds in the reference assembly that could not be mapped to chromosomes.

Figure 3: Probability of accurate (100% length and sequence identity) TE assembly with respect to significant predictor variables: TE length ($b = -1.633$, $Z = -20.766$, $P < 2 \times 10^{-16}$), GC content ($b = 0.186$, $Z = 3.171$, $P = 0.00152$), divergence ($b = 0.692$, $Z = 7.501$, $P = 6.35 \times 10^{-14}$), and number of high identity (< 0.01 substitutions per base compared to the canonical sequence) copies within family ($b = -0.529$, $Z = -2.936$, $P = 0.00333$). Black lines represent predicted values from the GLMM fit to the binary data (colored points). The upper sets of points represent TEs which were perfectly assembled, while the lower set of points represent TEs which are absent from the assembly or were mis-assembled with respect to the reference. The exact positions of the colored points along the Y-axis should therefore be disregarded. Colors indicate different TE families (122 total). To visualize the interaction between divergence and the number of high identity copies ($b = 0.382$, $Z = 3.921$, $P = 8.81 \times 10^{-5}$), we plotted predicted values for both families with low numbers of high identity copies (dashed line) as well as families with high numbers of high identity copies (solid line).

Figure 4: Assembly metrics as a function of depth of coverage of TruSeq synthetic long-reads. **A:** NG(X)

contig length for full and down-sampled coverage data sets. This metric represents the size of the contig for which X% of the genome length (180 Mbp) lies in contigs of that size or longer. **B:** The proportion of genes and transposable elements accurately assembled (100% length and sequence identity) for full and down-sampled coverage data sets.

Figure S1: Diagram of the TruSeq synthetic long-read library preparation protocol.

Figure S2: Dot plots depicting NUCmer (Delcher et al., 2002) alignment between assembled contigs and the reference genome. Segments off of the diagonal represent various classes of mis-assembly (insertions, deletions, or translocations with respect to the reference sequence). Red segments represent forward alignments, while blue segments indicate an inversion with respect to the rest of the contig alignment. Dot plots were generated using the mummerplot feature of MUMmer (Kurtz et al., 2004).

Figure S3: IGV screenshot (Robinson et al., 2011; Thorvaldsdóttir et al., 2013) of a representative case where assembly fails due to a deficiency of long-read data derived from a long transposable element sequence. The upper-most track (blue) represents the NUCmer alignment of assembled contigs to the reference genome. The middle track represents the BWA alignment of the underlying TruSeq synthetic long-reads. For each of these tracks, blue and red shading indicate the orientation of the alignment (i.e. whether the sequence is reverse complemented). The bottom tracks (blue) indicates the boundaries of genes and transposable elements.

Figure S4: IGV screenshots (Robinson et al., 2011; Thorvaldsdóttir et al., 2013) of representative cases where assembly succeeds or fails based on characteristics of TEs in the genomic region. See the legend of Figure S4 for descriptions of each of the alignment tracks. **A:** A case where assembly fails in the presence of tandem repeats of elements from the Dm88 family. **B:** A case where assembly succeeds in a repeat-dense region of chromosome arm 2R.

Figures

656

Figure 1

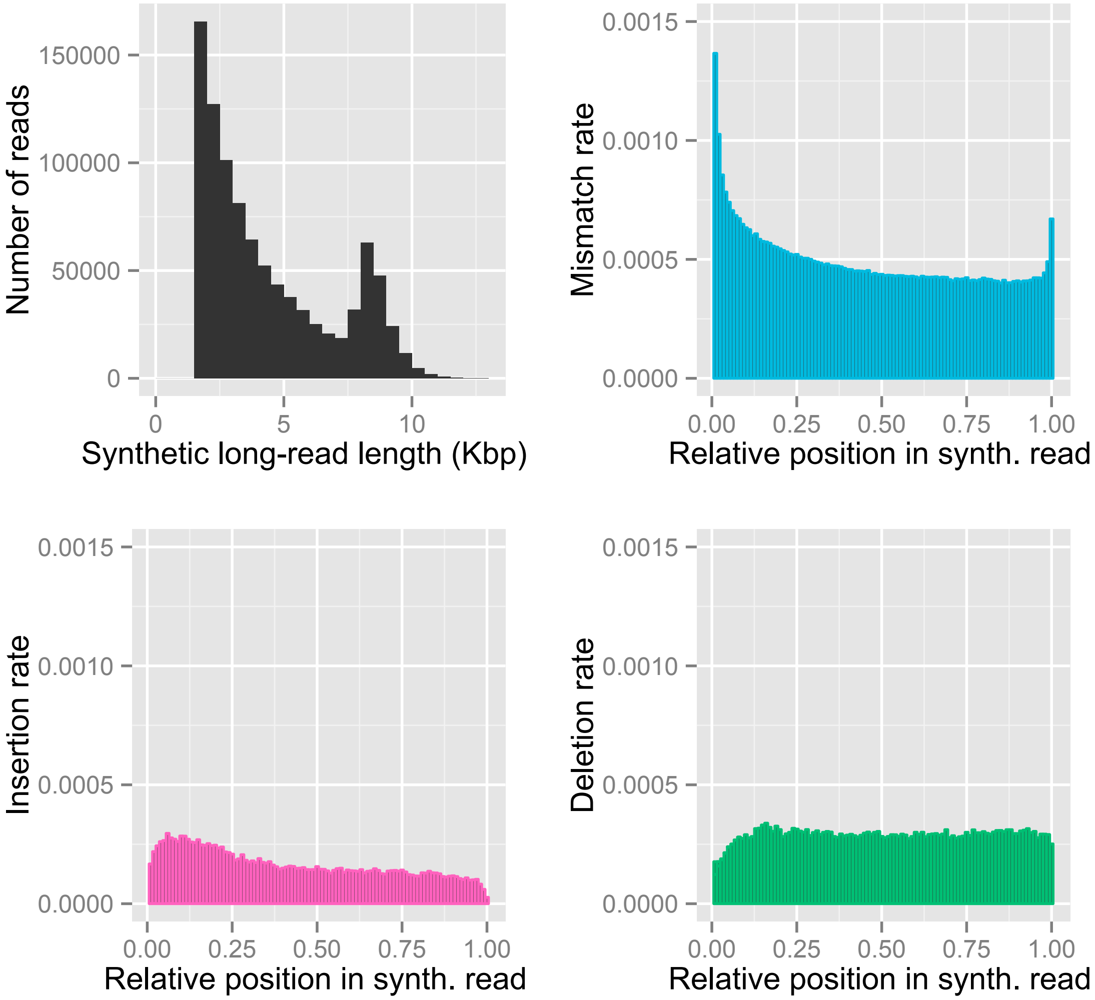


Figure 2

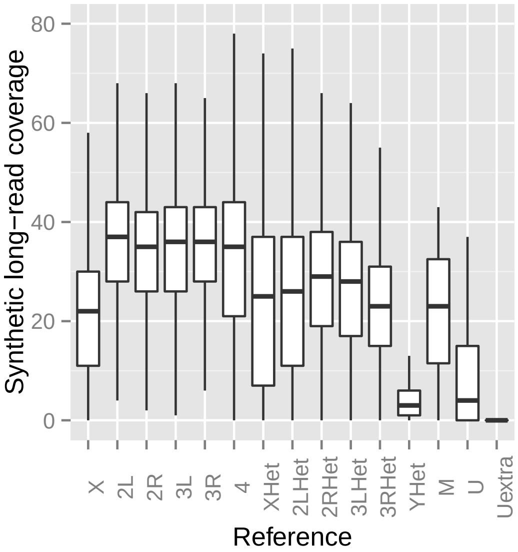

Synthetic long−read coverage
80
60
40
20
0
X
2L
2R
3L
3R
4
XHet
2LHet
2RHet
3LHet
3RHet
YHet
M
U
Uextra
Reference

