A practical guide to *de novo* genome assembly using long reads

Mahul Chakraborty*, James G. Baldwin-Brown*, Anthony D. Long, J.J. Emerson†

* These authors contributed equally

† To whom correspondence should be addressed: jje@uci.edu

Department of Ecology and Evolutionary Biology, University of California Irvine, Irvine, California, United States of America

Abstract:

Genome assemblies that are accurate, complete, and contiguous are essential for identifying important structural and functional elements of genomes and for identifying genetic variation. Nevertheless, most recent genome assemblies remain incomplete and fragmented. While long molecule sequencing promises to deliver more complete genome assemblies with fewer gaps, concerns about error rates, low yields, stringent DNA requirements, and uncertainty about best practices may discourage many investigators from adopting this technology. Here, in conjunction with the gold standard *Drosophila melanogaster* reference genome, we analyze recently published long molecule sequencing data to identify what governs completeness and contiguity of genome assemblies. We also present a meta-assembly tool for improving contiguity of final assemblies constructed via different methods. Our results motivate a set of preliminary best practices for assembly, a “missing manual” that guides key decisions in building high quality *de novo* genome assemblies, from DNA isolation to polishing the assembly.

Introduction:

De novo genome assembly is the process of stitching DNA fragments together into contiguous segments (contigs) representing an organism's chromosomes(Simpson and Pop 2015). Until recently, genomes tended to be assembled using fragments shorter than 1,000 bp. However, such assemblies tend to be highly fragmented when they are generated using sequencing reads shorter than common repeats(Baker 2012; Bradnam, et al. 2013; Myers 1995; Simpson and Pop 2015). Longer reads can circumvent this problem, even when such reads exhibit errors rates as high as 20%(Koren and Phillippy 2015; Lam, et al. 2014; Lander and Waterman 1988; Motahari, et al. 2013; Shomorony, et al. 2015). Importantly, error-prone reads can be corrected, provided there is sufficient coverage and the errors are approximately uniformly distributed(Lander and Waterman 1988). Single molecule sequencing, like that offered by Pacific Biosciences (PacBio), meets this criterion with reads that are routinely tens of kilobases in length(Kim, et al. 2014; Koren, et al. 2013; Koren and Phillippy 2015; Pendleton, et al. 2015). While PacBio sequences have high error rates (~15%), errors are nearly uniformly distributed across sequences(Koren and Phillippy 2015). With sufficient coverage, these sequences can be used to correct themselves(Churchill and Waterman 1992). Assemblies using such correction are referred to as PacBio only assembly(Berlin, et al. 2015). Alternatively, researchers can perform a hybrid assembly using a combination of noisy PacBio long molecules and high quality short reads (e.g. Illumina).(Koren, et al. 2012) (Pendleton, et al. 2015)

Recently, the value of long molecule sequencing has been definitively demonstrated with the release of several high quality reference-grade genomes assembled from PacBio sequencing data (Berlin, et al. 2015; Kim, et al. 2014). Despite these successes, shepherding a genome project through the process of DNA isolation, sequencing, and assembly still poses many uncertainties and challenges, especially for research groups who see genomes as a means to another goal rather than the goal itself. For example, because high quality genome assembly relies upon long sequencing reads to bridge repetitive genomic regions(Bresler, et al. 2013; Lam, et al. 2014; Lander

and Waterman 1988; Myers, et al. 2000) and high coverage to circumvent read errors(Baker 2012; Churchill and Waterman 1992; Motahari, et al. 2013), the stringent DNA isolation requirements (size, quantity, and purity) for PacBio sequencing(Kim, et al. 2014) intended for genome assembly are different than those typically employed. Moreover, at present, the low average read quality produced by PacBio sequencing causes coverage requirements to be at least 50-fold (Berlin, et al. 2015; Koren and Phillippy 2015; Sakai, et al. 2015). This, combined with its comparatively expensive price, makes striking the right balance between price and assembly quality important. Exacerbating the problem is the fact that rediscovering the optimal approach for a genome project is itself expensive and time consuming. As a consequence of these challenges and uncertainties, many groups may opt out of a long molecule approach, or worse, sink scarce resources into an approach ill-suited for their goals because the consequences of many decisions involved in long molecule sequencing projects have not been synthesized.

In order to derive an optimal strategy for genome assembly we investigated sample handling (i.e. DNA isolation, quality control, shearing, library loading, etc.), assembly strategies, and properties of the data (i.e. read quality, length, and read filtering). We first evaluate strategies for assembling PacBio reads, and how they perform with differing amounts of sequence coverage. Then, we assess the contribution of read length and read quality to assembly contiguity. We also introduce quickmerge, a simple, fast, and general meta-assembler that merges assemblies to generate a more contiguous assembly. We also describe the protocols, quality-control practices, and size selection strategies that consistently yield high quality DNA reads required for reference grade genome assemblies. Finally, we recommend a strategy flexible enough to yield high quality assemblies from as little as 25X long molecule coverage to as much as >100X.

Results

Long read assembly

PacBio self correction has been used to assemble the *D. melanogaster* reference strain (ISO1) genome so contiguously that most chromosome arms were represented by fewer than 10 contigs(Berlin, et al. 2015). This assembly was generated by using the PBcR pipeline(Berlin, et al. 2015) and 121X (15.8 Gb), or 42 SMRTcells' worth, of PacBio long molecule sequences(Kim, et al. 2014). However, currently, such high coverage may be too expensive for many projects, especially when the genome of the target organism is large. Consequently, we set out to determine how much sequence data is required to obtain assemblies of desired contiguity. We first selected reads from 15, 20, 25, 30, and 35 randomly chosen SMRTcells (5.16Gb, 6.87Gb, 8.12Gb, 10.06Gb, 12.85Gb) from the 42 SMRTcells of ISO1 PacBio reads(Kim, et al. 2014). Our sampling method was inclusive and additive: to obtain 20 SMRTcells, we took the 15 previously randomly chosen SMRTcells and then added 5 more randomly selected SMRTcells to it. We then assembled these datasets using the PBcR pipeline. As shown in Fig. 1, the contig NG50 (NG50; G =130×10^6 bp) improves until it plateaus at 77X coverage (30 SMRTcells). At extremely high coverage (42 SMRTcells), the NG50 surges again. Notably, despite the extreme contiguity of these sequences, we are still discussing complete contigs, not gap containing scaffolds.

Hybrid assembly

As Fig. 1 makes clear, PB only assembly leads to relatively fragmented genomes at lower coverage (Fig. 1), we investigated whether another assembly strategy could perform better with similar amounts of long molecule data. We chose DBG2OLC(Ye, et al. 2014) for its speed and its ability to assemble using less than 30X of long molecule coverage (*cf.* PacBio only methods, which typically require higher coverage(Koren and Phillippy 2015)). DBG2OLC is a hybrid method, which uses both long read data and contigs obtained from a De Bruijn graph assembly. We used contigs from a single Illumina assembly generated using 64X of Illumina paired end reads (Langley, et al. 2012). As shown in Fig. 1, the assembly NG50 increases dramatically as PacBio coverage increased, plateauing near 10 SMRT cells (26X). Beyond this point, NG50 remained relatively constant. Alignment of the test assemblies to the ISO1 reference genome showed that the high level of contiguity in the 26X hybrid assembly without

downsampling was due to chimeric contigs, and that these errors are fixed as coverage increases (supplementary Fig. 1-2). Chimeras were also absent when only the longest 50% or 75% of reads from the 26X dataset are used.

To measure the impact of read length on hybrid assembly contiguity, we down-sampled the datasets by discarding the shortest reads such that the resulting datasets contained 50% and 75% of initial total basepairs of data. We then ran the same assembly pipelines using these downsampled datasets and compared to the assemblies constructed from their counterparts that were not downsampled. Our downsampling shows that with high levels of PacBio coverage, modest gains in assembly contiguity can be obtained by simply discarding the shortest reads (Fig. 1, red lines). Our hybrid assembly results indicate that improvements in contiguity above 30X are modest, though hybrid assemblies remain more contiguous than PacBio-only assemblies up until above 60X coverage. For projects limited by the cost of long molecule sequencing, a hybrid approach using ~30X PacBio sequence coverage is an attractive target that minimizes sequencing in exchange for modest sacrifices in contiguity.

Assembly merging

With modest PacBio sequence coverage (≤50X), hybrid assemblies are less fragmented than their self corrected counterparts, but more fragmented than self corrected assemblies generated from higher read coverage (Fig. 1). Despite this, for lower coverage, many contigs exhibit complementary contiguity, as observed in alignments (e.g. Supplementary Fig. 3a) between a PB only assembly (20 SMRT cells or 52X reads; NG50 1.98 Mb) and a hybrid assembly (longest 30X from 20 SMRTcells reads; NG50 3.2 Mb). For example, the longest contig (16.8 Mb) in the PB only assembly, which aligns to the chromosome 3R of the reference sequence (Supplementary Fig. 3c), is spanned by 5 contigs in the hybrid assembly (Supplementary Fig. 3b). This complementarity suggests that merging might improve the overall assembly.

