GigaScience

Assessment of human diploid genome assembly with 10x Linked-Reads data

--Manuscript Draft--

Manuscript Number:	GIGA-D-19-00092
Full Title:	Assessment of human diploid genome assembly with 10x Linked-Reads data
Article Type:	Data Note
Funding Information:	
Abstract:	Background: Producing cost-effective haplotype-resolved personal genomes remains challenging. 10x Linked-Read sequencing, with its high base quality and long-range information, has been demonstrated to facilitate de novo assembly of human genomes and variant detection. In this study, we investigate in depth how the parameter space of 10x library preparation and sequencing affects assembly quality, on the basis of both simulated and real libraries. Findings: We prepared and sequenced eight 10x libraries with a diverse set of parameters from standard cell lines NA12878 and NA24385 and performed whole genome assembly on the data. We also developed the simulator LRTK-SIM to follow the workflow of 10x data generation and produce realistic simulated Linked-Read data sets. We found that assembly quality could be improved by increasing the total sequencing coverage (C) and keeping physical coverage of DNA fragments (CF) or read coverage per fragment (CR) within broad ranges. The optimal physical coverage was between 332X and 823X and assembly quality worsened if it increased to greater than 1,000X for a given C. Long DNA fragments could significantly extend phase blocks, but decreased contig contiguity. The optimal length-weighted fragment length (Wμ_FL) was around 50 – 150kb. When broadly optimal parameters were used for library preparation and sequencing, ca. 80% of the genome was assembled in a diploid state. Conclusion: The Linked-Read libraries we generated and the parameter space we identified provide theoretical considerations and practical guidelines for personal genome assemblies based on 10x Linked-Read sequencing. Keywords: 10x Linked-Read sequencing, de novo assembly, diploid human genome, library preparation
Corresponding Author:	arend sidow UNITED STATES
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	
Corresponding Author's Secondary Institution:	
First Author:	Lu Zhang
First Author Secondary Information:	
Order of Authors:	Lu Zhang
	Xin Zhou
	Ziming Weng
	arend sidow
Order of Authors Secondary Information:	
Additional Information:	
Question	**Response**
Are you submitting this manuscript to a special series or article collection?	No

Experimental design and statistics Full details of the experimental design and statistical methods used should be given in the Methods section, as detailed in our Minimum Standards Reporting Checklist. Information essential to interpreting the data presented should be made available in the figure legends. Have you included all the information requested in your manuscript?	Yes
Resources A description of all resources used, including antibodies, cell lines, animals and software tools, with enough information to allow them to be uniquely identified, should be included in the Methods section. Authors are strongly encouraged to cite Research Resource Identifiers (RRIDs) for antibodies, model organisms and tools, where possible. Have you included the information requested as detailed in our Minimum Standards Reporting Checklist?	Yes
Availability of data and materials All datasets and code on which the conclusions of the paper rely must be either included in your submission or deposited in publicly available repositories (where available and ethically appropriate), referencing such data using a unique identifier in the references and in the "Availability of Data and Materials" section of your manuscript. Have you have met the above requirement as detailed in our Minimum Standards Reporting Checklist?	Yes

Assessment of human diploid genome assembly with 10x Linked-Reads data

Lu Zhang[1,2,3, *], Xin Zhou[3, *], Ziming Weng[2], Arend Sidow[2,4,†]

[1]Department of Computer Science, Hong Kong Baptist University

[2]Department of Pathology, Stanford University

[3]Department of Computer Science, Stanford University

[4]Department of Genetics, Stanford University

*These authors contributed equally to this work. †Correspondence and requests for materials should be addressed to Arend Sidow (email: arend@stanford.edu)

Abstract

Background: Producing cost-effective haplotype-resolved personal genomes remains challenging. 10x Linked-Read sequencing, with its high base quality and long-range information, has been demonstrated to facilitate *de novo* assembly of human genomes and variant detection. In this study, we investigate in depth how the parameter space of 10x library preparation and sequencing affects assembly quality, on the basis of both simulated and real libraries.

Findings: We prepared and sequenced eight 10x libraries with a diverse set of parameters from standard cell lines NA12878 and NA24385 and performed whole genome assembly on the data. We also developed the simulator LRTK-SIM to follow the workflow of 10x data generation and produce realistic simulated Linked-Read data sets. We found that assembly quality could be improved by increasing the total sequencing coverage (*C*) and keeping physical coverage of DNA fragments (C_F) or read coverage per fragment (C_R) within broad ranges. The optimal physical coverage was between 332X and 823X and assembly quality worsened if it increased to greater than 1,000X for a given *C*. Long DNA fragments could significantly extend phase blocks, but decreased contig contiguity. The optimal length-weighted fragment length ($W\mu_{FL}$) was around 50 – 150kb. When broadly optimal parameters were used for library preparation and sequencing, ca. 80% of the genome was assembled in a diploid state.

Conclusion: The Linked-Read libraries we generated and the parameter space we identified provide theoretical considerations and practical guidelines for personal genome assemblies based on 10x Linked-Read sequencing.

Keywords: 10x Linked-Read sequencing, *de novo* assembly, diploid human genome, library preparation

Data description

Introduction

The human genome holds the key for understanding the genetic basis of human evolution, hereditary illnesses and many phenotypes. Whole-genome reconstruction and variant discovery, accomplished by analysis of data from whole-genome sequencing experiments, are foundational for the study of human genomic variation and analysis of genotype-phenotype relationships. Over the past decades, cost-effective whole-genome sequencing has been revolutionized by short-fragment approaches, the most widespread of which have been the consistently improving generations of the original Solexa technology [1, 2], now referred to as Illumina sequencing. Illumina's strengths and weaknesses are inherent in the sample prep and sequencing chemistry. Illumina generates short paired reads (2x150 base pairs for the highest-throughput platforms) from short fragments (usually 400-500 base pairs) [3]. Because many clonally amplified molecules generate a robust signal during the sequencing reaction, Illumina's average per-base error rates are very low.

The lack of long-range contiguity between end-sequenced short fragments limits their application for reconstructing personal genomes. Long-range contiguity is important for phasing variants and dealing with genomic complex regions. For haplotyping, variants can be phased by population-based methods [4, 5] or family-based recombination inference [6, 7], but such approaches are only feasible for common variants or large pedigrees. Furthermore, highly polymorphic regions such as the HLA in which the reference sequence does not adequately capture the diversity segregating in the population are refractory to mapping-based approaches and require *de novo* assembly to reconstruct [8]. Short-read/short-fragment data are challenged by interspersed repetitive sequences from mobile elements and by segmental duplications, and only support highly fragmented genome reconstruction [9, 10].

In principle, many of these challenges can be overcome by long-read/long-fragment sequencing [11, 12]. Assembly of Pacific Biosciences (PacBio) or Oxford Nanopore (ONT) data can yield impressive contiguity of contigs and scaffolds. In one study, scaffold N50 reached 31.1Mb by hierarchically integrating PacBio long reads and BioNano for a hybrid assembly, which also uncovered novel tandem repeats and replicated the structural variants that were newly included in the updated hg38 human reference sequence. Another study [13] produced human genome assemblies with ONT data, in which a contig N50 ~3Mb was achieved, and long contigs covered all class I HLA regions. However, long-fragment sequencing suffers from low throughput, high cost, and low base quality, hampering its usefulness for personal genome assembly.

Hierarchical assembly pipelines in which multiple data types are used [14] as another approach. For example, in the reconstruction of an Asian personal genome, fosmid clone pools and Illumina data were merged, but because fosmid libraries are highly labor intensive to generate and sequence, this approach is not generalizable to personal genomes. The "Long Fragment Read" (LFR) approach, where a long fragment is sequenced at high depth via single-molecule fragmented amplification, reported promising personal genome assembly and variant phasing by attaching a barcode to the short reads derived from the same long fragment. However, because LFR is implemented in a 384 well plate, many long fragments would be labelled by the same barcodes, making it difficult for binning short-reads, and the great sequencing depth required rendered LFR not cost-effective.

An alternative approach is offered by the 10x Genomics Chromium system, which distributes the DNA prep into millions of partitions where partition-specific barcode sequences are attached to short amplification products that are templated off the input fragments. Because of the limited reaction efficiency in each partition, the sequencing depth for each fragment is too shallow to reconstruct the original long-fragment, distinguishing this approach from LFR [15]. However, to compensate for the low read coverage of each fragment, each genomic region is covered by hundreds of DNA fragments, giving overall sequence coverage that is in a range comparable to standard Illumina short-fragment sequencing while providing very high physical coverage. Novel computational approaches leveraging the special characteristics of 10x Genomics data have already generated significant advances in power and accuracy of haplotyping [16, 17], cancer genome reconstruction [18, 19], metagenomic assemblies [20] , and *de novo* assembly of human and other genomes [21-23], compared to standard Illumina short-fragment sequencing. While the uniformity of sequence coverage is not as good as with PCR-free Illumina libraries, 10x Linked-Read sequencing is a promising technology that combines low per-base error and good small-variant discovery with long-range information for much improved SV detection in mapping-based approaches [19, 24], and the possibility of long-range contiguity in *de novo* assembly [21, 23, 25].

Practical advantages of the technology include the low DNA input mass requirement (1ng per library, or approximately 300 haploid human genome equivalents). Real input quantities can vary, along with other factors, to influence an interconnected array of parameters that are relevant to genome assembly and reconstruction. The parameters over which the experimenter has influence are (**Figure 1**): i). C_R: average **C**overage of short **R**eads per fragment; ii). C_F: average physical **C**overage of the genome by long DNA **F**ragments; iii). $N_{F/P}$: **N**umber of **F**ragments per **P**artition; iv). Fragment length distribution, several parameters of which are used, specifically μ_{FL}: Average Unweighted DNA **F**ragment **L**ength and Wμ_{FL}: Length-**W**eighted average of DNA **F**ragment **L**ength. Note that several parameters depend on each other. For example, a greater amount of input DNA will increase $N_{F/P}$; shorter fragments increase $N_{F/P}$ at the same DNA input amount compared to longer fragments; less input DNA will (within practical constraints) increase C_R and decrease C_F; and their absolute values are set by how much total sequence coverage is generated because C_R x C_F = C.

Our goal in this study was to experimentally explore the 10x parameter space and evaluate the quality of *de novo* diploid assembly as a function of the parameter values. For example, we set out to ask whether longer input fragments produce better assemblies, or what the effect of sequencing vs. physical coverage is on contiguity of assembly. In order to constrain the parameter space, we first performed computer simulations with reasonably realistic synthetic data. The simulation results suggested certain parameter combinations that we then approximated in the generation of real, high-depth, sequence data on two human reference genome cell lines, NA12878 and NA24385. These simulated and real data sets were then used to produce *de novo* assemblies, with an emphasis on the performance of 10x's Supernova2 [21]. We finally assessed the quality of the assemblies using standard metrics of contiguity and accuracy, facilitated by the existence of a gold standard (in the case of simulations) and comparisons to the reference genome (in the case of real data).

Library preparation, physical parameters and sequencing coverage

We made six DNA preparations that varied in fragment size distribution and amount of input DNA, three each from NA12878 and NA24385. From these, we prepared eight libraries, five from NA12878 and three from NA24385 (**Table S1**). To generate library L_{1L}, L_{1M} and L_{1H},, genomic DNA was extracted from ca. 1 million cultured NA12878 cells using the Gentra Puregene Blood Kit following manufacturer's instructions (Qiagen, Cat. No 158467). The GEMs were divided into 3 tubes with 5%, 20%, and 75% to generate libraries L_{1L}, L_{1M} and L_{1H}, respectively (**Figure S1-S3**). For the other libraries, to generate longer DNA fragments (Wμ_{FL}=150kb and longer, **Figure S4-S8**), a modified protocol was applied. Two-hundred thousand NA12878 or NA24385 cells of fresh culture were added to 1mL cold 1x PBS in a 1.5 ml tube and pelleted for 5 minutes at 300g. The cell pellets were completely resuspended in the residual supernatant by vortexing and then lysed by adding 200ul Cell Lysis Solution and 1ul of RNaseA Solution (Qiagen, Cat. No 158467), mixing by gentle inversion, and incubating at 37°C for 15-30 minutes. This cell lysis solution is used immediately as input for the 10x Chromium prep (ChromiumTM Genome Library & Gel Bead Kit v2, PN-120258; ChromiumTM i7 Multiplex Kit, PN-120262). Fragment size of the input DNA can be controlled by gentle handling during lysis and DNA preparation for Chromium. The amount of input DNA (between 1.25 and 4 ng) was varied to achieve a wide range of physical coverage (C_F).The Chromium Controller was operated and the GEM prep was performed as instructed by the manufacturer. Individual libraries were then constructed by end repairing, A-tailing, adapter ligation and PCR amplification. All libraries were sequenced with three lanes of paired-end 150bp runs on the Illumina HiSeqX to obtain very high coverage (C=94x-192x), though the two with the fewest number of gel beads (L_{1L} and L_{1M}) exhibited high PCR duplication rates because of the reduced complexity of the libraries (**Table S1**).

Linked-Reads subsampling

The high sequencing coverage in the libraries allowed subsampling to facilitate the matching of parameters among the different libraries, for purposes of comparability; these subsampled Linked-Read sets are denoted R_{id} (**Figure 1**). We aligned the 10x Linked-Reads to human reference genome (hg38) followed by removing PCR duplication by barcode-aware analysis in Long Ranger[18]. Original input DNA fragments were inferred by collecting the read-pairs with the same barcode that were aligned in proximity to each other. A fragment was terminated if the distance between two consecutive reads with the identical barcode larger than 50kb. Fragments were required to have at least two read pairs with the same barcode and a length of at least 2 kb. Partitions with fewer than three fragments were removed. We subsampled short-reads for each fragment to satisfy the expected C_R.

Generating 10x simulated libraries by LRTK-SIM

To compare the observations from real data with a known truth set, we developed LRTK-SIM, a simulator that follows the workflow of the 10x Chromium system and generates synthetic Linked-Reads like those produced by an Illumina HiSeqX machine (**Supplementary Information** and **Figure S9**). Based on the parameters commonly employed by 10x Genomics Linked-Read sequencing and the characteristics of our libraries, LRTK-SIM generated simulated datasets from the human reference (hg38), explicitly modeling the five key steps in real data generation. Parameters in parentheses are from the standard 10x Genomics protocol: 1. Shearing genomic DNA into long fragments (Wμ_{FL} from 50kb to 100kb); 2. Loading DNA to the 10x Chromium instrument (~1.25ng DNA); 3. Allocating DNA fragments into partitions which are attached the unique barcodes (~10 fragments per partition); 4. Generating short fragments; 5. Generating Illumina paired-end short reads (800M~1200M reads). LRTK-SIM first generated a diploid reference genome as a template by duplicating the human reference genome (hg38) into two haplotypes and inserting SNVs from high-confidence regions in GIAB of NA12878; For low-confidence regions we randomly simulated 1 SNV per 1 kb. The ratio was 2:1 for heterozygous and homozygous SNVs. From this diploid reference genome, LRTK-SIM generated long DNA fragments by randomly shearing each haplotype with multiple copies into pieces whose lengths were sampled from an exponential distribution with mean of μ_{FL}. These fragments were then allocated to pseudo-partitions, and all the fragments within each partition were assigned the same barcode. The number of fragments for each partition was randomly picked from a Poisson distribution with mean of $N_{F/P}$. Finally, paired-end short reads were generated according to C_R and replaced the first 16bp of the reads from forward strand to the assigned barcodes followed by 7 Ns. More information about implementation can be found in **Supplementary Information**. From that diploid genome, Linked-Read datasets were generated that varied in C_R, C_F and μ_{FL} (Wμ_{FL}) (**Table S2**-**S3**). Varying $N_{F/P}$ was only done for chromosome 19 because of the infeasibility of running Supernova2 on whole genome assemblies with large $N_{F/P}$; within practically reasonable values, $N_{F/P}$ does not appear to influence assembly quality (**Figure S10**). In total, we generated 17 simulated Linked-Read datasets to explore the overall parameter space (**Table S2**-**S3**) and 11 to match the parameters of the abovementioned real libraries (**Figure 1**).

Human genome diploid assembly and evaluation

The scaffolds were generated by the “pseudohap2” output of Supernova2, which explicitly generated two haploid scaffolds, simultaneously. Contigs were generated by breaking the scaffolds if at least 10 consecutive ‘N’s appeared, per definition by Supernova2. For the simulations of human chromosome 19, we used the scaffolds from the “megabubbles” output. Contig and scaffold N50 and NA50 were used to evaluate assembly quality. Contigs longer than 500bp were aligned to hg38 by Minimap2[30]. We calculated contig NA50 on the basis of contig misassemblies reported by QUAST-LG [31]. For scaffolds (longer than 1kb), we calculated the NA50 following Assemblathon 1's procedure [32] (**Supplementary Information**).