Figure 3

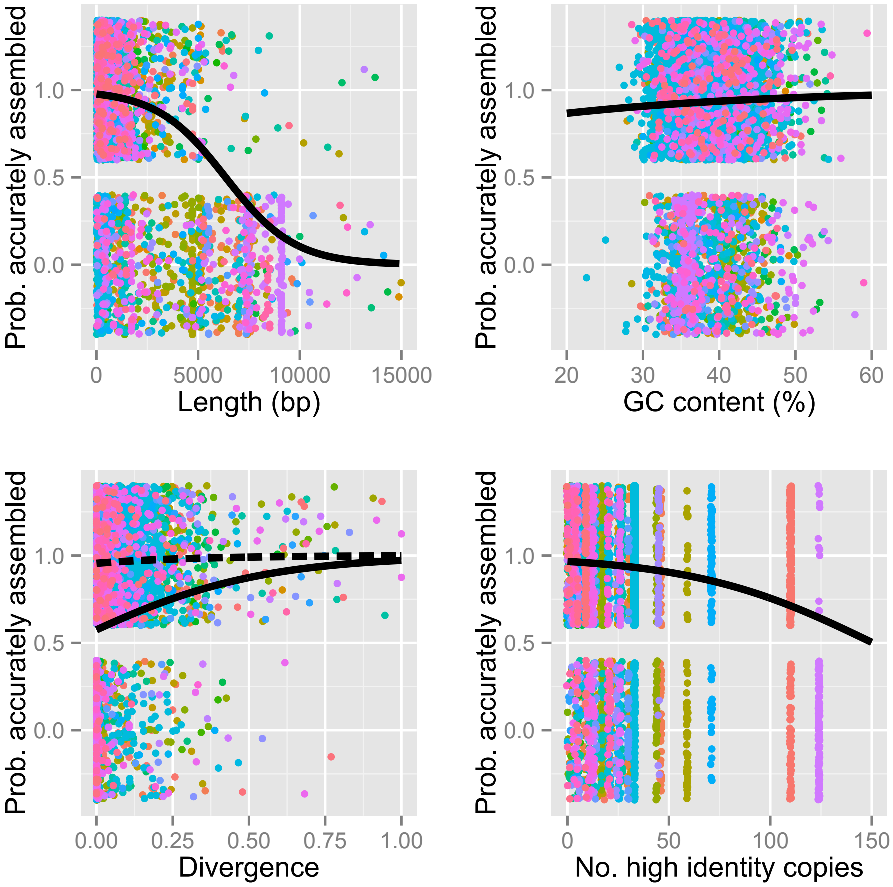


Figure 4

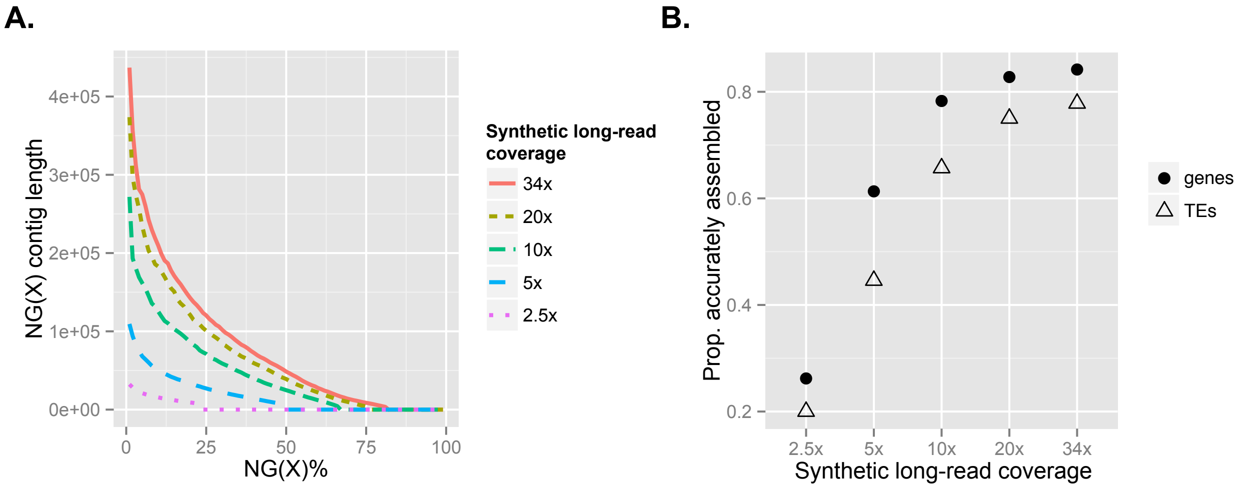


Figure S1

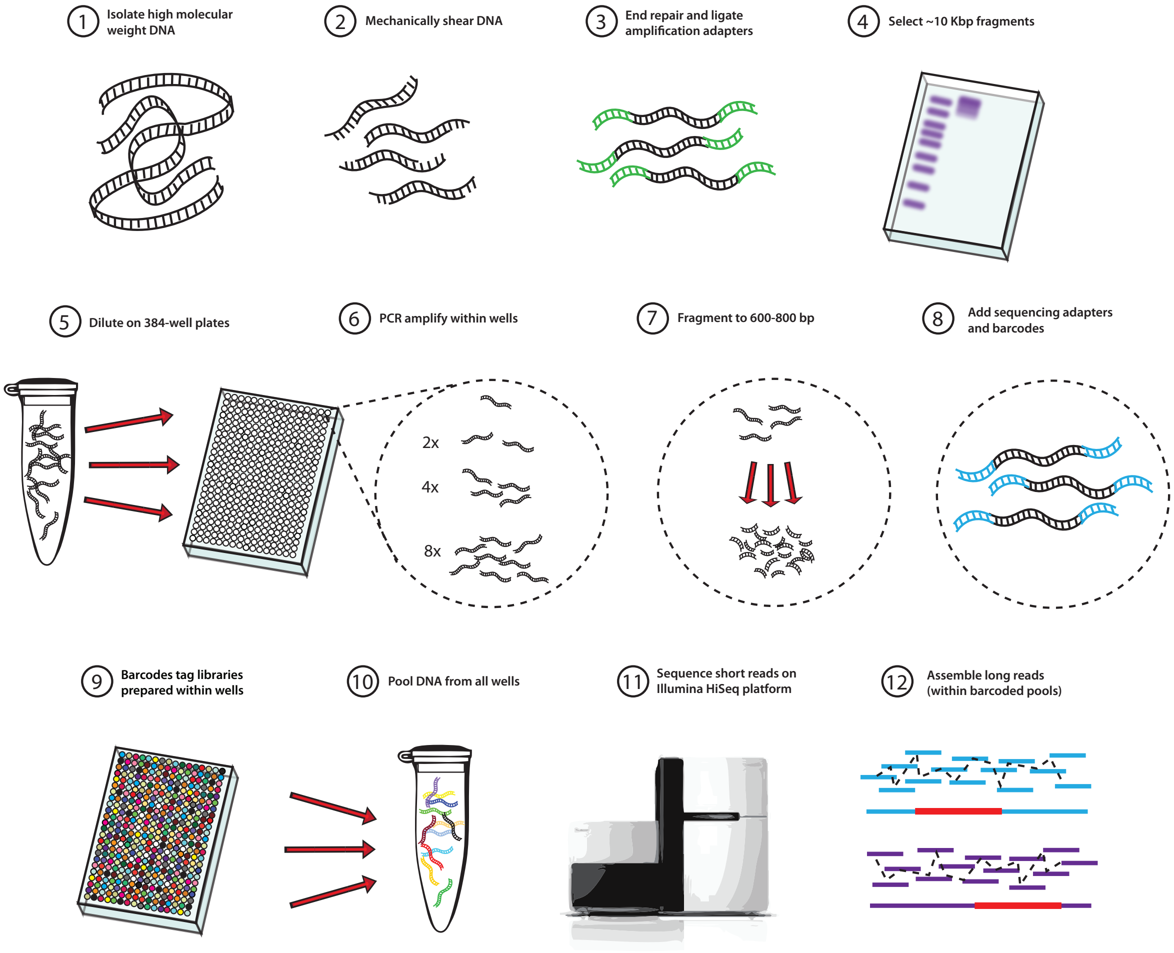
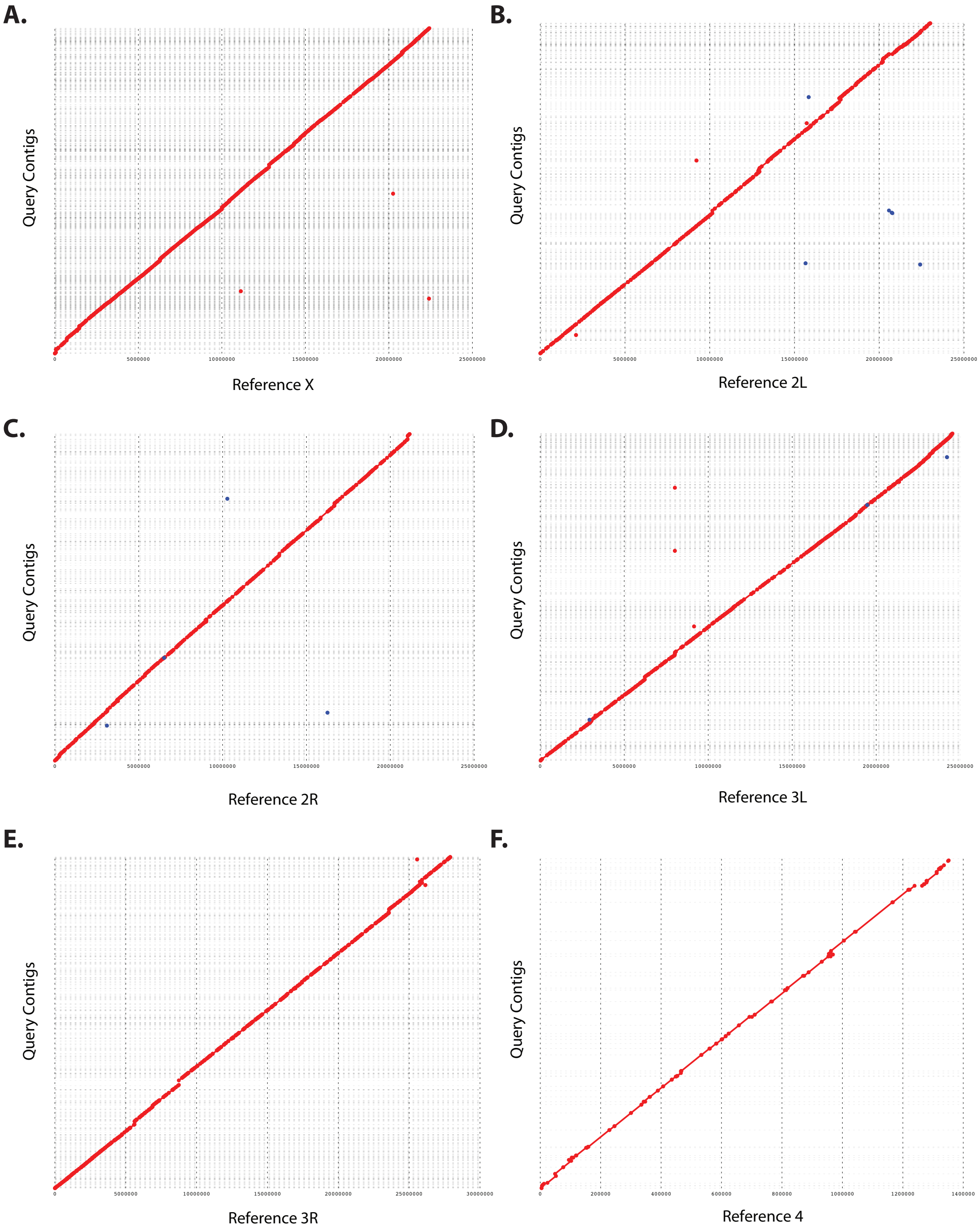

1 Isolate high molecular weight DNA
2 Mechanically shear DNA
3 End repair and ligate amplification adapters
4 Select ~10 Kbp fragments
5 Dilute on 384-well plates
6 PCR amplify within wells
7 Fragment to 600-800 bp
8 Add sequencing adapters and barcodes
2x
4x
8x
9 Barcodes tag libraries prepared within wells
10 Pool DNA from all wells
11 Sequence short reads on Illumina HiSeq platform
12 Assemble long reads (within barcoded pools)

Figure S2


A.
Query Contigs
0
5000000
10000000
15000000
20000000
25000000
Reference X
B.
Query Contigs
0
5000000
10000000
15000000
20000000
25000000
Reference 2L
C.
Query Contigs
0
5000000
10000000
15000000
20000000
25000000
Reference 2R
D.
Query Contigs
0
5000000
10000000
15000000
20000000
25000000
Reference 3L
E.
Query Contigs
0
5000000
10000000
15000000
20000000
25000000
30000000
Reference 3R
F.
Query Contigs
0
200000
400000
600000
800000
1000000
1200000
1400000
Reference 4

