bioRxiv preprint doi: https://doi.org/10.1101/001834; this version posted January 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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*

bioRxiv preprint doi: https://doi.org/10.1101/001834; this version posted January 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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, mostly due to the presence of repeats, which cannot be reconstructed unambiguously with short read data alone. One class of repeats, called transposable elements (TEs), is particularly problematic due to high sequence identity, high copy number, and a capacity to induce complex genomic rearrangements. Despite their importance to genome function and evolution, most current *de novo* assembly approaches cannot resolve 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 2-15 Kbp with an extremely low error rate (<0.05%). To test the utility of this technology, we sequenced and assembled the genome of the model organism *Drosophila melanogaster* (reference genome strain *yw;cn,bw,sp*) achieving an NG50 contig size of 77.9 Kbp and covering 97.2% of the current reference genome (including heterochromatin). 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 recover and accurately place 80.4% of annotated transposable elements with perfect identity to the current reference genome. As TEs are complex and highly repetitive features that are ubiquitous in genomes across the tree of life, TruSeq synthetic long-read technology offers a powerful approach to drastically improve *de novo* assemblies of whole genomes.

bioRxiv preprint doi: https://doi.org/10.1101/001834; this version posted January 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Introduction

Despite tremendous advances in DNA sequencing technology, computing power, and assembly approaches, *de novo* assembly of whole eukaryotic genomes using high-throughput sequencing data remains a challenge owing largely to the presence of repetitive DNA (Alkan et al., 2010; Treangen and Salzberg, 2012). 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). These elements, a common feature of almost all eukaryotic genomes sequenced to date, are particularly difficult to assemble accurately due to high sequence identity among multiple copies within a genome. In addition to spanning up to tens of kilobases, TEs from a single family can be present in thousands of copies. Consequently, TEs can dramatically affect genome size and structure, as well as genome function; transposition can 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), for example, by creating new transcription factor binding sites (Rebollo et al., 2012) or otherwise affecting expression of nearby genes (González et al., 2009).

Though TEs play a key role in genome evolution, many approaches to *de novo* assembly start by masking TEs and other repeats in order to simplify the assembly of non-repetitive DNA. The end result is a set of disjointed contigs (which may be oriented relative to one another by other means) along with a set of reads or small contigs that were deemed repetitive and could not be placed with respect to the rest of the assembly. For example, the *Drosophila* 12 Genomes Consortium (Clark et al., 2007) did not attempt to place individual TE sequences into the finished genomes. Instead, they attempted to estimate the abundance of TEs with resulting upper and lower-bounds differing by more than three fold.

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 sequence 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 low (Ye et al., 2011). Current approaches to deal with repeats such as TEs generally rely on depth of coverage and paired-end data (Alkan et al., 2010; Miller et al., 2010; Li et al., 2010). Depth of coverage is informative of copy-number, but unfortunately cannot guide accurate placement of repeats. Paired-end data can help resolve the orientation

bioRxiv preprint doi: https://doi.org/10.1101/001834; this version posted January 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

and distance between assembled flanking sequences, but do not resolve the repeat sequence itself. Likewise, if read pairs do not completely span a repeat, anchored in unique sequence, it is impossible to assemble the data unambiguously. Long inserts, commonly referred to as mate-pair libraries, are therefore useful to bridge across long TEs, but are labor-intensive and expensive to construct.

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. Several high-throughput long read (>1 Kbp) technologies have been developed, but most of these technologies have exceptionally high sequencing error rates (although error-correction strategies have been developed in some cases (Schatz et al., 2012)) and are low throughput. High error rates limit the specificity of long reads, meaning that assemblers cannot distinguish between sequencing errors and differences between slightly diverged copies of TEs. For instance, PacBio RS II (Pacific Biosciences) provides average read lengths of greater than 5 Kbp, but with a 15-18% error rate (Schatz et al., 2012). Meanwhile, other established sequencing technologies, such as Illumina, 454 (Roche), and Ion Torrent (Life Technologies), offer lower error rates of 0.1-1%, but relatively shorter read lengths (Glenn, 2011). Illumina has recently introduced a novel technology called TruSeq™ synthetic long-reads <`http://www.illumina.com/services/long-read-sequencing-service.ilmn`>, which builds upon underlying Illumina short read data to generate highly accurate synthetic reads up to 15 Kbp in length. This technology promises to dramatically advance a wide range of genomic applications.

Using a pipeline of standard existing tools, we showcase 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 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 quantify the substantial improvements to assembly enabled by synthetic long read technology. Because *D. melanogaster* harbors a large number (∼100) of families of active TEs, assembly of these repeats is particularly challenging

bioRxiv preprint doi: https://doi.org/10.1101/001834; this version posted January 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

due to the presence of long TE copies with high sequence identity. This is distinct from other species, including humans, which have TE copies that are shorter are more diverged, and therefore easier to assemble. Our demonstration of accurate TE assembly in *D. melanogaster* should therefore translate favorably to many other systems.

bioRxiv preprint doi: https://doi.org/10.1101/001834; this version posted January 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Results

TruSeq synthetic long-reads

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 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. Long reads are then assembled from the short reads using a proprietary assembler that accounts for properties of the molecular biology steps used in the library preparation. By reducing genome representation by approximately 50- to 100-fold, even abundant and identical repeats can be resolved so long as they are not 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 latest version of the reference genome assembly (Release 5; BDGP v3) contains a total of 168.7 Mbp of sequence, 120 Mbp of which is considered to lie in the euchromatin, which is less repeat dense than heterochromatic regions. This genome release also includes 10.0 Mbp of additional 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). Approximately 50 adult individuals from the *yw;cn,bw,sp* strain of *D. melanogaster* were pooled for the isolation of high molecular weight DNA, which was used to generate TruSeq long-read libraries using the aforementioned protocol (Figure S1). The *yw;cn,bw,sp* strain is the same strain which was used to generate the *D. melanogaster* reference genome (Adams et al., 2000). A total of 523,583 synthetic long reads exceeding 1.5 Kbp (an arbitrary length cutoff) were generated with four libraries (one Illumina HiSeq lane per library), comprising a total of 2.52 Gbp. Reads averaged 4,813 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).

bioRxiv preprint doi: https://doi.org/10.1101/001834; this version posted January 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

We first searched for and eliminated possible contaminants by comparing the reads to the NCBI nucleotide database (`http://www.ncbi.nlm.nih.gov/nuccore`) using BLASTN (Altschul et al., 1997) (see Methods; Table S1). The degree of contamination in the TruSeq synthetic long-read libraries prepared by Illumina was extremely low. Of 523,583 total reads, only 0.104% (544 reads) had top hits to non-insect species, and only 0.105% (549 reads) had top hits to species outside of genus *Drosophila*. Of the 523,034 hits to *Drosophila* species, 99.950% (522,772 reads) had top hits to *D. melanogaster*, while only 0.0501% (262 reads) had top matches to other *Drosophila* species. The most abundant contaminant reads had top matches to known symbionts of *D. melanogaster*, including acetic acid bacteria from the genera *Gluconacetobacter*, *Gluconobacter*, and *Acetobacter* (Table S1). Because we could not exclude that the few sequences with no BLAST results may correspond to fly-derived sequences not previously assembled in the reference genome, we included all sequences except those with top matches to non-insect species (523,039 total; e-value threshold 1e-09) in downstream analyses.