Performance of diploid assembly: influence of total coverage Diploid assembly by Linked-Reads requires sufficient total read coverage (C=C_R×C_F) to generate long contigs and scaffolds. In this experiment, to explore the roles of both physical coverage (C_F) and per-fragment read coverage (C_R), we first generated eight simulated libraries whose total coverage C ranged from 16x to 78x: four with C_R fixed and increasing C_F and four with fixed C_F, and increasing C_R (**Table S2**). Contig and scaffold N50s increased along with increasing either C_F or C_R (**Figure 2A** and **2B**). To investigate whether the trend was also present in the real datasets, we analyzed six real libraries (three by varying C_F, and the other three by varying C_R; **Figure 1**): as C increased, we varied C_F and C_R independently by fixing the other parameter. Contig and scaffold N50s also increased in in these real Linked-Read sets (**Figure 2C**, **2D**, **2E** and **2F**) as a function of total coverage *C*. Contig lengths did increase but not dramatically so when *C* was increased beyond 56X. Accuracy, which we define as the ratio between NA50 (N50 after breaking contigs or scaffolds at assembly errors) and N50 (**Figure 2C** and **2E**), did not change appreciably. For scaffolds in the real data sets, when *C* increased from 48X (R_3) to 67X (R_4), both scaffold N50 and NA50 were significantly improved (N50: 13.4Mb to 30.6Mb; NA50: 6.3Mb to 12.0Mb), but the accuracy dropped slightly from 46.6% to 39.1%, which indicated that scaffold accuracy may be refractory to extremely high *C* (**Figure 2F**). These results indicated that assembly length and accuracy were comparable over a broad range of C_F and C_R at constant *C*, which implied that assembly quality was mainly determined by *C*.

Performance of diploid assembly: influence of fragment length and physical coverage. To investigate if input weighted fragment length (as measured by Wμ_{FL}) influenced assembly quality, we generated four simulated libraries (**Table S3**) with fixed C_F and C_R and a range of fragment lengths (**Figure 3A**). Contig length decreased with increasing fragment length, a trend that was also seen in six real libraries (**Figure 3B;** *C*=56X; R_6 to R_{11}in **Figure 1**). We then simulated another six libraries with the same parameters as the real ones to explore the effects of physical coverage at constant *C*=56x (**Figure 3C**). Contig lengths decreased as a function of increasing physical coverage, a trend that is somewhat less clear in real data possibly due to confounding other parameters such as fragment length (**Figure 3D**).

Performance of diploid assembly: nature of the source genome. Assembly errors may occur because of heterozygosity, repetitive sequences, or sequencing error. To illuminate possible sources of assembly error, we performed simulations by generating 10x-like Linked-Reads as above from human chromosome 19, and then quantified assembly error against these synthetic gold standards. Removal of interspersed repeat sequences from the source genome resulted in better contigs with no loss of accuracy in experiments by varying C_F, C_R and μ_{FL}(**Figure 4A**, **4C** and **4E**) and better scaffolds only if C_R was above 1X (**Figure 4D**). Removal of variation had little effect on contigs and only gave rise to longer scaffolds if C_R was above 0.8X (**Figure S11**), which is difficult to achieve with real libraries. Finally, a 1% uniform sequencing error had no discernible effect (**Figure S12**).

<u>Performance of diploid assembly: fraction of genome in diploid state.</u> While contiguity is an important parameter for any whole genome assembly, evaluation of diploid assemblies necessitates estimating the fraction of the genome in which the assembly recovered the diploid state. To this end, we divided the contigs generated by Supernova2 into “diploid contigs”, which were extracted from its megabubble structures, and “haploid contigs” from non-megabubble structures. Pairs of scaffolds were extracted as the two haplotypes from megabubble structures if they shared the same start and end nodes in the assembly graph. Diploid contigs were generated by breaking the candidate scaffolds at the sequences with least 10 consecutive ‘N’s and were aligned to human reference genome (hg38) by Minimap2. The genome was split into 500bp windows and diploid regions were defined as the maximum extent of successive windows covered by two contigs, each from one haplotype._Alignment against the human reference genome revealed the overall genome coverages of the six assemblies to be around 91%. For most assemblies, 70%-80% of the genome was covered by two homologous contigs (**Table 1**), with R_6 only reaching 58.9%, probably due to the short fragments of the DNA prep (μ_{FL}=24kb). We also analyzed another seven assemblies produced by 10x Genomics, all of which had diploid fractions of about 80% as well (**Table S4**). In the male NA24385, non-pseudoautosomal regions of the X chromosome are hemizygous and should therefore be recovered as haploid regions. Between 79.9% and 87.6% of these regions were covered by one contig exactly depending on the assembled library. Library construction parameters other than fragment length appeared to have had little impact on the proportion of diploid regions (**Tables 1** and **Table S4**).

Overlapping the diploid regions from the assemblies of the same individual revealed that 50.24% and 67.27% of the genome for NA12878 and NA24385 (**Figure S13)**, respectively, were diploid in all the three assemblies. NA12878 was lower because of the low percentage of diploid regions in assembly R_6 (**Table 1)**. The overlaps were significantly greater than expected by chance (NA12878: 33.3%, p-value=0.0049; NA24385: 45.4%, p-value=0.0029. Chi square test). These observations were consistent with heterozygous variants being enriched in certain genomic segments, in which two haplotypes were more easily differentiated by Supernova2. Phase block lengths were mainly determined by total coverage *C* and increased in real data with increasing fragment length (**Figure S14**).

Discussion

In this study, we investigated human diploid assembly using 10x Linked-Read sequencing data on both simulated and real libraries. We developed the simulator LRTK-SIM to examine the likely impact of parameters in diploid assembly and compared results from simulated reads to those from real libraries. We thus determined the impact of key parameters (C_R, C_F, $N_{F/P}$ and μ_{FL}/Wμ_{FL}) with respect to assembly continuity and accuracy. Our study provides a general strategy to evaluate assemblies of 10x data and may have implications for the evaluation of other barcode-based sequencing technologies such as CPTv2-seq [26] or stLRF [27] in the future when they become commercially available.

10x Practicalities

For standard Illumina sequencing, library complexity is usually sufficient to generate tremendous numbers of reads from unique templates and read coverage can be increased simply by sequencing more. However, the 10x Chromium system performs amplification in each partition, and generally only about 20% to 40% of the original long fragment sequence can be captured as short fragments and eventually as reads, resulting in shallow sequencing coverage per fragment. Sequencing more deeply does not increase the per-fragment coverage much as most of the extra reads are from PCR duplicates. The solution is to sequence multiple 10x libraries constructed from the same DNA prep and merge them for analysis. This means that C_R remains in the standard range where PCR duplicates are relatively rare, but C_F increases proportionally to the number of libraries used. A practical limitation to this approach is that Supernova2 limits the number of barcodes to 4.8 million.

Our results showed that in practice, C_F should be between 335X and 823X, but no larger than 1000X, given the optimal coverage of C=56X recommended by 10x and the requirement for sufficient per-fragment read coverage. Surprisingly, we observed that including more extremely long fragments was detrimental for assembly quality. This is possibly due to the loss of barcode specificity for fragments spanning repetitive sequences. From a computational perspective, too many long fragments are harmful to deconvolving the *de bruijn* graph, as more complex paths need to be picked out. In our experiments, $W\mu_{FL}$ between 50kb and 150kb is the best choice to generate reliable assemblies.

Parameters driving assembly quality

Our results regarding assembly quality, and the 10x parameters that influence it, may be useful for efforts in which *de novo* assemblies are important for generation of an initial reference sequence. We show that maximization of N50 does not necessarily reflect assembly quality, which we were able to compare to NA50 because there exists a high-quality human reference genome. Contig and scaffold lengths mostly increased with ascending sequencing coverage, and at sufficient overall sequence coverage it did not matter much whether the increasing coverage C was accomplished by increasing C_R or C_F. However, both contig and scaffold accuracy decreased with increasing *C*. We also found, counterintuitively, that contig and scaffold length mostly decreased with increasing fragment length, a phenomenon that may be due to the specific implementation; however, until there is another assembler that can be compared to Supernova2 it will not be possible to reason about this effect. In addition, intrinsic properties of the genome matter greatly, as removal of repeats or lack of variation dramatically improves assembly quality.

Diploid assembly is the appropriate approach for assembly of genomes of diploid organisms that harbor variation. Therefore, an important metric to evaluate diploid assembly is the fraction of the genome that is assembled in a diploid state. The short input fragment length of R_6 resulted in roughly 20% less of the genome in a diploid state (<60% vs <80%) compared to the other libraries of the same individual. This observation suggests that in addition to metrics such as N50, evaluation of assembly quality should also include the fraction of the genome (or the assembly) that is in a diploid state.

Cost-benefit analysis

Overall, we have attempted to give practical guidelines to assembly of 10x data with Supernova2 and evaluate the performance across a wide range of metrics. Arguably, the metric that matters most in the context of a personal genome is the discovery of variation that lower-cost approaches do not enable. We estimate that the cost increase over standard Illumina sequencing is about 2x, given the 10X prep cost and the higher level of sequence coverage required. There may be many applications for which this combination of excellent single nucleotide variant detection (via barcode-aware read mapping) and precise structural variant discovery (via assembly), achieved by the same data set, is worth the price.

Comparison with hybrid assemblies

Hybrid assembly strategies have been applied successfully to produce human genome assembly of long contiguity [13, 28, 29]. In these studies, long contigs are first produced by single-molecule long-reads, such as PacBio (NG50=1.1Mb; [28]) or Nanopore (NG50=3.21Mb; [13]) comparing favorably to our best results for Linked-Reads assemblies (NG50=236kb). Scaffolding is then performed with complementary technologies such as BioNano to capture chromosomal level long-range information. It promoted the scaffold N50 of PacBio to 31.1Mb [28] and Illumina mate-pair sequencing with 10x data to 33.5Mb [22]. Using SuperNova2, the scaffold N50 from our studies reached ~27.86Mb (R_6) on the basis of 10x data alone, suggesting that 10x technology gives broadly comparable results at a fraction of the price of long-read-based hybrid assemblies.

Availability of supporting data

The raw sequencing data are deposited in the Sequence Read Archive and the corresponding BioProject accession number is PRJNA527321. Diploid assemblies and the codes for comparison are currently available at http://mendel.stanford.edu/supplementarydata/zhang_SN2_2019 and https://github.com/zhanglu295/Evaluate_diploid_assembly. LRTK-SIM is publicly available at https://github.com/zhanglu295/LRTK-SIM.

Additional files

Table S1. Parameters of libraries prepared for NA12878 and NA24385.

Table S2. Parameters used to generate linked-reads sets for evaluating the impact of C_F and C_R on assemblies.

Table S3. Parameters used to generate linked-reads sets for evaluating the impact of μ_{FL} and $N_{F/P}$ on assemblies.

Table S4. Genomic coverage and fraction of contigs in diploid state generated by Supernova2 for the seven libraries prepared by 10x Genomics. Non-PAR: non-pseudoautosomal regions of X chromosome. WFU, YOR, YORM, PR are female; HGP, ASH and CHI are male.

Figure S1. Basic statistics for L_{1L}. The distributions of **A**. the number of fragments per partition; **B**. sequencing depth per fragment; **C**. probability density function of unweighted fragment lengths; **D**. cumulative density function of unweighted fragment lengths; **E**. reversed cumulative density function of unweighted fragment lengths; **F**. reversed cumulative density function of weighted fragment lengths.

Figure S2. Basic statistics for L_{1M}. The distributions of **A**. number of fragments per partition; **B**. sequencing depth per fragment; **C**. probability density function of unweighted fragment lengths; **D**. cumulative density function of unweighted fragment lengths; **E**. reversed cumulative density function of unweighted fragment lengths; **F**. reversed cumulative density function of weighted fragment lengths.

Figure S3. Basic statistics for L_{1H}. The distributions of **A**. number of fragments per partition; **B**. sequencing depth per fragment; **C**. probability density function of unweighted fragment lengths; **D**. cumulative density function of unweighted fragment lengths; **E**. reversed cumulative density function of unweighted fragment lengths; **F**. reversed cumulative density function of weighted fragment lengths.

Figure S4. Basic statistics for $\boldsymbol{L_2}$. The distributions of **A**. number of fragments per partition; **B**. sequencing depth per fragment; **C**. probability density function of unweighted fragment lengths; **D**. cumulative density function of unweighted fragment lengths; **E**. reversed cumulative density function of unweighted fragment lengths; **F**. reversed cumulative density function of weighted fragment lengths.

Figure S5. Basic statistics for $\boldsymbol{L_3}$. The distributions of **A**. number of fragments per partition; **B**. sequencing depth per fragment; **C**. probability density function of unweighted fragment lengths; **D**. cumulative density function of unweighted fragment lengths; **E**. reversed cumulative density function of unweighted fragment lengths; **F**. reversed cumulative density function of weighted fragment lengths.

Figure S6. Basic statistics for $\boldsymbol{L_4}$. The distributions of **A**. number of fragments per partition; **B**. sequencing depth per fragment; **C**. probability density function of unweighted fragment lengths; **D**. cumulative density function of unweighted fragment lengths; **E**. reversed cumulative density function of unweighted fragment lengths; **F**. reversed cumulative density function of weighted fragment lengths.

Figure S7. Basic statistics for $\boldsymbol{L_5}$. The distributions of **A**. number of fragments per partition; **B**. sequencing depth per fragment; **C**. probability density function of unweighted fragment lengths; **D**. cumulative density function of unweighted fragment lengths; **E**. reversed cumulative density function of unweighted fragment lengths; **F**. reversed cumulative density function of weighted fragment lengths.

Figure S8. Basic statistics for $\boldsymbol{L_6}$. The distributions of **A**. number of fragments per partition; **B**. sequencing depth per fragment; **C**. probability density function of unweighted fragment lengths; **D**. cumulative density function of unweighted fragment lengths; **E**. reversed cumulative density function of unweighted fragment lengths; **F**. reversed cumulative density function of weighted fragment lengths.

Figure S9. The workflow of LRTK-SIM to simulate linked-reads

Figure S10. The effect of $N_{F/P}$ on human diploid assembly of chromosome 19 by Supernova2, where C (C=60X; C_F=300X and C_R=0.2X) and μ_{FL} (μ_{FL}=37kb) are fixed.

Figure S11. Comparison of assembly qualities from 10x data with and without single nucleotide variants by changing C_F, C_R and μ_{FL}. C_R was fixed to 0.2X in **A** and **B**; C_F was fixed to 300X in **C** and **D**; C_R was fixed 0.2X and C_F was fixed 300X in **E** and **F**.

Figure S12. Comparison of assembly qualities from 10x data with (1% uniform) and without sequencing error by changing C_F, C_R and μ_{FL}. C_R was fixed to 0.2X in **A** and **B**; C_F was fixed to 300X in **C** and **D**; C_R was fixed 0.2X and C_F was fixed 300X in **E** and **F**.

Figure S13. Overlaps of diploid regions for the three libraries from the same sample. Diploid regions for NA12878 (**A**) and NA24385 (**B**). The percentages denote the proportion of genome is diploid.

Figure S14. Phase block N50s as a function of different parameter combinations. **A**. simulated linked-reads with predefined parameters (**Table S4**) by changing C_F and C_R; **B**. simulated linked-reads with matched parameters of real linked-read sets (**Table S2**) by changing C_F and C_R; **C**. real linked-read sets (**Table S2**) by changing C_F and C_R; **D**. simulated linked-read sets (**Table S3**) with different $W\mu_{FL}$; **E.** simulated linked-read sets with matched parameters (**Table S3**) with real linked-read sets as C=56X; **F.** real linked-read sets with C=56X (**Table S3**).

Competing interest

Arend Sidow is a consultant and shareholder of DNAnexus, Inc.

Author Contributions

AS conceived the study. LZ and XZ wrote LRTK-SIM and performed the analyses. ZMW prepared the genomic DNA and 10x libraries. LZ, XZ, ZMW and AS analyzed the results and wrote the paper. All the authors read and approaved the final manuscript.

Acknowledgements

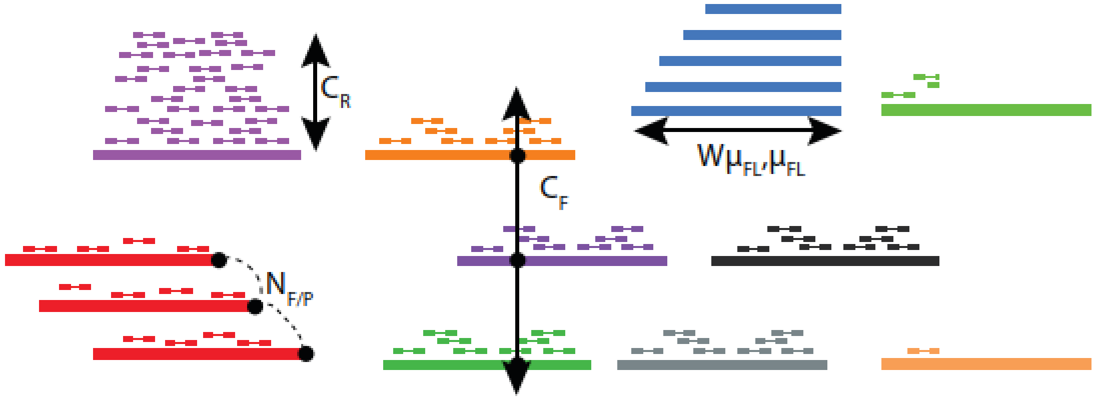
This research was supported by training and research grants from the National Institute of Standards and Technology. We would like to thank Justin Zook, Marc Salit, Alex Bishara, Noah Spies, Nancy Hansen, David Jaffe, and Deanna Church for informative discussions.

