

ILLUMINA TRUSEQ SYNTHETIC LONG-READS EMPOWER *DE NOVO* ASSEMBLY AND RESOLVE COMPLEX, HIGHLY REPETITIVE TRANSPOSABLE ELEMENTS

Rajiv C. McCoy¹, Ryan W. Taylor¹, Timothy A. Blauwkamp², Joanna L. Kelley³, Michael Kertesz⁴, Dmitry Pushkarev⁵, Dmitri A. Petrov^{*1} and Anna-Sophie Fiston-Lavier^{*1,6}

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

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

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

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

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

⁶Institut 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*

1 **Abstract**

2 High-throughput DNA sequencing technologies have revolutionized genomic analysis, including the *de novo*
3 assembly of whole genomes. Nevertheless, assembly of complex genomes remains challenging, mostly due to
4 the presence of repeats, which cannot be reconstructed unambiguously with short read data alone. One class
5 of repeats, called transposable elements (TEs), is particularly problematic due to high sequence identity, high
6 copy number, and a capacity to induce complex genomic rearrangements. Despite their importance to genome
7 function and evolution, most current *de novo* assembly approaches cannot resolve TEs. Here, we applied a
8 novel Illumina technology called TruSeq synthetic long-reads, which are generated through highly parallel
9 library preparation and local assembly of short read data and achieve lengths of 2-15 Kbp with an extremely
10 low error rate (<0.05%). To test the utility of this technology, we sequenced and assembled the genome
11 of the model organism *Drosophila melanogaster* (reference genome strain *yw;cn,bw,sp*) achieving an NG50
12 contig size of 77.9 Kbp and covering 97.2% of the current reference genome (including heterochromatin).
13 TruSeq synthetic long-read technology enables placement of individual TE copies in their proper genomic
14 locations as well as accurate reconstruction of TE sequences. We entirely recover and accurately place
15 80.4% of annotated transposable elements with perfect identity to the current reference genome. As TEs are
16 complex and highly repetitive features that are ubiquitous in genomes across the tree of life, TruSeq synthetic
17 long-read technology offers a powerful approach to drastically improve *de novo* assemblies of whole genomes.

18 Introduction

19 Despite tremendous advances in DNA sequencing technology, computing power, and assembly approaches, *de*
20 *novo* assembly of whole eukaryotic genomes using high-throughput sequencing data remains a challenge owing
21 largely to the presence of repetitive DNA (Alkan et al., 2010; Treangen and Salzberg, 2012). In some species,
22 repetitive DNA accounts for a large proportion of the total genome size, for example comprising more than half
23 of the human genome (Lander et al., 2001; de Koning et al., 2011) and 80% of some plant genomes (Feschotte
24 et al., 2002). Here, we focus on one class of dynamic repeats, called transposable elements (TEs). These
25 elements, a common feature of almost all eukaryotic genomes sequenced to date, are particularly difficult to
26 assemble accurately due to high sequence identity among multiple copies within a genome. In addition to
27 spanning up to tens of kilobases, TEs from a single family can be present in thousands of copies. Consequently,
28 TEs can dramatically affect genome size and structure, as well as genome function; transposition can induce
29 complex genomic rearrangements that detrimentally affect the host, but can also provide the raw material
30 for adaptive evolution (González et al., 2008; González and Petrov, 2009), for example, by creating new
31 transcription factor binding sites (Rebollo et al., 2012) or otherwise affecting expression of nearby genes
32 (González et al., 2009).

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

40 TEs, as with other classes of repeats, may also induce mis-assembly. For example, TEs that lie in tandem
41 may be erroneously collapsed, and unique interspersed sequence may be left out or appear as isolated contigs.
42 Several studies have assessed the impact of repeat elements on *de novo* genome assembly. For example, Alkan
43 et al. (2010) showed that the human assemblies are on average 16.2% shorter than expected, mainly due to
44 failure to assemble repeats, especially TEs and segmental duplications. A similar observation was made for
45 the chicken genome, despite the fact that repeat density in this genome is low (Ye et al., 2011). Current
46 approaches to deal with repeats such as TEs generally rely on depth of coverage and paired-end data (Alkan
47 et al., 2010; Miller et al., 2010; Li et al., 2010). Depth of coverage is informative of copy-number, but
48 unfortunately cannot guide accurate placement of repeats. Paired-end data can help resolve the orientation

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

53 A superior way to resolve TEs is to generate reads that exceed TE length, obviating assembly and allowing
54 TEs to be unambiguously placed based on unique flanking sequence. Several high-throughput long read (>1
55 Kbp) technologies have been developed, but most of these technologies have exceptionally high sequencing
56 error rates (although error-correction strategies have been developed in some cases (Schatz et al., 2012)) and
57 are low throughput. High error rates limit the specificity of long reads, meaning that assemblers cannot
58 distinguish between sequencing errors and differences between slightly diverged copies of TEs. For instance,
59 PacBio RS II (Pacific Biosciences) provides average read lengths of greater than 5 Kbp, but with a 15-18%
60 error rate (Schatz et al., 2012). Meanwhile, other established sequencing technologies, such as Illumina,
61 454 (Roche), and Ion Torrent (Life Technologies), offer lower error rates of 0.1-1%, but relatively shorter
62 read lengths (Glenn, 2011). Illumina has recently introduced a novel technology called TruSeq™ synthetic
63 long-reads <<http://www.illumina.com/services/long-read-sequencing-service.ilmn>>, which builds
64 upon underlying Illumina short read data to generate highly accurate synthetic reads up to 15 Kbp in length.
65 This technology promises to dramatically advance a wide range of genomic applications.

66 Using a pipeline of standard existing tools, we showcase the ability of TruSeq synthetic long-reads to
67 facilitate *de novo* assembly and resolve TE sequences in the genome of the fruit fly *Drosophila melanogaster*,
68 a key model organism in both classical genetics and molecular biology. We further investigate how coverage
69 of long reads affects assembly results, an important practical consideration for experimental design. While
70 the *D. melanogaster* genome is moderately large (~180 Mbp) and complex, it has already been assembled to
71 unprecedented accuracy. Through a massive collaborative effort, the initial genome project (Adams et al.,
72 2000) recovered nearly all of the 120 Mbp euchromatic sequence using a whole-genome shotgun approach that
73 involved painstaking molecular cloning and the generation of a bacterial artificial chromosome physical map.
74 Since that publication, the reference genome has been extensively annotated and improved using several
75 resequencing, gap-filling, and mapping strategies, and currently represents a gold standard for the genomics
76 community (Osoegawa et al., 2007; Celniker et al., 2002; Hoskins et al., 2007). By performing the assembly in
77 this model system with a high quality reference genome, our study is the first to systematically quantify the
78 substantial improvements to assembly enabled by synthetic long read technology. Because *D. melanogaster*
79 harbors a large number (~100) of families of active TEs, assembly of these repeats is particularly challenging

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

84 Results

85 TruSeq synthetic long-reads

86 This study used Illumina TruSeq synthetic long-read technology generated with a novel highly-parallel
87 next-generation library preparation method (Figure S1). The basic protocol was previously presented by
88 Voskoboynik et al. (2013) (who referred to it as LR-seq) and was patented by Molecuro, which was later
89 acquired by Illumina. The protocol (see Methods) involves initial mechanical fragmentation of gDNA into
90 ~10 Kbp fragments. These fragments then undergo end-repair and ligation of amplification adapters, before
91 being diluted onto 384-well plates so that each well contains DNA representing approximately 1-2% of the
92 genome (~200 molecules, in the case of *Drosophila melanogaster*). Polymerase chain reaction (PCR) is used
93 to amplify molecules within wells, followed by highly parallel Nextera-based fragmentation and barcoding of
94 individual wells. DNA from all wells is then pooled and sequenced on the Illumina HiSeq 2000 platform. Data
95 from individual wells are demultiplexed *in silico* according to the barcode sequences. Long reads are then
96 assembled from the short reads using a proprietary assembler that accounts for properties of the molecular
97 biology steps used in the library preparation. By reducing genome representation by approximately 50- to
98 100-fold, even abundant and identical repeats can be resolved so long as they are not represented multiple
99 times within a single well.

100 We applied TruSeq synthetic long-read technology to the fruit fly *D. melanogaster*, a model organism with
101 a high quality reference genome, including extensive repeat annotation (Fiston-Lavier et al., 2007; Quesneville
102 et al., 2003, 2005). The latest version of the reference genome assembly (Release 5; BDGP v3) contains a
103 total of 168.7 Mbp of sequence, 120 Mbp of which is considered to lie in the euchromatin, which is less repeat
104 dense than heterochromatic regions. This genome release also includes 10.0 Mbp of additional scaffolds (U)
105 which could not be mapped to chromosomes, as well as 29.0 Mbp of additional small scaffolds that could
106 not be joined to the rest of the assembly (Uextra). Approximately 50 adult individuals from the *yw;cn,bw,sp*
107 strain of *D. melanogaster* were pooled for the isolation of high molecular weight DNA, which was used to
108 generate TruSeq long-read libraries using the aforementioned protocol (Figure S1). The *yw;cn,bw,sp* strain
109 is the same strain which was used to generate the *D. melanogaster* reference genome (Adams et al., 2000).
110 A total of 523,583 synthetic long reads exceeding 1.5 Kbp (an arbitrary length cutoff) were generated with
111 four libraries (one Illumina HiSeq lane per library), comprising a total of 2.52 Gbp. Reads averaged 4,813
112 bp in length, but have a local maximum near 8.5 Kbp, slightly smaller than the ~10 Kbp DNA fragments
113 used as input for the protocol (Figure 1A).