In order to evaluate the accuracy of TruSeq synthetic long-reads, we mapped reads to the reference genome of *D. melanogaster*, identifying differences between the mapped reads and the reference sequence (see Methods). Of 523,039 input reads passing our contamination filter, 99.97% (522,901 reads) were successfully mapped to the reference genome, with 92.44% (483,514) mapping uniquely and 95.17% (497,751) having at least one alignment with a MAPQ score $\geq$20. TruSeq long-reads had very few mismatches to the reference at 0.0418% per base (0.0325% for reads with MAPQ $\geq$20) as well as a very low insertion rate of 0.0163% per base (0.0112% for reads with MAPQ $\geq$20) and a deletion rate of 0.0277% per base (0.0209% for reads with MAPQ $\geq$20). Error rates estimated with this mapping approach are conservative, as residual heterozygosity in the sequenced line mimics errors. We therefore used two approaches (see Methods) to calculate corrected error rates. The first approach uses the number of mismatches overlapping known SNPs to correct the error rate to 0.0192%. Along with this estimate, we also estimate that 2.99% of the sites segregating in the *Drosophila* Genetic Reference Panel (DGRP) (Mackay et al., 2012) remain polymorphic in the sequenced strain (i.e. constitute residual heterozygosity). The second approach assumes that all non-singleton mismatches represent true polymorphism, yielding an estimated long-read error rate of 0.0183%. Both estimates are nearly an order of magnitude lower than other high-throughput sequencing technologies ($\sim$0.1% for Illumina, $\sim$1% for 454 and Ion Torrent (Glenn, 2011), 15-18% for PacBio (Schatz et al., 2012)). The reason that TruSeq synthetic long-read achieve such low error rates is that they are built as consensuses of multiple overlapping short Illumina reads. We further observed that all types of errors are more frequent at the beginning of reads, though the pattern is more pronounced for mismatches and deletions (Figures 1B, 1C, & 1D). Minor imprecision in the

bioRxiv preprint doi: https://doi.org/10.1101/001834; this version posted January 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

trimming of adapter sequence is likely responsible for this distinct error profile. Based on the observation of low error rates, no pre-processing steps were necessary to perform in preparation for assembly (though overlap-based trimming and detection of chimeric and spurious reads are performed by default by the Celera Assembler that we used in this study).

We then quantified the average depth of coverage of the mapped long-reads for each reference chromosome arm. We found consistent and uniform coverage of ∼21× of the euchromatin of each of the major autosomes (2L,2R,3L,3R; Figure 2). Coverage of the heterochromatic portions of autosomes was generally lower (∼11-14×), and also varied marginally both within and between chromosomes. This is explained by the fact that heterochromatin has high repeat content relative to euchromatin, making it more difficult to assemble into long reads. Consequently, the fourth chromosome showed relatively lower average read depth (15.8× compare to 21×), likely due to enrichment of heterochromatic islands on this chromosome (Haynes et al., 2006). Read depth on the sex chromosomes is also 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 read depth was lower still at 12.6× for the X chromosome (12.6×) as well as the Y chromosome (2.7×), which is entirely heterochromatic. Read depth for the mitochondrial genome was also relatively low (5.6×) 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

Li and Waterman (2003) showed that in addition to flow cytometry and other molecular biology approaches, genome size can be roughly estimated from raw sequence data by counting the occurrences of distinct k-mers (i.e., unique k-length subsequences of reads) prior to assembly. We used the k-mer counting software KmerGenie (Chikhi and Medvedev, 2014) to produce a k-mer abundance histogram, which depicts the number of occurrences of each unique k-mer within the TruSeq synthetic long-read dataset (Figure S2). The characteristic spike at low coverage represents a combination of errors and residual polymorphism while the high coverage tail represents genomic repeats. The observed abundance peak at approximately 21× provides an independent, reference-free estimate of the average depth of coverage. The relatively small peak at low coverage provides reference-free evidence that the error rate of TruSeq synthetic long-reads is extremely low. By modeling errors, polymorphism, and repeats, the program also estimates an optimal value of k for the assembly (155, here). Based on 155-mer abundance, the program estimated a total assembly length of 120.4 Mbp, in line with the 120 Mbp length of the euchromatic reference, but substantially lower than the 180 Mbp

bioRxiv preprint doi: https://doi.org/10.1101/001834; this version posted January 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

estimate based on flow cytometry. This discrepancy is likely due to lower coverage of the heterochromatin (as reported above) as well as a decision of the KmerGenie program to ignore highly repetitive k-mers in the genome size estimate (Chikhi and Medvedev, 2014).

To perform *de novo* assembly, we used the Celera Assembler, an overlap-layout-consensus assembler developed and used to generate 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 assembly contains 5,066 contigs of lengths ranging from 1,831 bp to 925.7 Kbp. The N50 contig length, the length of the contig for which half of the total assembly length is contained in contigs of that size or larger, is 109.2 Kbp, while the NG50 contig length (analogous to N50, but normalized to the expected genome size of 180 Mbp) is 77.9 Kbp (Table 1). 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) is 153.3 Mbp, with a GC content of 42.18% (compared to 41.74% GC content in the reference genome).

The Assemblathon 2 competition (Bradnam et al., 2013) introduced four simple statistics to assess the quality of a *de novo* assembly given a trusted reference genome sequence, which they term: coverage, validity, multiplicity, and parsimony. The coverage of our assembly, the proportion of the reference sequence (excluding U and Uextra) reconstructed in some form, was 0.9741. Validity, the proportion of the assembly that could be validated through alignment to the reference, was 0.8586. Upon inclusion of unmapped scaffolds U and Uextra, this metric increased to 0.9932, demonstrating that there is very little novel sequence in our assembly. Our assembly did show slight redundancy, with a multiplicity of 1.0419, calculated as the total length of all alignments divided by the total length of the reference sequence to which there is at least one alignment. Multiplicity may have been increased by the decision to set the assembler error rate parameter very low (based on high read accuracy). A low error rate means greater specificity to distinguish closely related repeats, but can also induce redundancy in the assembly in the face of even low rates of polymorphism and sequence error. Finally, the parsimony of our assembly (the multiplicity divided by the validity) was 1.2134. This metric effectively quantifies the average number of assembled bases that must be inspected in order to identify a reference-validated base. Each of these results compared favorably to the results from Assemblathon 2, albeit for a much smaller and simpler genome compared to the vertebrate species used in that competition. Likewise, because of the availability of the entire reference genome to which to compare (versus a small number of verified fosmid regions in the case of Assemblathon 2), we achieved much higher rates of validity, which in turn affects parsimony as well.

bioRxiv preprint doi: https://doi.org/10.1101/001834; this version posted January 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

In order to assess the presence or absence as well as the accuracy of the assembly of various genomic features, we developed a 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 used NUCmer (Delcher et al., 2002; Kurtz et al., 2004) to align assembled contigs to the reference genome, extracting the longest increasing subset of alignments with respect to the reference (weighted by length $\times$ sequence identity). 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. The presence of duplicated and repetitive sequences in introns complicates gene assembly and annotation, potentially causing genes to be fragmented. Nevertheless, we determined that 15,534 of 16,656 (93.2%) FlyBase-annotated genes have start and stop boundaries contained in a single aligned contig within our assembly. A total of 14,206 genes (85.3%) have their entire sequence reconstructed with perfect identity to the reference sequence, while 15,252 genes have the entire length aligned with >99% sequence identity. For the remaining 1,122 genes whose boundaries were not contained in a single contig, we found that 878 were partially reconstructed as part of one or more contigs.