Table

Linked-reads set	Overall (%)	Diploid regions (%)	Haploid regions (%)	Non-PAR (%)	Total contig length (contig>500bp)	Length of contigs from megabubble (contig>500bp)	Percentage (%)
R_6	91.9	58.9	27.7	-	5,632,483,053	3,758,345,846	66.73
R_7	91.1	73.3	11.3	-	5,613,140,437	4,668,186,478	83.17
R_8	91.7	77.2	9.2	-	5,635,127,471	4,896,821,850	86.90
R_9	91.3	73.4	12.2	85.9	5,637,615,919	4,438,175,621	78.72
R_{10}	91.7	79.2	5.8	79.9	5,749,001,471	4,793,226,150	83.37
R_{11}	91.7	78.1	7.9	87.6	5,677,566,094	4,723,083,367	83.19

Table 1. Genomic coverage of contigs generated by Supernova2. Non-PAR: non-pseudoautosomal regions of X chromosome. R_6, R_7 and R_8 are female; R_9, R_{10} and R_{11} are male.

Figures



Parameter	Typical values
$N_{F/P}$ = Number of fragments per partition	10 - 100
μ_{FL} = Mean fragment length	μ_{FL} = 10-100kb
$W\mu_{FL}$ = Weighted mean fragment length	$W\mu_{FL}$ = 20-400kb
C_R = Read coverage per fragment	C_R = 0.1x - 0.4x
C_F = Physical (fragment) coverage	C_F = 200x - 1000x
C = total coverage	$C = C_R * C_F$ = 40x - 80x

Linked-read set R (Real) / S (Simulated)	Sequenced Library	μ_{FL} (kb)	$W\mu_{FL}$ (kb)	C_F (X)	C_R (X)	C (X)
R_1 / S_1	L_{1L}	21.6	38.6/35.7	19	0.2	4
R_2 / S_2	L_{1M}	22.4	39.7/37.4	117	0.2	24
R_3 / S_3	L_{1M}	22.4	39.7/36.8	117	0.4	48
R_4 / S_4	L_{1H}	24.0	41.1/40.7	334	0.2	67
R_5 / S_5	L_{1M}	22.4	39.7/36.8	117	0.6	72
R_6 / S_6	L_{1H}	24.0	41.1/40.6	334	0.17	56
R_7 / S_7	L_2	79.0	304.3/131.8	123	0.45	56
R_8 / S_8	L_3	99.2	214.5/168.3	958	0.058	56
R_9 / S_9	L_4	92.1	216.9/154.1	1504	0.036	56
R_{10}/ S_{10}	L_5	120.8	267.4/203.7	208	0.27	56
R_{11}/ S_{11}	L_6	64.2	151.7/107.6	803	0.07	56

Figure 1. The Linked-Read sets prepared to evaluate the impact of C_F, C_R, μ_{FL} and $W\mu_{FL}$ on human diploid assembly.

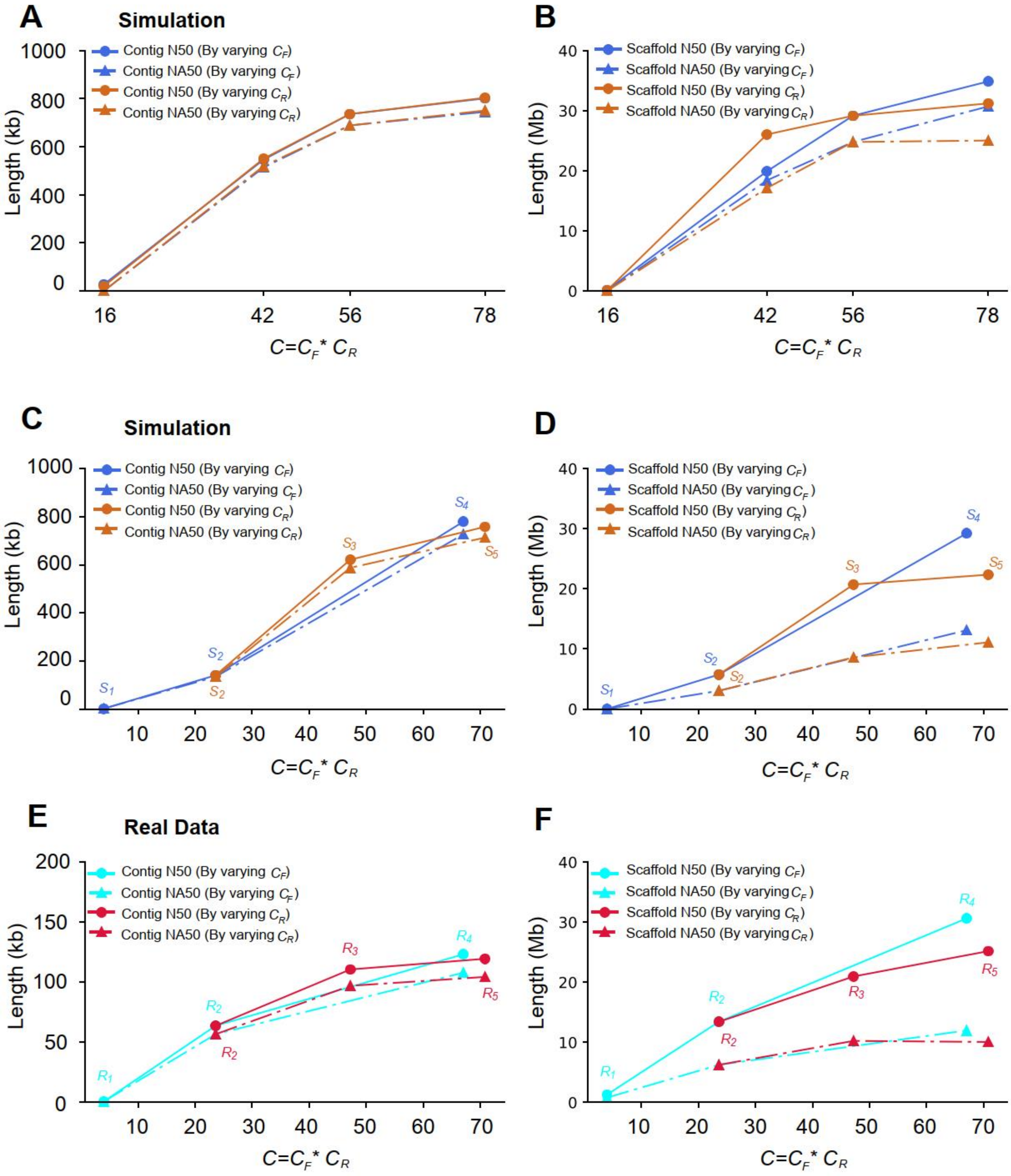


Figure 2. Contig and scaffold lengths (N50 and NA50) as a function of C_F or C_R. **A** and **B**: Simulated Linked-Reads with predefined parameters (**Table S2**); **C** and **D**: Simulated Linked-

reads with matched parameters of real Linked-Read data sets (**Figure 1**); **E** and **F**: Real Linked-Read sets (**Figure 1**).

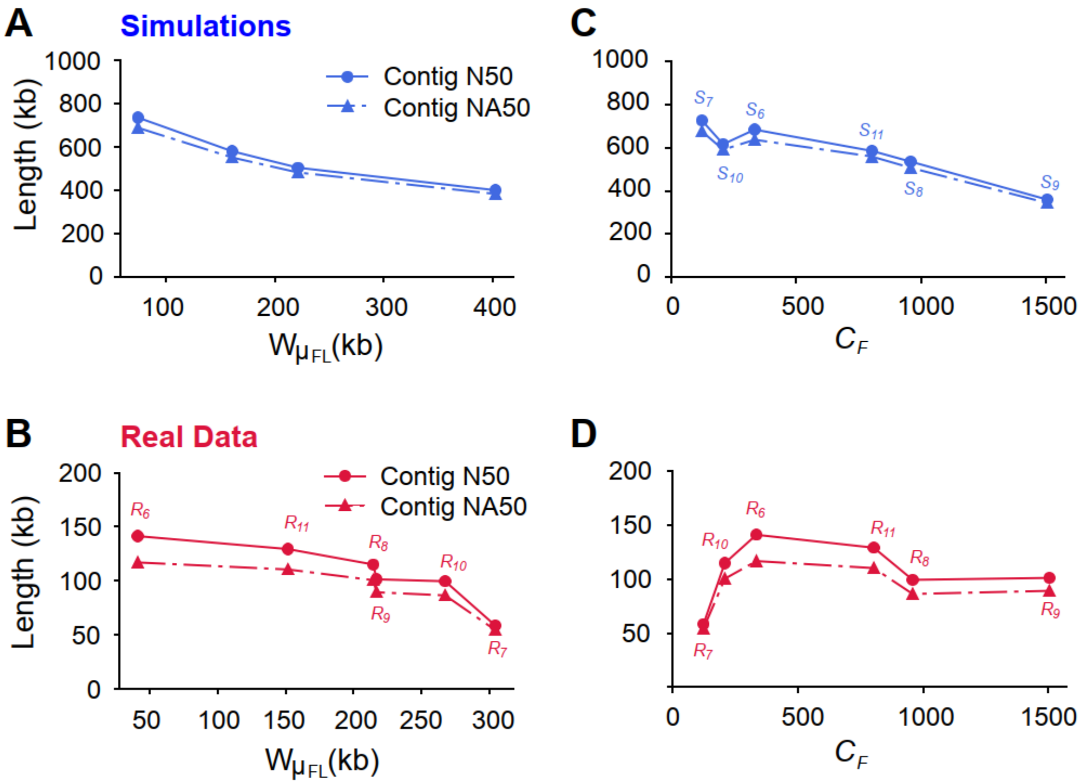


Figure 3. Contig qualities (N50 and NA50) as a function of fragment length $W\mu_{FL}$ or physical coverage C_F, at C=56X. **A** and **C**, results from simulations; **B** and **D**, results from real data.

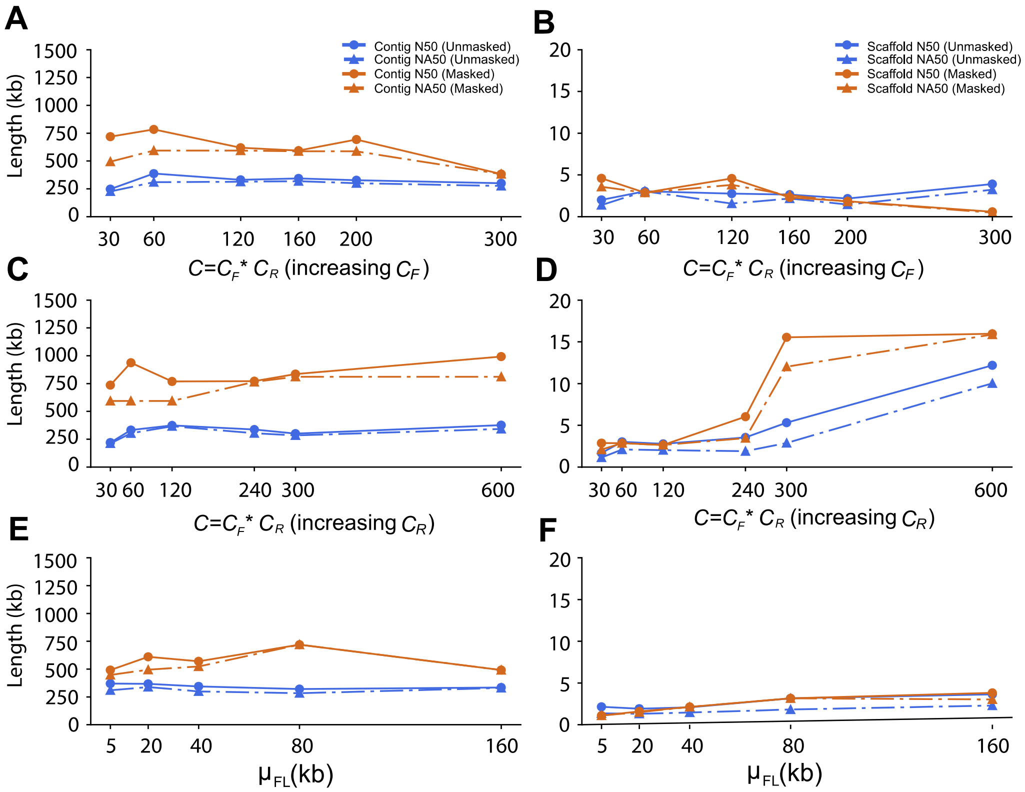


Figure 4. Comparison of contig and scaffold lengths from 10x data with masked and unmasked repetitive sequences by changing *C_F*, *C_R* and μ_{FL}. *C_R* was fixed to 0.2X in **A** and **B**; *C_F* was fixed to 300X in **C** and **D**; *C_R* was fixed to 0.2X and *C_F* was fixed to 300X in **E** and **F**.

References

1. Metzker ML. Sequencing technologies - the next generation. Nat Rev Genet. 2010;11 1:31-46. doi:10.1038/nrg2626.

2. Shendure J, Balasubramanian S, Church GM, Gilbert W, Rogers J, Schloss JA, et al. DNA sequencing at 40: past, present and future. Nature. 2017;550 7676:345-53. doi:10.1038/nature24286.

3. Head SR, Komori HK, LaMere SA, Whisenant T, Van Nieuwerburgh F, Salomon DR, et al. Library construction for next-generation sequencing: overviews and challenges. Biotechniques. 2014;56 2:61-4, 6, 8, passim. doi:10.2144/000114133.

4. O'Connell J, Sharp K, Shrine N, Wain L, Hall I, Tobin M, et al. Haplotype estimation for biobank-scale data sets. Nat Genet. 2016;48 7:817-20. doi:10.1038/ng.3583.

5. Delaneau O, Zagury JF and Marchini J. Improved whole-chromosome phasing for disease and population genetic studies. Nat Methods. 2013;10 1:5-6. doi:10.1038/nmeth.2307.

6. O'Connell J, Gurdasani D, Delaneau O, Pirastu N, Ulivi S, Cocca M, et al. A general approach for haplotype phasing across the full spectrum of relatedness. PLoS Genet. 2014;10 4:e1004234. doi:10.1371/journal.pgen.1004234.

7. Roach JC, Glusman G, Hubley R, Montsaroff SZ, Holloway AK, Mauldin DE, et al. Chromosomal haplotypes by genetic phasing of human families. Am J Hum Genet. 2011;89 3:382-97. doi:10.1016/j.ajhg.2011.07.023.

8. Kajitani R, Toshimoto K, Noguchi H, Toyoda A, Ogura Y, Okuno M, et al. Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. Genome Res. 2014;24 8:1384-95. doi:10.1101/gr.170720.113.

9. Alkan C, Sajjadian S and Eichler EE. Limitations of next-generation genome sequence assembly. Nat Methods. 2011;8 1:61-5. doi:10.1038/nmeth.1527.

10. Treangen TJ and Salzberg SL. Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat Rev Genet. 2011;13 1:36-46. doi:10.1038/nrg3117.

11. Huddleston J, Ranade S, Malig M, Antonacci F, Chaisson M, Hon L, et al. Reconstructing complex regions of genomes using long-read sequencing technology. Genome Res. 2014;24 4:688-96. doi:10.1101/gr.168450.113.

12. Lu H, Giordano F and Ning Z. Oxford Nanopore MinION Sequencing and Genome Assembly. Genomics Proteomics Bioinformatics. 2016;14 5:265-79. doi:10.1016/j.gpb.2016.05.004.

13. Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. Nat Biotechnol. 2018;36 4:338-45. doi:10.1038/nbt.4060.

14. Cao H, Wu H, Luo R, Huang S, Sun Y, Tong X, et al. De novo assembly of a haplotype-resolved human genome. Nat Biotechnol. 2015;33 6:617-22. doi:10.1038/nbt.3200.

15. Peters BA, Kermani BG, Sparks AB, Alferov O, Hong P, Alexeev A, et al. Accurate whole-genome sequencing and haplotyping from 10 to 20 human cells. Nature. 2012;487 7406:190-5. doi:10.1038/nature11236.