114 We first searched for and eliminated possible contaminants by comparing the reads to the NCBI nucleotide
115 database (<http://www.ncbi.nlm.nih.gov/nucleotide>) using BLASTN (Altschul et al., 1997) (see Methods;
116 Table S1). The degree of contamination in the TruSeq synthetic long-read libraries prepared by Illumina
117 was extremely low. Of 523,583 total reads, only 0.104% (544 reads) had top hits to non-insect species,
118 and only 0.105% (549 reads) had top hits to species outside of genus *Drosophila*. Of the 523,034 hits to
119 *Drosophila* species, 99.950% (522,772 reads) had top hits to *D. melanogaster*, while only 0.0501% (262 reads)
120 had top matches to other *Drosophila* species. The most abundant contaminant reads had top matches
121 to known symbionts of *D. melanogaster*, including acetic acid bacteria from the genera *Gluconacetobacter*,
122 *Gluconobacter*, and *Acetobacter* (Table S1). Because we could not exclude that the few sequences with no
123 BLAST results may correspond to fly-derived sequences not previously assembled in the reference genome, we
124 included all sequences except those with top matches to non-insect species (523,039 total; e-value threshold
125 $1e-09$) in downstream analyses.

126 In order to evaluate the accuracy of TruSeq synthetic long-reads, we mapped reads to the reference
127 genome of *D. melanogaster*, identifying differences between the mapped reads and the reference sequence (see
128 Methods). Of 523,039 input reads passing our contamination filter, 99.97% (522,901 reads) were successfully
129 mapped to the reference genome, with 92.44% (483,514) mapping uniquely and 95.17% (497,751) having at
130 least one alignment with a MAPQ score ≥ 20 . TruSeq long-reads had very few mismatches to the reference at
131 0.0418% per base (0.0325% for reads with MAPQ ≥ 20) as well as a very low insertion rate of 0.0163% per base
132 (0.0112% for reads with MAPQ ≥ 20) and a deletion rate of 0.0277% per base (0.0209% for reads with MAPQ
133 ≥ 20). Error rates estimated with this mapping approach are conservative, as residual heterozygosity in the
134 sequenced line mimics errors. We therefore used two approaches (see Methods) to calculate corrected error
135 rates. The first approach uses the number of mismatches overlapping known SNPs to correct the error rate
136 to 0.0192%. Along with this estimate, we also estimate that 2.99% of the sites segregating in the *Drosophila*
137 Genetic Reference Panel (DGRP) (Mackay et al., 2012) remain polymorphic in the sequenced strain (i.e.
138 constitute residual heterozygosity). The second approach assumes that all non-singleton mismatches represent
139 true polymorphism, yielding an estimated long-read error rate of 0.0183%. Both estimates are nearly an order
140 of magnitude lower than other high-throughput sequencing technologies ($\sim 0.1\%$ for Illumina, $\sim 1\%$ for 454 and
141 Ion Torrent (Glenn, 2011), 15-18% for PacBio (Schatz et al., 2012)). The reason that TruSeq synthetic long-
142 read achieve such low error rates is that they are built as consensus of multiple overlapping short Illumina
143 reads. We further observed that all types of errors are more frequent at the beginning of reads, though the
144 pattern is more pronounced for mismatches and deletions (Figures 1B, 1C, & 1D). Minor imprecision in the

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

149 We then quantified the average depth of coverage of the mapped long-reads for each reference chromosome
150 arm. We found consistent and uniform coverage of $\sim 21\times$ of the euchromatin of each of the major autosomes
151 (2L,2R,3L,3R; Figure 2). Coverage of the heterochromatic portions of autosomes was generally lower (~ 11 -
152 $14\times$), and also varied marginally both within and between chromosomes. This is explained by the fact that
153 heterochromatin has high repeat content relative to euchromatin, making it more difficult to assemble into
154 long reads. Consequently, the fourth chromosome showed relatively lower average read depth ($15.8\times$ compare
155 to $21\times$), likely due to enrichment of heterochromatic islands on this chromosome (Haynes et al., 2006). Read
156 depth on the sex chromosomes is also expected to be lower: 75% relative to the autosomes for the X and 25%
157 relative to the autosomes for the Y, assuming equal numbers of males and females in the pool. Observed read
158 depth was lower still at $12.6\times$ for the X chromosome ($12.6\times$) as well as the Y chromosome ($2.7\times$), which is
159 entirely heterochromatic. Read depth for the mitochondrial genome was also relatively low ($5.6\times$) in contrast
160 to high mtDNA representation in short read genomic libraries, which we suspect to be a consequence of the
161 fragmentation and size selection steps of the library preparation protocol.

162 **Assessment of assembly content and accuracy**

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

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

178 To perform *de novo* assembly, we used the Celera Assembler, an overlap-layout-consensus assembler
179 developed and used to generate the first genome sequence of a multicellular organism, *D. melanogaster*
180 (Adams et al., 2000), as well as one of the first diploid human genome sequences (Levy et al., 2007). Our
181 assembly contains 5,066 contigs of lengths ranging from 1,831 bp to 925.7 Kbp. The N50 contig length, the
182 length of the contig for which half of the total assembly length is contained in contigs of that size or larger,
183 is 109.2 Kbp, while the NG50 contig length (analogous to N50, but normalized to the expected genome size
184 of 180 Mbp) is 77.9 Kbp (Table 1). Note that because the TruSeq synthetic long-read data are effectively
185 single end reads, only contig rather than scaffold metrics are reported. The total length of the assembly (i.e.
186 the sum of all contig lengths) is 153.3 Mbp, with a GC content of 42.18% (compared to 41.74% GC content
187 in the reference genome).

188 The Assemblathon 2 competition (Bradnam et al., 2013) introduced four simple statistics to assess the
189 quality of a *de novo* assembly given a trusted reference genome sequence, which they term: coverage, validity,
190 multiplicity, and parsimony. The coverage of our assembly, the proportion of the reference sequence (excluding
191 U and Uextra) reconstructed in some form, was 0.9741. Validity, the proportion of the assembly that could
192 be validated through alignment to the reference, was 0.8586. Upon inclusion of unmapped scaffolds U and
193 Uextra, this metric increased to 0.9932, demonstrating that there is very little novel sequence in our assembly.
194 Our assembly did show slight redundancy, with a multiplicity of 1.0419, calculated as the total length of all
195 alignments divided by the total length of the reference sequence to which there is at least one alignment.
196 Multiplicity may have been increased by the decision to set the assembler error rate parameter very low
197 (based on high read accuracy). A low error rate means greater specificity to distinguish closely related
198 repeats, but can also induce redundancy in the assembly in the face of even low rates of polymorphism
199 and sequence error. Finally, the parsimony of our assembly (the multiplicity divided by the validity) was
200 1.2134. This metric effectively quantifies the average number of assembled bases that must be inspected in
201 order to identify a reference-validated base. Each of these results compared favorably to the results from
202 Assemblathon 2, albeit for a much smaller and simpler genome compared to the vertebrate species used in
203 that competition. Likewise, because of the availability of the entire reference genome to which to compare
204 (versus a small number of verified fosmid regions in the case of Assemblathon 2), we achieved much higher
205 rates of validity, which in turn affects parsimony as well.

206 In order to assess the presence or absence as well as the accuracy of the assembly of various genomic
207 features, we developed a pipeline that reads in coordinates of generic annotations and compares the reference
208 and assembly for these sequences (see Methods). As a first step in the pipeline, we used NUCmer (Delcher
209 et al., 2002; Kurtz et al., 2004) to align assembled contigs to the reference genome, extracting the longest
210 increasing subset of alignments with respect to the reference (weighted by length \times sequence identity).
211 We then tested whether both boundaries of a given genomic feature were present within the same aligned
212 contig. For features that met this criterion, we performed local alignment of the reference sequence to the
213 corresponding contig using BLASTN (Altschul et al., 1997), evaluating the results to calculate the proportion
214 of the sequence aligned as well as the percent identity of the alignment. The presence of duplicated and
215 repetitive sequences in introns complicates gene assembly and annotation, potentially causing genes to be
216 fragmented. Nevertheless, we determined that 15,534 of 16,656 (93.2%) FlyBase-annotated genes have start
217 and stop boundaries contained in a single aligned contig within our assembly. A total of 14,206 genes (85.3%)
218 have their entire sequence reconstructed with perfect identity to the reference sequence, while 15,252 genes
219 have the entire length aligned with $>99\%$ sequence identity. For the remaining 1,122 genes whose boundaries
220 were not contained in a single contig, we found that 878 were partially reconstructed as part of one or more
221 contigs.

222 To gain more insight about the alignment on a per-chromosome basis, we further investigated the NUCmer
223 alignment of the 5,066 assembled contigs to the reference genome. Upon requiring high stringency alignment
224 ($>99\%$ sequence identity and >1 Kbp aligned), there were 1,973 alignments of our contigs to the euchromatic
225 portions of chromosomes X, 2, 3, and 4, covering a total of 117.8 Mbp (97.9%) of the euchromatin (Table
226 2). For the heterochromatic sequence (XHet, 2Het, 3Het, and YHet), there were 523 alignments at this
227 same threshold, covering 8.2 Mbp (88.1%) of the reference. Of the 2,820 remaining contigs that were not
228 represented by these alignments, 1,082 aligned with the same stringency to portions of the unmapped scaffolds
229 (U and Uextra).

