Mate-pair Library Construction with Controlled Polymerization Enables Comprehensive Structural Rearrangement Detection

Mate-pair Library Construction with Controlled Polymerization Enables Comprehensive Structural Rearrangement Detection

Zirui Dong,[1,2#] Xia Zhao,[3,4#] Qiaoling Li,[3,4#] Zhenjun Yang,[1,2#] Yang Xi[3,4], Andrei Alexeev[5], Hanjie Shen,[3,4] Ou Wang,[3,4] Jie Ruan,[3,4] Han Ren,[3,4] Hanmin Wei,[7] Xiaojuan Qi,[3,4] Jiguang Li,[3,4] Xiaofan Zhu,[1,2] Dai Peng,[6] Kong Xiangdong,[6] Killeen Kirkconnell,[5] Oleg Alferov,[5] Shane Giles,[5] Jennifer Yamtich,[5] Bahram G. Kermani,[5] Chao Dong,[3,4] Pengjuan Liu[3,4], Zilan Mi,[3,4] Wenwei Zhang,[3,4] Xun Xu,[3,4,9] Radoje Drmanac,[3,4,5]* Kwong Wai Choy,[1,2,8*] Yuan Jiang[5]*

1. Department of Obstetrics and Gynaecology, The Chinese University of Hong Kong, Hong Kong, China
2. Shenzhen Research Institute, The Chinese University of Hong Kong, Shenzhen, China
3. BGI-Shenzhen, Shenzhen 518083, China
4. China National Genebank, BGI-Shenzhen, Shenzhen 518120, China
5. Advanced Genomics Technology Lab, Complete Genomics Inc., 2904 Orchard Pkwy, San Jose, California 95134, USA.
6. Genetics and Prenatal Diagnosis Center, the First Affiliated Hospital of Zhengzhou University, Zhengzhou 450052, China
7. MGI, BGI-Shenzhen, Shenzhen 518083, China.
8. The Chinese University of Hong Kong-Baylor College of Medicine Joint Center For Medical Genetics, Hong Kong, China

9. James D. Watson Institute of Genome Sciences, Hangzhou 310058, China

# These authors contributed equally to this work.

*Correspondence:

Dr. Yuan Jiang: yjiang@completegenomics.com, Dr. Kwong Wai Choy: richardchoy@cuhk.edu.hk, and Dr. Radoje Drmanac: rade@completegenomics.com

Abstract

It is important, yet difficult, to identify genomic structural rearrangements associated with congenital diseases or tumors. Mate-pair sequencing enables the positioning of a long DNA fragment with complete and precise breakpoints and has therefore become a common diagnostic approach for identifying chromosomal aberrations. Several methods are currently used for detection. However, due to cost, the need for large input quantity, and operation complexity, existing workflows are unsuitable for large-scale clinical studies. Herein, we describe a new process that couples advanced controlled polymerization with a non-conventional adapter ligation to generate mate-pairs with desirable length that yield minimal GC bias and improved coverage uniformity. Compared to other methods, our strategy can achieve 8-fold improved DNA circularization efficiency, a 39.3-fold reduction of read-pairs that do not cross the circularization junction, and the lowest chimeric rate, collectively producing an ~50% increase of physical coverage. In a proof-of-concept study using five insertion translocations, the structural rearrangements were comprehensively detected using longer 100-bp reads enabled by this approach. Based on its ability to identify single-nucleotide-resolution changes, this approach shows promise as an integrated method for the comprehensive detection of genomic variants at a fraction of current cost.

Keywords

Mate-pair Library, Controlled Polymerization, Chromosomal Structural Rearrangement

Introduction

Human structural variants (SVs) account for up to 300 Mb of sequence variation among individual human genomes and affect a variety of lengths ranging from 50 bp to millions of base pairs (1,2). SVs (i.e., deletions, duplications, and structural rearrangements) are commonly implicated in genetic disorders due to gene-dosage changes, gene(s) disruption, gene fusion or the dysregulation of disease-causing gene(s) (2-4). Thus, the identification of SVs, particularly for structural rearrangements, and the delineation of breakpoints at single-nucleotide resolution are essential for disease diagnosis and further management.