16. Edge P, Bafna V and Bansal V. HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies. Genome Res. 2017;27 5:801-12. doi:10.1101/gr.213462.116.

17. Patterson M, Marschall T, Pisanti N, van Iersel L, Stougie L, Klau GW, et al. WhatsHap: Weighted Haplotype Assembly for Future-Generation Sequencing Reads. J Comput Biol. 2015;22 6:498-509. doi:10.1089/cmb.2014.0157.

18. Zheng GX, Lau BT, Schnall-Levin M, Jarosz M, Bell JM, Hindson CM, et al. Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. Nat Biotechnol. 2016;34 3:303-11. doi:10.1038/nbt.3432.

19. Spies N, Weng Z, Bishara A, McDaniel J, Catoe D, Zook JM, et al. Genome-wide reconstruction of complex structural variants using read clouds. Nat Methods. 2017;14 9:915-20. doi:10.1038/nmeth.4366.

20. Bishara A, Moss EL, Kolmogorov M, Parada AE, Weng Z, Sidow A, et al. High-quality genome sequences of uncultured microbes by assembly of read clouds. Nat Biotechnol. 2018; doi:10.1038/nbt.4266.

21. Weisenfeld NI, Kumar V, Shah P, Church DM and Jaffe DB. Direct determination of diploid genome sequences. Genome Res. 2017;27 5:757-67. doi:10.1101/gr.214874.116.

22. Mostovoy Y, Levy-Sakin M, Lam J, Lam ET, Hastie AR, Marks P, et al. A hybrid approach for de novo human genome sequence assembly and phasing. Nat Methods. 2016;13 7:587-90. doi:10.1038/nmeth.3865.

23. Hulse-Kemp AM, Maheshwari S, Stoffel K, Hill TA, Jaffe D, Williams SR, et al. Reference quality assembly of the 3.5-Gb genome of Capsicum annuum from a single linked-read library. Hortic Res. 2018;5:4. doi:10.1038/s41438-017-0011-0.

24. Elyanow R, Wu HT and Raphael BJ. Identifying structural variants using linked-read sequencing data. Bioinformatics. 2017; doi:10.1093/bioinformatics/btx712.

25. Jones SJ, Haulena M, Taylor GA, Chan S, Bilobram S, Warren RL, et al. The Genome of the Northern Sea Otter (Enhydra lutris kenyoni). Genes (Basel). 2017;8 12 doi:10.3390/genes8120379.

26. Zhang F, Christiansen L, Thomas J, Pokholok D, Jackson R, Morrell N, et al. Haplotype phasing of whole human genomes using bead-based barcode partitioning in a single tube. Nat Biotechnol. 2017;35 9:852-7. doi:10.1038/nbt.3897.

27. Peters BA, Liu J and Drmanac R. Co-barcoded sequence reads from long DNA fragments: a cost-effective solution for "perfect genome" sequencing. Front Genet. 2014;5:466. doi:10.3389/fgene.2014.00466.

28. Pendleton M, Sebra R, Pang AW, Ummat A, Franzen O, Rausch T, et al. Assembly and diploid architecture of an individual human genome via single-molecule technologies. Nat Methods. 2015;12 8:780-6. doi:10.1038/nmeth.3454.

29. Ma ZS, Li L, Ye C, Peng M and Zhang YP. Hybrid assembly of ultra-long Nanopore reads augmented with 10x-Genomics contigs: Demonstrated with a human genome. Genomics. 2018; doi:10.1016/j.ygeno.2018.12.013.

30. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34 18:3094-100. doi:10.1093/bioinformatics/bty191.

31. Mikheenko A, Prjibelski A, Saveliev V, Antipov D and Gurevich A. Versatile genome assembly evaluation with QUAST-LG. Bioinformatics. 2018;34 13:i142-i50. doi:10.1093/bioinformatics/bty266.

32. Earl D, Bradnam K, St John J, Darling A, Lin D, Fass J, et al. Assemblathon 1: a competitive assessment of de novo short read assembly methods. Genome Res. 2011;21 12:2224-41. doi:10.1101/gr.126599.111.

Supplementary Notes

LRTK-SIM

LRTK-SIM was designed to simulate linked-reads generated by the 10X Chromium instrument and Illumina sequencer, which consists of four key steps:

1. **Generate diploid reference genome of NA12878.** (See Methods in main manuscript.)
2. **Simulate long DNA fragments.** Assuming the DNA fragment physical coverage was C_F, the total volume of DNA was represented as $V=C_F*L$, where L represented the length of reference genome. Equal physical coverage was assumed for two haplotypes, i.e. $\frac{C_F}{2}$. The start point for each fragment was randomly picked across the reference and its unweighted length is generated from an exponential distribution with mean of μ_{FL}.
3. **DNA fragment barcoding.** About 4.79 million 16-mer barcode sequences were available in the Long Ranger whitelist. LRTK-SIM generated 'pseudo' partitions with unique barcodes and allocated long DNA fragments to them. The number of fragments per partition was sampled from a Poisson distribution with mean of $N_{F/P}$.
4. **Simulate Illumina short reads.** LRTK-SIM assumed short reads cover the long DNA fragments uniformly. The number of reads to be generated were calculated as $\frac{C_R \times V}{R_L}$, where R_L is the length for short reads (150bp by default). The insert size for paired-end reads follows a normal distribution with mean of 400bp. The empirical base quality was learned from linked-reads that were sequenced by Illumina HiSeq X and the positions of sequencing errors were selected randomly.

LRTK-SIM has two advantages: 1. LRTK-SIM is an all-in-one package and implemented in python; it does not rely on any external packages or programs. 2. It explicitly imitates the generation of linked-reads and includes four key parameters: C_F, C_R, μ_{FL} and $N_{F/P}$ to fine-tune modeling of the experimental workflow. LRTK-SIM is speeded up by multithreading and publicly available at https://github.com/LRTK/LRTK-SIM.

Evaluation of diploid assembly

The performance of assembly was evaluated by QUAST-LG and our in-house programs. The scaffolds were broken into contigs when encountering at least 10 consecutive 'N's. Minimap2 was applied to align contigs against reference genome and QUAST-LG was used to post-process the alignments. Contigs shorter than 500bp and scaffolds shorter than 1kb were removed from evaluation. Contig-aligned blocks were generated by breaking contigs at misassemblies, including relocation, inversion and translocation (according to QUAST-LG).

Similar to contigs, for calculating scaffold NA50 scaffold-aligned blocks were obtained by breaking scaffolds at contig misjoins. For example, if there are three contigs ordered as {*a*, *b*, *c*} in the scaffold, misjoins were classified into four types: 1) Relocation, if the alignment-based order of the contigs was different but on the same chromosome; for example if the contigs were aligned in the order {*a*, *c*, *b*} to the hg38 assembly. 2) Inversion, if two connected contigs were in the same order but one of them had reversed orientation in the alignment. 3) Translocation, if two sequences from different chromosomes were merged into adjacent contigs. 4) Indel, if an internal contig such as *b* was unaligned or removed due to insufficient contig length (<500bp). Indels were not considered misjoins if *b* was shorter than 200bp or the gap between *a* and *c* was within 1000bp of *b*'s length.

Supplementary Tables

Raw DNA Preparation	Sequenced Library	Sample id	Raw coverage (X)	μ_{FL}/Wμ_{FL} (kb)	PCR duplication (%)	C_F (X)	C_R (X)
$Prep_1$	L_{1L}	NA12878	94	21.6/38.7	53.08	19.3	0.83
$Prep_1$	L_{1M}	NA12878	175	22.4/39.7	29.24	117.6	0.54
$Prep_1$	L_{1H}	NA12878	192	24.0/41.1	10.92	334	0.27
$Prep_2$	L_2	NA12878	103	79.0/304.3	19.97	123.2	0.41
$Prep_3$	L_3	NA12878	106	99.2/214.5	11.09	958.7	0.07
$Prep_4$	L_4	NA24385	117	92.1/216.9	10.88	1504.6	0.05
$Prep_5$	L_5	NA24385	100	120.8/267.4	18.51	208.4	0.25
$Prep_6$	L_6	NA24385	100	64.2/151.7	12.39	803.3	0.08

Table S1. Parameters of libraries prepared for NA12878 and NA24385.

Parameters	Liked-Read set	$\mu_{FL}/W\mu_{FL}$ (kb)	$N_{F/P}$	C_F (X)	C_R (X)	C (X)
C_F	C_{F1}	37/75	10	156	0.10	16
	C_{F2}	37/75	10	156	0.27	42
	C_{F3}	37/75	10	156	0.36	56
	C_{F4}	37/75	10	156	0.5	78
C_R	C_{R1}	37/75	10	44	0.36	16
	C_{R2}	37/75	10	117	0.36	42
	C_{R3}	37/75	10	156	0.36	56
	C_{R4}	37/75	10	217	0.26	78

Table S2. Parameters used to generate Linked-Read sets for evaluating the impact of C_F and C_R on assemblies.

Parameters	**Liked-Read set**	$\boldsymbol{\mu_{FL}/W\mu_{FL}}$ **(kb)**	***N*** $\boldsymbol{N_{F/P}}$	$\boldsymbol{C_F}$ **(X)**	$\boldsymbol{C_R}$ **(X)**	$\boldsymbol{C}$ **(X)**
μ_{FL}	$\mu_{FL1}/W\mu_{FL1}$	37/75	10	156	0.36	56
	$\mu_{FL2}/W\mu_{FL2}$	80/161	10	156	0.36	56
	$\mu_{FL3}/W\mu_{FL3}$	110/221	10	156	0.36	56
	$\mu_{FL4}/W\mu_{FL4}$	200/402	10	156	0.36	56
$N_{F/P}$	$N_{F/P1}$	37/75	1	156	0.36	56
	$N_{F/P2}$	37/75	2	156	0.36	56
	$N_{F/P3}$	37/75	4	156	0.36	56
	$N_{F/P4}$	37/75	8	156	0.36	56
	$N_{F/P5}$	37/75	16	156	0.36	56

Table S3. Parameters used to generate Linked-Read sets for evaluating the impact of μ_{FL} and $N_{F/P}$ on assemblies.

Library	Overall (%)	Diploid regions (%)	Haploid regions (%)	Non-PAR (%)
HGP	91.9%	79.7%	4.59%	88.8%
ASH	91.8%	79.5%	5.26%	88.0%
WFU	91.9%	76.5%	8.59%	-
CHI	91.8%	78.3%	7.50%	87.7%
YOR	91.8%	80.1%	2.27%	-
YORM	91.9%	76.7%	2.68%	-
PR	91.9%	77.2%	7.89%	-

Table S4. Genomic coverage and fraction of contigs in diploid state generated by Supernova2 for the seven libraries prepared by 10x Genomics. Non-PAR: non-pseudoautosomal regions of X chromosome. WFU, YOR, YORM, PR are female; HGP, ASH and CHI are male.

Supplementary Figures

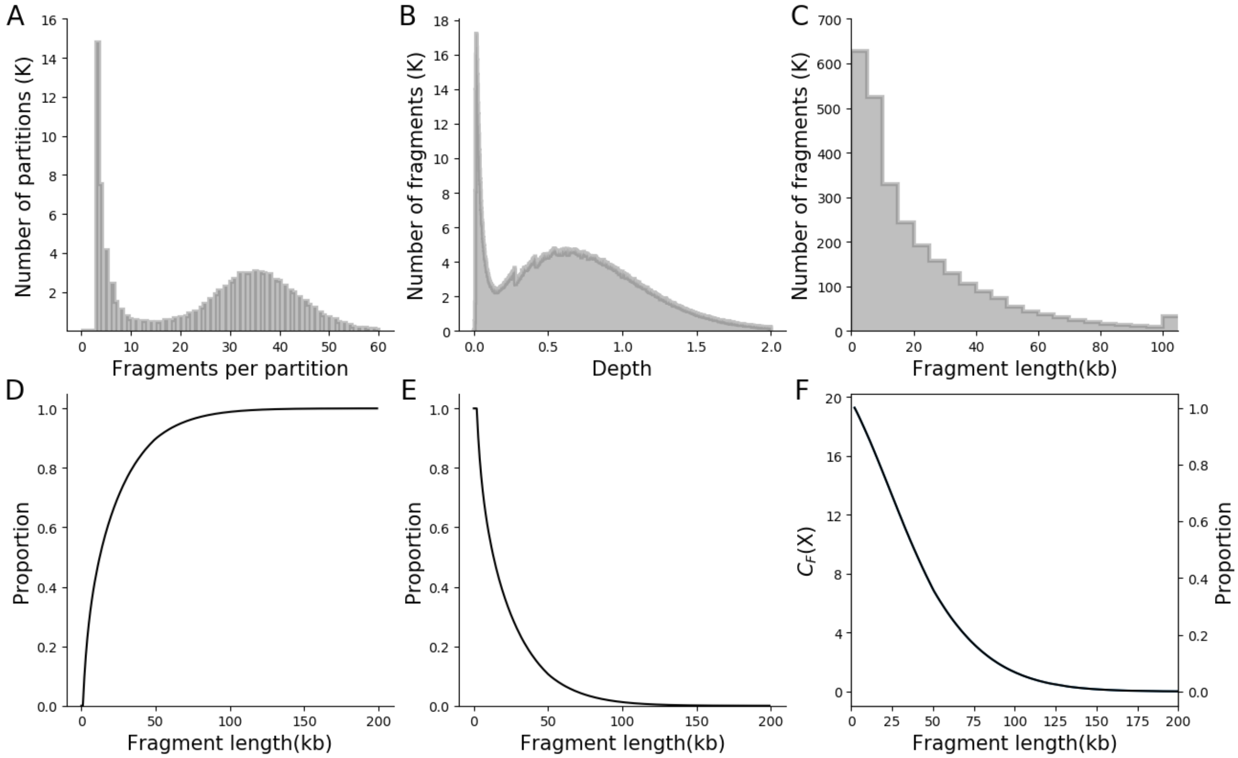


Figure S1. Basic statistics for L_{1L}. The distributions of **A**. the number of fragments per partition; **B**. sequencing depth per fragment; **C**. probability density function of unweighted fragment lengths; **D**. cumulative density function of unweighted fragment lengths; **E**. reversed cumulative density function of unweighted fragment lengths; **F**. reversed cumulative density function of weighted fragment lengths.

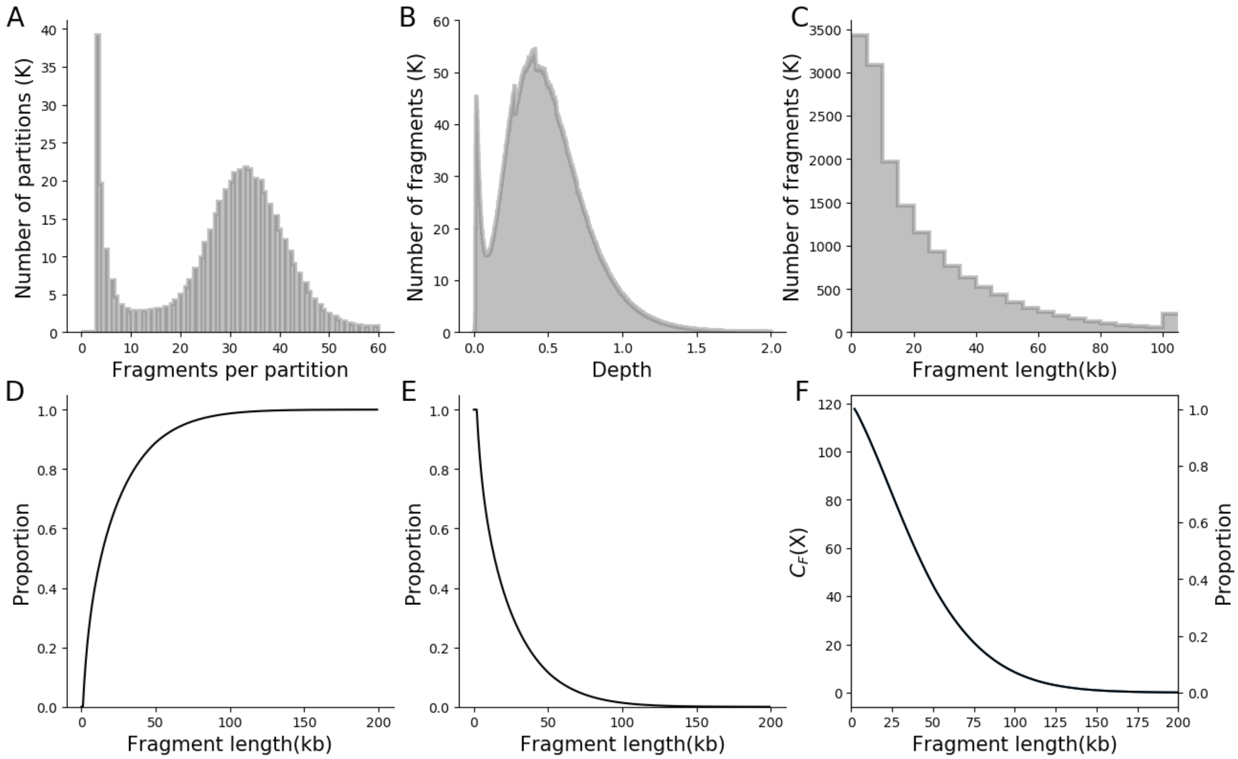


Figure S2. Basic statistics for L_{1M}. The distributions of **A**. number of fragments per partition; **B**. sequencing depth per fragment; **C**. probability density function of unweighted fragment lengths; **D**. cumulative density function of unweighted fragment lengths; **E**. reversed cumulative density function of unweighted fragment lengths; **F**. reversed cumulative density function of weighted fragment lengths.