230 Because repeats are a common cause of assembly failure, we hypothesized that gaps in the alignment
231 of our assembly to the reference genome would overlap known repeats. We therefore analyzed the content
232 of the 3,758 gaps in the high-stringency NUCmer global alignment, which represent failures of sequencing,
233 library preparation, or assembly. We applied RepeatMasker (Smit, Hubley, & Green. RepeatMasker Open-
234 3.0. 1996-2010. <http://www.repeatmasker.org>) to the reference sequences corresponding to alignment
235 gaps, revealing that 42.71% of gap sequence are comprised of TEs, 17.38% of satellites, 2.66% of simple
236 repeats, and 0.09% of other low complexity sequence. These proportions of gap sequences composed of TEs

237 and satellites exceed the overall genome proportions of 20.35% and 4.00%, while the proportions composed
238 of simple repeats and low complexity sequences are comparable to the overall genome proportions of 2.50%
239 and 0.30%. Because a large proportion of the gap sequence was comprised of TEs, we investigated which TE
240 families were most responsible for these assembly failures. A total of 587 of the 3,758 gaps overlapped the
241 coordinates of annotated TEs, with young TE families being the most highly represented. For example, LTR
242 elements from the *roo* family were the most common, with 129 copies (of only 136 copies in the genome)
243 overlapping gap coordinates. TEs from this family are long (canonical length of 9,092 bp) and recently
244 diverged (mean of 0.0086 substitutions per base), and are therefore difficult to assemble. In-depth analysis of
245 TruSeq synthetic long-reads alignments to the locations of *roo* elements revealed that coverage was generally
246 lower within the boundaries of TE insertion sites, likely due to failure to assemble long-reads from underlying
247 short read data (e.g., Figure S4). Conversely, elements of the high-copy number (2,235 copies) INE-1 family
248 were underrepresented among gaps in the alignment, with only 54 copies overlapping gaps. INE-1 elements
249 tend to be short (611 bp canonical length) and represent older transposition with greater divergence among
250 copies.

251 Manual curation of the alignment also revealed that assembly is particularly poor in regions of tandem
252 arrangement of TE copies from the same family, a result that is expected because repeats will be present within
253 individual wells during library preparation (Figure S5A). In contrast, assembly can be successful in regions
254 with high-repeat density, provided that the TEs are from different families (Figure S5B). Together, these
255 observations about the assembly of particular TE families motivated formal investigation of the characteristics
256 of particular TE copies and TE families that affect their assembly, as we describe in the following section.

257 **Assessment of TE assembly**

258 Repeats can induce three common classes of mis-assembly. First, tandem repeats may be erroneously col-
259 lapsed into a single copy. While the accuracy of TruSeq synthetic long-reads are advantageous in this case,
260 such elements may still complicate assembly because they are likely to be present within a single molecule
261 (and therefore a single well) during library preparation. Second, large repeats may fail to be assembled
262 because reads do not span the repeat anchored in unique sequence, a situation where TruSeq long-reads are
263 clearly beneficial. Finally, highly identical repeat copies introduce ambiguity into the assembly graph, which
264 can result in repeats being placed in the wrong location in the assembly. As TEs are diverse in their organi-
265 zation, length, copy number, and divergence, we decided to assess the accuracy of TE assembly with respect
266 to each of these factors. We therefore compared reference TE sequences to the corresponding sequences in

267 our assembly. Because a naive mapping approach could result in multiple reference TE copies mapping to
268 the same location in the assembly, our approach was specifically designed to restrict the search space within
269 the assembly based on the NUCmer global alignment (see Methods).

270 Of the 5,425 TE copies annotated in the *D. melanogaster* reference genome, 4,588 (84.6%) had both
271 boundaries contained in a single contig of our assembly aligned to the reference genome, with 4,362 (80.4%)
272 perfectly reconstructed based on length and sequence identity.

273 In order to test which properties of TE copies affected faithful reconstruction, we fit a generalized linear
274 mixed model (GLMM) with a binary response variable indicating whether or not each TE copy was perfectly
275 assembled. For the fixed effects, we first included TE length, as we expect assembly to be less likely in
276 cases where individual reads do not span the length of the entire TE copy. We also included TE divergence
277 estimates (FlyTE database. Fiston-Lavier, pers. comm.), as low divergence (corresponding to high sequence
278 identity) can cause TEs to be misplaced or mis-assembled. Average coverage of the chromosome on which
279 the TE copy is located was also included, as higher coverage generally improves assembly results. Finally,
280 we included a random effect of TE family, which accounts for various family-specific factors not represented
281 by the fixed effects, such as sequence complexity. We found that length ($b = -5.588 \times 10^{-4}$, $Z = -20.294$,
282 $P < 2 \times 10^{-16}$), divergence ($b = 4.073$, $Z = 6.864$, $P = 6.69 \times 10^{-12}$), and coverage ($b = 5.474 \times 10^{-2}$,
283 $Z = 3.493$, $P = 0.000478$) were significant predictors of accurate TE assembly (Figure 3; Table S3). Longer
284 and less divergent TE copies, as well as those falling on chromosomes with lower depth of coverage, resulted
285 in a lower probability of accurate assembly (Figure 3).

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

294 In spite of the limitations revealed by this analysis, we observed several remarkable cases where accurate
295 assembly was achieved, distinguishing the sequences of TEs from a single family with few substitutions among
296 the set. For example, the 10 of the 11 elements in the *Juan* family have less than 0.1% divergence with respect
297 to the canonical sequence, yet all 11 copies were assembled with 100% accuracy in separate contigs.

298 **Impact of the coverage on assembly results**

299 Due to high read quality, *de novo* assembly with TruSeq synthetic long-reads requires lower depth of coverage
300 compared to assembly with short reads. However, the relationship between coverage and assembly quality
301 is complex, as we expect a plateau in assembly quality at the point where the assembly is no longer limited
302 by data quantity. To evaluate the impact of depth of coverage on the quality of the resulting assembly, we
303 randomly down-sampled the full $21\times$ dataset to $15\times$, $10\times$, $5\times$, and $2.5\times$. We then performed separate *de novo*
304 assemblies for each of these down-sampled datasets, evaluating and comparing assemblies using the same size
305 and correctness metrics previously reported for the full-coverage assembly. We observed an expected nonlinear
306 pattern for several important assembly metrics, which begin to plateau as depth of coverage increases. NG50
307 contig length increases rapidly with coverage up to approximately $10\times$, increasing only marginally at higher
308 coverage (Figure 4A). We do not expect the monotonic increase to continue indefinitely, as very high coverage
309 can overwhelm OLC assemblers such as Celera (see documentation, which recommends no more than $25\times$).
310 Gene content also increases only marginally as coverage increases above approximately $10\times$, but TE content
311 does not saturate as rapidly (Figure 4B). Our results likewise suggest that even very low coverage assemblies
312 ($2.5\times$) using TruSeq synthetic long-reads can accurately recover more than half of all annotated genes as well
313 as nearly 40% of annotated TEs.

314 Discussion

315 Rapid technological advances and plummeting costs of DNA sequencing technologies have allowed biologists
316 to explore the genomes of species across the tree of life. However, translating the massive amounts of sequence
317 data into a high quality reference genome ripe for biological insight represents a substantial technical hurdle.
318 Repeat elements, which are diverse in their structure and copy number, are the main reason for this technical
319 bottleneck. While not every repeat causes assembly failure, Phillippy et al. (2008) appropriately noted that
320 nearly every assembly failure is caused by repeats. Consequently, many assemblers attempt to mask repeat
321 elements prior to assembly, thereby removing them from the final genome sequence. While this approach
322 may improve assembly contiguity and accuracy, diverse classes of repeats represent an important feature
323 of species' genomes across the tree of life, fundamentally affecting genome size structure as well as genome
324 function (González and Petrov, 2009; Feschotte et al., 2002; Kidwell and Lisch, 2001; Cordaux and Batzer,
325 2009; Nekrutenko and Li, 2001).

326 Despite their importance to genome content and function, few tools (e.g. T-lex2 (Fiston-Lavier et al.,
327 2011), RetroSeq (Keane et al., 2013), Tea (Lee et al., 2012)) are currently available for discovery and anno-
328 tation of TE sequences in high-throughput sequencing data. Because these tools depend on the quality of
329 the assembly to which they are applied, annotation is generally limited to short and divergent TE families,
330 biasing our current view of TE organization. Accurate assembly and annotation of TEs and other repeats
331 will dramatically enrich our understanding of the complex interactions between TEs and host genomes as
332 well as genome evolution in general.