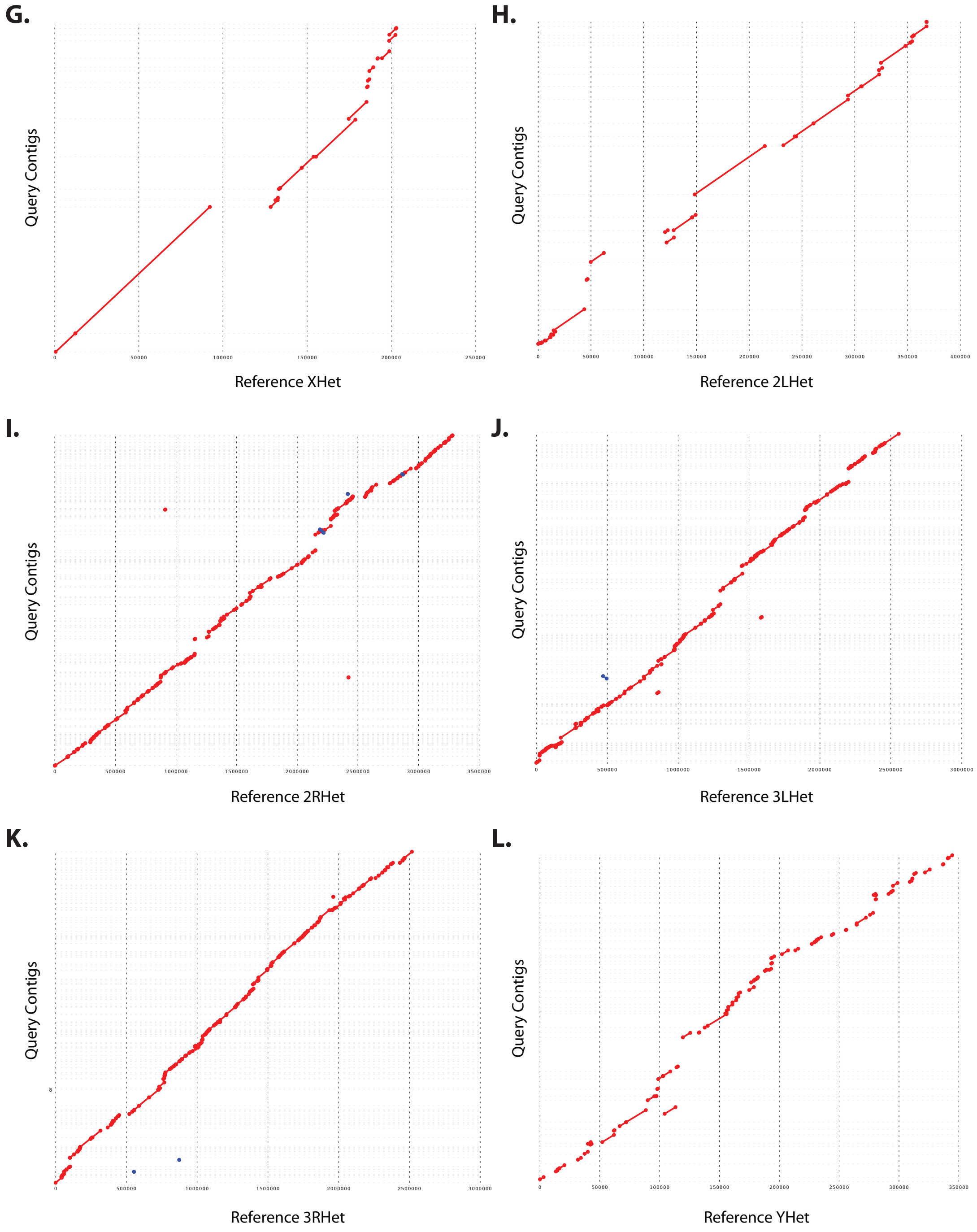

G.
Query Contigs
0
50000
100000
150000
200000
250000
Reference XHet
H.
Query Contigs
0
50000
100000
150000
200000
250000
300000
350000
400000
Reference 2LHet
I.
Query Contigs
0
500000
1000000
1500000
2000000
2500000
3000000
3500000
Reference 2RHet
J.
Query Contigs
0
500000
1000000
1500000
2000000
2500000
3000000
Reference 3LHet
K.
Query Contigs
0
500000
1000000
1500000
2000000
2500000
3000000
Reference 3RHet
L.
Query Contigs
0
50000
100000
150000
200000
250000
300000
350000
Reference YHet

Figure S3

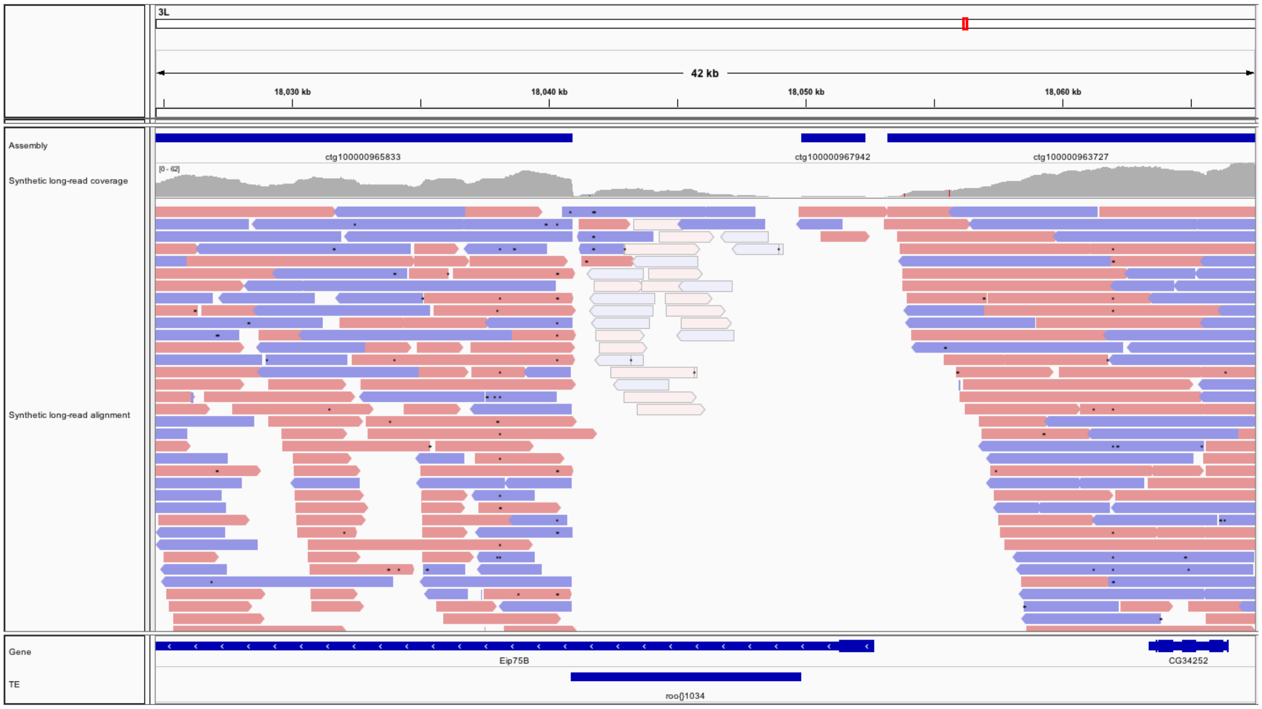
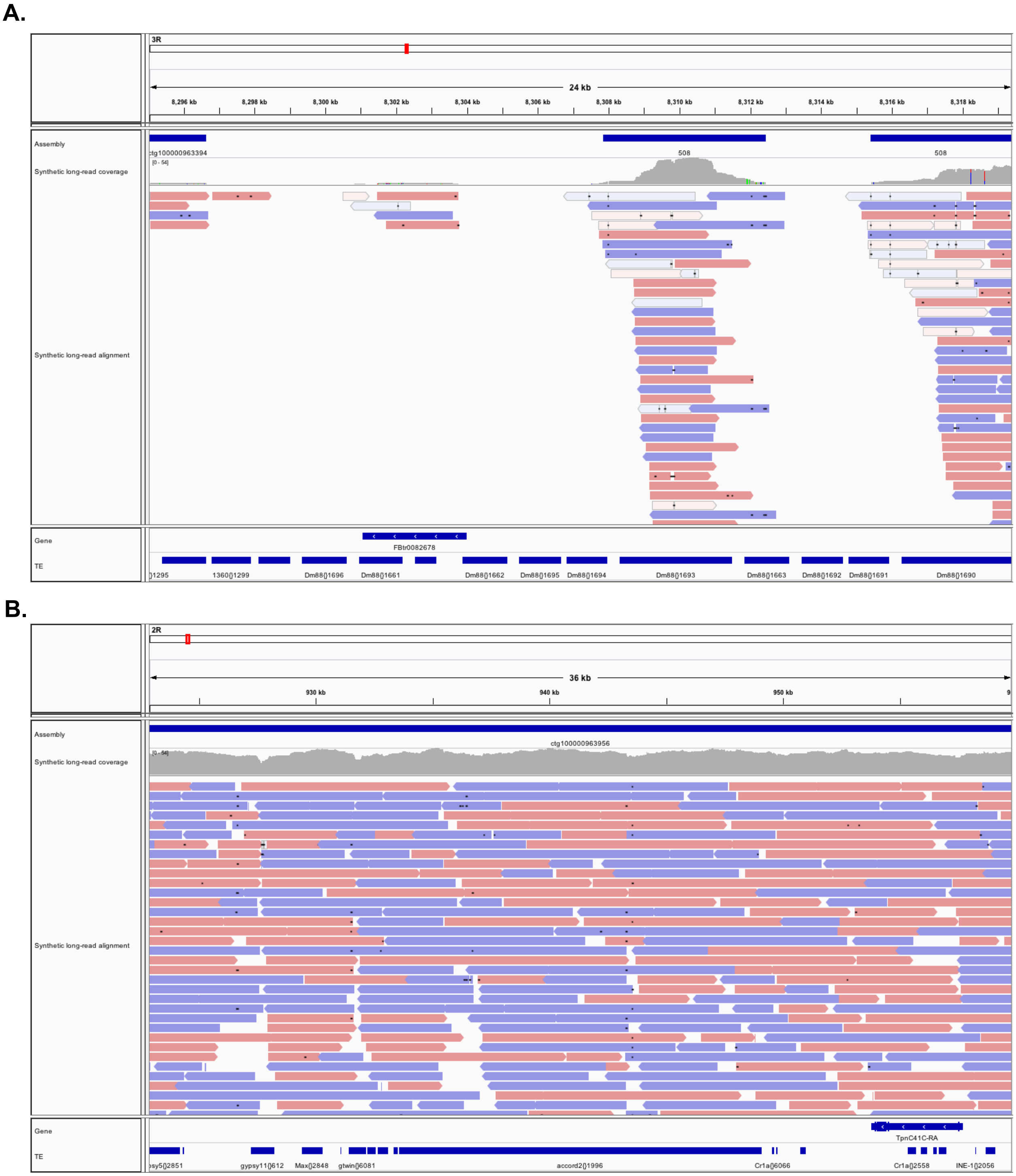