We first attempted to merge the hybrid assembly and the PB only assembly using the existing meta assembler minimus2(Treangen, et al. 2011), but the program

often failed to run to completion when merging a hybrid assembly and a PB only assembly, and when it did finish, the run times were measured in days. We therefore developed a program, quickmerge, that merges assemblies using the MUMmer(Kurtz, et al. 2004) alignment between the assemblies. Assembly contiguity improved dramatically when we merged the above hybrid and PB-only assemblies (assembly NG50 9.1 Mb; Fig. 1, supplementary Fig. 4). Further, assembly merging closed gaps present in the published ISO1 PacBio genome assembly (supplementary Fig. 5)(Berlin, et al. 2015) . The longest merged contig (27.5Mb), which aligns to the chromosome arm 3R of the reference sequence (supplementary Fig. 5), was longer than PacBio assembly based on 42 SMRTcells (25.4Mb)(Berlin, et al. 2015) (supplementary Fig. 5). This indicates that the contiguity of even high coverage PB-only assemblies can be increased by addition of inexpensive Illumina reads, and gaps in hybrid assembly can be closed by PB-only assembly even when the PB-only assembly quality is suboptimal.

Assessment of assembly quality

We assessed assembly quality using the *Quast* software package(Gurevich, et al. 2013). We confined our assessment to assemblies related to application of the quickmerge meta assembler, leaving the assessment of PBcR and DBG2OLC assemblies to their respective publications (Berlin, et al. 2015; Ye, et al. 2014). Quast quantifies assembly contiguity and additionally identifies misassemblies, indels, gaps, and substitutions in an assembly when compared to a known reference. We found that, compared to the *D. melanogaster* reference, all assemblies had relatively few errors, with the primary difference among the assemblies being genome contiguity (NG50). Hybrid assemblies tended to have fewer assembly errors than PB-only assemblies: the total number of misassemblies and the total number of contigs with misassemblies tended to be higher in PB only assemblies compared to hybrid assemblies. Still, PB-only assemblies tended to have slightly fewer mismatched bases compared to the reference, and slightly fewer small indels. Merged assemblies, being a mix of PB-only and hybrid assemblies, tended to have intermediate *Quast* statistics, although the merged assemblies improved upon the source assemblies in terms of misassemblies and misassembled contigs. Overall, the rate of mismatches was low at an average

(across all assemblies) of 47 errors per 100kb (Supplementary Table 1, Supplementary Fig. 12). Mismatches and indels can be further reduced using existing programs, such as *Quiver*(Chin, et al. 2013). We used *Quiver* to polish all non-downsampled hybrid, self, and merged assemblies that used at least 15 SMRTcells of data. After *Quiver*, the average mismatch rate of the selected assemblies decreased from 24 per 100kb to 15, while the average indel rate decreased from 180 per 100kb to 32 (Supplementary Fig. 13).

Size selection and assembly contiguity

Long reads generated by library preparation with aggressive size selection(Kim, et al. 2014) can generate extremely contiguous and accurate *de novo* assemblies(Berlin, et al. 2015). Genomic DNA libraries prepared with less stringent size selection (see Methods) can generate reads that are substantially shorter than the reads that have been shown to assemble into nearly gapless contigs (Kim, et al. 2014) (Fig. 2a). Longer reads are predicted to generate more contiguous genomes (Lander and Waterman 1988; Motahari, et al. 2013). We measured this by assembling genomes using randomly sampled whole reads (see Materials and Methods) from the ISO1 dataset to simulate a read length distribution comparable to, but slightly longer than is typical when size selection is not aggressive. Due to the long read length distribution of the ISO1 dataset relative to the shorter target distribution above, a maximum of 52X of ISO1 data could be sampled.

Consistent with the theoretical prediction that, all else being equal, shorter reads produce more fragmented assemblies(Lander and Waterman 1988; Motahari, et al. 2013), reads from the downsampled 20 SMRTcell ISO1 data produced a PB-only assembly with an NG50 of 1.38 Mb, which is shorter than the NG50 (1.98 Mb) of the assembly from the same amount of ISO1 long read data (Fig. 2c). In addition, nearly all long contigs present in the original 20 SMRTcell assembly are fragmented in the assembly from the shorter reads (Supplementary Fig. 6), although the amount of sequence data (52X) used to build the assemblies is the same.

For hybrid assembly, the shorter dataset also produced significantly less contiguous assemblies, consistent with predictions from theory(Motahari, et al. 2013) (Fig. 2b). The NG50 achieved with 26X coverage of the shorter dataset was 1.62Mb, compared to an NG50 of 3.58Mb with the original ISO1 data. This is consistent with the PB-only result – longer read lengths lead to higher assembly contiguity. Thus, a library preparation procedure that aggressively size selects DNA is crucial in delivering long contigs.

The effects of read quality on assembly

As with reduction in read length, increased read errors are predicted to worsen assembly quality because noisier reads increase the required read length and coverage to attain a high quality assembly (Churchill and Waterman 1992; Shomorony, et al. 2015). When a PacBio sequencing experiment is pushed for high yield through either high polymerase or template concentration, the data exhibits lower quality scores (Fig. 3). Thus, with equal coverage and read length distribution, reads with higher error rates should result in a more fragmented assembly. To measure this effect, we partitioned the ISO1 PacBio read data into three groups with equal amounts of sequence (Supplementary Fig. 7). For the first two groups, the data was split in half, with one half comprising the reads from the bottom 50% of phred scores and the other comprising the top 50%. Cutoffs were chosen for individual 100bp length bins, so the resulting datasets maintained the length distribution of the original data. The third dataset was generated by randomly selecting 50% of the reads in the full dataset. We then performed PacBio-only and hybrid assemblies with these data.

Low read quality had a particularly dramatic effect on assembly by self correction (Fig. 4): the high quality and randomly sampled reads produced substantially better assemblies (6.23 Mb and 6.15 Mb, respectively) than the assembly made from low quality reads (NG50 146 kb). Hybrid assembly contiguity was far more robust to low quality reads (Fig. 4: NG50 of 3.1Mb for the high quality reads, 2.5Mb for the unfiltered reads, and 2.2Mb for the low quality reads), showing only moderate variation amongst different quality datasets.

DNA isolation for long reads

As shown in the previous sections, read length is an important determinant of genome assembly contiguity. The method used for DNA isolation to generate the published PacBio Drosophila assembly involved DNA extraction by CsCl density gradient centrifugation and g-Tube (Covaris, Woburn, MA) based DNA shearing(Kim, et al. 2014). CsCl gradient centrifugation is a time-consuming method that requires expensive equipment that is not routinely found in most labs. Additionally, g-Tubes are expensive, require specific centrifuges, and are extremely sensitive to both the total mass of DNA input and to its length. These problems can be circumvented by using a widely available DNA gravity flow anion exchange column extraction kit in concert with a blunt needle shearing method(Graham and Hill 2001). Because the DNA fragment size distribution is so important, field inversion gel electrophoresis (FIGE) is an essential quality control step to validate the length distribution of the input DNA (Fig. 5) (see Methods for details). Sequences generated from libraries constructed from this isolation method are comparable to or longer than the published Drosophila PacBio reads (Kim, et al. 2014) (Fig. 2a). The length distribution of the input DNA can potentially be improved further by using needles that generate even longer DNA fragments after shearing (supplementary Fig. 8).

Discussion:

Genome assembly projects must balance cost against genome contiguity and quality(Baker 2012). Self correction and assembly using only long reads clearly produces complete and contiguous genomes (Fig.1; supplementary Table 1). However, it is often impractical to collect the quantity of PacBio sequence data (>50X) necessary for high quality self correction either because of price or because of scarcity of appropriate biological material, especially when assembling very large genomes. For example, at least 40 µg of high quality genomic DNA is required to for us to generate 1.5 µg of PacBio library when we use two rounds of size selection in the library preparation protocol. A 1.5 µg library produces, on average, 15-20 Gb of long DNA molecules. This dramatic loss of DNA during library preparation limits the amount of

PacBio data that can be obtained for a given quantity of source tissue. When a project is limited by cost or tissue availability, a hybrid approach using a mix of short and long read sequences is an alternative to self corrected long read sequences.