333 One of the simplest ways to accurately resolve repeat sequences is to acquire reads longer than the length of
334 the repeats themselves. Here, we presented a novel sequencing approach (TruSeq) that allows the generation
335 of highly accurate synthetic reads up to 15 Kbp in length. We showcased the utility of this approach for
336 assembling highly repetitive, complex TEs with high accuracy, a feat that was not possible with short read
337 data alone. As a first step in our analysis, we analyzed the content of the long-read data, evaluating long-
338 read accuracy as well as uniformity of coverage of the *D. melanogaster* reference genome. We found that the
339 reads were highly accurate, with error rates lower than other current long-read sequencing technologies. We
340 also observed relatively uniform coverage across both the euchromatic and heterochromatic portions of the
341 autosomes, with somewhat reduced coverage of the heterochromatin, which can be explained by both the
342 fact that heterochromatin is more difficult to sequence as well as the fact that it is generally more repetitive
343 and therefore more difficult to assemble into long reads from underlying short read data. Low coverage of
344 the unmapped scaffolds U and Uextra may have a similar explanation, but the non-zero coverage of these

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

348 Our assembly achieved an NG50 contig length of 77.9 Kbp, covering 97.41% of the existing reference
349 genome, and assembling 85.3% of annotated genes with perfect sequence identity. Using both standard
350 assembly metrics (number of contigs, contig length, etc.) and new metrics introduced by Assemblathon 2
351 (Bradnam et al., 2013), we demonstrated that our assembly compares favorably to other *de novo* assemblies of
352 other large and complex genomes. Nevertheless, we expect that future methodological advances will unlock
353 the full utility of TruSeq synthetic long-read technology. We used a simple pipeline of existing tools to
354 investigate the advantages of TruSeq long-read technology, but new algorithms and assembly software will
355 be tailored specifically for this platform in the near future (J. Simpson, pers. comm.).

356 In addition to general improvements for *de novo* assembly, our study demonstrates that TruSeq synthetic
357 long-reads enable accurate assembly of complex, highly repetitive TE sequences. Our assembly contains
358 80.4% of annotated TEs perfectly identical in sequence to the current reference genome. Despite the high
359 quality of the current reference, errors undoubtedly exist in the current TE annotations, and it is likely that
360 there is some divergence between the sequenced strain and the reference strain from which it was derived,
361 making the estimate of the quality of TE assembly conservative. Likewise, we used a generalized linear
362 modeling approach to demonstrate that TE length is the main feature limiting the assembly of individual
363 TE copies, a limitation that could be partially overcome by future improvements to the library preparation
364 technology to achieve even longer synthetic reads. Finally, by performing this assessment in *D. melanogaster*,
365 a species with particularly active, abundant, and identical TEs, our results suggest that TruSeq technology
366 will empower studies of TE dynamics for many non-model species in the near future.

367 The TruSeq synthetic long-read approach represents a “generation 2.5” sequencing technology that builds
368 upon second-generation Illumina short read data. This new approach promises to dramatically advance a
369 wide range of genomic applications. Meanwhile, several third-generation sequencing platforms have been
370 developed to sequence long molecules directly. One such technology, Oxford Nanopore sequencing (Oxford,
371 UK) (Clarke et al., 2009), possesses several advantages over existing platforms, including the generation of
372 reads exceeding 5 Kbp at a speed 1 bp per nanosecond. Pacific Biosciences’ (Menlo Park, CA, USA) single-
373 molecule real-time (SMRT) sequencing likewise uses direct observation of enzymatic reactions to produce
374 base calls in real time with reads averaging $\sim 1,300$ bp in length, and fast sample preparation and sequencing
375 (1-2 days each) (Roberts et al., 2013). Perhaps most importantly, neither Nanopore nor SMRT sequencing

376 requires PCR amplification, which reduces biases and errors that place an upper limit on the sequencing
377 quality of most other platforms. By directly sequencing long molecules, these third-generation technologies
378 will likely outperform TruSeq synthetic long-reads in certain capacities, such as the accurate reconstruction
379 of highly-identical tandem repeats which could be collapsed within TruSeq long-reads.

380 Most current approaches to *de novo* assembly ignore repetitive elements such as TEs, focusing only on
381 the reconstruction of non-repetitive sequences. Such approaches bias perspectives of evolution of complex
382 genomes, which can be comprised of more than 50% repetitive DNA. In addition to accurately recovering
383 more than 97% of the current high quality reference genome, our assembly using TruSeq synthetic long-reads
384 accurately placed and perfectly reconstructed the sequence of 85.3% of genes and 80.4% of TEs, a result which
385 is unprecedented in the field of *de novo* genome assembly. These improvements to *de novo* assembly, facilitated
386 by TruSeq synthetic long-reads and other long-read technologies, will empower comparative analyses that
387 will enlighten the understanding of the dynamics of repeat elements and genome evolution in general.

388 **Methods**

389 **Reference genome and annotations**

390 The latest release of the *D. melanogaster* genome sequence at the time of the preparation of this manuscript
391 (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
392 were detected using the program BLASTER (Quesneville et al., 2003, 2005).
393

394 **Library preparation**

395 High molecular weight DNA was separately isolated from pooled samples of the *y;cn,bw,sp* strain of *Drosophila*
396 *melanogaster* using a standard ethanol precipitation-based protocol. Approximately 50-100 adult individuals,
397 both males and females, were pooled for the extraction to achieve sufficient gDNA quantity for preparation
398 of multiple TruSeq synthetic long-read libraries.

399 Four synthetic long read libraries were prepared by Illumina using a proprietary TruSeq synthetic long-
400 read protocol, previously known as Moleclo or LR-seq (Voskoboynik et al., 2013). To produce each library,
401 extracted gDNA is sheared into approximately 10 Kbp fragments, ligated to amplification adapters, and then
402 diluted to the point that each well on a 384-well plate contains approximately 200 molecules, representing
403 approximately 1.5% of the entire genome. These pools of DNA are then amplified by long range PCR.
404 Barcoded libraries are prepared within each well using Nextera-based fragmentation and PCR-mediated
405 barcode and sequencing adapter addition. The libraries undergo additional PCR amplification if necessary,
406 followed by paired-end sequencing on the Illumina HiSeq 2000 platform. Assembly is parallelized into many
407 local assemblies, which means that the likelihood of individual assemblies containing multiple members of
408 gene families (that are difficult to distinguish from one another and from polymorphism within individual
409 genes) is greatly reduced. These local assemblies are performed using a proprietary short read assembler that
410 accounts for particular molecular biology aspects of the library preparation.

411 **Assessment of long read quality**

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

416 NCBI nucleotide database (<http://www.ncbi.nlm.nih.gov/nuccore>). The TruSeq synthetic long-reads
417 were mapped to a repeat-masked version of the *D. melanogaster* reference genome as single-end reads using
418 BWA-MEM (Li and Durbin, 2009). Depth of coverage was estimated by applying the GATK DepthOfCov-
419 erage tool to the resulting alignment.

420 To estimate error rates, we again mapped the data to the euchromatic arms of the *D. melanogaster*
421 reference genome using BWA-MEM (Li and Durbin, 2009), then parsed the resulting BAM file to calculate
422 position-dependent mismatch, insertion, and deletion profiles. Because a portion of this effect would result
423 from accurate sequencing of genomes harboring residual heterozygosity, we used data from the *Drosophila*
424 Genetic Reference Panel (DGRP) (Mackay et al., 2012) to estimate both the rate of residual heterozy-
425 gosity as well as a corrected error rate of the TruSeq synthetic long-reads. We applied the jvarkit util-
426 ity (<https://github.com/lindenb/jvarkit/wiki/Biostar59647>) to identify positions in the reference
427 genome where mismatches occurred. We then used the relationship that the total number sites with mis-
428 matches to the euchromatic reference chromosome arms (M) = 487,455 = $Lm + pL\theta$, where L is the
429 120,381,546 bp length of the reference sequence to which we aligned, m is the per base error rate, p is
430 the proportion of heterozygous sites still segregating in the inbred line, and θ is the average proportion of
431 pairwise differences between *D. melanogaster* genome sequences, estimated as 0.141 from DGRP. Meanwhile,
432 the number of mismatches that overlap with SNP sites in DGRP (M_{SNP}) = 28,657 = $Lm\theta_D + pL\theta$, where
433 θ_D is the proportion of sites that are known SNPs within DGRP (0.0404). Note that this formulation makes
434 the simplifying assumption that all segregating SNPs would have been previously observed in DGRP, which
435 makes the correction conservative. Solving for the unknown variables:

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

436
437
438
439 To convert m to the TruSeq synthetic long-read error rate, we simply divide by the average depth of cov-
440 erage of the euchromatic sequence (20.58 \times), estimating a corrected error rate of 0.0192% per base. This
441 estimate is still conservative in that it does not account for mismatches observed multiple times at a sin-
442 gle site, which should overwhelmingly represent residual polymorphism. We therefore additionally applied
443 a second approach where we assumed that all 12,064 sites with more than one mismatch represented true
444 polymorphism, calculating the error rate as above with only singleton mismatches. As expected, this method
445 yielded a slightly lower error rate of 0.0183%.

446 **Genome assembly**

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

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

464 For the down-sampled assemblies with lower coverage, we based the expected coverage on the average
465 mapped depth of coverage of $21\times$ for the full dataset. We randomly sampled reads from a concatenated
466 FASTQ of all four libraries until the total length of the resulting dataset was equal to the desired coverage.

467 **Assessment of assembly quality**

468 We aligned the contigs produced by the Celera Assembler to the reference genome sequence using the NUCmer
469 pipeline (version 3.23) (Delcher et al., 2002; Kurtz et al., 2004). From this alignment, we used the delta-filter
470 tool to extract the longest increasing subset of alignments to the reference (i.e. the longest consistent set of
471 alignments with respect to the reference sequence). We then used to coordinates of these alignments to both
472 measure overall assembly quality and investigate assembly of particular genomic features, including genes,
473 TEs, and segmental duplications.