Recently, whole-genome sequencing (WGS) using paired-end analysis coupled with Sanger sequencing has enabled the molecular delineation of SVs (5-7). However, SVs are predominantly mediated by repetitive elements, which are commonly larger than 1 kb and are difficult to identify using standard WGS assays such as those with a small-insert library (300-500 bp) (2,6,8). To overcome this challenge (9), the sequencing of two ends from a large DNA fragment (i.e., 3-8 kb) was developed and named mate-pair library construction (5,7). This method enables the identification of unique sequences in the flanking regions of repetitive elements that potentially reveal precise SV breakpoint(s). In addition, data generated using mate-pair library construction shows increased physical coverage and thus decreases the minimum read depth requirement (i.e., 1 to 2-fold read depth or 8.25-fold physical depth) and facilitates clinical use; this approach was introduced as low-pass WGS in our previous study (7,8).

Biotin-labeling and blunt-end circularization (BLBEC) followed by random fragmentation has been the most common protocol used for mate-pair library construction in the past decade (5,7,10). However, BLBEC has several limitations: (a) the efficiency

of circularization through blunt-end ligation is low, and the DNA input requirement is high (3-20 μg); (b) read utility is low due to a high percentage of read pairs that do not cross the circularization junction [“inward” read-pairs (5,11)] or from PCR duplication; and (c) the reagent cost is as high as US $350, thus making clinical implementation impractical (5,7). To address these challenges, researchers have utilized adapters with a 2-bp or longer overhang for circularization and created two ~26-bp sequencing templates flanking the junction using the type II restriction enzyme *EcoP*15I (5,12). Although these modifications reduce library cost from $350 to $188 (5), 9% of chromosomal rearrangements remain undetected due to the short-read alignment (~26 bp) (3,13), and a high quantity of DNA is still required as input (i.e., 20 μg) (5). To obtain longer pair-end read length, researchers performed DNA extension by nick translation with two nicks (or gaps) on the opposite strands in double-stranded DNA circles (dsCirs) (14). The length of DNA extension was restricted by adjusting the reaction temperature and duration and followed by enzyme digestion with exonucleases. However, gel purification was required because a large size range of DNA templates existed both after initial fragmentation and after enzyme digestion (14,15). In contrast, to obtain longer read-length (i.e., 100-bp) and increase library process efficiency, researchers started to optimize the original BLBEC method by introducing Tn5 transposase for fragmentation and circularizing DNAs with sticky ends (16-19), but this resulted in a broad range of fragment sizes (16) and created a GC bias, thus reducing coverage uniformity (20,21). In addition, random fragmentation after DNA circularization commonly results in a high percentage of reads harboring adapter sequence (i.e., “adapter-contaminated”) further lowering down read utility. Finally, all of these protocols require the enrichment of biotin-labeled sequences,

resulting in poor process efficiency, high reagent cost and a high percentage of "inward" read-pairs (5,22).

Herein, we describe a cost-effective and high-efficiency method of mate-pair library construction by controlled polymerization (MpCP) during nick translation and primer extension followed by 3' branch ligation (3'BL), a special ligation reaction approach discovered in our previous study (23). Compared with previous approaches, our method produces 8-fold increased efficiency for fragment circularization and 50% increased read-pair utility. Given its capacity for the comprehensive identification of SVs, this workflow could make low-pass WGS feasible and robust for SV detection in the clinical setting.

Materials and Methods

Sample and Data Enrollment

The study protocol was approved by the Ethics Committee of BGI-Shenzhen and The Chinese University of Hong Kong. Informed consent was obtained from six subjects, each of whom have G-banded chromosome analysis results available (**Table 1**). G-banded chromosome analysis results indicated that five subjects were insertion translocation carriers, whereas the other individual had a normal karyotype (**Table 1**). For each subject, ~400 μl of peripheral blood was collected in an EDTA anticoagulant tube.