Our results show that when 64.3X of 100bp paired end Illumina reads is used in combination with 10X –30X of PacBio sequences, reasonably high quality hybrid assemblies can be obtained, with 30X of PacBio sequences yielding the best assembly. In fact, as our results show, a 30X hybrid assembly is less fragmented and hence of higher quality than even a 50X self-corrected assembly (Fig. 1). However, our results also show that with the same long molecule data, PB only and hybrid assemblies often assemble complementary regions of the genome. Hence merging of a PB only and a hybrid assembly results in a better assembly than either of the two (supplementary table 1), regardless of the total amount of long molecule sequences (≥30X) used. Thus, projects for which ≥30X of single molecule sequence can be generated will better served by collecting an additional 50-100X of Illumina data. These data can then be used to generate both a self-corrected assembly and a hybrid assembly, which can then be merged to obtain an assembly of comparable contiguity to PB only assemblies using twice the amount of PacBio data (Fig. 1). This merged assembly approach produced the highest NG50 of any assembly at all coverage levels at which it could be tested, with little or no tradeoff in base accuracy or misassemblies (Supplementary Fig. 12-13).

Nonetheless, it is clear that the tools available for genomic assembly have inherent technical limitations: DBG2OLC assembly contiguity asymptotes as PacBio read coverage passes about 30X, and the PBcR pipeline produces the best assembly when the longest reads that make up 40X (of genome size) data are corrected and only the longest 25X from the corrected sequences are assembled(Berlin, et al. 2015). Indeed, when coverage greater than 25X is used for PacBio only assembly, there is a real loss of assembly quality as coverage increases (data not shown). This may be because an increase in coverage leads to the stochastic accumulation of contradictory reads that cannot be easily reconciled, a limitation of the overlap-layout-consensus (OLC) algorithm used in assembling the long reads(Miller, et al. 2010; Myers 1995).

Single molecule sequencing technologies, as offered by PacBio and Oxford Nanopore(Goodwin, et al. 2015), promise to improve the quality of *de novo* genome assemblies substantially. However, as we have shown using PacBio sequences as example, not all LMS data is equally useful when assembling genomes. We provide empirical validation, perhaps for the first time, of length and quality on assembly contiguity. Additionally, our results provide a novel insight: high throughput short reads can still be useful in improving contiguity of assemblies created with LMS, even when LMS coverage is high. In light of our results, we have a compiled a list of best practices for DNA isolation, sequencing, and assembly (Supplementary Fig. 10 and Supplementary Fig. 11). Particularly important for DNA isolation is quality control of read length via pulsed field gel electrophoresis. Regarding assembly, we recommend that researchers obtain between 50x and 100x Illumina sequence. Next is to determine how much long molecule coverage to obtain: between 25x and 35x, or greater than 35x. With coverage below 35X, PB only methods often fail to assemble, and produce low contiguity when they do assemble, and thus, we can only confidently recommend a hybrid assembly. Above 35X, we recommend meta assembly of a hybrid and a PB only assembly. In this case, we recommend downsampling to the 35X longest PacBio reads when generating the hybrid assembly may be helpful because hybrid assembly contiguity decreases above this coverage level, but this has not been extensively tested. For the last several years, the rapid development of short read sequencing has fostered an explosion of genome sequencing. However, as a result of the popularity of short read technologies, the average quality and contiguity of published genomes has plummeted(Alkan, et al. 2011). Indeed, short read sequences are poorly suited to the task of assembly, especially when compared with long molecule alternatives. While long molecule sequencing has rekindled the promise of high quality reference genomes for any organism, it is substantially more expensive than short read alternatives. In order to mitigate uncertainties inherent in adopting new technology, we have outlined the most salient features to consider when planning a genome assembly project. We have recommended effective DNA isolation and preparation practices that result in long reads that take advantage of what the PacBio technology has to offer. We have also provided a guide for assembly that leads to extremely contiguous genomes even when

circumstances prevent the collection of large quantities of long molecule sequence data recommended by current methods.

Methods:

PB-only Assembly

For PacBio sequences, the assembly pipeline is divided into three parts: correction, assembly, and polishing. Correction reduces the error rate in the reads to 0.5-1%(Berlin, et al. 2015), and is necessary because reads with a high (~15%) error rate are extremely difficult to assemble(Myers, et al. 2000). Correction is facilitated by high PacBio coverage, which allows the error corrector to successfully 'vote out' errors in the PacBio reads. For self correction, we used the PBcR pipeline(Berlin, et al. 2015) as implemented in wgs8.3rc1 which, by default, corrects the longest 40X reads. The second step involves assembling the corrected reads into contigs. We used the Celera assembler (Myers, et al. 2000), included in the same wgs package, for assembly. A third optional step involves polishing the contigs using Quiver(Chin, et al. 2013), which brings the error rate down to 0.01% or lower. All of the assemblies described in this paper were generated with the same PBcR command and spec file (commands and settings, Supplementary materials).

For PB only assembly of *D. melanogaster* ISO1 sequences, we used a publicly available PacBio sequence dataset (Kim, et al. 2014). We chose the *D. melanogaster* dataset for our experiments and simulations because *D. melanogaster* is widely used in genetics and genomics research and its reference sequence (release 5.57,http://www.fruitfly.org) is one of the best, if not the best, eukaryotic multicellular genome assemblies in terms of assembly contiguity. This is true for both the PacBio generated assembly (21Mb contig N50)[13] and the Sanger assembly (14Mb scaffold N50) of ISO1. A high quality reference assembly serves as a great positive control and a reference.

We evaluated assembly qualities using the standard assembly statistics (average contig size, number of contigs, assembled genome size, N50, etc.) using the package Quast(Gurevich, et al. 2013).

Hybrid Assembly

PB only assembly of high error, long molecule sequences depends upon redundancy between the various low quality reads to 'vote out' errors and identify the true sequence in the sequenced individual. An alternative approach to this problem is to use known high quality sequencing reads to correctly call the bases in the sequence, and then to use PacBio reads to identify the connectivity of the genome. In order to achieve the best possible assembly results, we tested several different hybrid assembly pipelines before choosing *DBG2OLC* and *Platanus* (Kajitani, et al. 2014). In our early tests, the next highest performing hybrid assembler, a combination of *ECTools* (Lee, et al. 2014) and *Celera*, achieved a highest N50 of 616kb in *Arabidopsis thaliana* using 19 SMRT cells of data(Lee, et al. 2014); in contrast, using 20 SMRT cells of the same data, the *DBG2OLC* and *Platanus* pipeline produced an N50 of 4.8Mb. We thus disregarded *ECTools* and focused on *DBG2OLC*. We tested the alternative error corrector, *LorDEC* (Salmela and Rivals 2014), along with the *Celera* assembler, but found that the *Lordec*-corrected Celera assembly of our standard *D. melanogaster* dataset (26X of PacBio data and 64.3X of Illumina data) produced an NG50 of only 109KB; thus, we also discarded *Lordec* as a viable assembly choice compared to *DBG2OLC*. Using the standard 64.3X of Illumina data discussed above and 26X of PacBio data, we compared *DBG2OLC* runs using three different De Bruijn graph assemblers: *SOAP*(Luo, et al. 2012), *ABySS* (Simpson, et al. 2009), and *Platanus*. The NG50s for the three assemblies were, respectively, 2.43Mb, 0.167Mb, and 3.59Mb. Based on this result, we chose to use *Platanus* for the remainder of the assemblies.

We used the pipeline recommended by *DBG2OLC*(Ye, et al. 2014) to perform hybrid assemblies. In this pipeline, we used *Platanus* to perform De Bruijn graph assembly on the Illumina reads. We used 8.36 Gb (64.3X) of Illumina sequence data of the ISO1 *D. melanogaster* inbred line generated by the DPGP project (Langley, et al. 2012) to generate a De Bruijn graph assembly using *Platanus*. We used *DBG2OLC* to align our PacBio reads to the De Bruijn graph assembly to produce a 'backbone', then, according to the DBG2OLC standard pipeline, used the backbone generate the consensus using the programs *Blasr* (Chaisson and Tesler 2012) and *PBDagCon*

(https://github.com/PacificBiosciences/pbdagcon). As with the PB only assemblies above, we evaluated assembly quality using the *Quast* package.