474 Using this alignment, we identified the locations of reference-annotated gene and TE sequences in our
475 assembly and used local alignment with BLASTN (Altschul et al., 1997) to determine sequence identity and

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

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

480 The GLMM and GLM used to test the characteristics of TEs that affected accurate assembly were built
481 using the *lme4* package (Bates et al., 2013) within the R statistical computing environment (R Core Team,
482 2013). In the GLMM, the response variable was represented by a binary indicator denoting whether or not
483 the entire length of the TE was accurately assembled. This model assumed a binomial error distribution
484 with a logit link function. TE copy length, divergence (number of substitutions per base compared to the
485 canonical sequence of the TE family), and average coverage of the corresponding chromosome were included
486 as fixed effects, while TE family was included as a random effect. For the GLM, we aggregated assembly
487 results by family, with the proportion of copies in the family accurately assembled included as the response
488 variable. This allowed us to include copy number as a fixed effect, along with the average length, average
489 divergence, and average depth of coverage of the corresponding chromosomes for each TE family. In both
490 models, all predictor variables were standardized to zero mean and unit variance prior to fitting, in order to
491 compare the magnitude of the effects.

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

494 **Data access**

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

498 **Acknowledgements**

499 Thank you to Alan Bergland for performing the DNA extractions and to Anthony Long for providing the
500 strain. Thanks also to Julie Collens and Courtney McCormick for preparing and delivering the long read
501 libraries.

502 **Author contributions**

503 RCM, RWT, and ASFL contributed to the data analysis. RCM prepared the manuscript, ASFL also con-
504 tributed to the writing of the manuscript, and all other authors contributed comments and revisions. TAB
505 and MK contributed to the data generation and provided guidance during planning stages of the experiment.
506 DAP helped design the experiment and provided guidance on analyses throughout. All authors read and
507 approved the final manuscript.

508 **Disclosure declaration**

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

513 **Figure legends**

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

517 in the reads, mapping errors, errors in the reference genome, or accurate sequencing of residual polymorphism.

518

519 Figure 2: Depth of coverage per chromosome arm. The suffix “Het” indicates the heterochromatic portion
520 of the corresponding chromosome. M.iso1 is the mitochondrial genome of the *yw;cn,bw,sp* strain. U and
521 Uextra are additional scaffolds in the reference assembly that could not be mapped to chromosomes.

522

523 Figure 3: Probability of accurate (100% length and sequence identity) assembly with respect to significant
524 predictor variables: TE length ($b = -5.588 \times 10^{-4}$, $Z = -20.294$, $P < 2 \times 10^{-16}$), divergence ($b = 4.073$,
525 $Z = 6.864$, $P = 6.69 \times 10^{-12}$), and average depth of coverage of the chromosome on which the TE is found
526 ($b = 5.474 \times 10^{-2}$, $Z = 3.493$, $P = 0.000478$). The upper sets of points represent TEs which were perfectly
527 assembled, while the lower set of points represents TEs which are absent from the assembly or were mis-
528 assembled with respect to the reference. Lines represent predicted values from the GLMM fit to these data.
529 Colors indicate different TE families (126 total).

530

531 Figure 4: Assembly metrics as a function of depth of coverage of TruSeq synthetic long-reads. **A:** NG(X)
532 contig length for full and down-sampled coverage data sets. This metric represents the size of the contig
533 for which X% of the genome length (180 Mbp) lies in contigs of that size or longer. **B:** The proportion
534 of genes and transposable elements accurately assembled (100% length and sequence identity) for full and
535 down-sampled coverage data sets.

536

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

538

539 Figure S2: K-mer coverage histogram generated by KmerGenie (Chikhi and Medvedev, 2014), depicting the
540 number of occurrences of each unique k-mer within the TruSeq synthetic long-read dataset. Low coverage
541 k-mers represent long-read errors and polymorphism (or errors in the reference genome) while high coverage
542 k-mers (x-axis truncated at $60\times$) represent genomic repeats.

543

544 Figure S3: Dot plots depicting NUCmer (Delcher et al., 2002) global alignment between assembled contigs
545 and the reference genome. Segments off of the diagonal represent various classes of mis-assembly (insertions,
546 deletions, or translocations with respect to the reference sequence). Blue segments represent forward align-
547 ments, while red segments indicate an inversion with respect to the rest of the contig alignment. Dot plots

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

549

550 Figure S4: IGV screenshot (Robinson et al., 2011; Thorvaldsdóttir et al., 2013) of a representative case where
551 assembly fails due to a deficiency of long-read data derived from a long transposable element sequence. The
552 upper-most track (blue) represents the NUCmer alignment of assembled contigs to the reference genome.
553 The middle track represents the BWA alignment of the underlying TruSeq synthetic long-reads. For each
554 of these tracks, blue and red shading indicate the orientation of the alignment (i.e. whether the sequence is
555 reverse complemented). The bottom track (green) indicates the boundaries of transposable elements.

556

557 Figure S5: IGV screenshots (Robinson et al., 2011; Thorvaldsdóttir et al., 2013) of representative cases where
558 assembly succeeds or fails based on characteristics of TEs in the genomic region. See the legend of Figure S4
559 for descriptions of each of the alignment tracks. **A:** A case where assembly fails in the presence of tandem
560 repeats of elements from the Dm88 family. **B:** A case where assembly succeeds in a repeat-dense region of
561 chromosome arm 2R.