NA19240 genomic DNA was obtained from the Coriell Institute (Camden, NJ) (24) and used for method evaluation and optimization. In addition, whole-genome sequencing data from four reciprocal balanced translocations using BLBEC in our pilot study (**Table 1**) and six cases prepared with the Nextera Mate-pair Library Construction Kit (Illumina) obtained from the 1000 Genomes Project (25) were also used.

DNA Preparation and Qualification

Genomic DNA from peripheral blood was extracted with the QIAamp DNA blood mini kit (Qiagen, Hilden, Germany), quantified with the Quant-iT dsDNA HS Assay kit (Invitrogen, Carlsbad, CA, USA) and QC with gel electrophoresis. An aliquot of 1 μg DNA (OD260/OD280 > 1.8; OD260/OD230 > 1.5) from each sample was further sheared to a fragment size ranging from 3-8 kb by HydroShear (Digilab, Inc., Hopkinton, MA) using parameters reported in our pilot study (7).

Nick Translation by Limiting Nucleotide Quantities

After end-repair and the creation of an A-overhang, genomic DNA samples were ligated with adapter, Ad 1 (**Figure 1**). Subsequently, after annealing, PCR amplification was

performed with a pair of primers to generate double-stranded DNA (dsDNA) with adapters incorporating dUTPs in both strands (**Figure 2**). PCR products were then treated at 37°C for 60 min with 10 units of an UDG/EndoVIII cocktail [USER; New England Biolabs (NEB)] to create a 14-bp complementary overhang (**Supplementary Table 1**). Double-stranded DNA circles (dsCirs) were formed with two extendable gaps located 89 bp apart on opposite strands (**Figure 2**), and linear DNA was digested by plasmid safe (Epicentre) at 37°C for 1.5 hours. For comparison, five replicates were circularized by BLBEC using the method described in our pilot study (7). Circularization efficiency was calculated as the ratio of the DNA amount remaining (after linear DNA digestion) to the original DNA input.

Nick translation by Pol I (DNA polymerase I) was performed in a 20-μl reaction containing 0.23 pmol dsCirs, NEBuffer 2, 10 U DNA Polymerase I (NEB), and different concentrations (10, 5, or 3.33 μM) of dNTPs (Affymetrix, Inc, Santa Clara, CA) (**Figure 2**). The dNTP mix contained equimolar quantities of each dNTP (dATP, dTTP, dCTP, and dGTP). The reaction mixture was incubated at 10°C for 20 min. The products were purified with the MinElute PCR Purification Kit (Qiagen). Exonucleases were used for digestion and to form linear DNA (14). Size analysis was performed for the MinElute purified Exo-treated products on 6% polyacrylamide TBE gels (Invitrogen, Thermo Fisher, Waltham, MA) (**Figure 2**).

Primer Extension

In brief, the 3'-ends of genomic DNA (ranging from 700 to 3,000 bp) were ligated to the 3'-end of adapter, Ad2 (ON5 and ON6) and subsequently denatured at 96°C, annealed with a primer (ON10) at 56°C, and extended with Taq polymerase and serial dilutions of

dNTPs (33.3, 25, 20, 16.7 and 14.3 pmol) at 72°C for 10 min. Products were purified with Axygen beads and treated with Exonuclease VII, and bead purification was performed. Size analysis was performed with equivalent volumes on a 6% TBE gel (**Figure 3**).

Mate-pair Library Construction and Optimization

Mate-pair library construction on NA19240 DNA was conducted by combining controlled nick translocation and primer extension. To avoid repeating the amplification of each DNA fragment by nick translation, adapters with only one gap (in one strand instead of in both strands) were formed (**Figure 1**). After adapter ligation (Ad1) and PCR amplification, dsCir DNA was subjected to nick translation and primer extension in different conditions. In brief, 1.5 pmol of dsCirs for library #1, #2, and #3 were incubated with Polymerase I and different amounts of dNTP mix; for library #4, 0.45 U of Pol I and 0.45 U of Taq Polymerase were used instead.