To gain more insight about the alignment on a per-chromosome basis, we further investigated the NUCmer alignment of the 5,066 assembled contigs to the reference genome. Upon requiring high stringency alignment (>99% sequence identity and >1 Kbp aligned), there were 1,973 alignments of our contigs to the euchromatic portions of chromosomes X, 2, 3, and 4, covering a total of 117.8 Mbp (97.9%) of the euchromatin (Table 2). For the heterochromatic sequence (XHet, 2Het, 3Het, and YHet), there were 523 alignments at this same threshold, covering 8.2 Mbp (88.1%) of the reference. Of the 2,820 remaining contigs that were not represented by these alignments, 1,082 aligned with the same stringency to portions of the unmapped scaffolds (U and Uextra).

Because repeats are a common cause of assembly failure, we hypothesized that gaps in the alignment of our assembly to the reference genome would overlap known repeats. We therefore analyzed the content of the 3,758 gaps in the high-stringency NUCmer global alignment, which represent failures of sequencing, library preparation, or assembly. We applied RepeatMasker (Smit, Hubley, & Green. RepeatMasker Open-3.0. 1996-2010. <`http://www.repeatmasker.org`>) to the reference sequences corresponding to alignment gaps, revealing that 42.71% of gap sequence are comprised of TEs, 17.38% of satellites, 2.66% of simple repeats, and 0.09% of other low complexity sequence. These proportions of gap sequences composed of TEs

bioRxiv preprint doi: https://doi.org/10.1101/001834; this version posted January 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

and satellites exceed the overall genome proportions of 20.35% and 4.00%, while the proportions composed of simple repeats and low complexity sequences are comparable to the overall genome proportions of 2.50% and 0.30%. Because a large proportion of the gap sequence was comprised of TEs, we investigated which TE families were most responsible for these assembly failures. A total of 587 of the 3,758 gaps overlapped the coordinates of annotated TEs, with young TE families being the most highly represented. For example, LTR elements from the *roo* family were the most common, with 129 copies (of only 136 copies in the genome) overlapping gap coordinates. TEs from this 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. In-depth analysis of TruSeq synthetic long-reads alignments to the locations of *roo* elements revealed that coverage was generally lower within the boundaries of TE insertion sites, likely due to failure to assemble long-reads from underlying short read data (e.g., Figure S4). Conversely, elements of the high-copy number (2,235 copies) INE-1 family were underrepresented among gaps in the alignment, with only 54 copies overlapping gaps. 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 S5A). In contrast, assembly can be successful in regions with high-repeat density, provided that the TEs are from different families (Figure S5B). Together, these observations about the assembly of particular TE families motivated formal investigation of the characteristics of particular TE copies and TE families that affect their assembly, as we describe in the following section.

Assessment of TE 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 long-reads are clearly beneficial. Finally, highly identical repeat copies introduce ambiguity into the assembly graph, which can result in repeats being placed in the wrong location in the assembly. As TEs are diverse in their organization, length, copy number, 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

bioRxiv preprint doi: https://doi.org/10.1101/001834; this version posted January 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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 5,425 TE copies annotated in the *D. melanogaster* reference genome, 4,588 (84.6%) had both boundaries contained in a single contig of our assembly aligned to the reference genome, with 4,362 (80.4%) 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. For the fixed effects, we first included TE length, as we expect assembly to be less likely in cases where individual reads do not span the length of the entire TE copy. We also included TE divergence estimates (FlyTE database. Fiston-Lavier, pers. comm.), as low divergence (corresponding to high sequence identity) can cause TEs to be misplaced or mis-assembled. Average coverage of the chromosome on which the TE copy is located was also included, as higher coverage generally improves assembly results. 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. We found that length ($b = -5.588 \times 10^{-4}$, $Z = -20.294$, $P < 2 \times 10^{-16}$), divergence ($b = 4.073$, $Z = 6.864$, $P = 6.69 \times 10^{-12}$), and coverage ($b = 5.474 \times 10^{-2}$, $Z = 3.493$, $P = 0.000478$) were significant predictors of accurate TE assembly (Figure 3; Table S3). Longer and less divergent TE copies, as well as those falling on chromosomes with lower depth of coverage, resulted in a lower probability of accurate assembly (Figure 3).

However, we also hypothesized that copy number (TE copies per family), could be important, as high copy number represents more opportunities for false joins which can break the assembly or generate chimeric contigs. Because copy number is a property of TE families (the random effect), it could not be incorporated using the GLMM framework. To test this effect, we fit a generalized linear model with the proportion of TE copies accurately assembled per TE family as the response variable. In this model, we included mean length, mean divergence, and mean copy number as predictors. This model indeed revealed that copy number is a significant predictor of TE assembly ($b = -0.04302$, $Z = -2.275$, $P = 0.0229$), with fewer TEs accurately assembled for high copy number families.

In spite of the limitations revealed by this 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, the 10 of the 11 elements in the *Juan* family have less than 0.1% divergence with respect to the canonical sequence, yet all 11 copies were assembled with 100% accuracy in separate contigs.

bioRxiv preprint doi: https://doi.org/10.1101/001834; this version posted January 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Impact of the coverage on assembly results

Due to high read quality, *de novo* assembly with TruSeq synthetic long-reads requires lower depth of coverage compared to assembly with short reads. However, 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 coverage on the quality of the resulting assembly, we randomly down-sampled the full 21× dataset to 15×, 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 depth of coverage increases. NG50 contig length increases rapidly with coverage up to approximately 10×, increasing only marginally at higher 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 recommends no more that 25×). Gene content also increases only marginally as coverage increases above approximately 10×, but TE content does not saturate as rapidly (Figure 4B). Our results likewise suggest that even very low coverage assemblies (2.5×) using TruSeq synthetic long-reads can accurately recover more than half of all annotated genes as well as nearly 40% of annotated TEs.

bioRxiv preprint doi: https://doi.org/10.1101/001834; this version posted January 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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 represents a substantial technical hurdle. Repeat elements, which are diverse in their structure and copy number, are the main reason for this technical bottleneck. While not every repeat causes assembly failure, Phillippy et al. (2008) appropriately noted that nearly every assembly failure is caused by repeats. Consequently, many assemblers attempt to mask repeat elements prior to assembly, thereby removing them from the final genome sequence. While this approach may improve assembly contiguity and accuracy, diverse classes of repeats represent an important feature of species' genomes across the tree of life, fundamentally affecting 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).