562

563 **Figures**

Figure 1

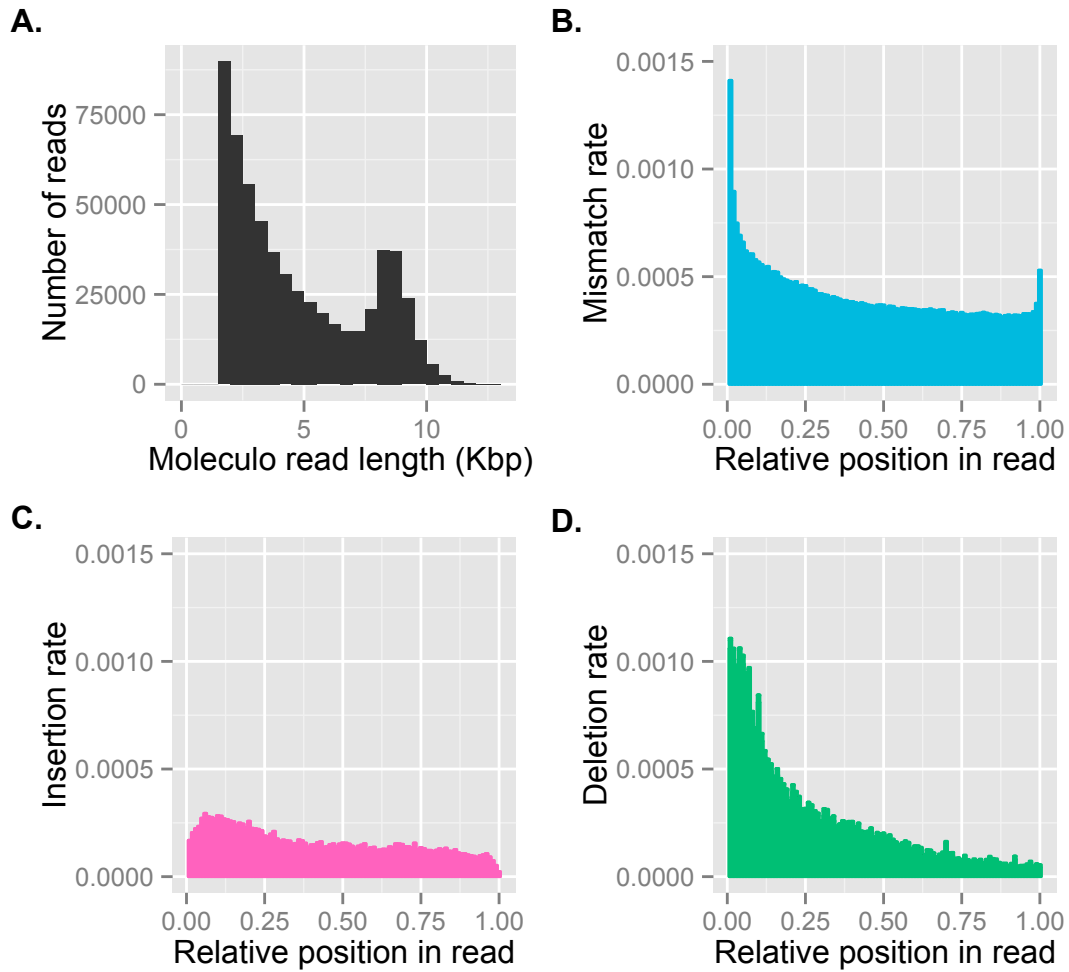


Figure 2

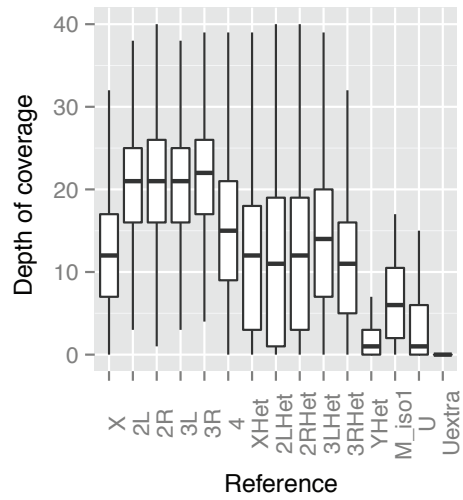


Figure 3

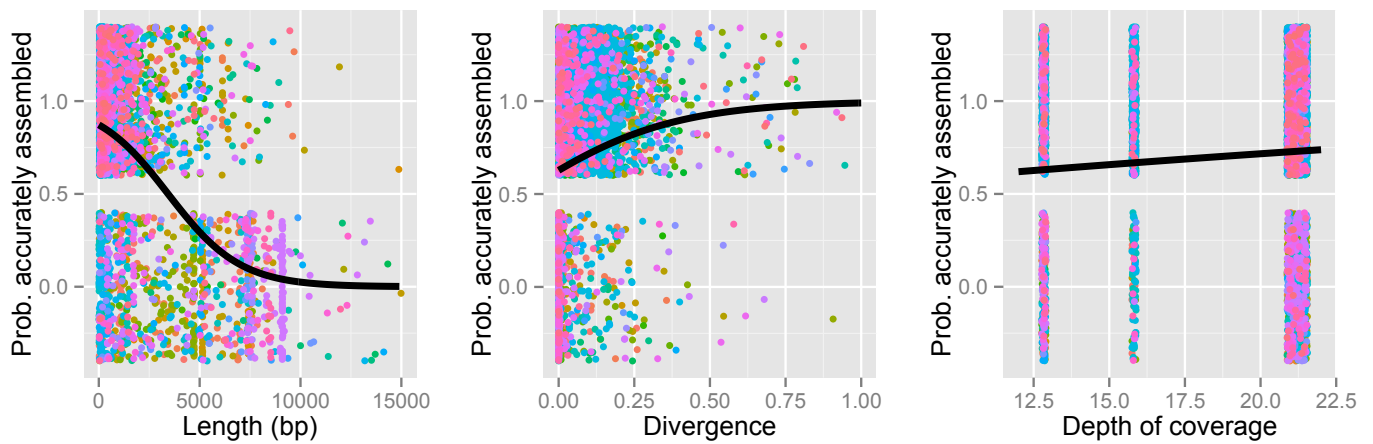


Figure 4

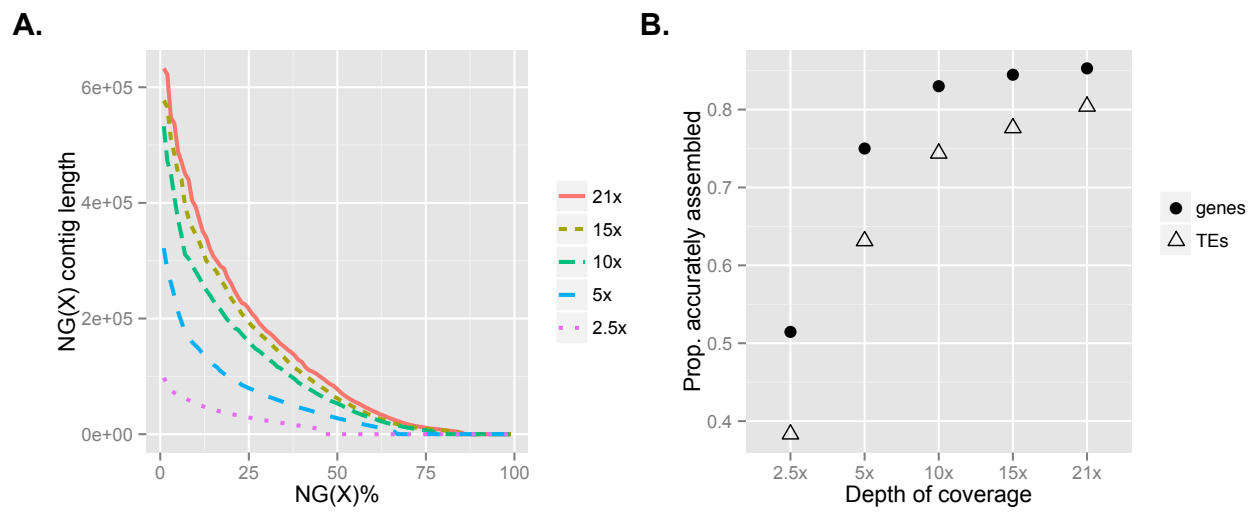


Figure S1