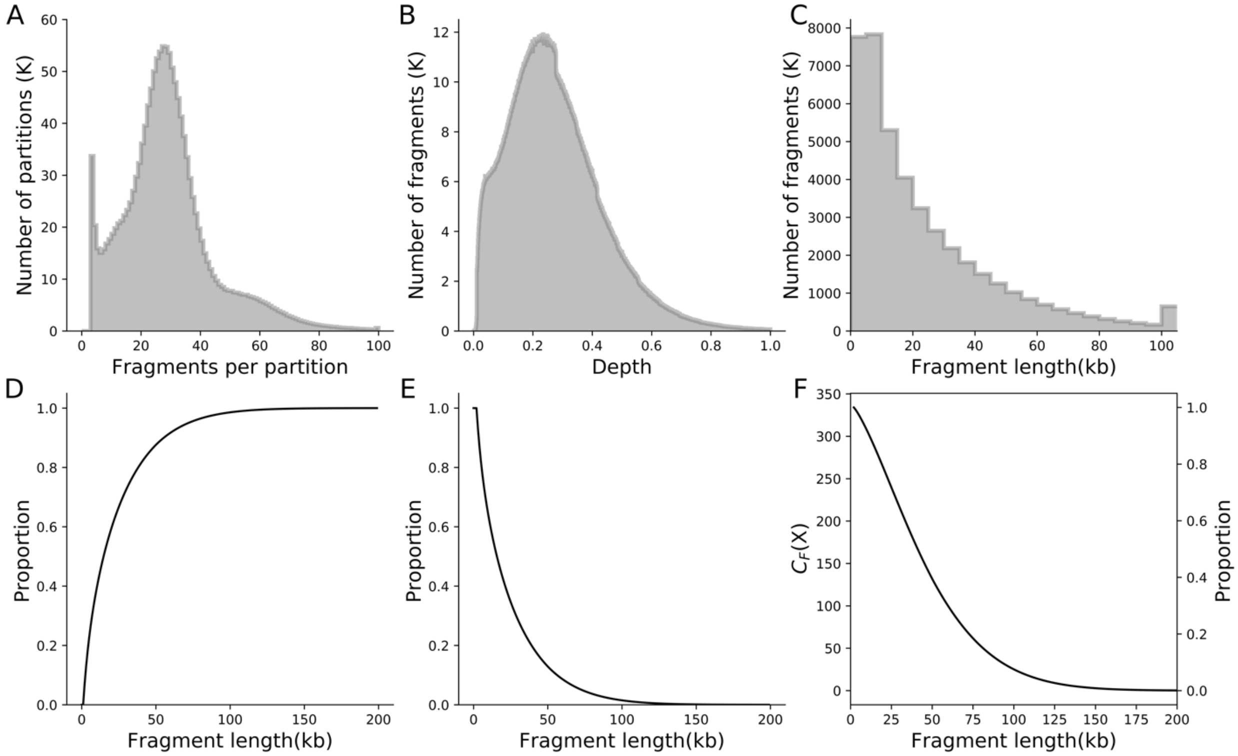


Figure S3. Basic statistics for L_{1H}. The distributions of **A**. number of fragments per partition; **B**. sequencing depth per fragment; **C**. probability density function of unweighted fragment lengths; **D**. cumulative density function of unweighted fragment lengths; **E**. reversed cumulative density function of unweighted fragment lengths; **F**. reversed cumulative density function of weighted fragment lengths.

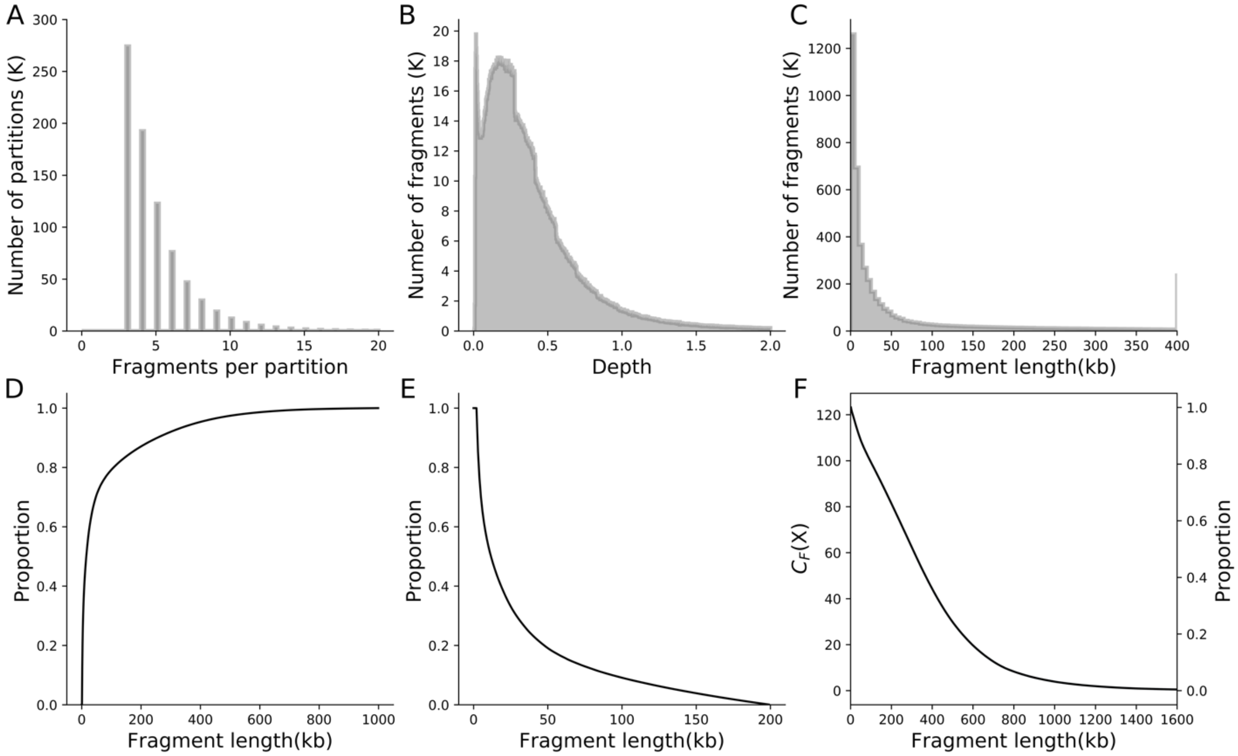


Figure S4. Basic statistics for L_2. The distributions of **A**. number of fragments per partition; **B**. sequencing depth per fragment; **C**. probability density function of unweighted fragment lengths; **D**. cumulative density function of unweighted fragment lengths; **E**. reversed cumulative density function of unweighted fragment lengths; **F**. reversed cumulative density function of weighted fragment lengths.

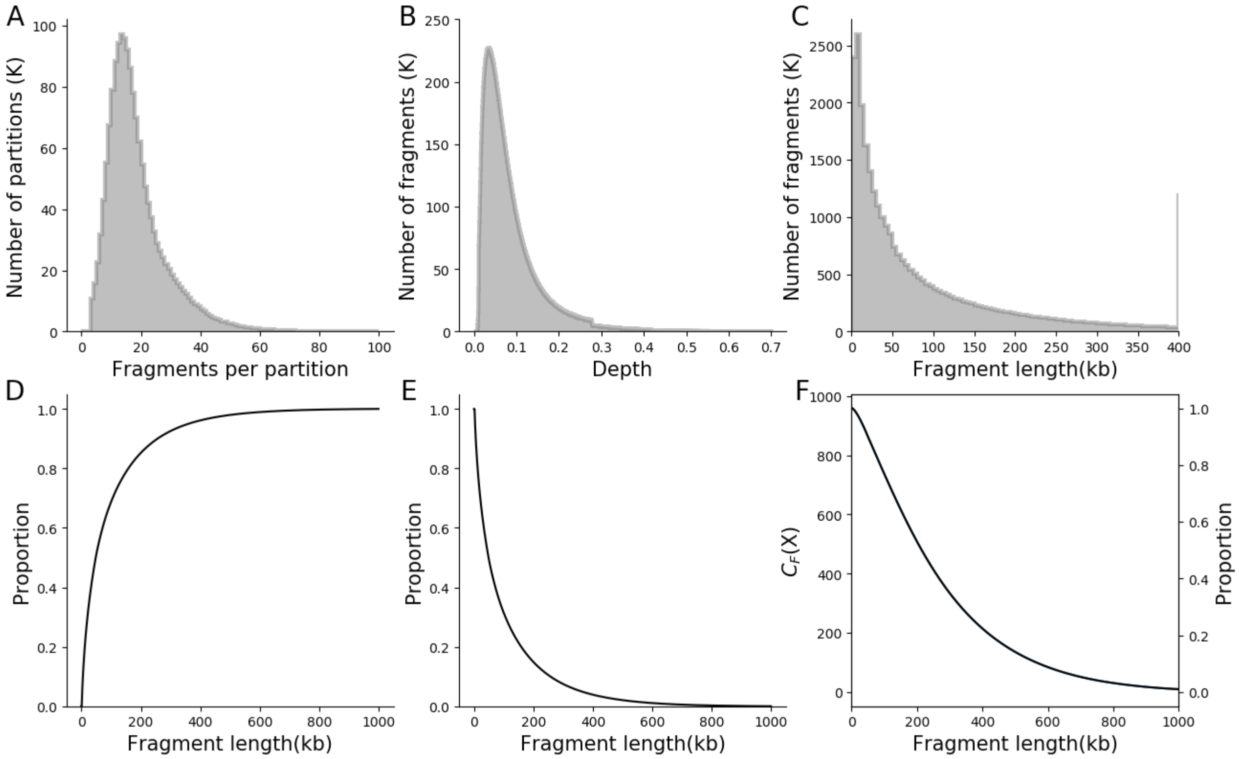


Figure S5. Basic statistics for L_3. The distributions of **A**. number of fragments per partition; **B**. sequencing depth per fragment; **C**. probability density function of unweighted fragment lengths; **D**. cumulative density function of unweighted fragment lengths; **E**. reversed cumulative density function of unweighted fragment lengths; **F**. reversed cumulative density function of weighted fragment lengths.

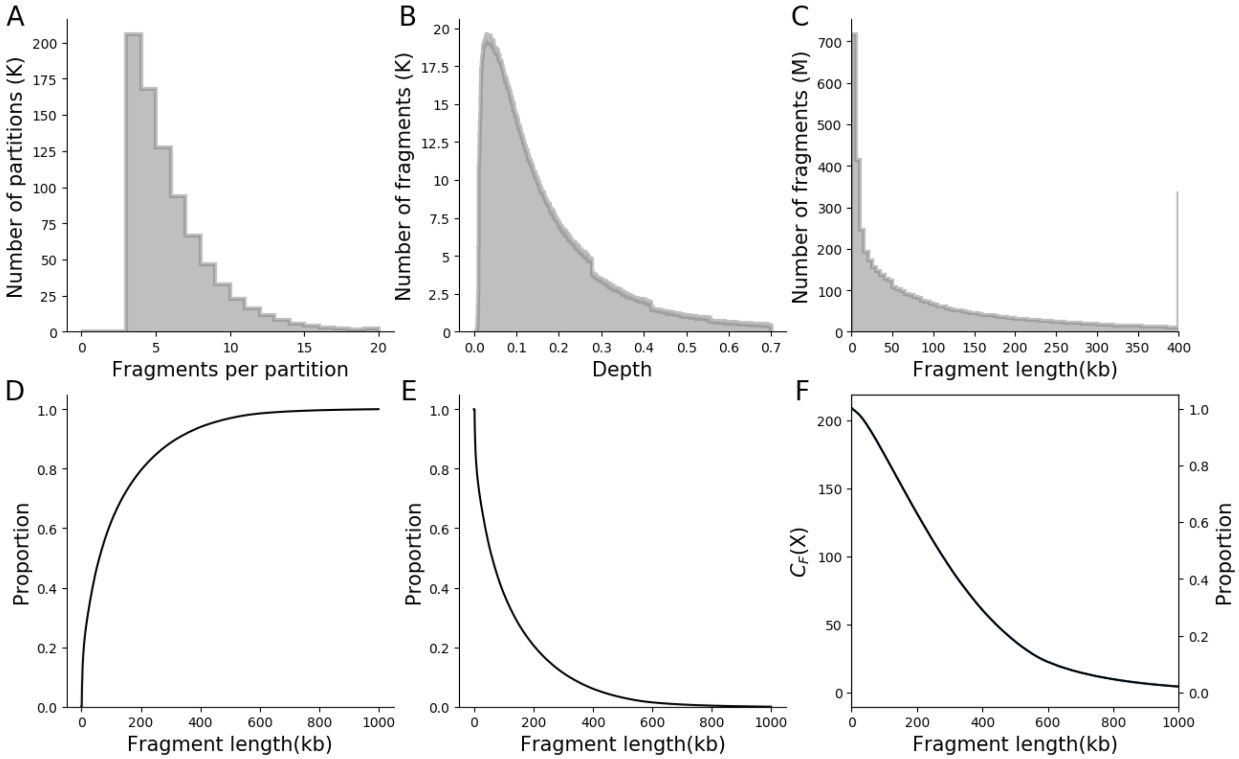


Figure S6. Basic statistics for L_4. The distributions of **A**. number of fragments per partition; **B**. sequencing depth per fragment; **C**. probability density function of unweighted fragment lengths; **D**. cumulative density function of unweighted fragment lengths; **E**. reversed cumulative density function of unweighted fragment lengths; **F**. reversed cumulative density function of weighted fragment lengths.

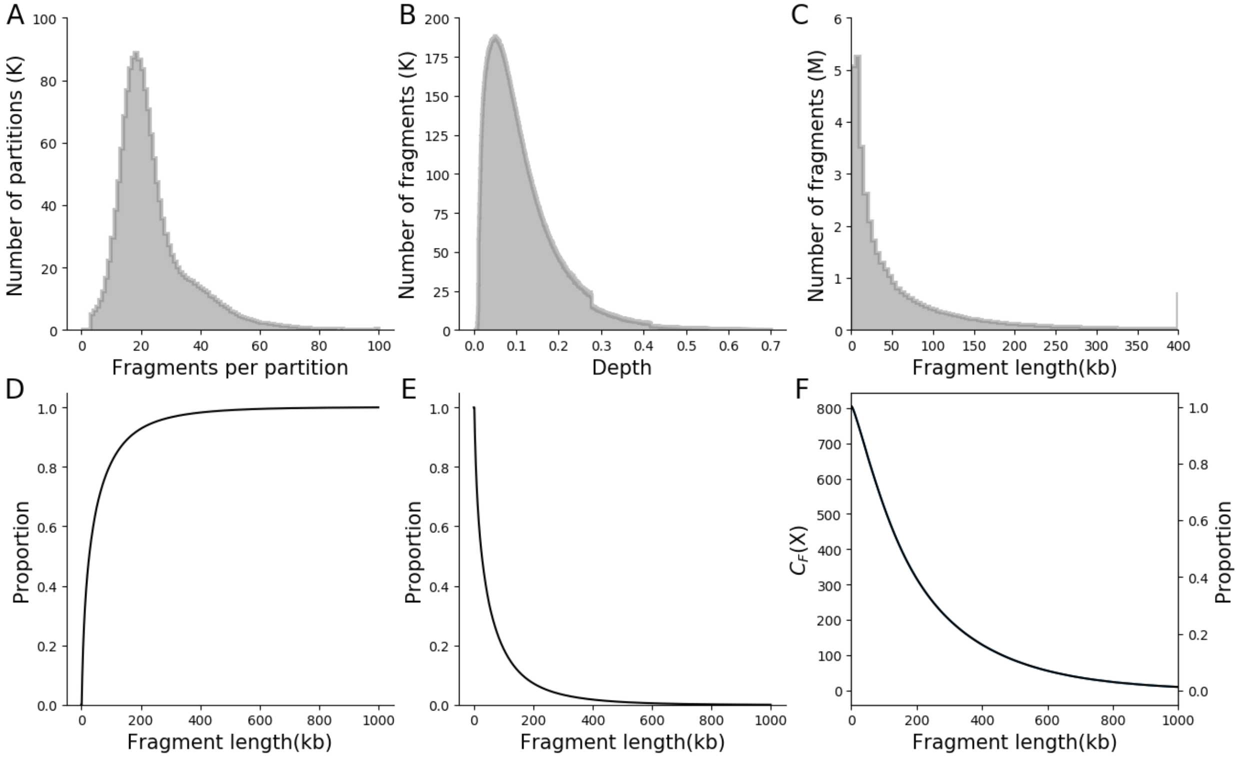


Figure S7. Basic statistics for L_5. The distributions of **A**. number of fragments per partition; **B**. sequencing depth per fragment; **C**. probability density function of unweighted fragment lengths; **D**. cumulative density function of unweighted fragment lengths; **E**. reversed cumulative density function of unweighted fragment lengths; **F**. reversed cumulative density function of weighted fragment lengths.