To determine whether providing extra dATP and dTTP would improve data performance, an equal quantity of each dNTP was added (33 pmol of each) for library #1, whereas 33 pmol of dCTP and dGTP and 100 pmol of dATP and dTTP (three-fold of dCTP/dGTP) were added for each of the other three libraries. After nick translation, 3’BL was performed by mixing the products with the 3’-end of Ad2 (, ON5 and ON6) in a 120-µl of reaction volume, which was followed by bead purification.

Primer extension was performed by two methods in different libraries. Controlled primer extension by adjusting reaction temperature and time (ttCPE) was used for libraries #1 and #2; the reaction contained the 3’-BL product, Pfu Turbo Cx (Agilent Technologies), and primer ON10. The mixture was incubated at 92°C for 5 min, 56°C for

30 sec, and 60°C for 30 sec. Controlled primer extension by adjustment of nucleotide amount (naCPE) was used in libraries #3 and #4. The reaction consisted of 3'-BL products, Taq polymerase, primer ON10, and different amounts of dNTPs. The naCPE program was 96°C for 5 min, 56°C for 1 min, and 72°C for 5 min. A sample of 40 µl purified product from either ttCPE or naCPE was ligated to the 5'-end of Ad2 (ON7 and ON8) and purified with Axygen beads.

Libraries #1 and #4 were amplified using Q5 high-fidelity DNA polymerase. Library #2 was amplified using Pfu Turbo Cx. Library #3 was amplified using pyrophage polymerase (Lucigen, Middleton, WI). For all four libraries, single-stranded circles (ssCirs) and DNA nanoballs (DNBs) were prepared and subjected to improved cPAL sequencing on a nanoball (DNB)-based sequencing platform (Complete Genomics, Inc., Mountain View, CA). We thus obtained 28 bases (2xSE28) from the 5' ends of each adaptor, Ad1 and Ad2 (26).

Mate-pair Library Construction for Low-Pass Whole-Genome Sequencing

A sample of 1 µg genomic DNA was sheared by Hydroshear to 3-8 kb according to reported procedures (7,27) and purified with Agencourt AmpureXP beads. After end-repair, A-tailing, and DNA ligation with Ad1 (with a 10-bp barcode sequence,), 320 ng of ligated DNA was amplified with Pfu Turbo Cx (Agilent Technologies) and extra dNTPs in a 200-µl of reaction volume according to the manufacturer's directions (Pfu Turbo Cx, Agilent Technologies) to produce longer DNA fragments. After purification with Agencourt AmpureXP beads, 3 µg of each amplified DNA (six samples in total) was selected and pooled as one library. PCR products from the whole library were then

treated with USER (NEB) and T4 DNA ligase (Enzymatics, Inc., Qiagen) to form dsCir DNA with only one gap as described above.

Nucleotide amount controlled nick translation (naCNT) was performed with 1 pmol dsCir DNA; Bst DNA Polymerase, Full Length (NEB); Klenow fragment (Enzymatics, Inc.); and limited dNTPs. 3'BL was performed to ligate the 3'-end of Ad2 to the naCNT products. Primer extension was accomplished using ttCPE as described above, and the reaction mixture was incubated at 92℃ for 5 min, 56℃ for 60 sec, and 60℃ for 40 sec; the sample was then purified with Agencourt AmpureXP beads. ttCPE products were ligated to the 5'-end of Ad2 and amplified with Pfu Turbo Cx. ssCirs and DNBs were prepared and sequenced for a minimum of 70 million read-pairs in each sample with reads of 2x50 bp and 2x100 bp, respectively, on the BGISEQ-500 platform (BGI-Wuhan, Wuhan, China).