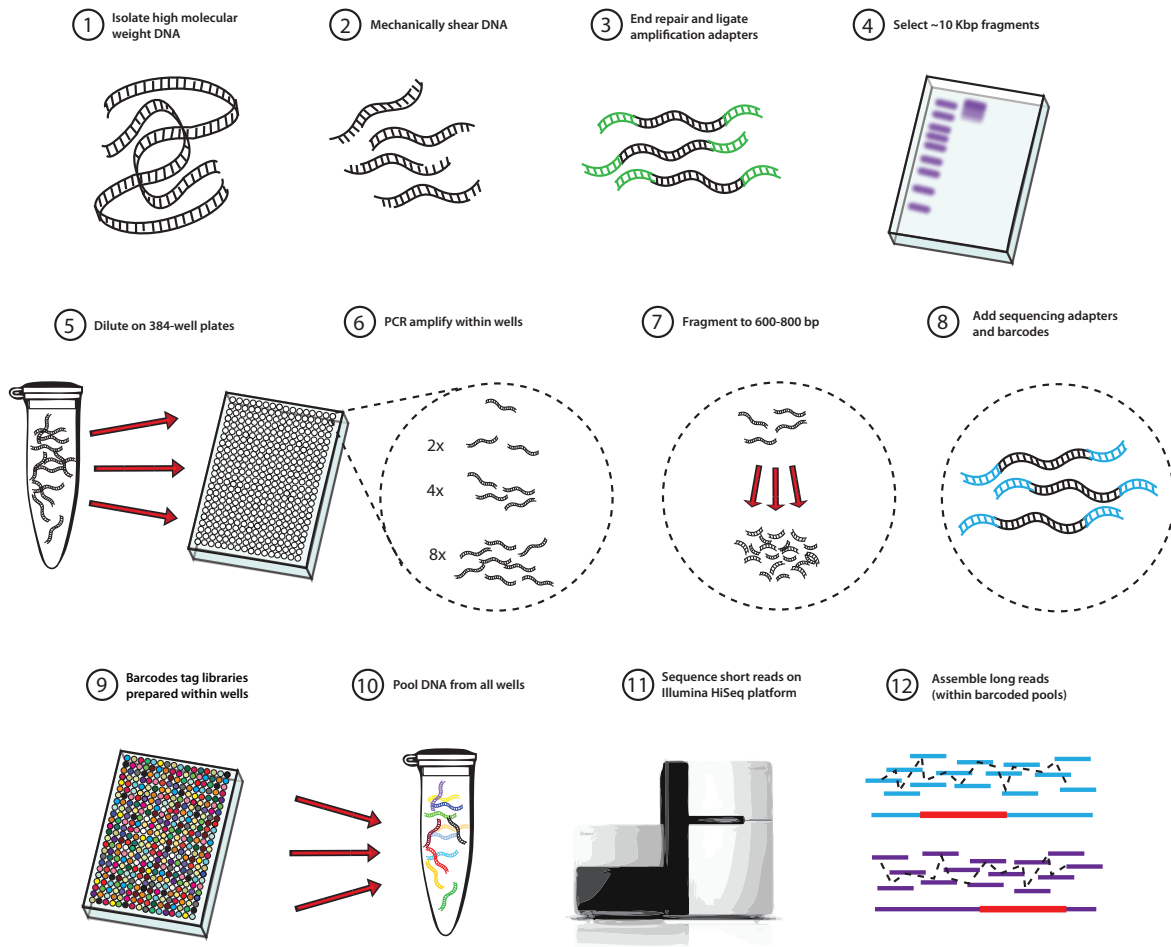


Figure S2

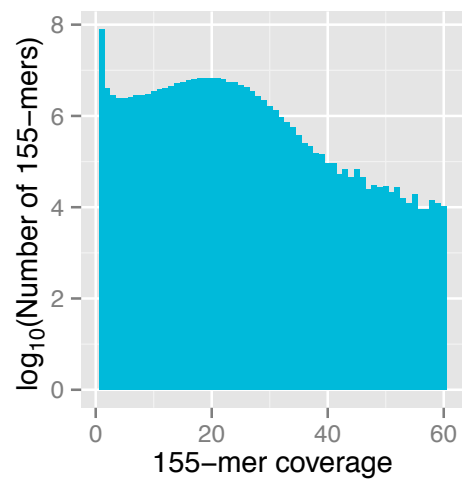
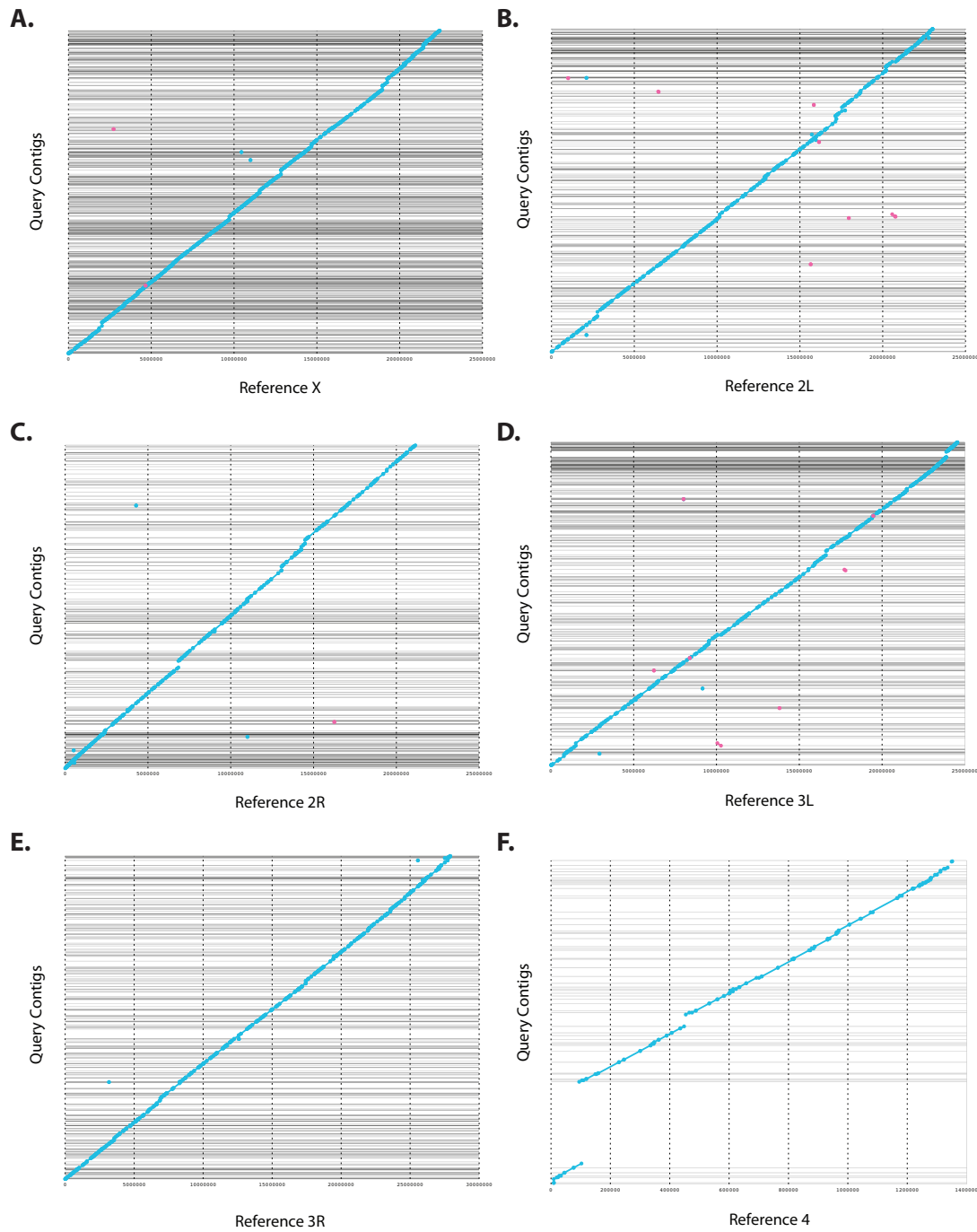


Figure S3



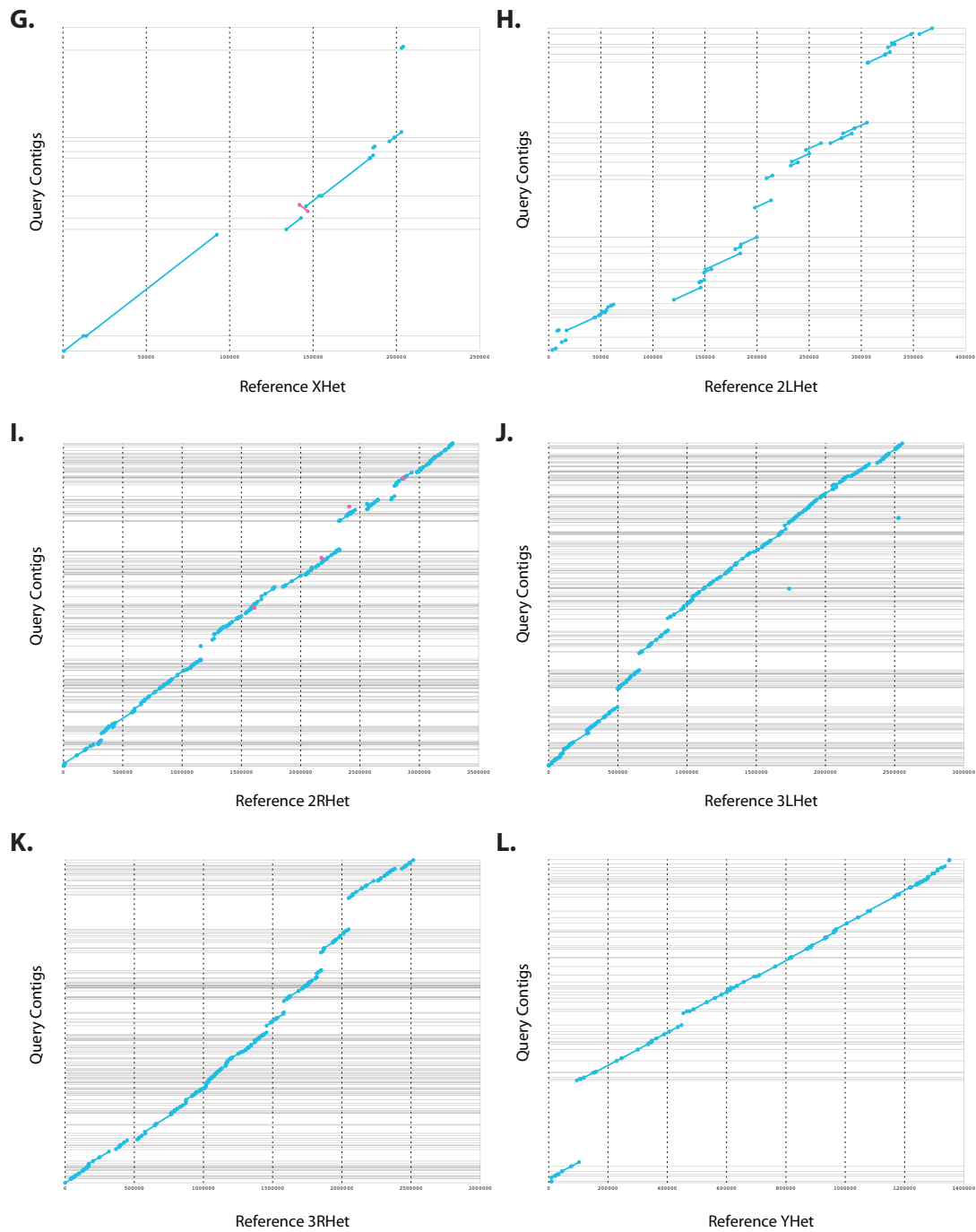


Figure S4

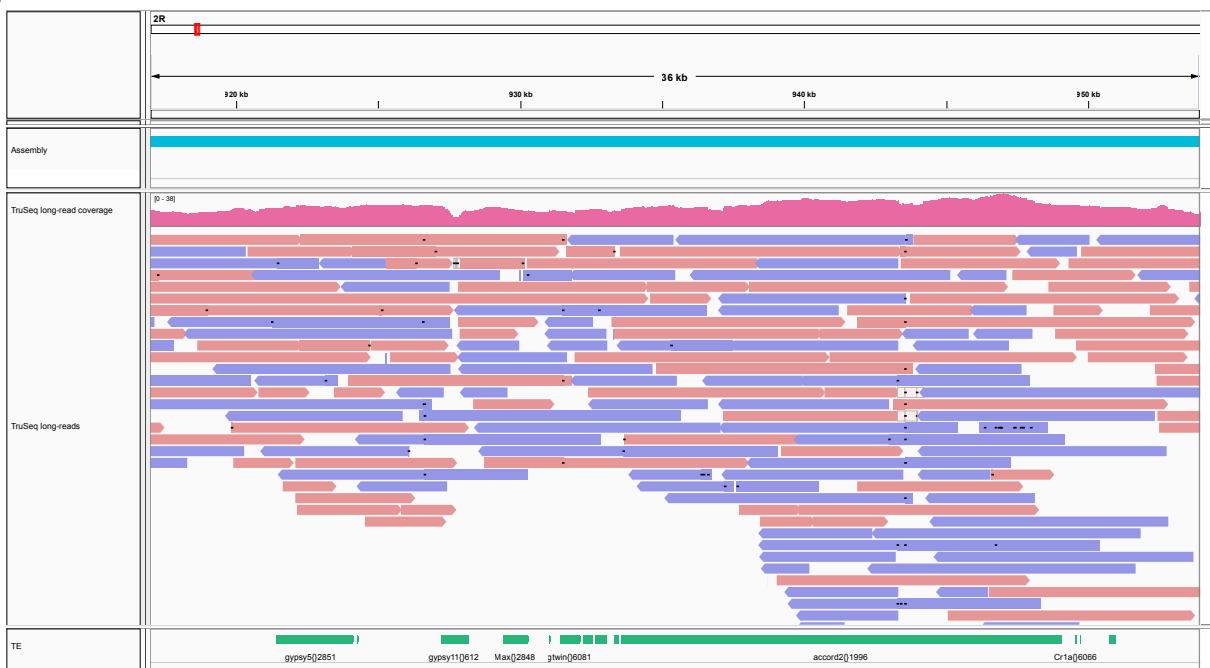


Figure S5

A.