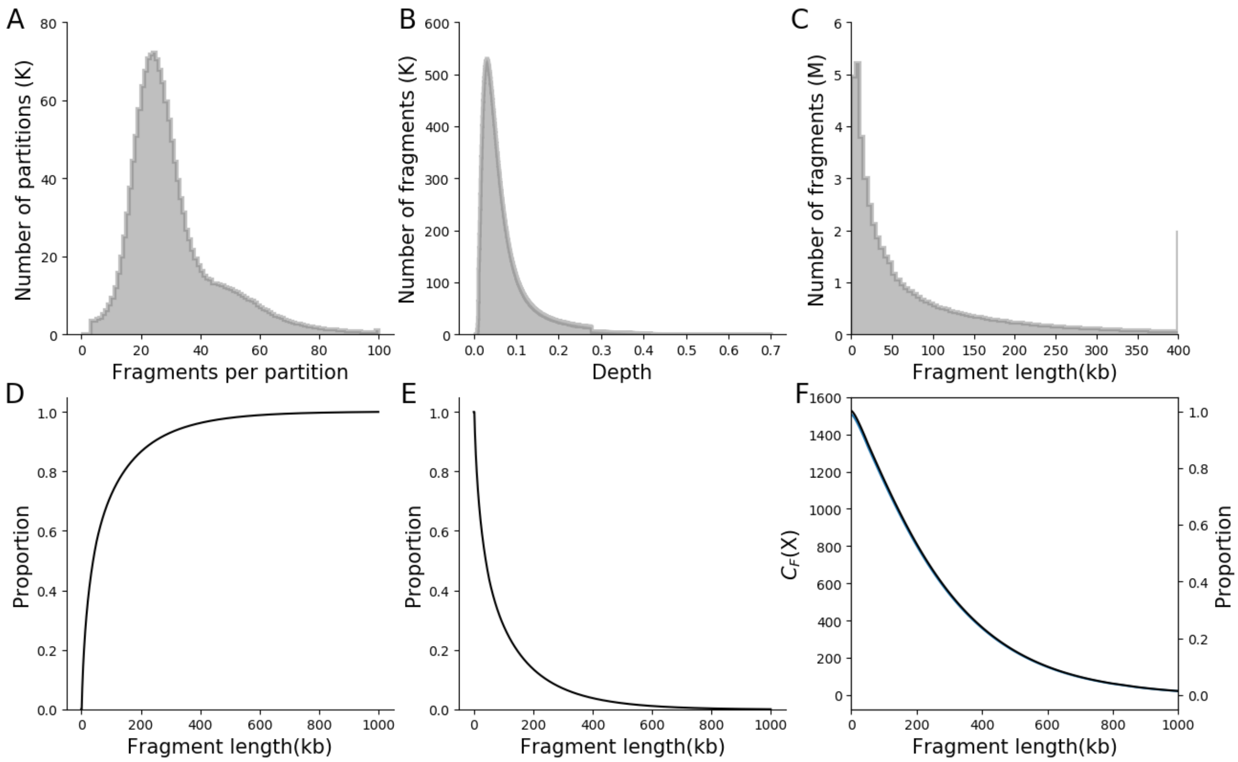


Figure S8. Basic statistics for L_6. The distributions of **A**. number of fragments per partition; **B**. sequencing depth per fragment; **C**. probability density function of unweighted fragment lengths; **D**. cumulative density function of unweighted fragment lengths; **E**. reversed cumulative density function of unweighted fragment lengths; **F**. reversed cumulative density function of weighted fragment lengths.

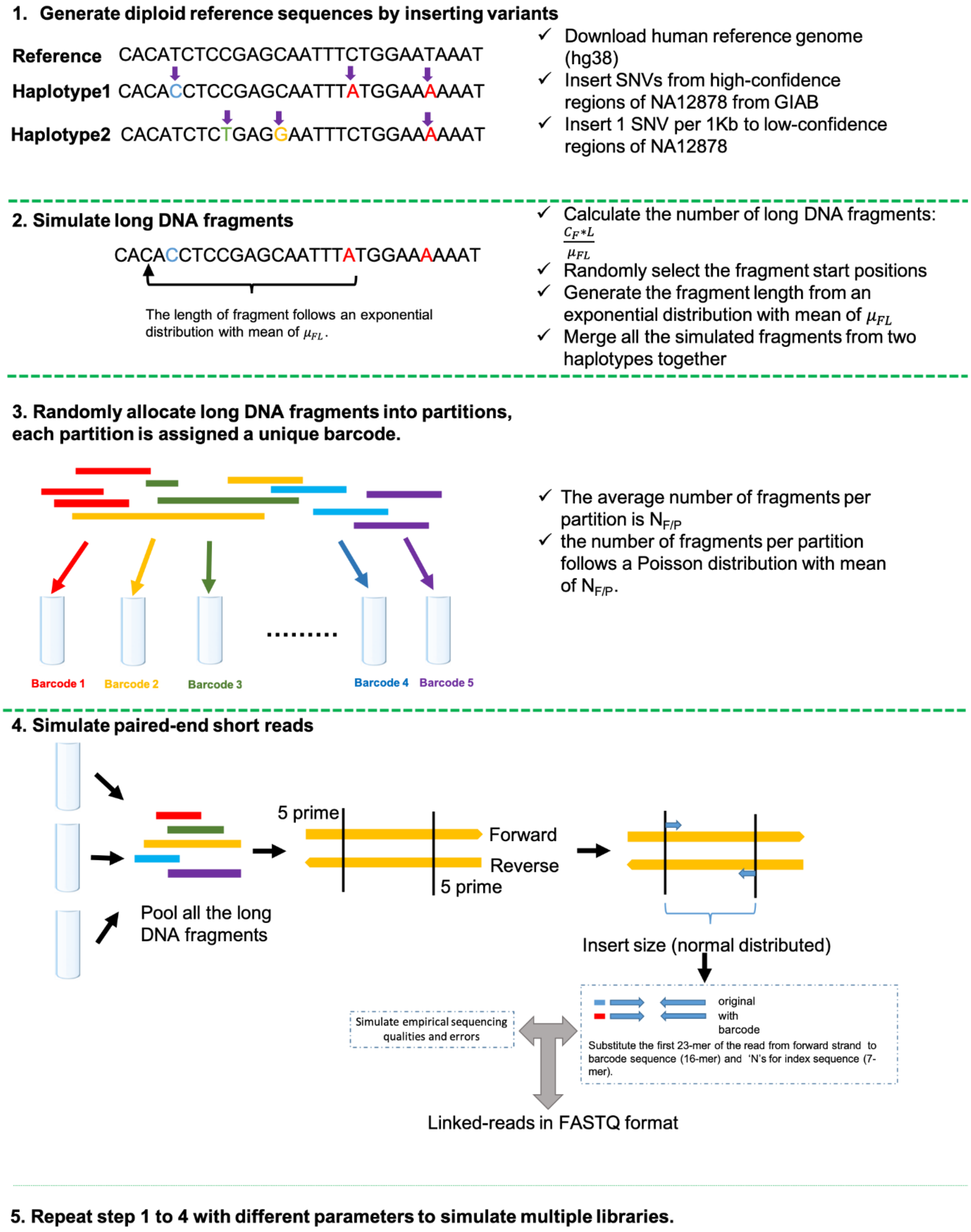

1. Generate diploid reference sequences by inserting variants
Reference CACATCTCCGAGCAATTTCTGGAATAAAT
Haplotype1 CACACCTCCGAGCAATTTATGGAAAAAAT
Haplotype2 CACATCTCTGAGGAATTTCTGGAAAAAAT
✓ Download human reference genome (hg38)
✓ Insert SNVs from high-confidence regions of NA12878 from GIAB
✓ Insert 1 SNV per 1Kb to low-confidence regions of NA12878
2. Simulate long DNA fragments
CACACCTCCGAGCAATTTATGGAAAAAAT
The length of fragment follows an exponential distribution with mean of μ_{FL}.
✓ Calculate the number of long DNA fragments: $\frac{C_F * L}{\mu_{FL}}$
✓ Randomly select the fragment start positions
✓ Generate the fragment length from an exponential distribution with mean of μ_{FL}
✓ Merge all the simulated fragments from two haplotypes together
3. Randomly allocate long DNA fragments into partitions, each partition is assigned a unique barcode.
Barcode 1
Barcode 2
Barcode 3
....
Barcode 4
Barcode 5
✓ The average number of fragments per partition is $N_{F/P}$
✓ the number of fragments per partition follows a Poisson distribution with mean of $N_{F/P}$.
4. Simulate paired-end short reads
Pool all the long DNA fragments
5 prime
Forward
Reverse
5 prime
Insert size (normal distributed)
original
with barcode
Substitute the first 23-mer of the read from forward strand to barcode sequence (16-mer) and 'N's for index sequence (7-mer).
Simulate empirical sequencing qualities and errors
Linked-reads in FASTQ format
5. Repeat step 1 to 4 with different parameters to simulate multiple libraries.

Figure S9. The workflow of LRTK-SIM to simulate linked-reads

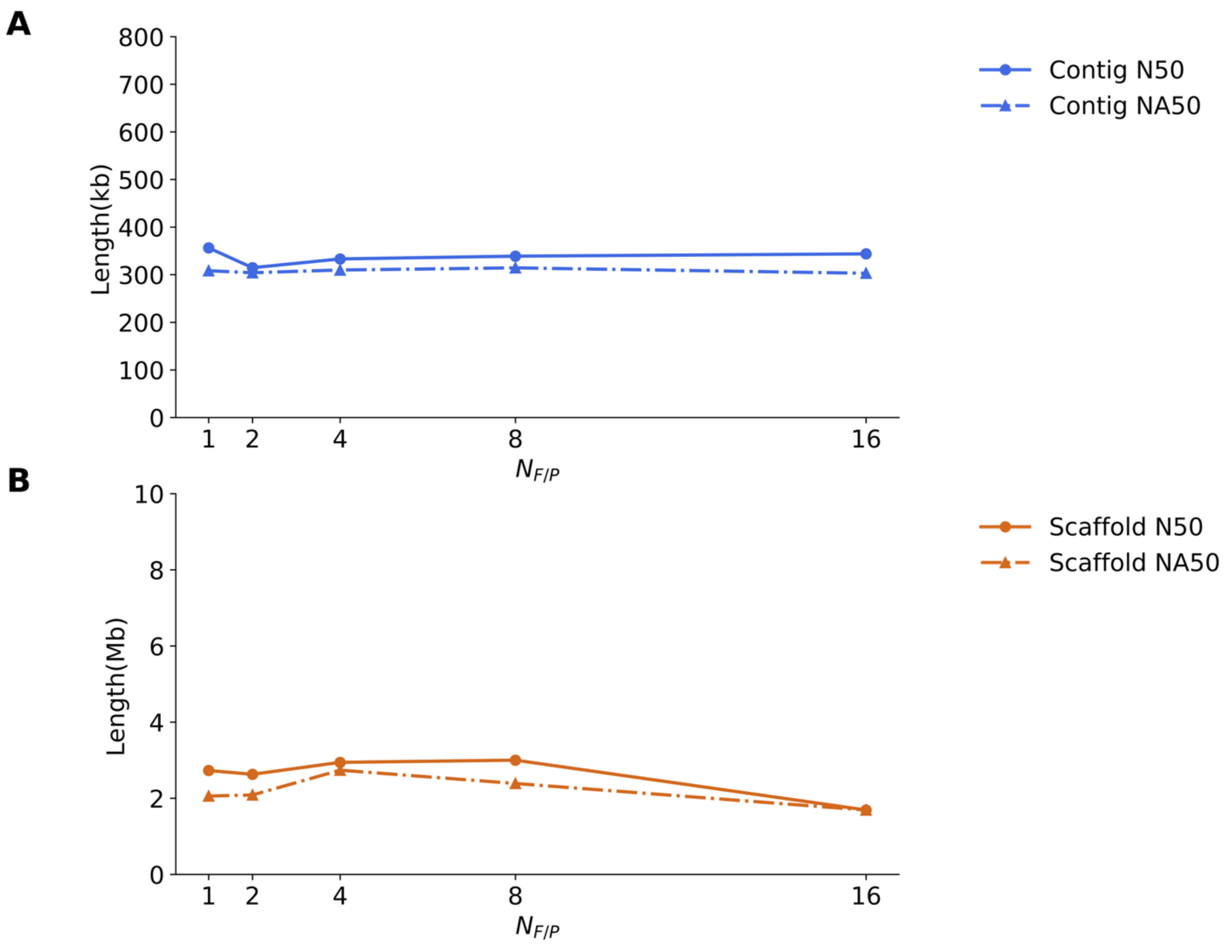


Figure S10. The effect of $N_{F/P}$ on human diploid assembly of chromosome 19 by Supernova2, where C (C=60X; C_F=300X and C_R=0.2X) and μ_{FL} (μ_{FL}=37kb) are fixed.

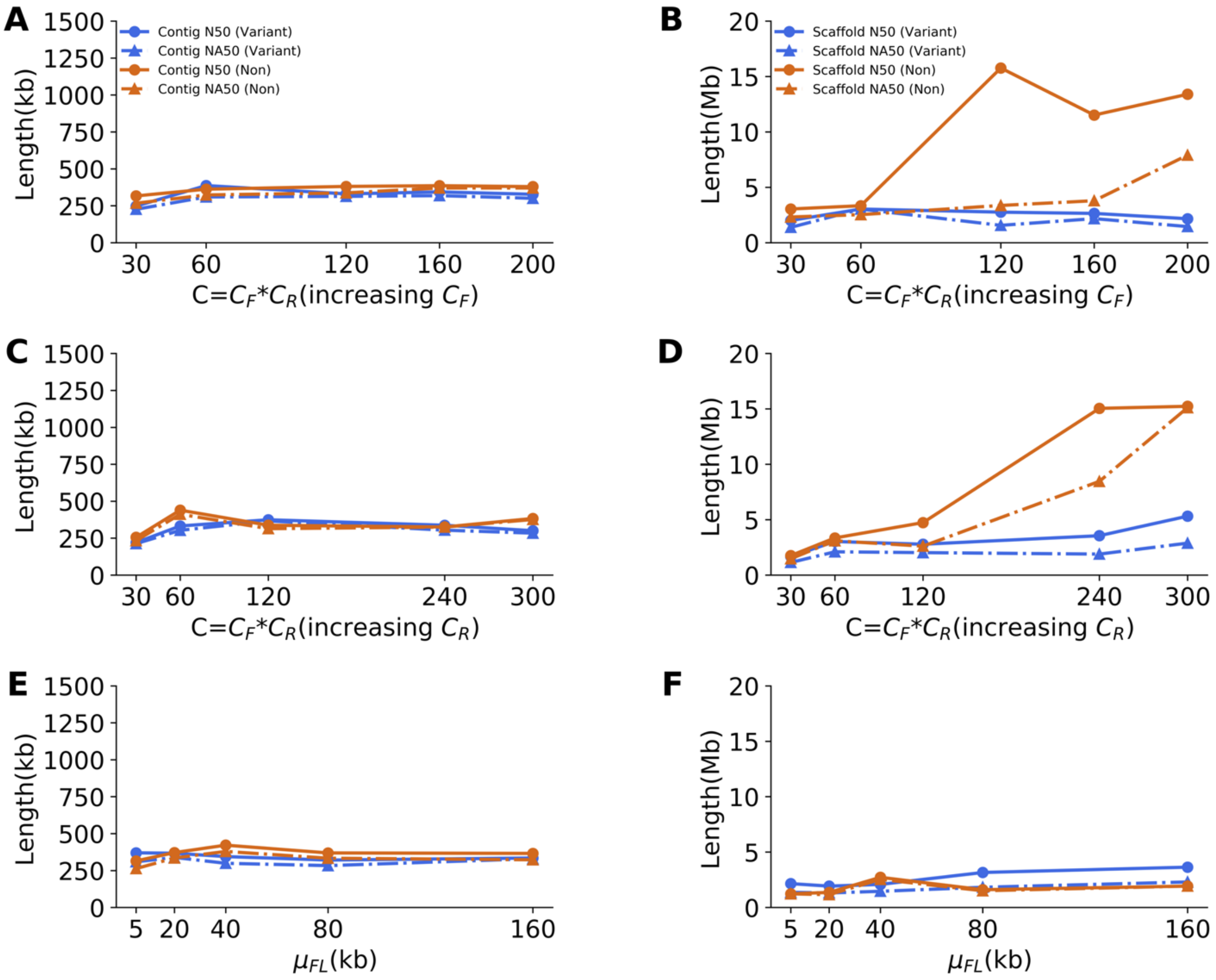


Figure S11. Comparison of assembly qualities from 10x data with and without single nucleotide variants by changing C_F, C_R and μ_{FL}. C_R was fixed to 0.2X in **A** and **B**; C_F was fixed to 300X in **C** and **D**; C_R was fixed 0.2X and C_F was fixed 300X in **E** and **F**.