Despite their importance to genome content and function, few tools (e.g. T-lex2 (Fiston-Lavier et al., 2011), RetroSeq (Keane et al., 2013), Tea (Lee et al., 2012)) are currently available for discovery and annotation of TE sequences in high-throughput sequencing data. Because these tools depend on the quality of the assembly to which they are applied, annotation is generally limited to short and divergent TE families, 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 length of the repeats themselves. Here, we presented a novel sequencing approach (TruSeq) that allows the generation of highly accurate synthetic reads up to 15 Kbp in length. We showcased the utility of this approach for assembling highly repetitive, complex TEs with high accuracy, a feat that was not possible with short read data alone. As a first step in our analysis, we analyzed the content of the long-read data, evaluating long-read accuracy as well as uniformity of coverage of the *D. melanogaster* reference genome. We found that the reads were highly accurate, with error rates lower than other current long-read sequencing technologies. We also observed relatively uniform coverage across both the euchromatic and heterochromatic portions of the autosomes, with somewhat reduced coverage of the heterochromatin, which can be explained by both 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. Low coverage of the unmapped scaffolds U and Uextra may have a similar explanation, but the non-zero coverage of these

bioRxiv preprint doi: https://doi.org/10.1101/001834; this version posted January 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

chromosomes within our dataset suggests that at least a portion represents true fly-derived sequences. Low coverage of the mitochondrial genome is likely a consequence of the size selection step used in the library preparation protocol.

Our assembly achieved an NG50 contig length of 77.9 Kbp, covering 97.41% of the existing reference genome, and assembling 85.3% of annotated genes with perfect sequence identity. Using both standard assembly metrics (number of contigs, contig length, etc.) and new metrics introduced by Assemblathon 2 (Bradnam et al., 2013), we demonstrated that our assembly compares favorably to other *de novo* assemblies of other large and complex genomes. 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 of TruSeq long-read technology, but new algorithms and assembly software will be tailored specifically for this platform in the near future (J. Simpson, pers. comm.).

In addition to general improvements for *de novo* assembly, our study demonstrates that TruSeq synthetic long-reads enable accurate assembly of complex, highly repetitive TE sequences. Our assembly contains 80.4% 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 the 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. Finally, by performing this assessment in *D. melanogaster*, a species with particularly active, abundant, and identical TEs, our results suggest that TruSeq technology will empower studies of TE dynamics for many non-model species in the near future.

The TruSeq synthetic long-read approach represents a "generation 2.5" sequencing technology that builds upon second-generation Illumina short read data. This new approach promises to dramatically advance a wide range of genomic applications. Meanwhile, several third-generation sequencing platforms have been developed to sequence long molecules directly. One such technology, Oxford Nanopore sequencing (Oxford, UK) (Clarke et al., 2009), possesses several advantages over existing platforms, including the generation of reads exceeding 5 Kbp at a speed 1 bp per nanosecond. Pacific Biosciences' (Menlo Park, CA, USA) single-molecule real-time (SMRT) sequencing likewise uses direct observation of enzymatic reactions to produce base calls in real time with reads averaging ~1,300 bp in length, and fast sample preparation and sequencing (1-2 days each) (Roberts et al., 2013). Perhaps most importantly, neither Nanopore nor SMRT sequencing

bioRxiv preprint doi: https://doi.org/10.1101/001834; this version posted January 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

requires PCR amplification, which reduces 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 the accurate reconstruction of highly-identical tandem repeats which could be collapsed within TruSeq long-reads.

Most current approaches to *de novo* assembly ignore repetitive elements such as TEs, focusing only on the reconstruction of non-repetitive sequences. Such approaches bias perspectives of evolution of complex genomes, which can be comprised of more than 50% repetitive DNA. In addition to accurately recovering more than 97% of the current high quality reference genome, our assembly using TruSeq synthetic long-reads accurately placed and perfectly reconstructed the sequence of 85.3% of genes and 80.4% of TEs, a result which is unprecedented in the field of *de novo* genome assembly. These 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.

bioRxiv preprint doi: https://doi.org/10.1101/001834; this version posted January 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Methods

Reference genome and annotations

The latest release of the *D. melanogaster* genome sequence at the time of the preparation of this manuscript (Release 5.53) and corresponding TE annotations were downloaded from FlyBase (`http://www.fruitfly.org/`). All TE features come from data stored in the FlyTE database (Fiston-Lavier, pers. comm.), 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 *Drosophila melanogaster* using a standard ethanol precipitation-based protocol. Approximately 50-100 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.

Four synthetic long read libraries were prepared by Illumina using a proprietary TruSeq synthetic long-read protocol, previously known as Moleculo or LR-seq (Voskoboynik et al., 2013). 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 is parallelized into many local assemblies, 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. These local assemblies are performed using a proprietary short read assembler that accounts for particular molecular biology aspects of the library preparation.

Assessment of long read quality

To estimate the degree of contamination of the *D. melanogaster* libraries prepared by Illumina, we used BLASTN (Altschul et al., 1997) to search the 523,583 total reads against the *D. melanogaster* reference sequences (including heterochromatic scaffolds and unmapped scaffolds U and Uextra) with a stringent cut-off of e-value $<$ 1e-12. We also used BLASTN to compare the reads against reference sequences from the

bioRxiv preprint doi: https://doi.org/10.1101/001834; this version posted January 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

NCBI nucleotide database (`http://www.ncbi.nlm.nih.gov/nuccore`). The TruSeq synthetic long-reads were mapped to a repeat-masked version of 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 again mapped the data the the euchromatic arms of the *D. melanogaster* reference genome using BWA-MEM (Li and Durbin, 2009), then parsed the resulting 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/Biostar59647`>) 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 $(M) = 487{,}455 = Lm + pL\theta$, 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 $(M_{SNP}) = 28,657 = Lm\theta_D + pL\theta$, 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 (20.58×), estimating a corrected error rate of 0.0192% 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. We therefore additionally applied a second approach where we assumed that all 12,064 sites with more than one mismatch represented true polymorphism, calculating the error rate as above with only singleton mismatches. As expected, this method yielded a slightly lower error rate of 0.0183%.

bioRxiv preprint doi: https://doi.org/10.1101/001834; this version posted January 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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 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, 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).

After testing the assembler using a range of parameters, we decided upon three modifications to the default assembly parameters to take advantage of unique aspects of the data (B. Walenz, pers. comm.): 1) we used the bogart unitigger, rather than the default utg algorithm, 2) we decreased the unitig graph error rate to 0.3% and unitig merge error rate to 0.45% based on the low observed error rate upon mapping data to the reference as well as the low level of residual heterozygosity in this inbred line, and 3) we increased the specificity of overlap seeds by increasing the k-mer size to 31 and doubling the overlap threshold. In the face of very high read quality, these modifications to increase assembler specificity should not substantially reduce sensitivity to detect true overlaps.

For the down-sampled assemblies with lower coverage, we based the expected coverage on the average mapped depth of coverage of 21× for the full dataset. We randomly sampled reads from a concatenated FASTQ of all four 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 longest increasing subset of alignments to the reference (i.e. the longest consistent set of alignments with respect to the reference sequence). We then used to coordinates of these alignments to both measure overall assembly quality and investigate assembly of particular genomic features, including genes, TEs, and segmental duplications.

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

bioRxiv preprint doi: https://doi.org/10.1101/001834; this version posted January 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

length ratio (assembled length/reference length) for each sequence.

To calculate Assemblathon 2 statistics, we used the COMPASS tool (Bradnam et al., 2013), modifying it to use the same NUCmer alignment rather than performing a new alignment with LASTZ (Harris, 2007). COMPASS and the modifications can be found at `https://github.com/rmccoy7541/compass`.