3L
42 kb
18,030 kb
18,040 kb
18,050 kb
18,060 kb
Assembly
ctg100000965833
ctg100000967942
ctg100000963727
Synthetic long-read coverage
[0 - 62]
Synthetic long-read alignment
Gene
Eip75B
CG34252
TE
roo{}1034

Figure S4


A.
3R
24 kb
8,296 kb
8,298 kb
8,300 kb
8,302 kb
8,304 kb
8,306 kb
8,308 kb
8,310 kb
8,312 kb
8,314 kb
8,316 kb
8,318 kb
Assembly
Synthetic long-read coverage
Synthetic long-read alignment
ctg100000963394
[0 - 54]
508
508
Gene
TE
FBtr0082678
1295
1360{}1299
Dm88{}1696
Dm88{}1661
Dm88{}1662
Dm88{}1695
Dm88{}1694
Dm88{}1693
Dm88{}1663
Dm88{}1692
Dm88{}1691
Dm88{}1690
B.
2R
36 kb
930 kb
940 kb
950 kb
Assembly
Synthetic long-read coverage
Synthetic long-read alignment
ctg100000963956
[0 - 54]
Gene
TE
TpnC41C-RA
sy5{}2851
gypsy11{}612
Max{}2848
gtwin{}6081
accord2{}1996
Cr1a{}6066
Cr1a{}2558
INE-1{}2056

657 Tables

Table 1: Size and correctness metrics for *de novo* assembly. The N50 length metric measures the length of the contig for which 50% of the total assembly length is contained in contigs of that size or larger, while the L50 metric is the rank order of that contig if all contigs are ordered from longest to shortest. NG50 and LG50 are similar, but based on the expected genome size of 180 Mbp rather than the assembly length. QUAST (Gurevich et al., 2013) metrics are based on alignment of contigs to the euchromatic reference chromosome arms (which also contain most of the centric heterochromatin). NA50 and LA50 are analogous to N50 and L50, respectively, but in this case the lengths of aligned blocks rather than contigs are considered.

Metric	Value
Number of contigs	5598
Total size of contigs	147445959
Longest contig	567504
Shortest contig	1506
Number of contigs > 10 Kbp	2805
Number of contigs > 100 Kbp	331
Mean contig size	26339
Median contig size	10079
N50 contig length	69692
L50 contig count	554
NG50 contig length	48552
LG50 contig count	833
Contig GC content	42.26%
Genome fraction	96.86% (92.24%)
Duplication ratio	1.15 (1.14)
NA50	60103 (63010)
LA50	623 (618)
Mismatches per 100 Kbp	7.77 (21.9)
Short indels (≤5 bp) per 100 Kbp	5.10 (7.93)
Long indels (>5 bp) per 100 Kbp	0.46 (1.05)
Fully unaligned contigs	377 (179)
Partially unaligned contigs	1214 (70)

*Values in parentheses represent metrics calculated upon inclusion of the heterochromatic reference scaffolds (XHet, 2LHet, 2RHet, 3LHet, 3RHet, YHet, and U), which contain gaps of arbitrary size and are in some cases not oriented with respect to one another (see Release Notes <`http://www.fruitfly.org/data/sequence/README.RELEASE5`>). Values outside of parentheses represent comparison of the assembly only to high quality reference scaffolds X, 2L, 2R, 3L, 3R, and 4.

Table 2: Alignment statistics for Celera Assembler contigs aligned to the reference genome with NUCmer (Delcher et al., 2002; Kurtz et al., 2004), filtered to extract only the optimal placement of each draft contig on the reference (see Supplemental Materials). Note that the number of gaps can be substantially fewer than the number of aligned contigs because alignments may partially overlap or be perfectly adjacent with respect to the reference. The number of gaps can also exceed the number of aligned contigs due to multiple partial alignments of contigs to the reference sequence.

Reference	Aligned contigs	Alignment gaps	Length aligned (bp)	Percent aligned
X	1141	797	20720725	92.4%
2L	547	271	22354714	97.1%
2R	586	291	20645481	97.6%
3L	712	349	23835623	97.1%
3R	657	304	27453817	98.3%
4	74	40	1232723	91.2%
XHet	32	8	153247	75.1%
2LHet	41	10	278753	75.6%
2RHet	278	68	2497813	75.9%
3LHet	206	75	2233661	87.4%
3RHet	231	74	2100876	83.5%
YHet	29	38	151545	43.7%
M	0	1	0	0%
U	1158	1198	4512500	44.9%

Table S1: Number of read pairs in Illumina short read libraries (2×100 bp) and corresponding TruSeq synthetic long-read libraries (1.5-15 Kbp). In the case of mol-32-2827 and mol-32-283d, short read data from separate flow cells were combined, as indicated.

Short read library ID	Flow cell & lane ID	No. read pairs	TruSeq library ID	No. synthetic long-reads
LP6005512-DNA_A01-LRAAA-05	D2672ACXX, 1	212463575	mol-32-281c	170951
LP6005512-DNA_A01-LRAAA-06	D2672ACXX, 2	203972521	mol-32-2827	240750
	D2B7LACXX, 7	82066168		
LP6005512-DNA_A01-LRAAA-07	D2672ACXX, 3	196599647	mol-32-2832	174387
LP6005512-DNA_A01-LRAAA-08	D2672ACXX, 4	154537575	mol-32-283d	254770
	D2B7LACXX, 8	175910619		
LP6005512-DNA_A01-LRAAA-09	C2A96ACXX, 3	174398573	mol-32-2f5f	59705
LP6005512-DNA_A01-LRAAA-10	C2A96ACXX, 4	182493763	mol-32-2f6a	55273

Table S2: Top BLAST hits to the NCBI nucleotide database for all TruSeq synthetic long-reads. Only species/strains with ≥6 hits are reported here.

No. long reads	Species/strain of top BLAST hit
953797	*Drosophila melanogaster*
214	*Gluconacetobacter diazotrophicus* PAl 5
175	*Enterobacteria* phage HK629
163	*Gluconacetobacter xylinus* E25
114	*Gluconacetobacter xylinus* NBRC 3288
97	*Gluconobacter oxydans* 621H
96	*Drosophila mauritiana*
83	*Gluconobacter oxydans* H24
76	*Acetobacter pasteurianus* 386B
58	Cloning vector pSport1
44	*Drosophila pseudoobscura pseudoobscura*
30	*Drosophila simulans*
30	synthetic construct
25	*Acetobacter pasteurianus* IFO 3283-01
14	*Drosophila sechellia*
10	*Burkholderia lata*
9	*Cloning vector* placZ.attB
8	*Acetobacter aceti* NBRC 14818
7	*Acetobacter pasteurianus* IFO 3283-01/12
7	*Agrobacterium fabrum* str. C58
7	*Azospirillum brasilense* Sp245
6	*Granulibacter bethesdensis* CGDNIH3
6	*Rhodomicrobium vannielii* ATCC 17100
6	*Zymomonas mobilis mobilis* ATCC 29191

Table S3: Family membership of TEs overlapping gaps in the alignment of the genome assembly to the high quality reference genome. Families with ≥10 overlaps are reported here.