Data Analysis, Detection of Structural Variants, and Validation

Paired-end reads generated from a DNA nanoarray-sequencing platform (26) (Complete Genomics, Inc.) were aligned to the human reference genome (hg19) with a reported alignment method (28), while sequencing data with longer sequencing read-length reported in our pilot study or generated in low-pass WGS (**Table 1**) were aligned with BWA (29). Detection of single-nucleotide variants (SNVs) in the deep-sequencing data was carried out based on our reported study (24) and the fraction of genome/exome with fully called SNVs was calculated as the length of region with both alleles confidently called SNVs comparing to our reported datasets (24) dividing by the full length of genome/exome.

For evaluation of the performance in low-pass WGS, uniquely aligned read-pairs in both ends were selected in each sample for further analysis. Physical coverage was calculated as the sum of the aligned distance between each pair of reads that were not duplicated due to PCR amplification.

To determine whether GC bias exists during library construction and sequencing, the consistency of the GC percentage between the genome sequence and sequencing data was assessed. In brief, an equal number (N=50 millions) of aligned read-pairs was randomly selected from each sample (30) that was sequenced on the BGISEQ-500 platform with our MpCP approach and merged into a total dataset to avoid bias generated by genomic alterations in each particular sample. Adjustable sliding windows (i.e., 50-kb and 5-kb increments) were determined using an equal number of aligned read-pairs, and the GC percentage of each window was calculated based on the human reference genome (hg19). For each sample, the GC percentage in each window was calculated as the median GC percentage of all the read-pairs aligned to this window. The consistency of the GC percentage between the genome sequence and sequencing data was then assessed for each sample for comparison. Copy-number variant analysis was performed according to our published method (30). The identification of SVs was accomplished for each sample using our reported method based on a four-step procedure, including event clustering, systematic error filtering, random error filtering, and aligned orientations of each event (7,8).

Genomic reference sequences (hg19) surrounding each putative breakpoint (both upstream and downstream) were used for primer design with Primer3 and Primer-Blast (Biotechnology Information). PCR amplification was performed, and DNA from the YH

cell line a well-established normal control (8,31) was used as a negative control for validation of such rare rearrangement event in case sample. PCR products were subjected to Sanger sequencing on an ABI 3730 machine (Applied Biosystems, Thermo Fisher Scientific, Wilmington, DE), and sequencing results were aligned with BLAT for the confirmation of SV.

Data Access

Whole-genome sequencing data used in this study has been made available in EBI-ENA with the BioProject number PRJEB28073 (ERP110239) and in the CNGB Nucleotide Sequence Archive (CNSA: https://db.cngb.org/cnsa) under the accession number CNP0000078.

Results

A New Mate-pair Library Workflow: MpCP

For clinical labs to conduct large-scale tests, an inexpensive and scalable mate-pair sequencing process is crucial. To achieve this goal, we recently designed a simplified mate-pair library construction workflow called mate-pair by controlled polymerization (MpCP), which shows higher process efficiency, lower cost, and better sequencing quality. In the new schema (**Figure 1**), genomic DNA was first fragmented to the size of interest, ligated to an adapter (Ad1, **Supplementary Figure S1**), and circularized to form double-stranded circles (dsCirs). To improve the efficiency of DNA circle formation, 14-bp overhangs were used. The longer overhang resulted in an average 32% circularization efficiency in five replicates (data not shown) compared to only ~4% from five replicates prepared with the BLBEC protocol described previously (8,27).

The Ad1 dsCirs contained a gap that acted as the initiation site for a controlled polymerization reaction termed Controlled Nick Translation (CNT). Then, the 3'-end of Ad2 (Ad2_3') was directly added to the 3'OH termini of duplex DNA containing gaps using 3'BL (23). After ligation, another controlled polymerization reaction was performed on the single-stranded DNA (ssDNA) template and is therefore named Controlled Primer Extension (CPE). This enzymatic target enrichment strategy can efficiently eliminate the "inward read-pairs" that occur with the biotin capture method of other reported protocols (Table 1). In CPE, a primer was hybridized to Ad2_3' and extended onto the genomic template on the other side of Ad1. The 5'-end of Ad2 (Ad2_5') was then added through another 3'BL to the newly synthesized CPE template, as shown in **Figure 1**. Each template of the mate-pair has a selected length (an example