The GLMM and GLM 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). In the GLMM, 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, divergence (number of substitutions per base compared to the canonical sequence of the TE family), and average coverage of the corresponding chromosome were included as fixed effects, while TE family was included as a random effect. For the GLM, we aggregated assembly results by family, with the proportion of copies in the family accurately assembled included as the response variable. This allowed us to include copy number as a fixed effect, along with the average length, average divergence, and average depth of coverage of the corresponding chromosomes for each TE family. In both models, 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).

bioRxiv preprint doi: https://doi.org/10.1101/001834; this version posted January 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Data access

Raw data, the genome assembly, and code used for the data analysis can be found at **XXX**. Scripts written for the assessment of presence or absence of genomic features in the *de novo* assembly can be found at https://github.com/rmccoy7541/assess-assembly.

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 long read libraries.

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. TAB and MK contributed to the data generation and provided guidance during planning stages of the experiment. DAP helped design the experiment and provided guidance on analyses throughout. 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

bioRxiv preprint doi: https://doi.org/10.1101/001834; this version posted January 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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_iso1 is the mitochondrial genome of the *yw;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) assembly with respect to significant predictor variables: TE length ($b = -5.588 \times 10^{-4}$, $Z = -20.294$, $P < 2 \times 10^{-16}$), divergence ($b = 4.073$, $Z = 6.864$, $P = 6.69 \times 10^{-12}$), and average depth of coverage of the chromosome on which the TE is found ($b = 5.474 \times 10^{-2}$, $Z = 3.493$, $P = 0.000478$). The upper sets of points represent TEs which were perfectly assembled, while the lower set of points represents TEs which are absent from the assembly or were mis-assembled with respect to the reference. Lines represent predicted values from the GLMM fit to these data. Colors indicate different TE families (126 total).

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: K-mer coverage histogram generated by KmerGenie (Chikhi and Medvedev, 2014), depicting the number of occurrences of each unique k-mer within the TruSeq synthetic long-read dataset. Low coverage k-mers represent long-read errors and polymorphism (or errors in the reference genome) while high coverage k-mers (x-axis truncated at 60×) represent genomic repeats.

Figure S3: Dot plots depicting NUCmer (Delcher et al., 2002) global 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). Blue segments represent forward alignments, while red segments indicate an inversion with respect to the rest of the contig alignment. Dot plots

bioRxiv preprint doi: https://doi.org/10.1101/001834; this version posted January 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

were generated using the mummerplot feature of MUMmer (Kurtz et al., 2004).

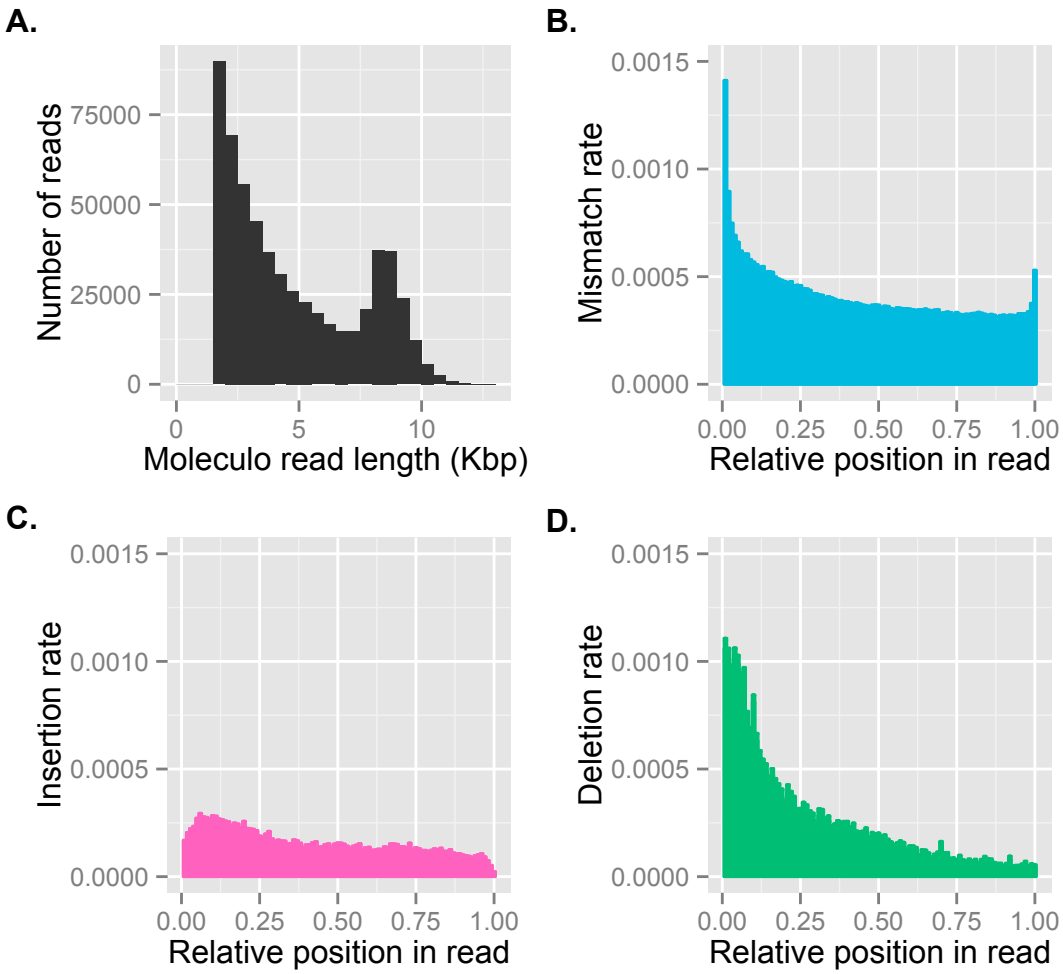
Figure S4: IGV screenshot (Robinson et al., 2011; Thorvaldsdóttir et al., 2013) of a representative case where assembly fails due to a deficiency of long-read data derived from a long transposable element sequence. The upper-most track (blue) represents the NUCmer alignment of assembled contigs to the 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 track (green) indicates the boundaries of transposable elements.