Family	No. TE copies
roo	117
INE-1	84
1360	34
F	26
FB	21
invader4	20
297	18
mdg1	16
Dm88	15
Doc	15
Tirant	14
HMS-Beagle	11
opus	11
copia	10
invader1	10
invader3	10

Table S4: Assembly results for all annotated transposable elements in the *D. melanogaster* genome. As inKaminker et al. (2002), we report the average length of TE copies within each family, the average divergence between each copy and the canonical sequence, and the number of elements that comprise each family. We then report the number of elements of each family entirely recovered in our assembly with perfect identity to the reference genome, as well as the number that are partially recovered, mis-assembled, or contain mismatches relative to the reference. Finally, we report the number of elements from each family that are entirely absent from the assembly (i.e., both start and end coordinates lie within alignment gaps).

Family	Length	Divergence	Total	Full length	Partial/Mis-assembled	Absent
1360	758	0.059	304	241	56	7
17.6	4852	0.014	20	6	14	0
1731	1112	0.109	13	10	3	0
297	3906	0.044	80	35	41	4
3S18	2816	0.070	17	11	2	4
412	5414	0.036	37	11	25	1
accord	1976	0.195	3	2	1	0
accord2	3707	0.089	7	6	1	0
aurora	3124	NA	1	1	0	0
baggins	1625	0.027	35	29	4	2
Bari1	1447	0.019	6	6	0	0
Bari2	663	0.103	5	5	0	0
blood	7121	0.008	25	1	24	0
BS	1074	0.040	43	37	6	0
BS3	703	0.037	29	28	0	1
BS4	749	NA	1	1	0	0
Burdock	3319	0.050	22	10	12	0
Circe	2473	0.122	5	4	1	0
copia	4233	0.020	35	6	29	0
Cr1a	1597	0.092	152	136	14	2
diver	5029	0.039	11	1	9	1
diver2	1231	0.107	47	39	5	3
Dm88	1698	0.144	31	9	10	12
Doc	3386	0.025	68	19	41	8
Doc2	1688	0.161	7	5	2	0
Doc3	1229	0.259	21	17	3	1
Doc4	1925	0.315	7	7	0	0
F	3025	0.108	70	30	39	1
FB	1063	0.129	60	37	21	2
flea	3358	0.077	29	11	17	1
frogger	1986	NA	2	1	1	0
Fw2	1683	0.196	9	8	1	0
Fw3	423	NA	7	6	1	0
G	916	0.227	17	12	5	0
G2	1051	0.067	22	20	2	0
G3	1996	0.095	7	6	1	0
G4	1212	0.038	28	27	1	0
G5	994	0.069	25	22	3	0
G5A	735	0.063	27	27	0	0
G6	1346	0.112	10	10	0	0
G7	553	0.048	4	4	0	0
GATE	2915	0.080	20	11	7	2

Table S4 continued: Assembly results for all annotated transposable elements in the *D. melanogaster* genome. As in Kaminker et al. (2002), we report the average length of TE copies within each family, the average divergence between each copy and the canonical sequence, and the number of elements that comprise each family. We then report the number of elements of each family entirely recovered in our assembly with perfect identity to the reference genome, as well as the number that are partially recovered, mis-assembled, or contain mismatches relative to the reference. Finally, we report the number of elements from each family that are entirely absent from the assembly (i.e., both start and end coordinates lie within alignment gaps).

Family	Length	Divergence	Total	Full length	Partial/Mis-assembled	Absent
gtwin	1559	0.084	19	17	1	1
gypsy	1514	0.147	18	17	0	1
gypsy2	2840	0.077	12	10	2	0
gypsy3	1629	0.126	15	13	2	0
gypsy4	1253	0.144	15	13	2	0
gypsy5	1879	0.144	10	7	3	0
gypsy6	1353	0.071	15	13	1	1
gypsy7	1292	0.126	4	4	0	0
gypsy8	980	0.103	57	54	1	2
gypsy9	1276	0.136	10	9	1	0
gypsy10	2886	0.086	7	7	0	0
gypsy11	1316	0.185	5	5	0	0
gypsy12	1391	0.103	50	45	4	1
H	1049	0.170	59	44	9	6
HB	1017	0.061	60	51	9	0
Helena	674	0.079	9	9	0	0
HeT-A	2436	0.036	25	8	17	0
HeT-Tag	21	0.012	23	1	22	0
HMS-Beagle	4610	0.043	23	7	14	2
HMS-Beagle2	2710	0.096	13	8	4	1
hopper	857	0.027	24	15	8	1
hopper2	1011	0.063	14	11	3	0
I	2350	0.113	38	24	8	6
Idefix	2169	0.114	17	12	5	0
INE-1	246	0.112	2235	2106	65	64
invader1	911	0.060	45	25	11	9
invader2	2196	0.063	19	12	6	1
invader3	1994	0.054	33	15	12	6
invader4	730	0.020	32	13	6	13
invader5	4175	0.106	3	2	1	0
invader6	1320	0.090	8	8	0	0
Ivk	2755	0.094	11	8	3	0
jockey	1605	0.040	96	76	16	4
jockey2	549	0.060	28	27	1	0
Juan	3272	0.037	11	9	2	0
looper1	1214	0.066	4	4	0	0
mariner2	627	0.064	23	22	1	0
Max	2393	0.302	21	17	4	0
McClintock	1781	0.046	8	5	2	1
mdg1	4894	0.052	41	12	25	4
mdg3	3254	0.034	21	9	10	2

Table S4 continued: Assembly results for all annotated transposable elements in the *D. melanogaster* genome. As in Kaminker et al. (2002), we report the average length of TE copies within each family, the average divergence between each copy and the canonical sequence, and the number of elements that comprise each family. We then report the number of elements of each family entirely recovered in our assembly with perfect identity to the reference genome, as well as the number that are partially recovered, mis-assembled, or contain mismatches relative to the reference. Finally, we report the number of elements from each family that are entirely absent from the assembly (i.e., both start and end coordinates lie within alignment gaps).

Family	Length	Divergence	Total	Full length	Partial/Mis-assembled	Absent
micropia	1771	0.133	13	8	4	1
ninja-Dsim-like	1390	0.315	19	15	1	3
NOF	2609	0.071	8	2	4	2
opus	4824	0.074	31	9	21	1
pogo	651	0.006	48	44	4	0
Porto1	1090	0.013	7	7	0	0
Q	124	0.277	5	5	0	0
Quasimodo	3922	0.089	29	16	12	1
R1-2	802	NA	2	2	0	0
R1A1	1169	0.256	27	18	8	1
roo	7411	0.009	136	12	111	13
rooA	3654	0.053	17	12	5	0
rover	4091	0.041	7	4	3	0
Rt1a	2132	0.048	26	23	2	1
Rt1b	2945	0.046	60	45	12	3
Rt1c	1050	0.084	34	24	7	3
S	1102	0.471	65	48	16	1
S2	575	0.054	14	10	1	3
springer	2836	0.067	24	16	7	1
Stalker	2748	0.025	18	9	8	1
Stalker2	5853	0.043	16	7	9	0
Stalker3	31	NA	1	1	0	0
Stalker4	2559	0.054	37	22	12	3
Tabor	2330	0.059	9	6	3	0
TART-A	2928	0.038	11	5	2	4
TART-B	258	NA	3	2	1	0
TART-C	987	NA	1	1	0	0
Tc1	947	0.039	26	25	1	0
Tc1-2	857	0.049	24	23	1	0
Tc3	447	0.096	19	17	2	0
Tirant	6401	0.084	25	4	18	3
Tom1	292	0.055	4	4	0	0
transib1	4581	0.075	3	1	2	0
transib2	918	0.029	24	19	4	1
transib3	1493	0.027	13	11	2	0
transib4	1946	0.049	8	7	1	0
Transpac	4394	0.038	6	1	5	0
X	1466	0.233	55	50	4	1
Xanthias	4533	NA	1	0	1	0
Y	NA	NA	4	1	3	0
ZAM	547	0.508	4	4	0	0

Table S5: Results of fitting a generalized linear mixed model with a binary response variable indicating whether individual TE copies are accurately assembled.