Assembly merging

Hybrid assembly and PacBio assembly were merged using a custom C++ program (https://github.com/mahulchak/quickmerge). The program takes two fasta files (containing contigs from a PB only assembly and contigs from a hybrid assembly) as inputs and splices contigs from the two assemblies together to produce an assembly with higher contiguity. First, the program MUMmer(Kurtz, et al. 2004) is used to compute the unique alignments between contigs from the two assemblies. Our program then uses these alignments and finds the high confidence overlaps (HCO) among them (supplementary Fig. s9). The program identifies HCOs by dividing the total alignment length between contigs by the length of unaligned but overlapping regions of the alignment partners (supplementary Fig. s9). The “HCO” parameter controls merging sensitivity at the cost of increased false positives: the higher the HCO parameter value, the more stringent the cutoff for HCO selection. For our assembly merging, we used 1.5 as the HCO cutoff. The program then searches amongst these HCOs to find alignments that involve long contigs in the linker assembly (here, the PB only assembly). The program then uses these long contigs as seeds to begin a search for contig length expansion. A higher HCO cutoff is used for seed contigs to avoid spurious seeding. We used an HCO value of 5.0 for seed contigs for all merged assemblies. The seed contigs are extended on both sides by looking for alignment partners in the alignment pool that passed the HCO >1.5 cutoff. Next, the ordered contigs are joined by crossing over from one assembly to another (supplementary Fig. s9). For 15, 20, 25, 30 SMRTcells datasets, merged assemblies were generated using the PB only assembly and their corresponding hybrid assemblies. For 35 and 42 (all reads) SMRTcells datasets, the PB only assemblies were merged with the hybrid assembly obtained from the 30 SMRT-cells dataset. All hybrid assemblies used for merging were generated without downsampling by read length or quality.

Downsampling

We used three different downsampling schemes on the *D. melanogaster* data: first, we randomly downsampled the data by drawing a random set of SMRTcells of data from the entire set of 42 SMRTcells; second, from those datasets, we downsampled the longest 50% and 75% of the reads. Finally, we downsampled the *D. melanogaster* data to match the read length distributions of PacBio reads from a pilot *Drosophila pseudoobscura* genome project that was produced using a standard protocol without aggressive size selection and generously made available by Stephen Richards. We used the *lowess* function in R with a smoother span (f) of 1/5 to generate curves representing the distribution of read lengths in the *D. melanogaster* and *D. pseudoobscura* datasets, then assigned a probability to each read length defined as the quotient of the melanogaster distribution and the pseudoobscura distribution at that read length. Reads were then randomly removed from the *D. melanogaster* dataset according to the assigned probabilities. This method was used for all numbers of SMRTcells up to 20. Thus, we generated a set of reads that relatively closely resembles the read length distribution of the original *D. pseudoobscura* data, but is made up of *D. melanogaster* sequence data, allowing for a comparison of assembly quality with regard to read length without differences in the genomes of the two species as a confounding factor. The lowess function resulted in a slightly over-smoothed distribution such that samples drawn from it were slightly longer than in *D. pseudoobscura*. Consequently, assemblies from these reads should be slightly better than if they exhibited the (shorter) distribution for *D. pseudoobscura*. As such, this choice is conservative and, if anything, underestimates the importance of size selection.

Additionally, we downsampled based on read quality to test the effect of read quality on assembly contiguity. We used a custom script to separate the entire 42 SMRTcell ISO1 dataset into two halves. One half contained the 50% of all reads with the lowest average base quality, while the other half contained the 50% of all reads with the highest average base quality. We also generated a dataset containing 50% of the data that consisted of randomly chosen reads (to preserve the quality distribution of the original data).

Preparing high quality DNA library for long reads

Obtaining high quality, high molecular weight (HMW) genomic DNA

We used Qiagen's Blood and Cell culture DNA Midi Kit for DNA extraction. As single molecule technologies (PacBio and Oxford Nanopore) do not require any sequence amplification step, a large amount of tissue is required to ensure enough DNA for library preparations that opt for no amplification (as is standard for genome assembly sequencing). For flies, 200 females or 250 males flies is sufficient for optimal yield (40-60ug ug DNA) from a single anion-exchange column. For other organisms, number of individuals need to be adjusted based on the tissue mass. A good rule of thumb is to keep the total amount of input tissue 100-150mg for optimal yield from each column.

To extract genomic DNA, 0-2 days old flies were starved for two hours, flash frozen in liquid nitrogen, and then ground into fine powder using a mortar and pestle pre-chilled with liquid nitrogen. The tissue powder is directly transferred into 9.5 ml of buffer G2 premixed with 38µl of RNaseA (100mg/ml) and then 250 µl (0.75U) of protease (Qiagen) is added to the tissue homogenate. The volume of protease can be increased to 500 µl to reduce the time of proteolysis. The tissue powder is mixed with the buffer by inverting the tube several times, ensuring that there are no large tissue clumps present in the solution. The homogenate is then incubated at 50˚C overnight with gentle shaking (with 500µl protease, this incubation time can be reduced to 2 hours or less).

The next day, the sample is taken out of the incubator shaker and centrifuged at 5000xg for 10 minutes at 4˚C to precipitate the tissue debris. The supernatant is decanted into a fresh 15ml tube. The little remaining particulate debris in the tube was removed with a 1 ml pipette. The sample is then vortexed for 5 seconds to increase the flow rate of the sample inside the column and then poured into the anion-exchange column. The column is washed and the DNA is eluted following the manufacturer's protocol. Genomic DNA is precipitated with 0.7 volumes of isopropanol and resuspended in Tris buffer (pH 8.0). For storage of one week or less, we kept the DNA at 4˚C to minimize freeze-thaw cycles; for longer storage, we kept the DNA at -20˚C.

Shearing the DNA

1.5" blunt end needles (Jensen Global, Santa Barbara, CA) were used to shear the DNA. The needle size can be varied to obtain DNA of different length distribution: 24 gauge needles produces a size range of 24-50 kb. To obtain larger fragments, <24 gauge needles need to be used. For the DNA we have sequenced, up to 200ug of high molecular weight raw genomic DNA was sheared using the 24 gauge needle (Fig. 5). Additionally, we have also sheared DNA with 21, 22, and 23 gauge needles to demonstrate the size distribution they generate (supplementary Fig. s8). In brief, the entire DNA solution is drawn into a 1ml Luer syringe and dispensed quickly through the needle. This step is repeated 20 times to obtain the desired distribution of fragment sizes.

Quality Control using FIGE

We verified the size distribution of unsheared and sheared genomic DNA using field inversion gel electrophoresis (FIGE), which allows separation of high molecular weight DNA. The DNA is run on a 1% agarose gel (0.5x TBE) with a pulse field gel ladder (New England Biolabs, Ipswich, MA). The gel is run at 4˚C overnight in 0.5 x TBE. To avoid temperature or pH gradient buildup, a pump is used to circulate the buffer. The FIGE is run using a BioRad Pulsewave 760 and a standard power supply with the following run conditions:

Initial time A: 0.6s, Final time B: 2.5 s, Ratio: 3, Run time: 8 h, MODE: 10, Initial time A: 2.5s, Final time B: 8s, Ratio: 3, Run time: 8 h, MODE: 11, Voltage: 135 V.

Library preparation

The needle sheared DNA is quantified with Qubit fluorometer (Life Technologies, Grand Island, NY) and NanoDrop (Thermo Scientific,,Wilmington, DE). Following quantification, 20 µg of sheared DNA is optionally run in four lanes of the Blue Pippin size selection instrument (Sage Science, Beverly, MA) using 15-50 kb as the cut-offs for size selection (Fig. 5). This optional size selection step increases final library yield at the cost of requiring more input DNA. This size selected DNA is then used to prepare SMRTbell template library following PacBio's protocol. A second round of size selection is performed on the SMRTbell template using a 15-50 kb cutoff to remove the smaller

fragments generated during the SMRTbell library preparation step (Fig. 5). The second step ensures that DNA molecule smaller than 15kb are not sequenced in zero mode waveguides.

DNA Sequencing

PacBio sequencing was conducted to demonstrate length distributions (*D. simulans* Fig. 2a) and evaluate the impact of library preparation on quality (Fig. 3), and was performed at the UCI High Throughput Core Facility using DNA isolated using the protocol described above. We sequenced one SMRTcell of *Drosophila* genomic DNA with the following conditions to obtain sequences with standard quality and length distribution: 10:1 polymerase to template ratio, 250 pM template concentration. To demonstrate the tradeoff between yield and quality, we sequenced one SMRTcell each for polymerase:template ratios of 40:1,80:1,100:1 with template concentration held constant at 200pM, and one SMRTcell each with 300pM and 400pM template concentration with the polymerase:template ratio being held constant at 10:1.