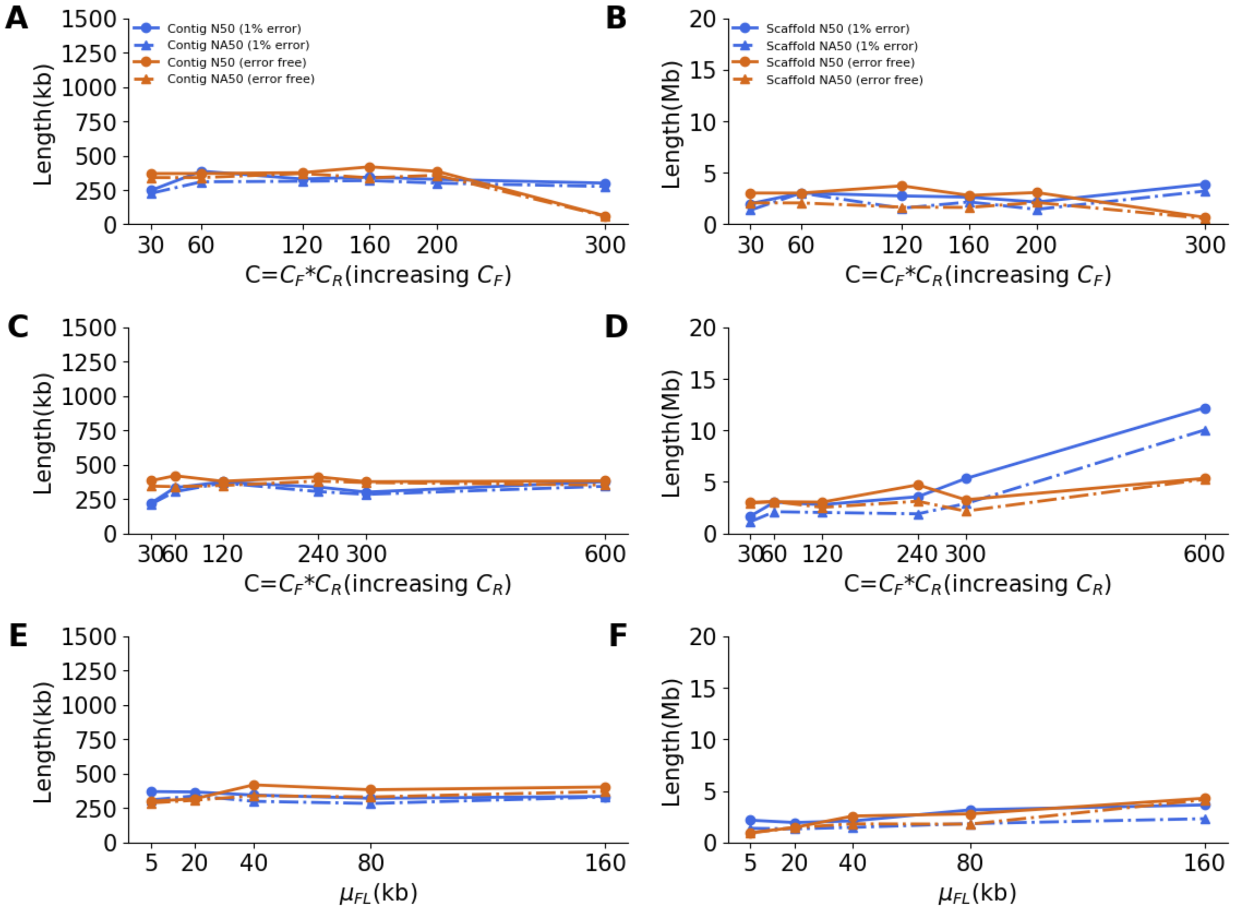


Figure S12. Comparison of assembly qualities from 10x data with (1% uniform) and without sequencing error by changing C_F, C_R and μ_{FL}. C_R was fixed to 0.2X in **A** and **B**; C_F was fixed to 300X in **C** and **D**; C_R was fixed 0.2X and C_F was fixed 300X in **E** and **F**.

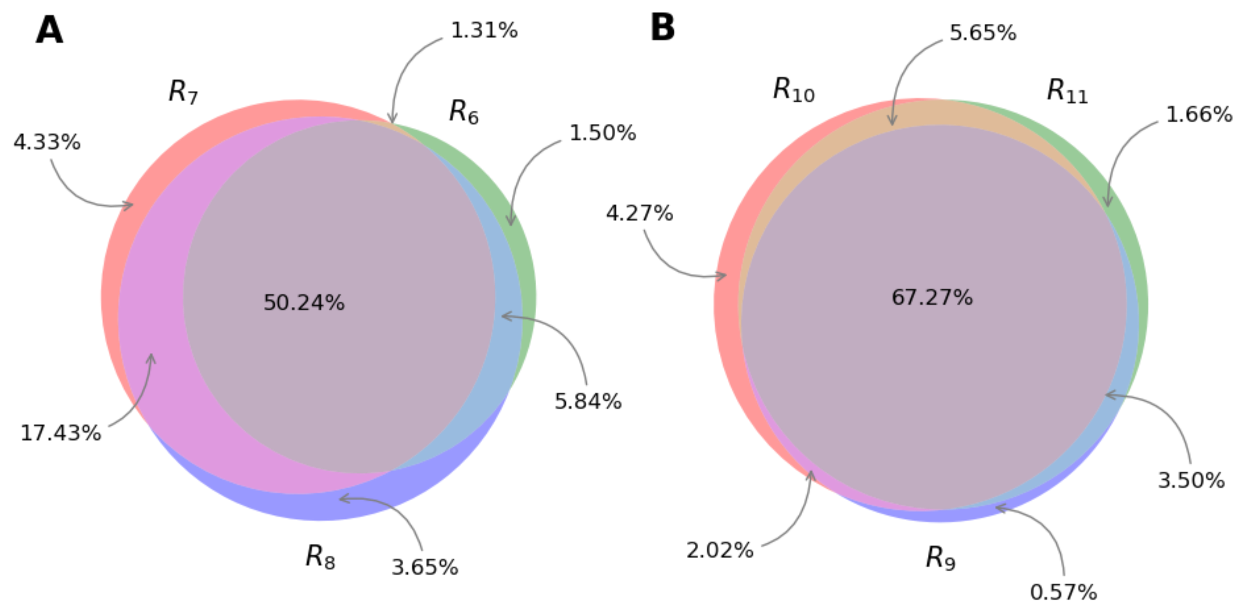


Figure S13. Overlaps of diploid regions for the three libraries from the same sample. Diploid regions for NA12878 (**A**) and NA24385 (**B**). The percentages denote the proportion of genome is diploid.

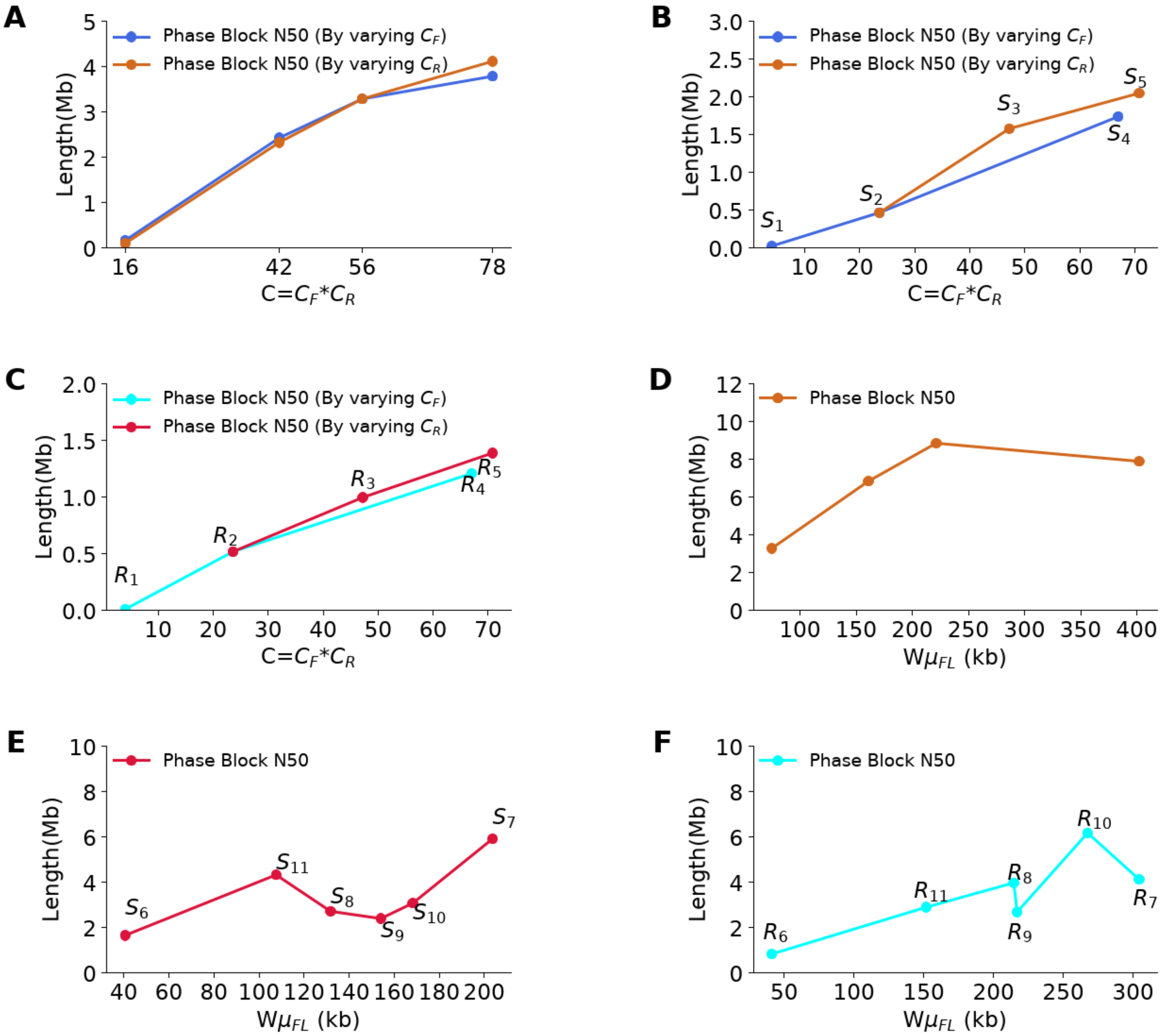
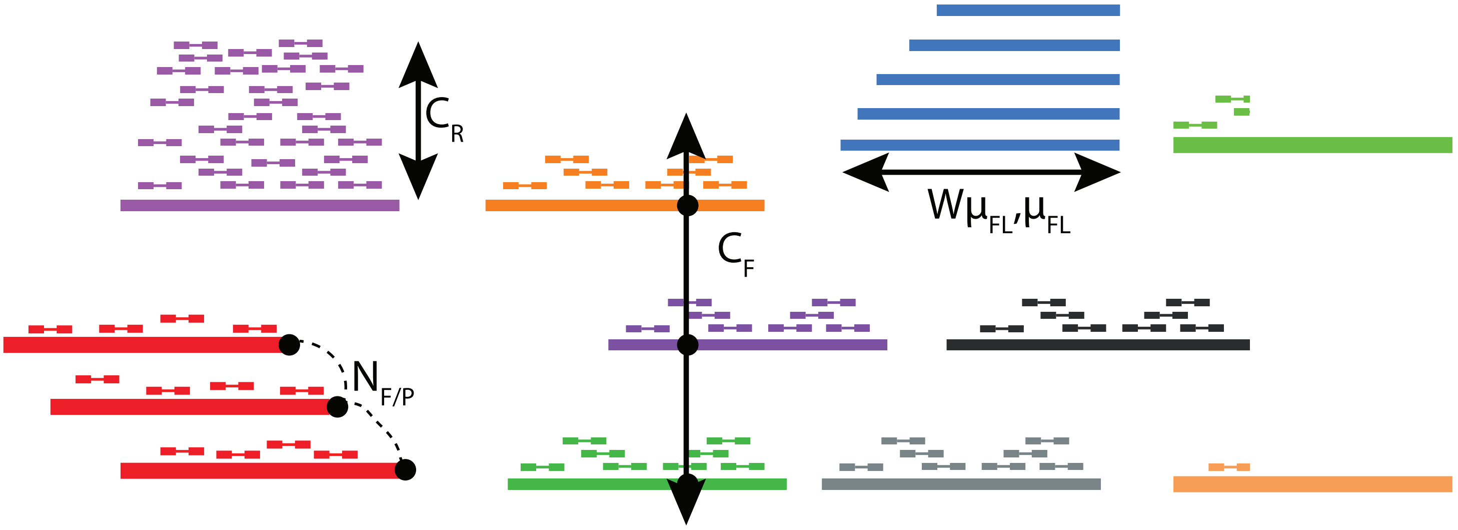


Figure S14. Phase block N50s as a function of different parameter combinations. **A.** simulated Linked-Reads with predefined parameters by changing C_F and C_R (**Table S4**); **B.** simulated Linked-Reads with matched parameters of real linked-read sets (**Table S2**) by changing C_F and C_R; **C.** real Linked-Read sets (**Table S2**) by changing C_F and C_R; **D.** simulated Linked-Read sets (**Table S3**) with different $W\mu_{FL}$; **E.** simulated Linked-Read sets with matched parameters (**Table S3**) with real Linked-Read sets as C=56X; **F.** real Linked-Read sets with C=56X (**Table S3**).