of a minimal pair-end length of 150 bp shown in **Supplementary Figure S2**), resulting from the CNT and CPE reactions, respectively and is separated by the Ad1 sequence with Ad2_5' and 3' sequences at either end. The ligation products were amplified using Ad2_5' and Ad2_3' primers resulting in a linear dsDNA library or a ssCir DNA library that can be sequenced by various technologies such as sequencing by synthesis (SBS) on the Illumina (32), Ion Torrent (33) or BGI's cPAS (34) platforms or sequencing by ligation (SBL) using the SOLiD (35) or first-generation CGI (26) platforms (**Supplementary Figure S3**).

Feasibility Testing of CP Techniques for MpCP Schema

As shown previously, polymerization speed is strongly dependent on polymerase Km and concentration, incubation time, and temperature (14). Here we tested another hypothesis that we could control DNA extension by limiting nucleotide amount. We first used dsCirs with two gaps flanking the adapter on opposite strands as templates to assay nick translation extension (**Figure 2a**). Three different amounts of dNTPs (2.6, 4.0 and 8.0 pmol) were supplied in the PolI-mediated CNT reactions, which resulted in different observed extension lengths (24, 30, and 62 bp, respectively) (**Figure 2b**), suggesting that nick translation can be controlled by nucleotide quantity. We calculated a predicted polymerization length per arm based on the ratio of total dNTPs to the number of naCNT initiation sites (ISs) (**Figure 2c**). For PolI-mediated reactions, observed extension lengths (24, 30, 62 bp) were close to predicted lengths (23, 35, 70 bp; **Figure** 2b, c). It appears that PolI can utilize most, if not all, of the dNTPs in the nucleotide amount controlled nick translation (naCNT) reactions. Based on this observation, if we normalize the dsCir quantity to 1 pmol in the MpCP schema, 100 pmol of total dNTPs will be needed in

naCNT for an extension of approximately 100 bp. We also found that naCNT products contain gaps of a few bp, because the 3' exonuclease activity of PolI can cleave a few nucleotides from the nicks in a 5' to 3' direction (data not shown).

To replace the complicated biotin capture method used to enrich targets in other protocols, we tried to use CPE to recover products of interest. We hypothesize that, as for CNT reactions, there are two ways to control primer extension length—by limiting the nucleotide amount (naCPE) or by managing the reaction time and temperature (ttCPE). We first assessed whether limiting dNTPs could control DNA extension length (**Figure 3**). The naCPE template was 0.16 pmol of an 700-3,000 bp genomic DNA fraction with 3'-ends ligated to an adapter sequence. **Figure 3b** demonstrates that different naCPE reactions generated different sizes of primer-extended products (lane 1-5 in Figure 3), and the size correlates with the dNTP amount. Excess dNTPs (5,000 pmol, lane 6) generated products of the original size range (700-3,000 bp). Reactions performed with the lowest quantity of dNTPs (14.3 pmol, lane 5 in Figure 3) generated the smallest products (210-300 bp). For ttCPE, the same amounts of MpCP intermediate products after naCNT and 3'BL were extended for different times in the presence of excess dNTPs. An extension of 300-450 bp was observed after a combination of 56℃ for one minute and 60℃ for 40 seconds with Pfu Turbo Cx (data not shown). After CPE, a partially duplex DNA with a 3' recessive end is produced. 3'BL can efficiently add adapters to gapped or 3'-recessed duplex DNA (23). Therefore, both naCNT and CPE products are perfect templates for 3'BL.

MpCP Library Optimization

We next compared different combinations of CP conditions and assayed the sequencing quality of the MpCP libraries using improved cPAL technology from previously published work (26). To minimize the coverage and GC bias in naCNT, two approaches were used (**Table 2**). First, more than one polymerase, such as a combination of PolI with Taq Pol, was included to process through most, if not all, polymerase-specific pausing sites to minimize bias. Second, dATP and dTTP (3xAT) were provided in a three-fold excess relative to dCTP and dGTP, because the human genome contains a higher percentage of A and T.