Acknowledgement:

The authors would like to thank Stephen Richards for sharing the length distribution from *Drosophila pseudoobscura* Pacific Biosciences data and Sergey Koren and Brian Walenz for their assistance with wgs. We would also like to thank Melanie Oakes and Valentina Ciobanu for assistance with sequencing. This work was made possible, in part, through access to the Genomic High Throughput Facility Shared Resource of the Cancer Center Support Grant (CA-62203) at the University of California, Irvine and NIH shared instrumentation grants 1S10RR025496-01 and 1S10OD010794-01

References

Alkan C, Coe BP, Eichler EE 2011. Genome structural variation discovery and genotyping. Nat Rev Genet 12: 363-376. doi: 10.1038/nrg2958

Baker M 2012. De novo genome assembly: what every biologist should know. Nat Methods 9: 333-337. doi: 10.1038/nmeth.1935

Berlin K, Koren S, Chin CS, Drake JP, Landolin JM, Phillippy AM 2015. Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. Nat Biotechnol 33: 623-630. doi: 10.1038/nbt.3238

Bradnam KR, Fass JN, Alexandrov A, Baranay P, Bechner M, Birol I, Boisvert S, Chapman JA, Chapuis G, Chikhi R, Chitsaz H, Chou WC, Corbeil J, Del Fabbro C, Docking TR, Durbin R, Earl D, Emrich S, Fedotov P, Fonseca NA, Ganapathy G, Gibbs RA, Gnerre S, Godzaridis E, Goldstein S, Haimel M, Hall G, Haussler D, Hiatt JB, Ho IY, Howard J, Hunt M, Jackman SD, Jaffe DB, Jarvis ED, Jiang H, Kazakov S, Kersey PJ, Kitzman JO, Knight JR, Koren S, Lam TW, Lavenier D, Laviolette F, Li Y, Li Z, Liu B, Liu Y, Luo R, Maccallum I, Macmanes MD, Maillet N, Melnikov S, Naquin D, Ning Z, Otto TD, Paten B, Paulo OS, Phillippy AM, Pina-Martins F, Place M, Przybylski D, Qin X, Qu C, Ribeiro FJ, Richards S, Rokhsar DS, Ruby JG, Scalabrin S, Schatz MC, Schwartz DC, Sergushichev A, Sharpe T, Shaw TI, Shendure J, Shi Y, Simpson JT, Song H, Tsarev F, Vezzi F, Vicedomini R, Vieira BM, Wang J, Worley KC, Yin S, Yiu SM, Yuan J, Zhang G, Zhang H, Zhou S, Korf IF 2013. Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. Gigascience 2: 10. doi: 10.1186/2047-217X-2-10

Bresler G, Bresler M, Tse D 2013. Optimal assembly for high throughput shotgun sequencing. Bmc Bioinformatics 14 Suppl 5: S18. doi: 10.1186/1471-2105-14-S5-S18

Chaisson MJ, Tesler G 2012. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. Bmc Bioinformatics 13: 238. doi: 10.1186/1471-2105-13-238

Chin CS, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, Clum A, Copeland A, Huddleston J, Eichler EE, Turner SW, Korlach J 2013. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat Methods 10: 563-569. doi: 10.1038/nmeth.2474

Churchill GA, Waterman MS 1992. The accuracy of DNA sequences: estimating sequence quality. Genomics 14: 89-98.

Goodwin S, Gurtowski J, Ethe-Sayers S, Deshpande P, Schatz MC, McCombie WR 2015. Oxford Nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome. Genome Res. doi: 10.1101/gr.191395.115

Graham CA, Hill AJ 2001. Introduction to DNA sequencing. Methods Mol Biol 167: 1-12. doi: 10.1385/1-59259-113-2:001

Gurevich A, Saveliev V, Vyahhi N, Tesler G 2013. QUAST: quality assessment tool for genome assemblies. Bioinformatics 29: 1072-1075. doi: 10.1093/bioinformatics/btt086

Kajitani R, Toshimoto K, Noguchi H, Toyoda A, Ogura Y, Okuno M, Yabana M, Harada M, Nagayasu E, Maruyama H, Kohara Y, Fujiyama A, Hayashi T, Itoh T 2014. Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. Genome Res 24: 1384-1395. doi: 10.1101/gr.170720.113

Kim KE, Peluso P, Babayan P, Yeadon PJ, Yu C, Fisher WW, Chin CS, Rapicavoli NA, Rank DR, Li J, Catcheside DE, Celniker SE, Phillippy AM, Bergman CM, Landolin JM 2014. Long-read, whole-genome shotgun sequence data for five model organisms. Sci Data 1: 140045. doi: 10.1038/sdata.2014.45

Koren S, Harhay GP, Smith TP, Bono JL, Harhay DM, McVey SD, Radune D, Bergman NH, Phillippy AM 2013. Reducing assembly complexity of microbial genomes with single-molecule sequencing. Genome Biol 14: R101. doi: 10.1186/gb-2013-14-9-r101

Koren S, Phillippy AM 2015. One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly. Curr Opin Microbiol 23: 110-120. doi: 10.1016/j.mib.2014.11.014

Koren S, Schatz MC, Walenz BP, Martin J, Howard JT, Ganapathy G, Wang Z, Rasko DA, McCombie WR, Jarvis ED, Adam MP 2012. Hybrid error correction and de novo assembly of single-molecule sequencing reads. Nat Biotechnol 30: 693-700. doi: 10.1038/nbt.2280

Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL 2004. Versatile and open software for comparing large genomes. Genome Biol 5: R12. doi: 10.1186/gb-2004-5-2-r12

Lam K-K, Khalak A, Tse D 2014. Near-optimal assembly for shotgun sequencing with noisy reads. Bmc Bioinformatics 15. doi: 10.1186/1471-2105-15-s9-s4
Lander ES, Waterman MS 1988. Genomic mapping by fingerprinting random clones: a mathematical analysis. Genomics 2: 231-239.

Langley CH, Stevens K, Cardeno C, Lee YC, Schrider DR, Pool JE, Langley SA, Suarez C, Corbett-Detig RB, Kolaczkowski B, Fang S, Nista PM, Holloway AK, Kern AD, Dewey CN, Song YS, Hahn MW, Begun DJ 2012. Genomic variation in natural populations of Drosophila melanogaster. Genetics 192: 533-598. doi: 10.1534/genetics.112.142018

Lee H, Gurtowski J, Yoo S, Marcus S, McCombie WR, Schatz M 2014. Error correction and assembly complexity of single molecule sequencing reads. bioRxiv.
Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y, Tang J, Wu G, Zhang H, Shi Y, Liu Y, Yu C, Wang B, Lu Y, Han C, Cheung DW, Yiu SM, Peng S,

Xiaoqian Z, Liu G, Liao X, Li Y, Yang H, Wang J, Lam TW, Wang J 2012. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. Gigascience 1: 18. doi: 10.1186/2047-217X-1-18

Miller JR, Koren S, Sutton G 2010. Assembly algorithms for next-generation sequencing data. Genomics 95: 315-327. doi: 10.1016/j.ygeno.2010.03.001

Motahari A, Ramchandran K, Tse D, Ma N, Ieee 2013. Optimal DNA shotgun sequencing: Noisy reads are as good as noiseless reads. 2013 Ieee International Symposium on Information Theory Proceedings (Isit): 1640-1644.

Myers EW 1995. Toward simplifying and accurately formulating fragment assembly. J Comput Biol 2: 275-290.

Myers EW, Sutton GG, Delcher AL, Dew IM, Fasulo DP, Flanigan MJ, Kravitz SA, Mobarry CM, Reinert KH, Remington KA, Anson EL, Bolanos RA, Chou HH, Jordan CM, Halpern AL, Lonardi S, Beasley EM, Brandon RC, Chen L, Dunn PJ, Lai Z, Liang Y, Nusskern DR, Zhan M, Zhang Q, Zheng X, Rubin GM, Adams MD, Venter JC 2000. A whole-genome assembly of Drosophila. Science 287: 2196-2204.

Pendleton M, Sebra R, Pang AW, Ummat A, Franzen O, Rausch T, Stutz AM, Stedman W, Anantharaman T, Hastie A, Dai H, Fritz MH, Cao H, Cohain A, Deikus G, Durrett RE, Blanchard SC, Altman R, Chin CS, Guo Y, Paxinos EE, Korbel JO, Darnell RB, McCombie WR, Kwok PY, Mason CE, Schadt EE, Bashir A 2015. Assembly and diploid architecture of an individual human genome via single-molecule technologies. Nat Methods 12: 780-786. doi: 10.1038/nmeth.3454

Sakai H, Naito K, Ogiso-Tanaka E, Takahashi Y, Iseki K, Muto C, Satou K, Teruya K, Shiroma A, Shimoji M, Hirano T, Itoh T, Kaga A, Tomooka N 2015. The power of single molecule real-time sequencing technology in the de novo assembly of a eukaryotic genome. bioRxiv.

Salmela L, Rivals E 2014. LoRDEC: accurate and efficient long read error correction. Bioinformatics 30: 3506-3514. doi: 10.1093/bioinformatics/btu538
Shomorony I, Courtade T, Tse D. 2015. Do Read Errors Matter for Genome Assembly? ArXiv e-prints.