Random effect		Variance	Std. Dev.
Family	(Intercept)	1.330	1.153

Fixed effect	Estimate	Std. Error	z value	$\Pr(>\lvert z\rvert)$
(Intercept)	1.216	0.170	7.135	9.70×10^{-13}
Length	-1.633	0.079	-20.766	$< 2 \times 10^{-16}$
GC content	0.186	0.059	3.171	0.001 52
Divergence	0.692	0.092	7.501	6.35×10^{-14}
High identity copies	-0.529	0.180	-2.936	0.003 33
Divergence × High identity copies	0.382	0.097	3.921	8.81×10^{-5}

Table S6: Contig IDs for sequences with no significant hit to the NCBI nucleotide database.

FASTA contig ID
ctg100000966696
ctg100000966814
ctg100000966837
ctg100000967379
ctg100000967449
ctg100000967457
ctg100000967511
ctg100000967560
ctg100000967605
ctg100000967626
ctg100000967687
ctg100000967750
ctg100000967783
ctg100000967784
ctg100000967787
ctg100000967852
ctg100000967896
ctg100000967928
ctg100000967969
ctg100000968010
ctg100000968064
ctg100000968094
ctg100000968196
ctg100000968200
ctg100000968250
ctg100000968272
ctg100000968281

References

Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., Scherer, S. E., Li, P. W., Hoskins, R. A., Galle, R. F., et al. (2000). The genome sequence of *Drosophila melanogaster*. *Science* **287**:2185–2195.

Alkan, C., Sajjadian, S., and Eichler, E. E. (2010). Limitations of next-generation genome sequence assembly. *Nature Methods* **8**:61–65.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25**:3389–3402.

Bates, D., Maechler, M., Bolker, B., and Walker, S. (2013). *lme4: Linear mixed-effects models using Eigen and S4*. R package version 1.0-5.
URL: *http://CRAN.R-project.org/package=lme4*

Benjamini, Y. and Speed, T. P. (2012). Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Research* **40**:e72.

Bradnam, K. R., Fass, J. N., Alexandrov, A., Baranay, P., Bechner, M., Birol, I., Boisvert, S., Chapman, J. A., Chapuis, G., Chikhi, R., et al. (2013). Assemblathon 2: evaluating *de novo* methods of genome assembly in three vertebrate species. *GigaScience* **2**:10.

Brown, C. T., Howe, A., Zhang, Q., Pyrkosz, A. B., and Brom, T. H. (2012). A reference-free algorithm for computational normalization of shotgun sequencing data. *arXiv.org* .

Casacuberta, E. and González, J. (2013). The impact of transposable elements in environmental adaptation. *Molecular ecology* **22**:1503–1517.

Celniker, S. E., Wheeler, D. A., Kronmiller, B., Carlson, J. W., Halpern, A., Patel, S., Adams, M., Champe, M., Dugan, S. P., Frise, E., et al. (2002). Finishing a whole-genome shotgun: release 3 of the *Drosophila melanogaster* euchromatic genome sequence. *Genome Biology* **3**:RESEARCH0079.

Chen, Z. X., Sturgill, D., Qu, J., Jiang, H., Park, S., Boley, N., Suzuki, A. M., Fletcher, A. R., Plachetzki, D., FitzGerald, P., et al. (in press). Comparative Analysis of the *D. melanogaster* modENCODE Transcriptome Annotation. *Genome Research* .

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., et al. (2007). Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature* **450**:203–218.

Clarke, J., Wu, H.-C., Jayasinghe, L., Patel, A., Reid, S., and Bayley, H. (2009). Continuous base identification for single-molecule nanopore DNA sequencing. *Nature Nanotechnology* **4**:265–270.

Cordaux, R. and Batzer, M. A. (2009). The impact of retrotransposons on human genome evolution. *Nature Reviews Genetics* **10**:691–703.

de Koning, A. P. J., Gu, W., Castoe, T. A., Batzer, M. A., and Pollock, D. D. (2011). Repetitive Elements May Comprise Over Two-Thirds of the Human Genome. *PLoS Genetics* **7**:e1002384.

Delcher, A. L., Phillippy, A., Carlton, J., and Salzberg, S. L. (2002). Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Research* **30**:2478–2483.

Duret, L. and Hurst, L. D. (2001). The elevated GC content at exonic third sites is not evidence against neutralist models of isochore evolution. *Molecular Biology and Evolution* **18**:757–762.

Feschotte, C., Jiang, N., and Wessler, S. R. (2002). Plant transposable elements: where genetics meets genomics. *Nature Reviews Genetics* **3**:329–341.

Fiston-Lavier, A.-S., Anxolabehere, D., and Quesneville, H. (2007). A model of segmental duplication formation in *Drosophila melanogaster*. *Genome Research* **17**:1458–1470.

Fiston-Lavier, A.-S., Carrigan, M., Petrov, D. A., and González, J. (2011). T-lex: a program for fast and accurate assessment of transposable element presence using next-generation sequencing data. *Nucleic Acids Research* **39**:e36.

Glenn, T. C. (2011). Field guide to next-generation DNA sequencers. *Molecular Ecology Resources* **11**:759–769.

González, J., Lenkov, K., Lipatov, M., Macpherson, J. M., and Petrov, D. A. (2008). High Rate of Recent Transposable Element–Induced Adaptation in *Drosophila melanogaster*. *PLoS Biology* **6**:e251.

González, J., Macpherson, J. M., and Petrov, D. A. (2009). A Recent Adaptive Transposable Element Insertion Near Highly Conserved Developmental Loci in *Drosophila melanogaster* .

González, J. and Petrov, D. A. (2009). The adaptive role of transposable elements in the *Drosophila* genome. *Gene* **448**:124–133.

Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**:1072–1075.

Haynes, K. A., Gracheva, E., and Elgin, S. C. R. (2006). A Distinct Type of Heterochromatin Within *Drosophila melanogaster* Chromosome 4. *Genetics* **175**:1539–1542.

Hoskins, R. A., Carlson, J. W., Kennedy, C., Acevedo, D., Evans-Holm, M., Frise, E., Wan, K. H., Park, S., Mendez-Lago, M., Rossi, F., et al. (2007). Sequence finishing and mapping of *Drosophila melanogaster* heterochromatin. *Science* **316**:1625–1628.

Hu, T. T., Eisen, M. B., Thornton, K. R., and Andolfatto, P. (2013). A second-generation assembly of the *Drosophila simulans* genome provides new insights into patterns of lineage-specific divergence. *Genome Research* **23**:89–98.

Huddleston, J., Ranade, S., Malig, M., Antonacci, F., Chaisson, M., Hon, L., Sudmant, P. H., Graves, T. A., Alkan, C., Dennis, M. Y., et al. (2014). Reconstructing complex regions of genomes using long-read sequencing technology. *Genome Research* .

Hurlbert, S. H. (1984). Pseudoreplication and the design of ecological field experiments. *Ecological monographs* **54**:187–211.

Jiao, X., Zheng, X., Ma, L., Kutty, G., Gogineni, E., Sun, Q., Sherman, B. T., Hu, X., Jones, K., Raley, C., et al. (2013). A Benchmark Study on Error Assessment and Quality Control of CCS Reads Derived from the PacBio RS. *Journal of data mining in genomics & proteomics* **4**.