B.



564 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).

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

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	<i>Drosophila melanogaster</i>
93	<i>Gluconacetobacter diazotrophicus</i> PA1 5
79	<i>Drosophila mauritiana</i>
64	<i>Enterobacteria</i> phage HK629
59	<i>Gluconacetobacter xylinus</i> NBRC 3288
42	<i>Acetobacter pasteurianus</i> 386B
37	<i>Gluconobacter oxydans</i> 621H
35	<i>Gluconobacter oxydans</i> H24
26	<i>Drosophila simulans</i>
18	Cloning vector pSport1
15	synthetic construct
13	<i>Acetobacter pasteurianus</i> IFO 3283-01
11	<i>Drosophila pseudoobscura</i>
9	<i>Acetobacter pasteurianus</i> IFO 3283-12
5	<i>Drosophila sechellia</i>
5	<i>Acetobacter aceti</i>
5	<i>Granulibacter bethesdensis</i> CGDNIH1
5	<i>Rhodomicrobium vanniellii</i> 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

References

- 565
- 566 Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., Scherer, S. E.,
567 Li, P. W., Hoskins, R. A., Galle, R. F., et al. (2000). The genome sequence of *Drosophila melanogaster*.
568 *Science* **287**:2185–2195.
- 569 Alkan, C., Sajjadian, S., and Eichler, E. E. (2010). Limitations of next-generation genome sequence assembly.
570 *Nature Methods* **8**:61–65.
- 571 Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997).
572 Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids*
573 *Research* **25**:3389–3402.
- 574 Bates, D., Maechler, M., Bolker, B., and Walker, S. (2013). *lme4: Linear mixed-effects models using Eigen*
575 *and S4*. R package version 1.0-5.
576 **URL:** <http://CRAN.R-project.org/package=lme4>
- 577 Bradnam, K. R., Fass, J. N., Alexandrov, A., Baranay, P., Bechner, M., Birol, I., Boisvert, S., Chapman,
578 J. A., Chapuis, G., Chikhi, R., et al. (2013). Assemblathon 2: evaluating *de novo* methods of genome
579 assembly in three vertebrate species. *GigaScience* **2**:10.
- 580 Celniker, S. E., Wheeler, D. A., Kronmiller, B., Carlson, J. W., Halpern, A., Patel, S., Adams, M., Champe,
581 M., Dugan, S. P., Frise, E., et al. (2002). Finishing a whole-genome shotgun: release 3 of the *Drosophila*
582 *melanogaster* euchromatic genome sequence. *Genome Biology* **3**:RESEARCH0079.
- 583 Chikhi, R. and Medvedev, P. (2014). Informed and automated k-mer size selection for genome assembly.
584 *Bioinformatics* **30**:31–37.
- 585 Clark, A. G., Eisen, M. B., Smith, D. R., Bergman, C. M., Oliver, B., Markow, T. A., Kaufman, T. C., Kellis,
586 M., Gelbart, W., Iyer, V. N., et al. (2007). Evolution of genes and genomes on the *Drosophila* phylogeny.
587 *Nature* **450**:203–218.
- 588 Clarke, J., Wu, H.-C., Jayasinghe, L., Patel, A., Reid, S., and Bayley, H. (2009). Continuous base identifica-
589 tion for single-molecule nanopore DNA sequencing. *Nature Nanotechnology* **4**:265–270.
- 590 Cordaux, R. and Batzer, M. A. (2009). The impact of retrotransposons on human genome evolution. *Nature*
591 *Reviews Genetics* **10**:691–703.

- 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.

- 618 Keane, T. M., Wong, K., and Adams, D. J. (2013). RetroSeq: transposable element discovery from next-
619 generation sequencing data. *Bioinformatics* **29**:389–390.
- 620 Kidwell, M. G. and Lisch, D. R. (2001). Perspective: transposable elements, parasitic DNA, and genome
621 evolution. *Evolution* **55**:1–24.
- 622 Kurtz, S., Phillippy, A., Delcher, A. L., Smoot, M., Shumway, M., Antonescu, C., and Salzberg, S. L. (2004).
623 Versatile and open software for comparing large genomes. *Genome Biology* **5**:R12.
- 624 Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K.,
625 Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. *Nature*
626 **409**:860–921.
- 627 Lee, E., Iskow, R., Yang, L., Gokcumen, O., Haseley, P., Luquette, L. J., Lohr, J. G., Harris, C. C., Ding,
628 L., Wilson, R. K., et al. (2012). Landscape of somatic retrotransposition in human cancers. *Science*
629 **337**:967–971.
- 630 Levy, S., Sutton, G., Ng, P. C., Feuk, L., Halpern, A. L., Walenz, B. P., Axelrod, N., Huang, J., Kirkness,
631 E. F., Denisov, G., et al. (2007). The diploid genome sequence of an individual human. *PLoS Biology*
632 **5**:e254.
- 633 Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform.
634 *Bioinformatics* **25**:1754–1760.
- 635 Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., Shi, Z., Li, Y., Li, S., Shan, G., Kristiansen, K., et al. (2010).
636 De novo assembly of human genomes with massively parallel short read sequencing. *Genome*
- 637 Li, X. and Waterman, M. S. (2003). Estimating the repeat structure and length of DNA sequences using
638 L-tuples. *Genome Research* **13**:1916–1922.
- 639 Mackay, T. F. C., Richards, S., Stone, E. A., Barbadilla, A., Ayroles, J. F., Zhu, D., Casillas, S., Han, Y.,
640 Magwire, M. M., Cridland, J. M., et al. (2012). The *Drosophila melanogaster* Genetic Reference Panel.
641 *Nature* **482**:173–178.
- 642 Miller, J. R., Koren, S., and Sutton, G. (2010). Assembly algorithms for next-generation sequencing data.
643 *Genomics* **95**:315–327.
- 644 Nekrutenko, A. and Li, W. H. (2001). Transposable elements are found in a large number of human protein-
645 coding genes. *Trends in genetics : TIG* **17**:619–621.

- 646 Osoegawa, K., Vessere, G. M., Li Shu, C., Hoskins, R. A., Abad, J. P., de Pablos, B., Villasante, A., and
647 de Jong, P. J. (2007). BAC clones generated from sheared DNA. *Genomics* **89**:291–299.
- 648 Phillippy, A. M., Schatz, M. C., and Pop, M. (2008). Genome assembly forensics: finding the elusive mis-
649 assembly. *Genome Biology* **9**:R55.
- 650 Quesneville, H., Bergman, C. M., Andrieu, O., Autard, D., Nouaud, D., Ashburner, M., and Anxolabehere,
651 D. (2005). Combined Evidence Annotation of Transposable Elements in Genome Sequences. *PLoS Com-
652 putational Biology* **1**:e22.
- 653 Quesneville, H., Nouaud, D., and Anxolabehere, D. (2003). Detection of New Transposable Element Families
654 in *Drosophila melanogaster* and *Anopheles gambiae* Genomes. *Journal of molecular evolution* **57**:S50–S59.
- 655 R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical
656 Computing, Vienna, Austria.
657 **URL:** <http://www.R-project.org/>
- 658 Rebollo, R., Romanish, M. T., and Mager, D. L. (2012). Transposable elements: an abundant and natural
659 source of regulatory sequences for host genes. *Annual Review of Genetics* **46**:21–42.
- 660 Roberts, R. J., Carneiro, M. O., and Schatz, M. C. (2013). The advantages of SMRT sequencing. *Genome
661 Biology* **14**:405.
- 662 Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., and Mesirov,
663 J. P. (2011). Integrative genomics viewer. *Nature Biotechnology* **29**:24–26.
- 664 Schatz, M. C., Walenz, B. P., Martin, J., Howard, J. T., Ganapathy, G., Wang, Z., Rasko, D. A., McCombie,
665 W. R., Jarvis, E. D., Koren, S., et al. (2012). Hybrid error correction and de novo assembly of single-
666 molecule sequencing reads. *Nature Biotechnology* **30**:693–700.
- 667 Thorvaldsdóttir, H., Robinson, J. T., and Mesirov, J. P. (2013). Integrative Genomics Viewer (IGV): high-
668 performance genomics data visualization and exploration. *Briefings in Bioinformatics* **14**:178–192.
- 669 Treangen, T. J. and Salzberg, S. L. (2012). Repetitive DNA and next-generation sequencing: computational
670 challenges and solutions. *Nature Reviews Genetics* **13**:36–46.
- 671 Voskoboinik, A., Neff, N. F., Sahoo, D., Newman, A. M., Pushkarev, D., Koh, W., Passarelli, B., Fan, H. C.,
672 Mantalas, G. L., Palmeri, K. J., et al. (2013). The genome sequence of the colonial chordate, *Botryllus
673 schlosseri*. *eLife* **2**:e00569.

674 Wickham, H. (2009). *ggplot2: elegant graphics for data analysis*. Springer New York.

675 **URL:** <http://had.co.nz/ggplot2/book>

676 Ye, L., Hillier, L. W., Minx, P., Thane, N., Locke, D. P., Martin, J. C., Chen, L., Mitreva, M., Miller, J. R.,

677 Haub, K. V., et al. (2011). A vertebrate case study of the quality of assemblies derived from next-generation

678 sequences. *Genome Biology* **12**:R31.