Figure S5: 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.

bioRxiv preprint doi: https://doi.org/10.1101/001834; this version posted January 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figures

Figure 1



bioRxiv preprint doi: https://doi.org/10.1101/001834; this version posted January 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure 2

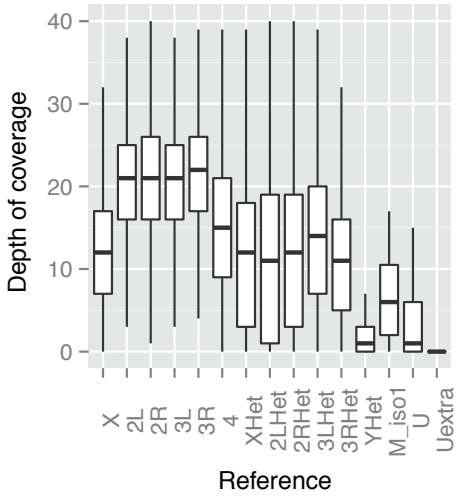
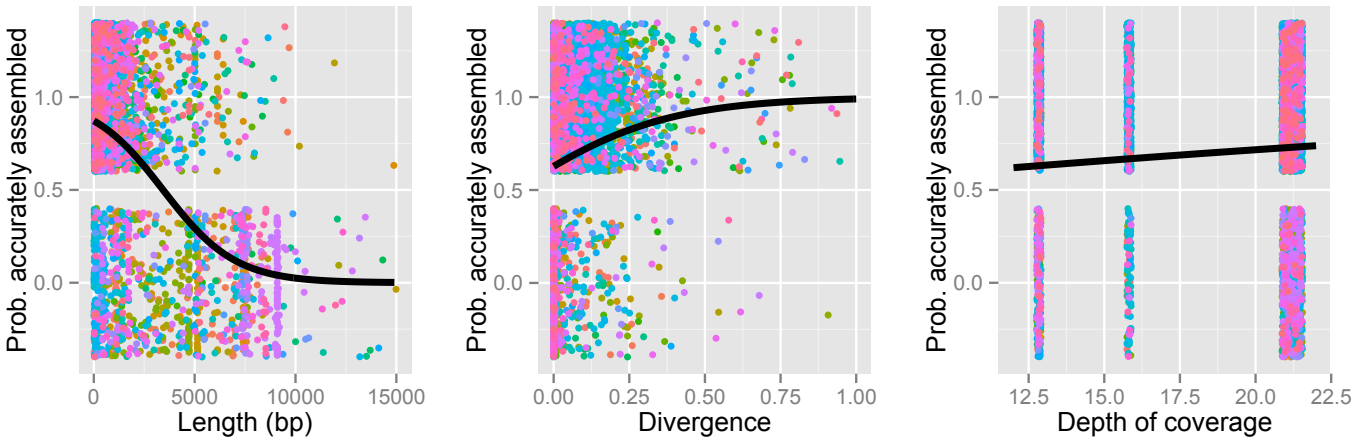
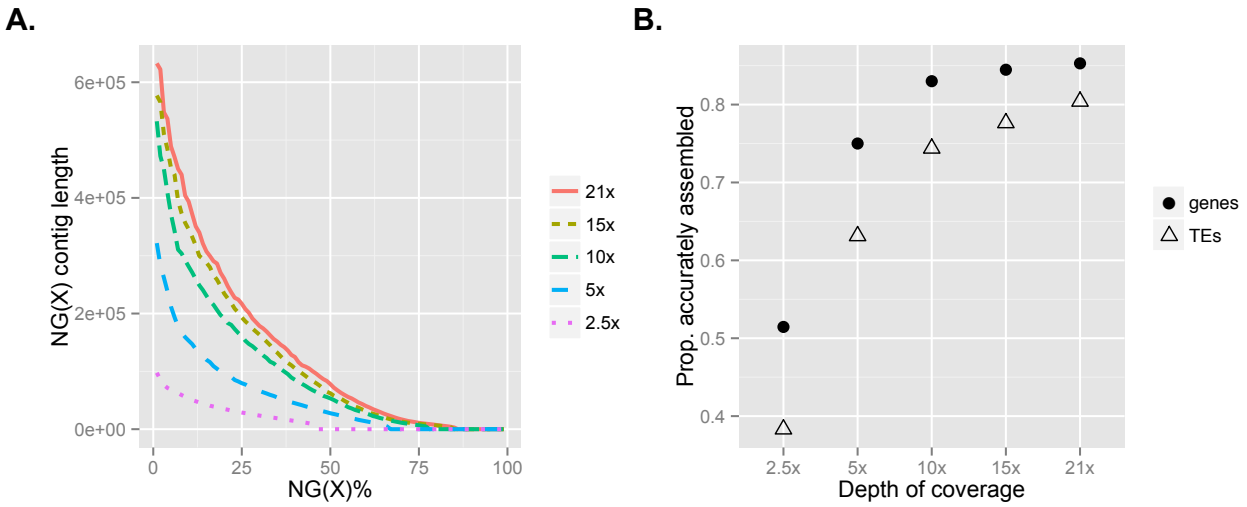


Figure 3



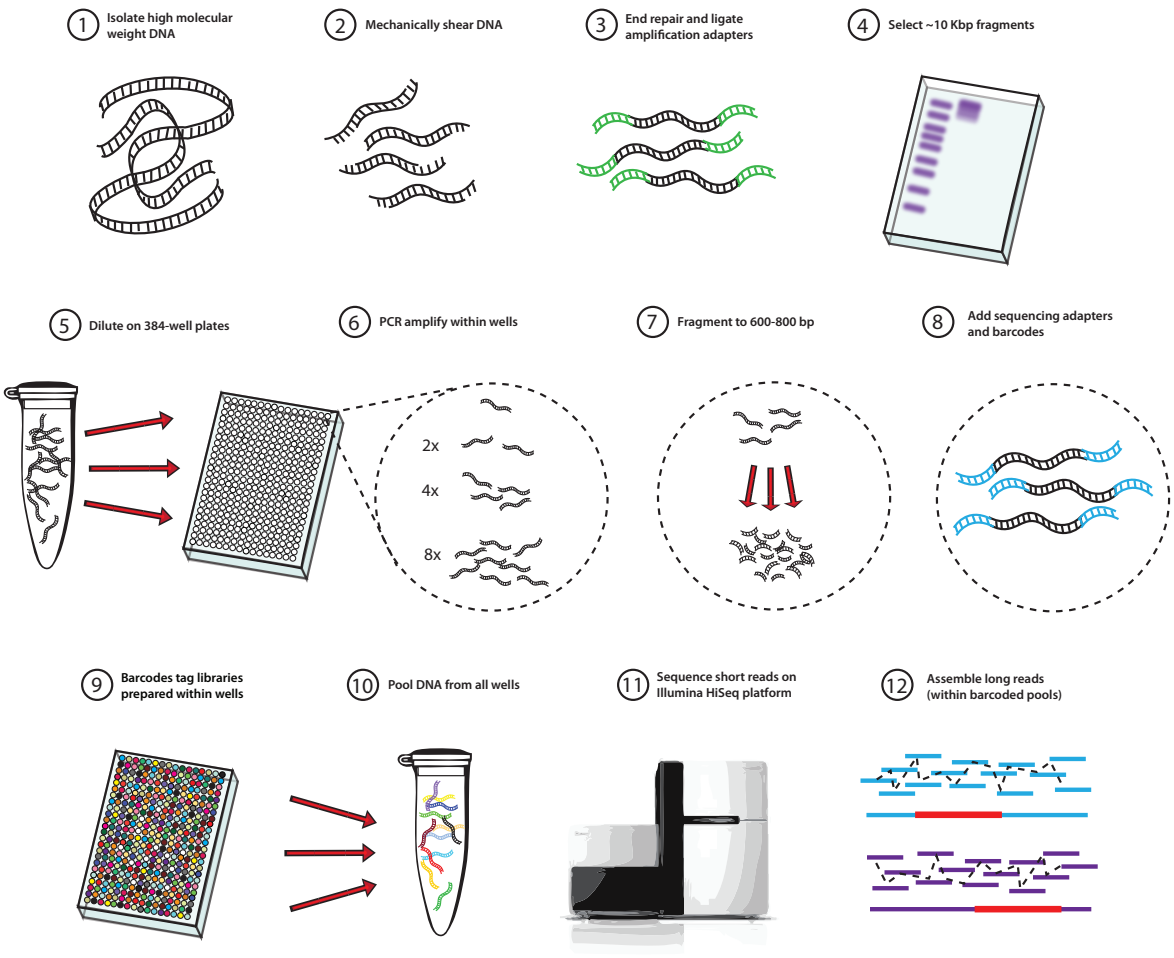
bioRxiv preprint doi: https://doi.org/10.1101/001834; this version posted January 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure 4