Simpson JT, Pop M 2015. The Theory and Practice of Genome Sequence Assembly. Annu Rev Genomics Hum Genet. doi: 10.1146/annurev-genom-090314-050032

Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I 2009. ABySS: a parallel assembler for short read sequence data. Genome Res 19: 1117-1123. doi: 10.1101/gr.089532.108
Treangen TJ, Sommer DD, Angly FE, Koren S, Pop M 2011. Next generation sequence assembly with AMOS. Curr Protoc Bioinformatics Chapter 11: Unit 11 18. doi: 10.1002/0471250953.bi1108s33

Ye C, Hill C, Ruan J, Zhanshan, Ma. 2014. DBG2OLC: Efficient Assembly of Large Genomes Using the Compressed Overlap Graph. ArXiv e-prints.

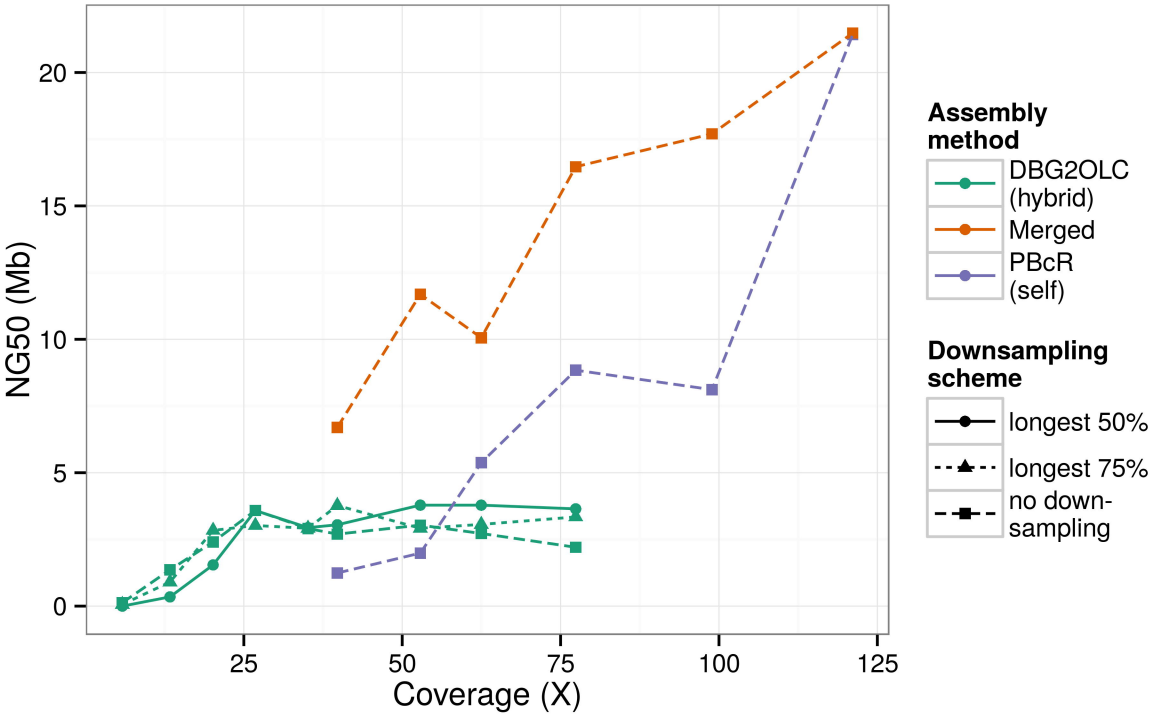
Figure 1. The NG50 of *D. melanogaster* assemblies produced using a variety of data sets. NG50 here is the contig size such that at least half of the 130Mb *D. melanogaster* genome (65Mb) is contained in contigs of that size or larger. "50% longest" and "75% longest", respectively, refer to datasets in which only the longest 50% or 75% of the available reads have been used. The coverage listed on the x-axis in this case refers to the total amount of available data (before downsampling). "ISO1 to Pseudo by removal" refers, to the downsampling scheme in which data was removed from the ISO1 dataset to cause the read lengths of the ISO1 data to resemble read lengths in the publically available *D. pseudoobscura* by removing reads differentially based upon length.

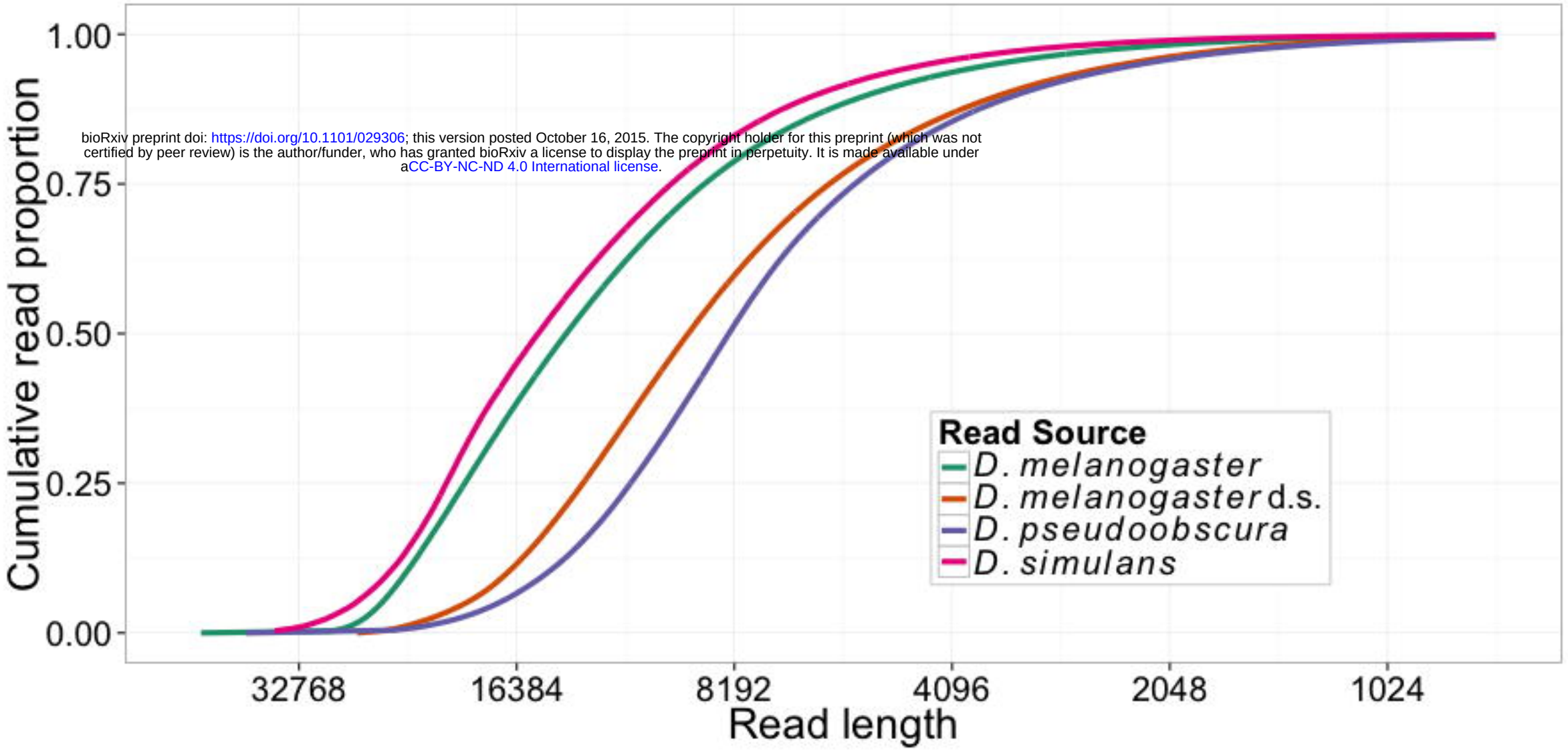
Figure 2. (a) The cumulative read length of various data sets, where *D. melanogaster* refers to the original ISO1 data set, *D. pseusoobscura* refers to a publicly available *D. pseudoobscura* dataset with a shorter average read length, *D. melanogaster* d.s. refers to the *D. melanogaster* data, downsampled to have read lengths resembling the *D. pseusoobscura* dataset, and *D. simulans* is a *D. simulans* dataset sequenced using our DNA preparation technique. (b) A plot of NG50 versus coverage of hybrid assemblies, as in Figure 1. This plot depicts the effect of reduced read length on NG50, while holding read quality and coverage constant. (c) Cumulative contig length distribution of 20 SMRTcells PB only assemblies created with the original ISO1 reads and the downsampled reads. Contig lengths in the shorter/downsampled reads assembly are considerably shorter than the contigs in the original reads assembly.