Kaminker, J. S., Bergman, C. M., Kronmiller, B., Carlson, J., Svirskas, R., Patel, S., Frise, E., Wheeler, D. A., Lewis, S. E., Rubin, G. M., et al. (2002). The transposable elements of the *Drosophila melanogaster* euchromatin: a genomics perspective. *Genome Biology* **3**:RESEARCH0084.

Keane, T. M., Wong, K., and Adams, D. J. (2013). RetroSeq: transposable element discovery from next-generation sequencing data. *Bioinformatics* **29**:389–390.

Kidwell, M. G. and Lisch, D. R. (2001). Perspective: transposable elements, parasitic DNA, and genome evolution. *Evolution* **55**:1–24.

Kofler, R., Betancourt, A. J., and Schlötterer, C. (2012). Sequencing of pooled DNA samples (Pool-Seq) uncovers complex dynamics of transposable element insertions in Drosophila melanogaster. *PLoS Genetics* **8**:e1002487.

Koren, S., Harhay, G. P., Smith, T. P., Bono, J. L., Harhay, D. M., McVey, S. D., Radune, D., Bergman, N. H., and Phillippy, A. M. (2013). Reducing assembly complexity of microbial genomes with single-molecule sequencing. *Genome Biology* **14**:R101.

Koren, S., Schatz, M. C., Walenz, B. P., Martin, J., Howard, J. T., Ganapathy, G., Wang, Z., Rasko, D. A., McCombie, W. R., Jarvis, E. D., et al. (2012). Hybrid error correction and *de novo* assembly of single-molecule sequencing reads. *Nature Biotechnology* **30**:693–700.

Kuleshov, V., Xie, D., Chen, R., Pushkarev, D., Ma, Z., Blauwkamp, T., Kertesz, M., and Snyder, M. (2014). Whole-genome haplotyping using long reads and statistical methods. *Nature Biotechnology* .

Kurtz, S., Phillippy, A., Delcher, A. L., Smoot, M., Shumway, M., Antonescu, C., and Salzberg, S. L. (2004). Versatile and open software for comparing large genomes. *Genome Biology* **5**:R12.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. *Nature* **409**:860–921.

Langley, C. H., Crepeau, M., Cardeno, C., Corbett-Detig, R., and Stevens, K. (2011). Circumventing heterozygosity: sequencing the amplified genome of a single haploid *Drosophila melanogaster* embryo. *Genetics* **188**:239–246.

Lee, E., Iskow, R., Yang, L., Gokcumen, O., Haseley, P., Luquette, L. J., Lohr, J. G., Harris, C. C., Ding, L., Wilson, R. K., et al. (2012). Landscape of somatic retrotransposition in human cancers. *Science* **337**:967–971.

Levy, S., Sutton, G., Ng, P. C., Feuk, L., Halpern, A. L., Walenz, B. P., Axelrod, N., Huang, J., Kirkness, E. F., Denisov, G., et al. (2007). The diploid genome sequence of an individual human. *PLoS Biology* **5**:e254.

Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**:1754–1760.

Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., Shi, Z., Li, Y., Li, S., Shan, G., Kristiansen, K., et al. (2010). *De novo* assembly of human genomes with massively parallel short read sequencing. *Genome Research* **20**:265–272.

Linheiro, R. S. and Bergman, C. M. (2012). Whole genome resequencing reveals natural target site preferences of transposable elements in Drosophila melanogaster. *PLoS ONE* **7**:e30008.

Mackay, T. F. C., Richards, S., Stone, E. A., Barbadilla, A., Ayroles, J. F., Zhu, D., Casillas, S., Han, Y., Magwire, M. M., Cridland, J. M., et al. (2012). The *Drosophila melanogaster* Genetic Reference Panel. *Nature* **482**:173–178.

Miller, J. R., Koren, S., and Sutton, G. (2010). Assembly algorithms for next-generation sequencing data. *Genomics* **95**:315–327.

Myers, E. W., Sutton, G. G., Delcher, A. L., Dew, I. M., Fasulo, D. P., Flanigan, M. J., Kravitz, S. A., Mobarry, C. M., Reinert, K. H., and Remington, K. A. (2000). A whole-genome assembly of *Drosophila*. *Science* **287**:2196–2204.

Nagarajan, N. and Pop, M. (2013). Sequence assembly demystified. *Nature Reviews Genetics* **14**:157–167.

Nekrutenko, A. and Li, W. H. (2001). Transposable elements are found in a large number of human protein-coding genes. *Trends in genetics : TIG* **17**:619–621.

Osoegawa, K., Vessere, G. M., Li Shu, C., Hoskins, R. A., Abad, J. P., de Pablos, B., Villasante, A., and de Jong, P. J. (2007). BAC clones generated from sheared DNA. *Genomics* **89**:291–299.

Platzer, A., Nizhynska, V., and Long, Q. (2012). TE-Locate: a tool to locate and group transposable element occurrences using paired-end next-generation sequencing data. *Biology* **1**:395–410.

Quesneville, H., Bergman, C. M., Andrieu, O., Autard, D., Nouaud, D., Ashburner, M., and Anxolabehere, D. (2005). Combined Evidence Annotation of Transposable Elements in Genome Sequences. *PLoS Computational Biology* **1**:e22.

Quesneville, H., Nouaud, D., and Anxolab h re, D. (2003). Detection of New Transposable Element Families in *Drosophila melanogaster* and *Anopheles gambiae* Genomes. *Journal of molecular evolution* **57**:S50–S59.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
URL: *http://www.R-project.org/*

Rebollo, R., Romanish, M. T., and Mager, D. L. (2012). Transposable elements: an abundant and natural source of regulatory sequences for host genes. *Annual Review of Genetics* **46**:21–42.

Robb, S. M. C., Lu, L., Valencia, E., Burnette, J. M., Okumoto, Y., Wessler, S. R., and Stajich, J. E. (2013). The use of RelocaTE and unassembled short reads to produce high-resolution snapshots of transposable element generated diversity in rice. *G3 (Bethesda, Md.)* **3**:949–957.

Roberts, R. J., Carneiro, M. O., and Schatz, M. C. (2013). The advantages of SMRT sequencing. *Genome Biology* **14**:405.

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., and Mesirov, J. P. (2011). Integrative genomics viewer. *Nature Biotechnology* **29**:24–26.

Salzberg, S. L., Phillippy, A. M., Zimin, A., Puiu, D., Magoc, T., Koren, S., Treangen, T. J., Schatz, M. C., Delcher, A. L., Roberts, M., et al. (2012). GAGE: A critical evaluation of genome assemblies and assembly algorithms. *Genome Research* **22**:557–567.

Simpson, J. T. and Durbin, R. (2012). Efficient de novo assembly of large genomes using compressed data structures. *Genome Research* **22**:549–556.

Sommer, D. D., Delcher, A. L., Salzberg, S. L., and Pop, M. (2007). Minimus: a fast, lightweight genome assembler. *BMC Bioinformatics* **8**:64.

Thorvaldsdóttir, H., Robinson, J. T., and Mesirov, J. P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics* **14**:178–192.

Voskoboynik, A., Neff, N. F., Sahoo, D., Newman, A. M., Pushkarev, D., Koh, W., Passarelli, B., Fan, H. C., Mantalas, G. L., Palmeri, K. J., et al. (2013). The genome sequence of the colonial chordate, *Botryllus schlosseri*. *eLife* **2**:e00569.

Wickham, H. (2009). *ggplot2: elegant graphics for data analysis.* Springer New York.
URL: *http://had.co.nz/ggplot2/book*

Ye, L., Hillier, L. W., Minx, P., Thane, N., Locke, D. P., Martin, J. C., Chen, L., Mitreva, M., Miller, J. R., Haub, K. V., et al. (2011). A vertebrate case study of the quality of assemblies derived from next-generation sequences. *Genome Biology* **12**:R31.

Zerbino, D. R. and Birney, E. (2008). Velvet: Algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Research* **18**:821–829.