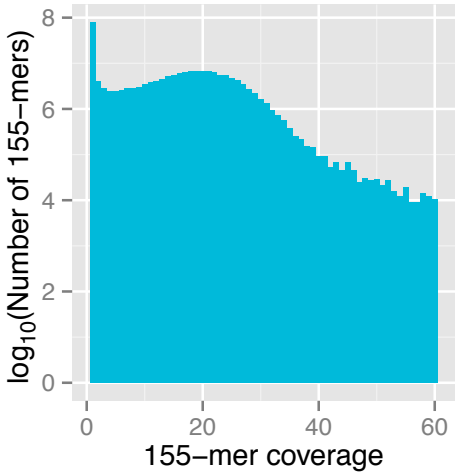
bioRxiv preprint doi: https://doi.org/10.1101/001834; this version posted January 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure S1



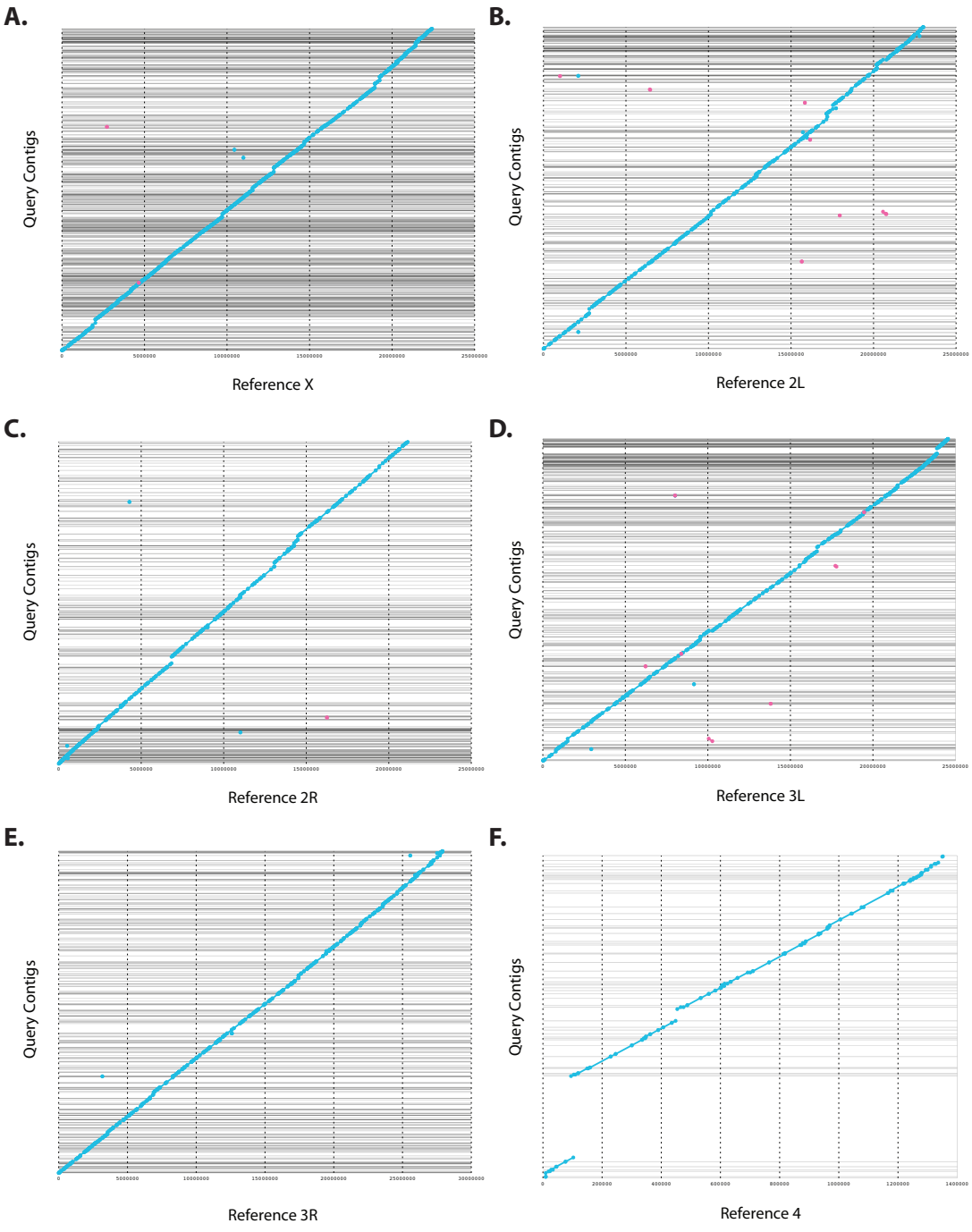
bioRxiv preprint doi: https://doi.org/10.1101/001834; this version posted January 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure S2