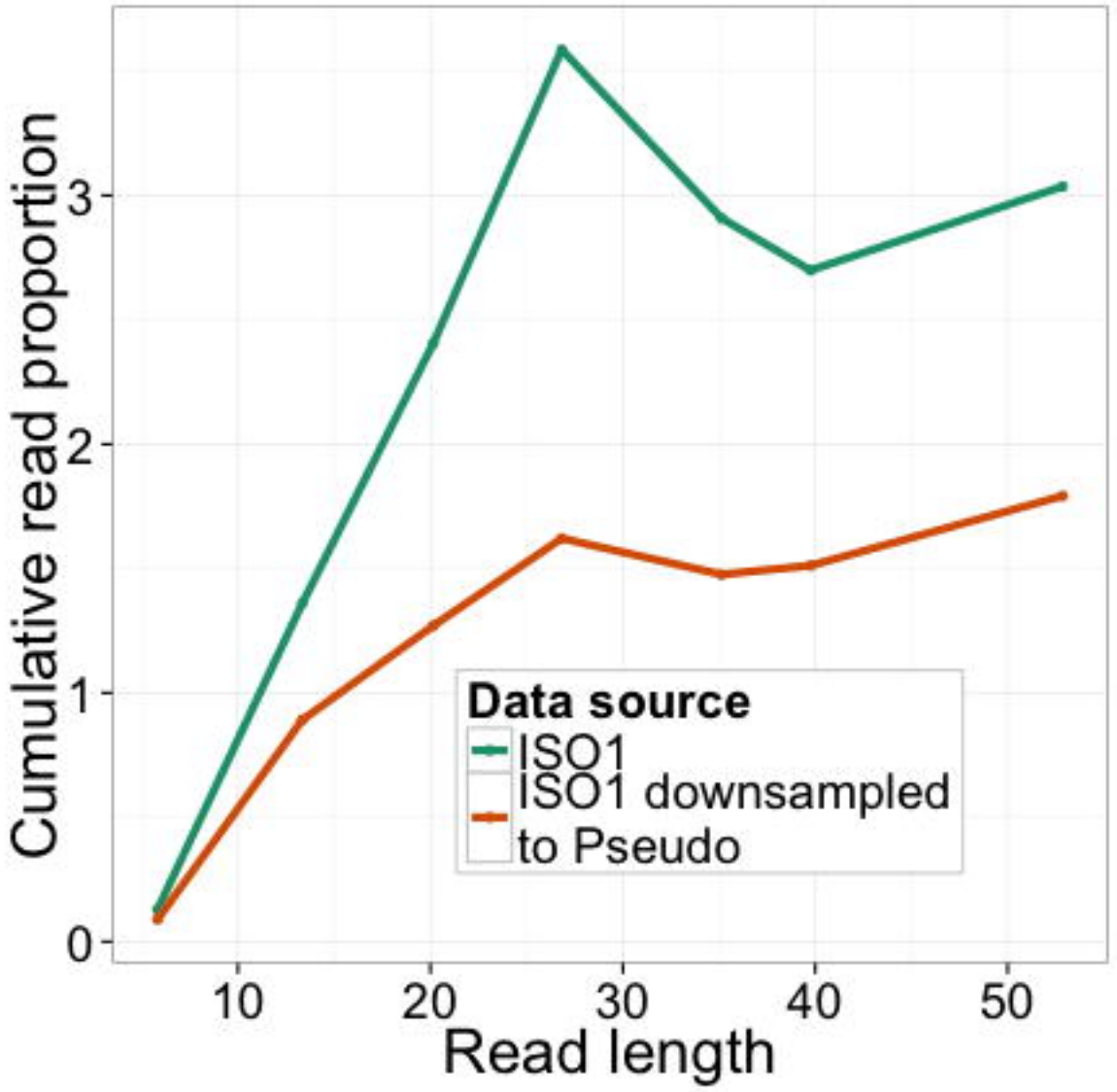
Figure 3. The distribution of read quality in sequencing runs performed at the UCI genomics core using our DNA preparation technique. "P" here refers to polymerase loading during sequencing (the proportion of polymerase to template, where 10 would indicate a 10:1 ratio of polymerase to template), while "T" refers to template loading concentration during sequencing (in picomolarity).

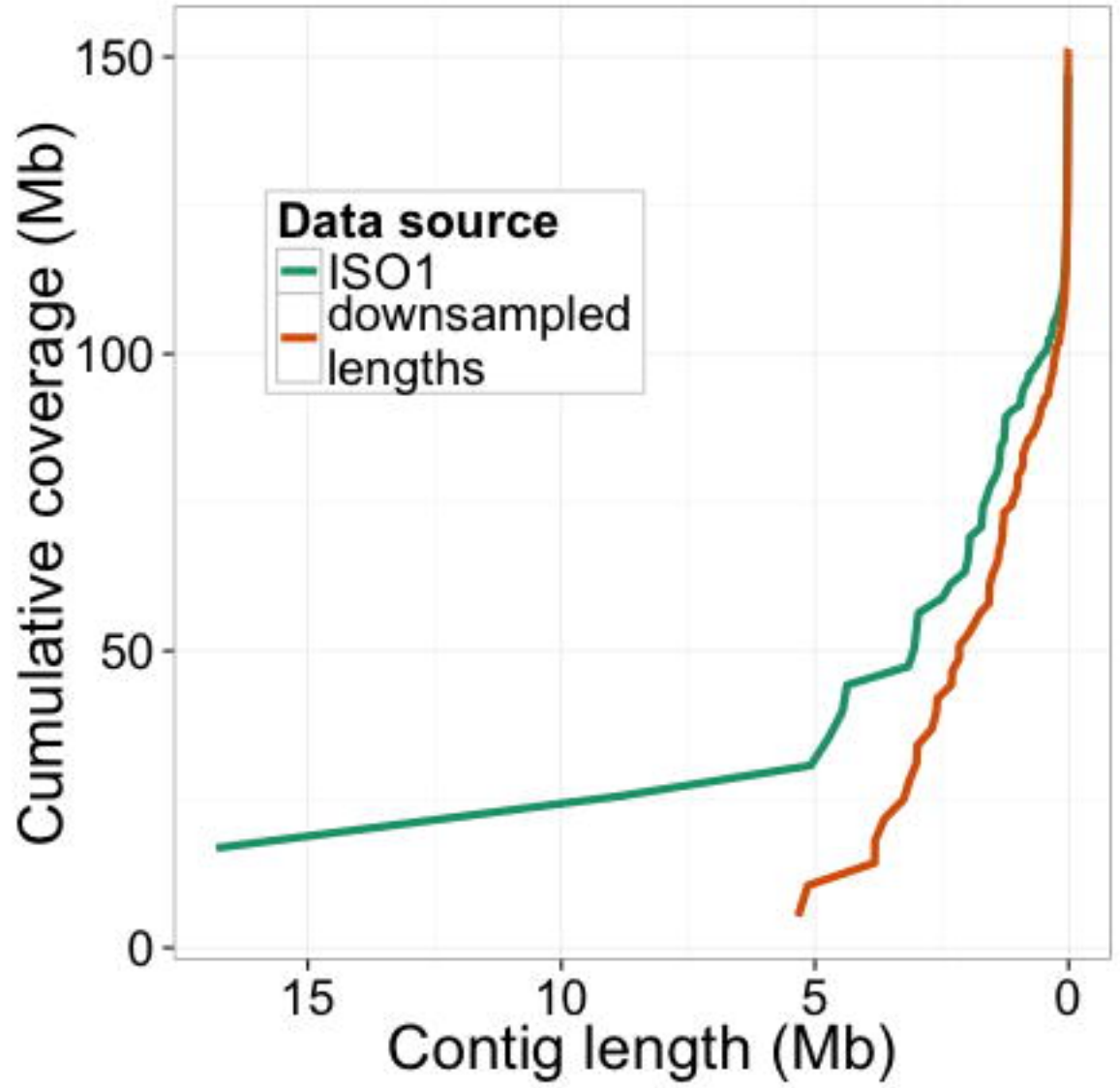
Figure 4. As in Figure 2a, a plot of cumulative length distribution. These curves represent the cumulative length distribution of final assemblies using low, medium, and high quality selected reads using either PB only assembly or hybrid assembly.