Table

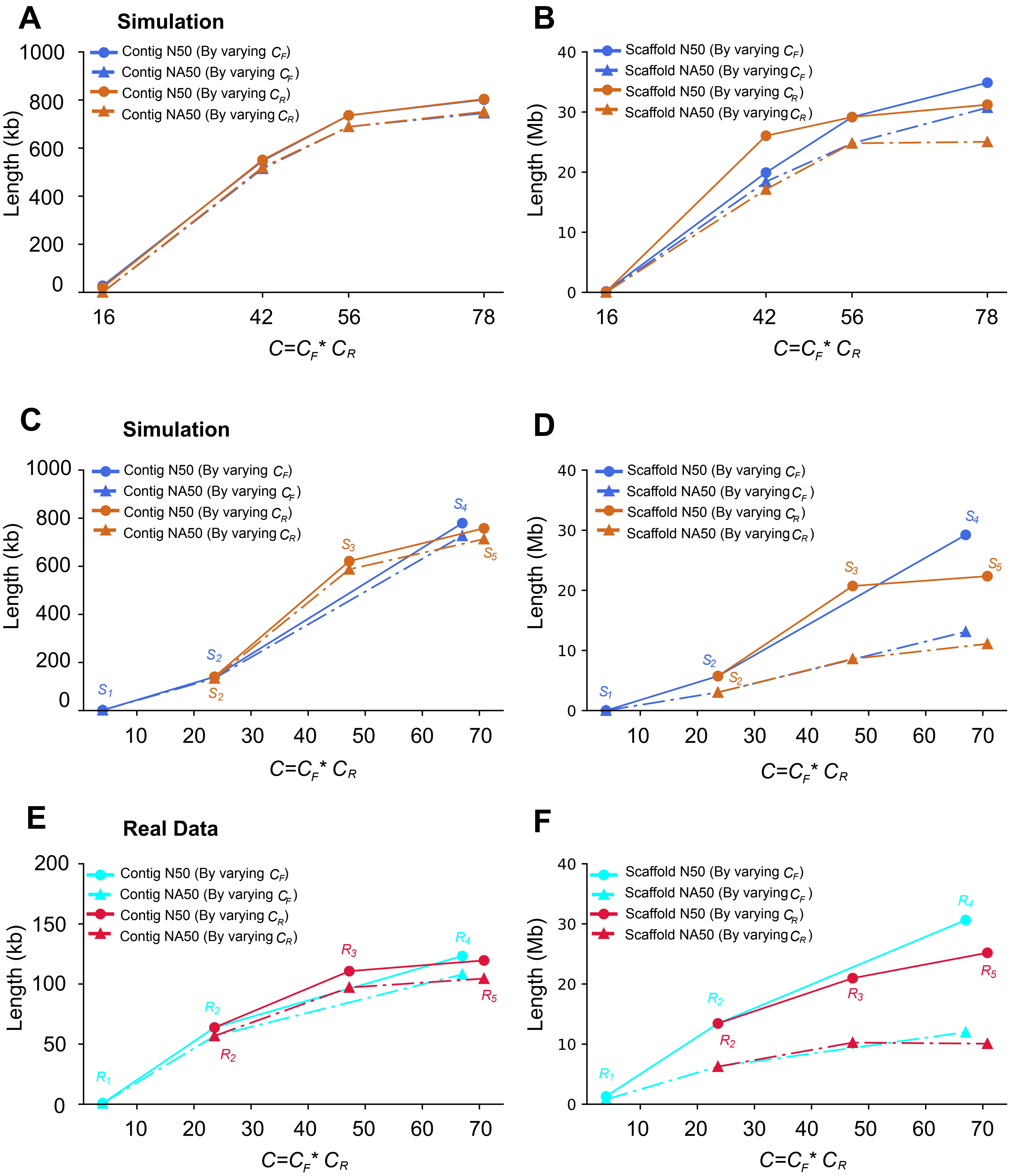
Linked-reads set	Overall (%)	Diploid regions (%)	Haploid regions (%)	Non-PAR (%)	Total contig length (contig>500bp)	Length of contigs from megabubble (contig>500bp)	Percentage (%)
R_6	91.9	58.9	27.7	-	5,632,483,053	3,758,345,846	66.73
R_7	91.1	73.3	11.3	-	5,613,140,437	4,668,186,478	83.17
R_8	91.7	77.2	9.2	-	5,635,127,471	4,896,821,850	86.90
R_9	91.3	73.4	12.2	85.9	5,637,615,919	4,438,175,621	78.72
R_{10}	91.7	79.2	5.8	79.9	5,749,001,471	4,793,226,150	83.37
R_{11}	91.7	78.1	7.9	87.6	5,677,566,094	4,723,083,367	83.19

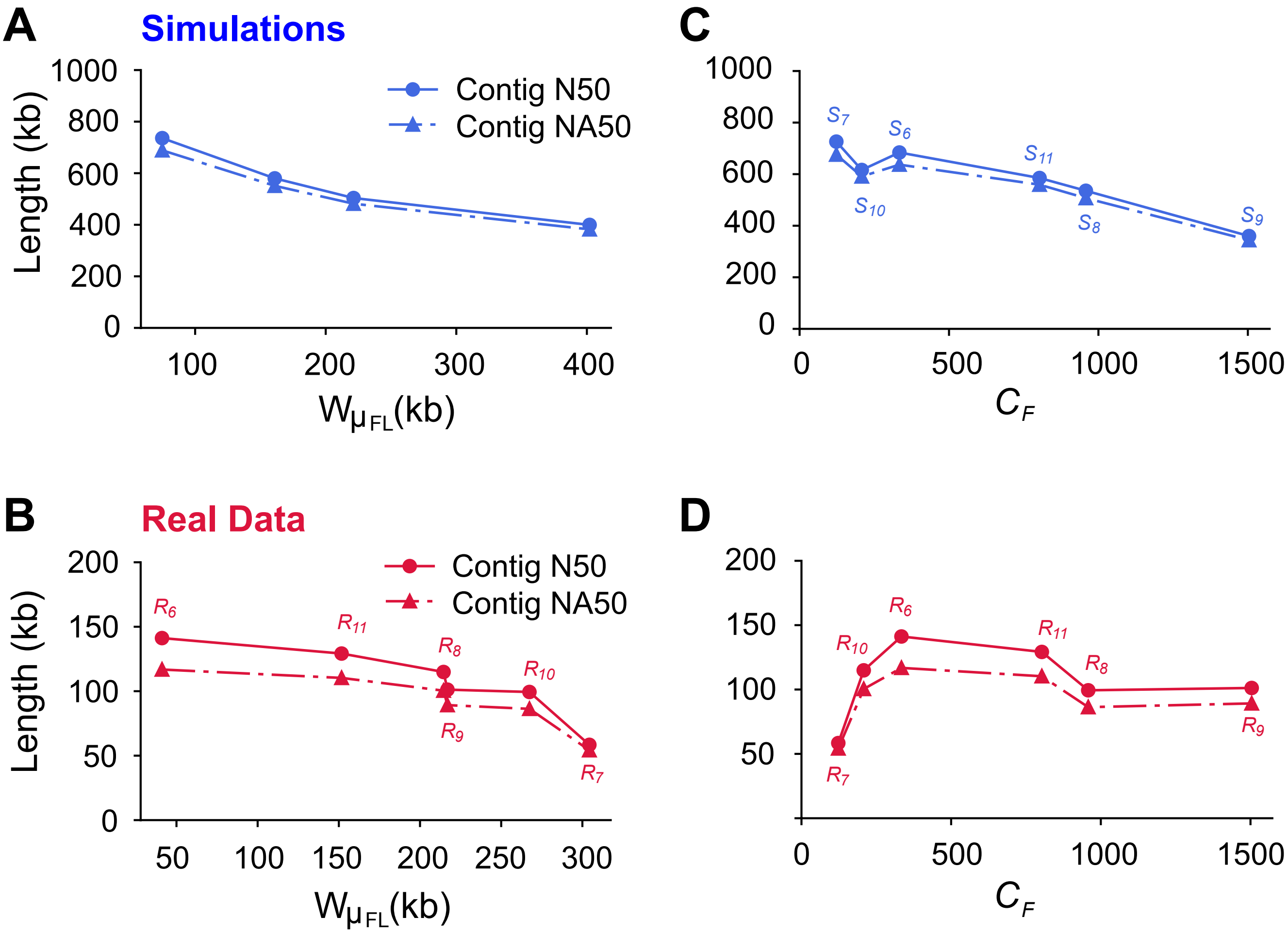
Table 1. Genomic coverage of contigs generated by Supernova2. Non-PAR: non-pseudoautosomal regions of X chromosome. R_6, R_7 and R_8 are female; R_9, R_{10} and R_{11} are male.

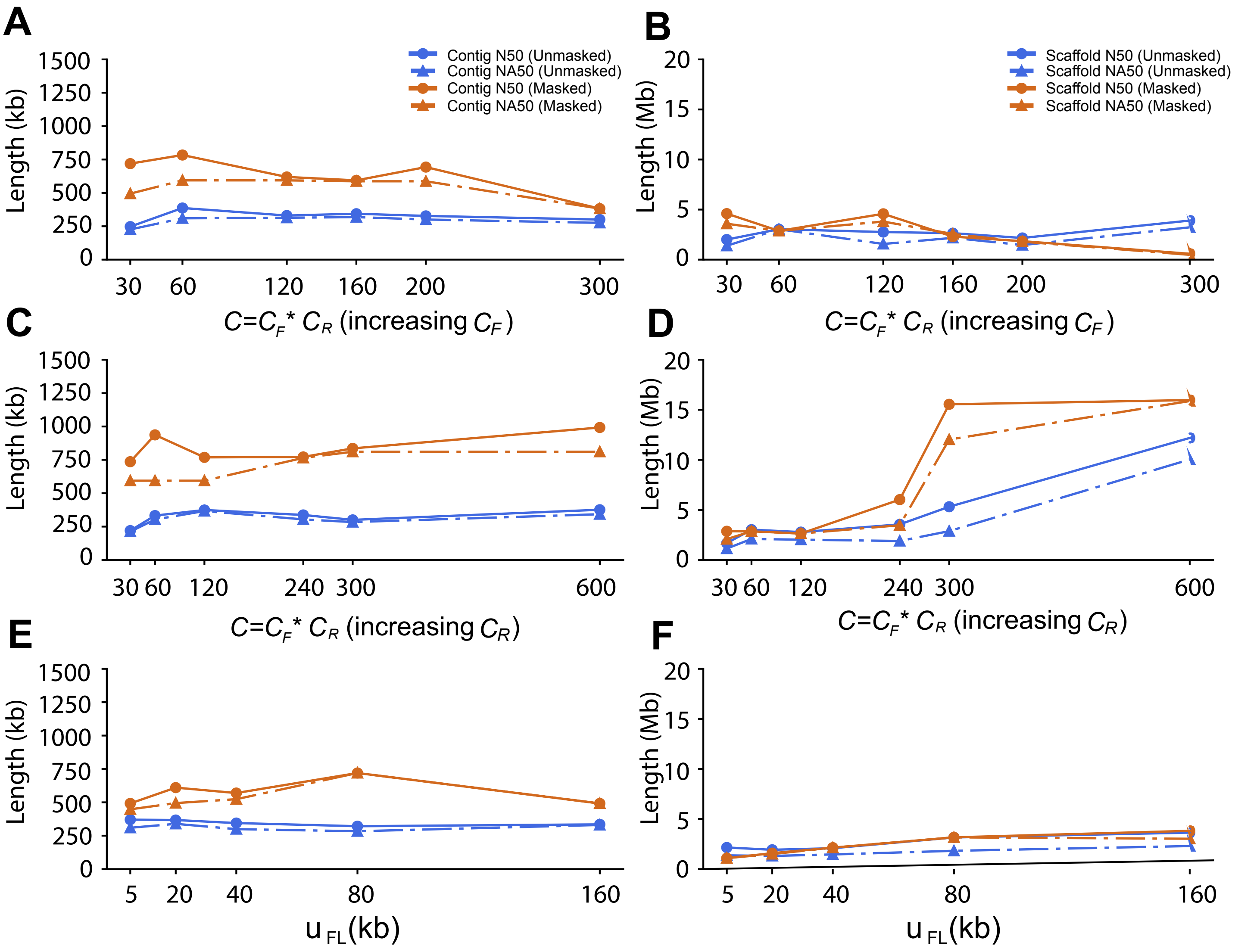


Parameter	Typical values
$N_{F/P}$ = Number of fragments per partition	10 - 100
μ_{FL} = Mean fragment length	μ_{FL} = 10-100kb
$W\mu_{FL}$ = Weighted mean fragment length	$W\mu_{FL}$ = 20-400kb
C_R = Read coverage per fragment	C_R = 0.1x - 0.4x
C_F = Physical (fragment) coverage	C_F = 200x - 1000x
C = total coverage	$C = C_R * C_F$ = 40x - 80x

Linked-read set R (Real) / S (Simulated)	Sequenced Library	μ_{FL} (kb)	$W\mu_{FL}$ (kb)	C_F (X)	C_R (X)	C (X)
R_1 / S_1	L_{1L}	21.6	38.6/35.7	19	0.2	4
R_2 / S_2	L_{1M}	22.4	39.7/37.4	117	0.2	24
R_3 / S_3	L_{1M}	22.4	39.7/36.8	117	0.4	48
R_4 / S_4	L_{1H}	24.0	41.1/40.7	334	0.2	67
R_5 / S_5	L_{1M}	22.4	39.7/36.8	117	0.6	72
R_6 / S_6	L_{1H}	24.0	41.1/40.6	334	0.17	56
R_7 / S_7	L_2	79.0	304.3/131.8	123	0.45	56
R_8 / S_8	L_3	99.2	214.5/168.3	958	0.058	56
R_9 / S_9	L_4	92.1	216.9/154.1	1504	0.036	56
R_{10}/ S_{10}	L_5	120.8	267.4/203.7	208	0.27	56
R_{11}/ S_{11}	L_6	64.2	151.7/107.6	803	0.07	56


A
Simulation
Contig N50 (By varying C_F)
Contig NA50 (By varying C_F)
Contig N50 (By varying C_R)
Contig NA50 (By varying C_R)
Length (kb)
1000
800
600
400
200
0
16
42
56
78
$C=C_F*C_R$
B
Scaffold N50 (By varying C_F)
Scaffold NA50 (By varying C_F)
Scaffold N50 (By varying C_R)
Scaffold NA50 (By varying C_R)
Length (Mb)
40
30
20
10
0
16
42
56
78
$C=C_F*C_R$
C
Simulation
Contig N50 (By varying C_F)
Contig NA50 (By varying C_F)
Contig N50 (By varying C_R)
Contig NA50 (By varying C_R)
S_1
S_2
S_2
S_3
S_4
S_5
Length (kb)
1000
800
600
400
200
0
10
20
30
40
50
60
70
$C=C_F*C_R$
D
Scaffold N50 (By varying C_F)
Scaffold NA50 (By varying C_F)
Scaffold N50 (By varying C_R)
Scaffold NA50 (By varying C_R)
S_1
S_2
S_2
S_3
S_4
S_5
Length (Mb)
40
30
20
10
0
10
20
30
40
50
60
70
$C=C_F*C_R$
E
Real Data
Contig N50 (By varying C_F)
Contig NA50 (By varying C_F)
Contig N50 (By varying C_R)
Contig NA50 (By varying C_R)
R_1
R_2
R_2
R_3
R_4
R_5
Length (kb)
200
150
100
50
0
10
20
30
40
50
60
70
$C=C_F*C_R$
F
Scaffold N50 (By varying C_F)
Scaffold NA50 (By varying C_F)
Scaffold N50 (By varying C_R)
Scaffold NA50 (By varying C_R)
R_1
R_2
R_2
R_3
R_4
R_5
Length (Mb)
40
30
20
10
0
10
20
30
40
50
60
70
$C=C_F*C_R$


A
Simulations
Length (kb)
1000
800
600
400
200
0
Contig N50
Contig NA50
100
200
300
400
$W_{\mu_{FL}}$(kb)
C
S_7
S_6
S_{10}
S_{11}
S_8
S_9
0
500
1000
1500
C_F
B
Real Data
Length (kb)
200
150
100
50
0
Contig N50
Contig NA50
R_6
R_{11}
R_8
R_{10}
R_9
R_7
50
100
150
200
250
300
$W_{\mu_{FL}}$(kb)
D
R_{10}
R_6
R_{11}
R_8
R_9
R_7
0
500
1000
1500
C_F


A
Length (kb)
1500
1250
1000
750
500
250
0
30
60
120
160
200
300
C=C_F* C_R (increasing C_F)
Contig N50 (Unmasked)
Contig NA50 (Unmasked)
Contig N50 (Masked)
Contig NA50 (Masked)
B
Length (Mb)
20
15
10
5
0
30
60
120
160
200
300
C=C_F* C_R (increasing C_F)
Scaffold N50 (Unmasked)
Scaffold NA50 (Unmasked)
Scaffold N50 (Masked)
Scaffold NA50 (Masked)
C
Length (kb)
30 60
120
240
300
600
C=C_F* C_R (increasing C_R)
D
Length (Mb)
30 60
120
240
300
600
C=C_F* C_R (increasing C_R)
E
Length (kb)
5
20
40
80
160
u_FL (kb)
F
Length (Mb)
5
20
40
80
160
u_FL (kb)

DEPARTMENT OF PATHOLOGY
DEPARTMENT OF GENETICS
STANFORD UNIVERSITY SCHOOL OF MEDICINE
STANFORD, CA 94305-5324

Stanford, March 20, 2019

To: GigaScience Editors

Dear Editors,

We are pleased to submit our comprehensive study on de novo human genome assembly as a Data Note to GigaScience. *De novo* assembly has received renewed attention for human genomes due to the availability of long-read and Linked-Read approaches. While long-read approaches will remain prohibitively expensive in the foreseeable future and only useful for highly specialized applications, linked-read approaches are cost-effective for larger human cohorts. We therefore explored the experimental and computational parameter space for 10x-based linked read assemblies with Supernova2, the only currently available *de novo* assembler that makes use of linked reads data to produce a diploid assembly.

Beyond the results, which will be useful for a broad audience of scientists who are sequencing human personal genomes, we generated two resources and therefore suggest that our paper clearly fits the purpose of GigaSciences' Data Note. (1) We generated extensive and very deep sequence data on two reference cell lines, NA128787 and NA24385, that will be useful for methods developers in assembly and variation detection. (2) We developed the Linked-Read data simulator, LRTK-SIM, which allows software and applications developers to generate highly realistic linked-read data from a template reference genome.

Linked-Read data sets are still rare but given the potential of the approach and its cost-effectiveness we believe that we are addressing a research direction whose importance will only grow in the future. As we are the first to comprehensively address *de novo* assembly with Linked-Reads, we believe our study will have considerable impact. We hope you will find it sufficiently compelling to send it out for review.

Sincerely,

Arend Sidow

Arend Sidow, Ph.D.
Professor of Pathology and of Genetics
SUMC R353
Stanford, CA 94305-5324

arend@stanford.edu
650-498-7024 (ph)
http://www.sidowlab.org
http://jimb.stanford.edu