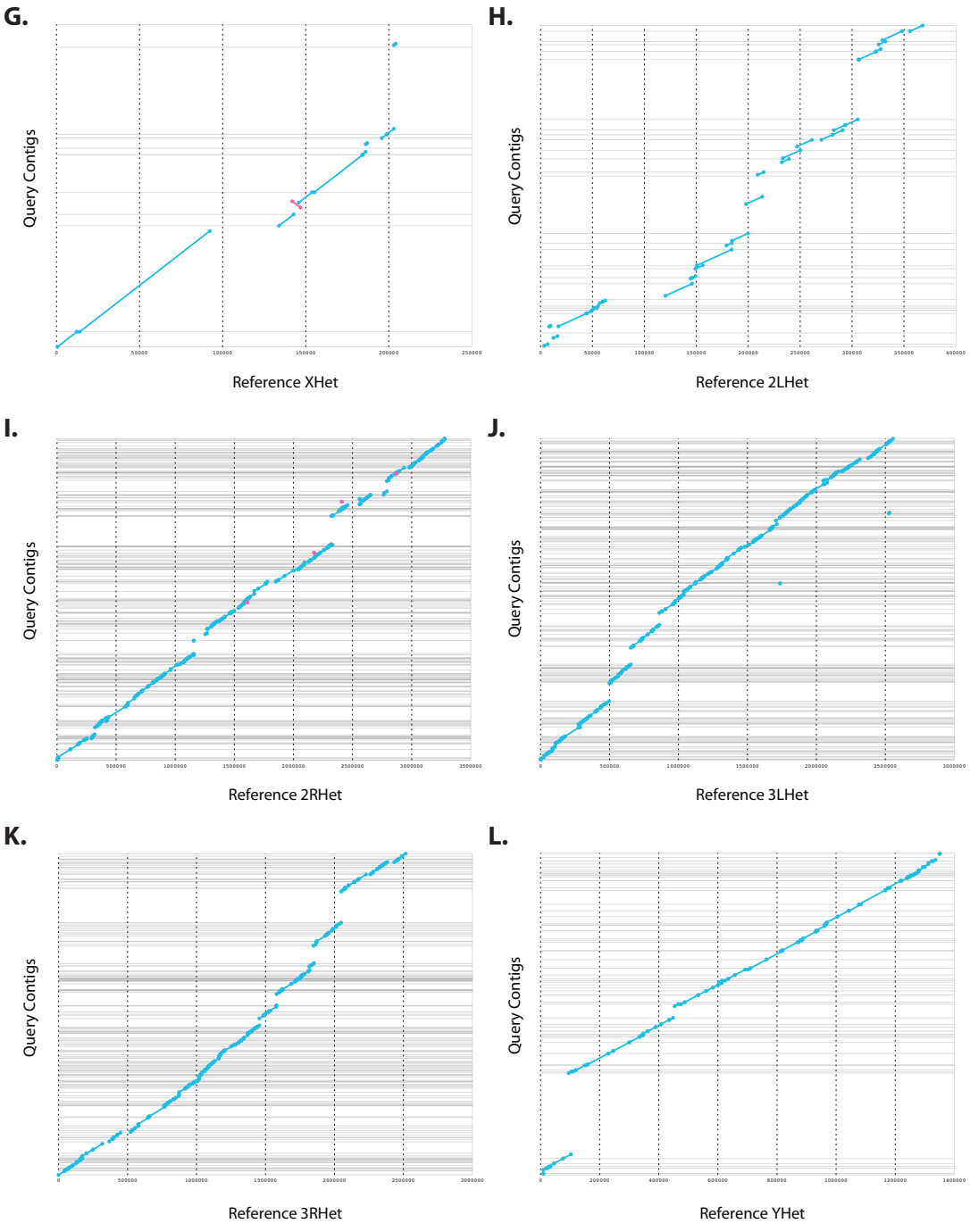


bioRxiv preprint doi: https://doi.org/10.1101/001834; this version posted January 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure S3



bioRxiv preprint doi: https://doi.org/10.1101/001834; this version posted January 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



bioRxiv preprint doi: https://doi.org/10.1101/001834; this version posted January 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure S4

bioRxiv preprint doi: https://doi.org/10.1101/001834; this version posted January 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure S5

bioRxiv preprint doi: https://doi.org/10.1101/001834; this version posted January 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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. Coverage, validity, multiplicity, and parsimony are metrics introduced in the Assemblathon 2 competition (Bradnam et al., 2013) and are also described in the main text (calculated using COMPASS <`https://github.com/jfass/compass`>).

Metric	Value
Number of contigs	5066
Total size of contigs	153282184
Longest contig	925686
Shortest contig	1831
Number of contigs > 1 Kbp	5066
Number of contigs > 10 Kbp	2377
Number of contigs > 100 Kbp	382
Mean contig size	30257
Median contig size	9388
N50 contig length	109246
L50 contig count	338
NG50 contig length	77878
LG50 contig count	481
Contig GC content	42.18%
Coverage	0.9741 (0.8540)
Validity	0.8586 (0.9932)
Multiplicity	1.0419 (1.0565)
Parsimony	1.2134 (1.0637)

*Values in parentheses represent Assemblathon 2 metrics calculated upon inclusion of reference sequences U and Uextra (unmapped scaffolds and additional contigs).

bioRxiv preprint doi: https://doi.org/10.1101/001834; this version posted January 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Table 2: Alignment statistics for Celera Assembler contigs aligned to the reference genome with high stringency (>99% sequence identity and >1 Kbp aligned). 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	633	493	21461492	95.71
2L	322	178	22565901	98.06
2R	299	144	20790246	98.31
3L	357	201	24158719	98.43
3R	307	162	27603229	98.92
4	55	42	1262176	93.37
XHet	8	8	148085	72.55
2LHet	25	15	253375	68.69
2RHet	166	81	2508722	76.28
3LHet	166	83	2238074	87.58
3RHet	132	70	2121303	84.26
YHet	26	28	127328	36.69
M	0	0	0	0
U	893	1215	4411149	43.90
Uextra	661	1038	1436060	4.95

bioRxiv preprint doi: https://doi.org/10.1101/001834; this version posted January 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Table S1: Top BLAST hits to the NCBI nucleotide database for all TruSeq synthetic long-reads. Only species/strains with ≥5 hits are reported here.

No. long reads	Species/strain of top BLAST hit
522772	*Drosophila melanogaster*
93	*Gluconacetobacter diazotrophicus* PAl 5
79	*Drosophila mauritiana*
64	*Enterobacteria* phage HK629
59	*Gluconacetobacter xylinus* NBRC 3288
42	*Acetobacter pasteurianus* 386B
37	*Gluconobacter oxydans* 621H
35	*Gluconobacter oxydans* H24
26	*Drosophila simulans*
18	Cloning vector pSport1
15	synthetic construct
13	*Acetobacter pasteurianus* IFO 3283-01
11	*Drosophila pseudoobscura*
9	*Acetobacter pasteurianus* IFO 3283-12
5	*Drosophila sechellia*
5	*Acetobacter aceti*
5	*Granulibacter bethesdensis* CGDNIH1
5	*Rhodomicrobium vannielii* ATCC 17100

Table S2: 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)	1.432	1.197	
Fixed effect	Estimate	Std. Error	z value	Pr($>\|z\|$)
(Intercept)	1.81722	0.13879	13.094	$<$ 2e-16
Length	-1.50683	0.07425	-20.294	$<$ 2e-16
Divergence	0.68216	0.09938	6.864	6.69e-12
Coverage	0.17921	0.05131	3.493	0.000478

bioRxiv preprint doi: https://doi.org/10.1101/001834; this version posted January 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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*

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.

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.

Chikhi, R. and Medvedev, P. (2014). Informed and automated k-mer size selection for genome assembly. *Bioinformatics* **30**:31–37.

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.

bioRxiv preprint doi: https://doi.org/10.1101/001834; this version posted January 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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

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

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

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

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

603 Glenn, T. C. (2011). Field guide to next-generation DNA sequencers. *Molecular Ecology Resources* **11**:759–
604 769.

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

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

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

611 Harris, R. S. (2007). *Improved pairwise alignment of genomic DNA*. Ph.D. thesis, Pennsylvania State
612 University.

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

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

bioRxiv preprint doi: https://doi.org/10.1101/001834; this version posted January 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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.

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.

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 ...* .

Li, X. and Waterman, M. S. (2003). Estimating the repeat structure and length of DNA sequences using L-tuples. *Genome Research* **13**:1916–1922.

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.

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.

bioRxiv preprint doi: https://doi.org/10.1101/001834; this version posted January 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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.

Phillippy, A. M., Schatz, M. C., and Pop, M. (2008). Genome assembly forensics: finding the elusive mis-assembly. *Genome Biology* **9**:R55.

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.

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.

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

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.

Treangen, T. J. and Salzberg, S. L. (2012). Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature Reviews Genetics* **13**:36–46.

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.

bioRxiv preprint doi: https://doi.org/10.1101/001834; this version posted January 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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.