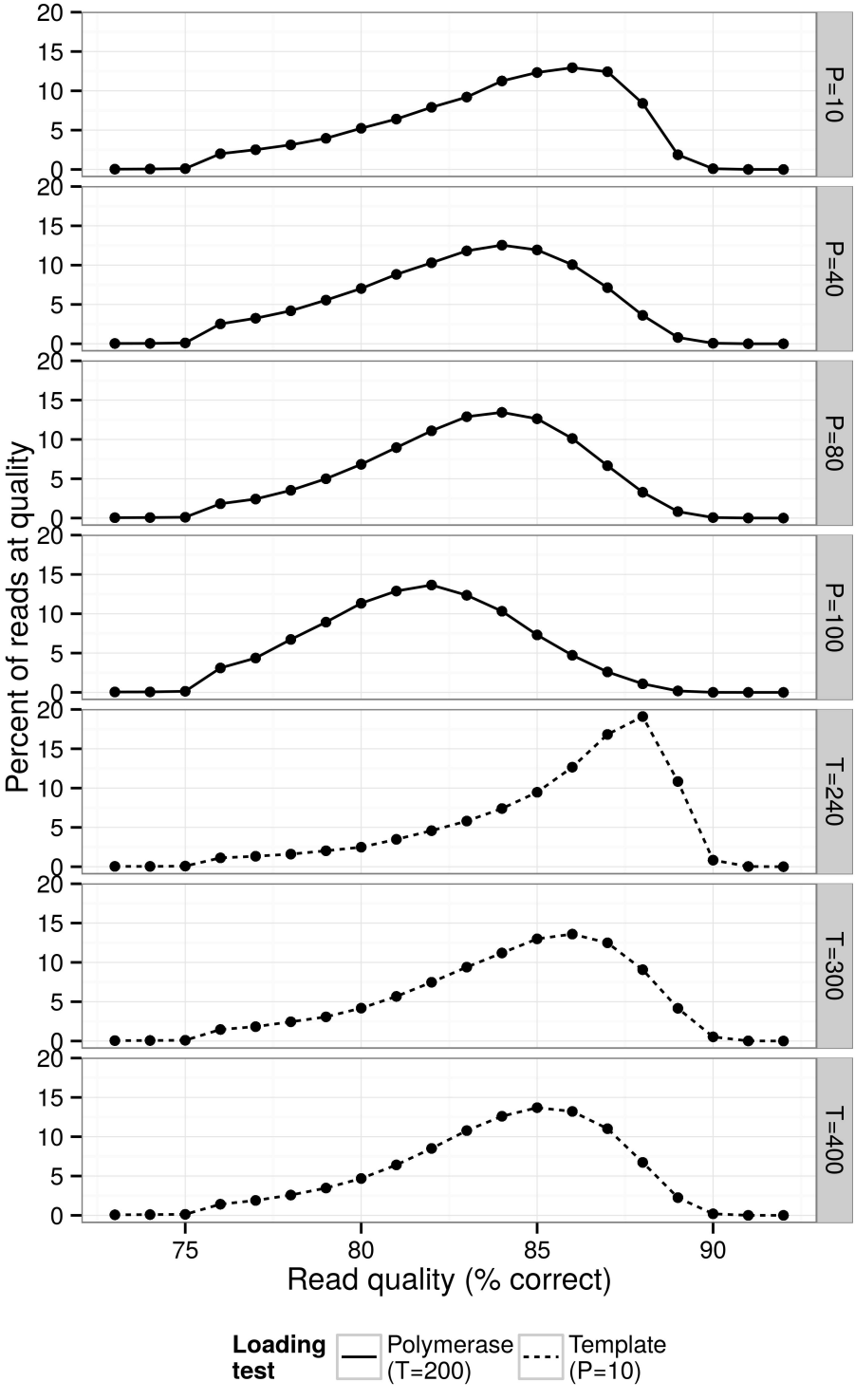
Figure 5: An example of correctly extracted and sheared DNA visualized using field inversion gel electrophoresis. The ladder is the NEB low range PFG marker (no longer produced). The lanes of the gel are as follows: ladder, unsheared DNA, DNA sheared with a 24 gauge needle, sheared DNA size selected with 15-50kb cut-off, SMRTbell template library after 15-50kb size selection. From the gel, it is evident that there is a minimal 'tail' of DNA below ~15kb, the preferred size selection minimum.

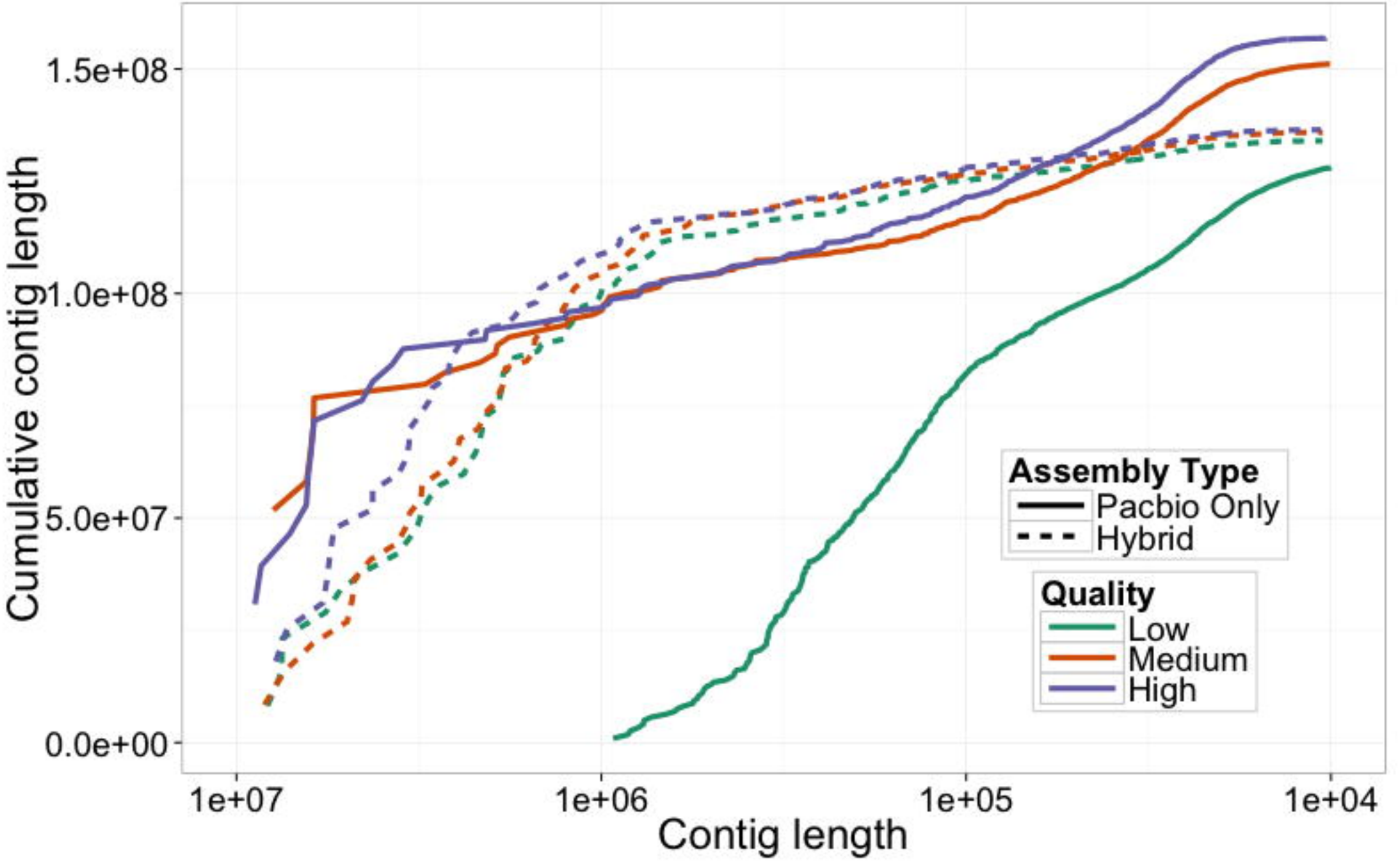

NG50 (Mb)
20
15
10
5
0
25
50
75
100
125
Coverage (X)
Assembly method
DBG2OLC (hybrid)
Merged
PBcR (self)
Downsampling scheme
longest 50%
longest 75%
no down-sampling








P=10
P=40
P=80
P=100
T=240
T=300
T=400
Percent of reads at quality
0
5
10
15
20
75
80
85
90
Read quality (% correct)
Loading test
Polymerase (T=200)
Template (P=10)


1.5e+08
1.0e+08
5.0e+07
0.0e+00
Cumulative contig length
1e+07
1e+06
1e+05
1e+04
Contig length
Assembly Type
Pacbio Only
Hybrid
Quality
Low
Medium
High