Four libraries with different CP conditions were constructed using NA19240 Coriell genomic DNA and sequenced at a read depth of 50-60 fold using two single-end sequences of 28 bp (2xSE28, See Material and Method). The mapping data are summarized in **Table 2**. Library #1 prepared with equal amounts of the four dNTPs during naCNT demonstrated the highest fraction (6.6%) of autosomal exome at <60% of the mean coverage and the lowest percentage (95.0%) of genome coverage with fully called SNVs (24) compared to the other three libraries prepared with 3-fold excess A and T during naCNT. This result indicated that the extra A and T provided during naCNT could improve the uniformity of coverage distribution by reducing the GC bias. No significant difference of genome or exome coverage was observed between library #3 and library #4, indicating no particular effect of using a combination of polymerases in naCNT. In addition, similar percentages of genome and exome coverage were observed for library #2 using ttCPE and libraries #3 and #4 using naCPE, suggesting that ttCPE and naCPE are interchangeable. However, because it is difficult to precisely quantify the branched templates after naCPE and 3' BL, ttCPE was adopted for further experiments.

Taken together, we incorporated naCNT with 3xAT and ttCPE into our optimized MpCP workflow because this workflow demonstrated minimal GC and coverage bias.

Validation of MpCP Workflow with Low-Pass WGS for SV Detection

One important application of sequencing with mate-pair libraries is the comprehensive detection of chromosomal structural rearrangements or variants using low-pass (or low-coverage) WGS. In this study, six samples for which G-banded chromosome analysis is available were used for MpCP library construction (1 µg input for fragmentation with end-products ranging from 3-8 kb) and sequencing for a minimum of 70 million paired-end reads (50 bp). Consistent with the previous performance of deep-sequencing, a minimal GC bias was indicated by the significantly high concordance (r>0.98) between the GC percentage in the human reference genome (hg19) and that of the aligned reads for each of the six samples prepared with MpCP (**Figure 4**).

To evaluate the performance of the MpCP method, WGS reads from four samples sequenced with non-size-selected BLBEC libraries on the HiSeq 2000 platform (7) to approximately 50 million read-pairs (50 bp, PE50) were used for comparison. First, we compared the read utility when MpCP was performed with the same number of read-pairs (50 million). Approximately 91.5% of the "inward" read-pairs in the WGS with BLBEC libraries were aligned to a distance <1 kb (**Supplementary Figure S4**), while, based on the schema design, MpCP would not generate such read-pairs with inconsistent aligned orientations (**Figure 1**). Here, we compared the percentages of read-pairs with an aligned distance <1 kb in each method. Consistent to our hypothesis, the results showed that MpCP produced only ~0.7% of read-pairs with an aligned distance <1 kb, compared to

the ~27.5% observed with BLBEC (**Table 1 and Figure 5A**). Even though only ~5.8% of these read-pairs remained for BLBEC when only considering uniquely aligned and non-duplicated read-pairs (**Figure 5B**), ~33% lower physical coverage was obtained using BLBEC compared to MpCP (**Figure 5C**), indicating a much higher utility of sequencing data using MpCP. Of note, a slightly higher duplication rate was observed in samples prepared with MpCP compared to those prepared with BLBEC, which might be due to the reduction of DNA input from 3 µg to 1 µg (**Table 1**). In addition, to evaluate performance using shorter reads such as the data generated using *EcoP*15I, the read-pairs from each sample were reduced to 26 bp (PE26) and subjected to analysis with the same parameters. Although no difference of the genome-wide physical coverage rate was seen between the two different library construction methods for the PE50 data, the coverage was 4% higher with MpCP compared to BLBEC with the 26-bp reads (**Table 1**); this increased coverage can be attributed primarily to the higher utility obtained from the MpCP workflow.

In addition, the same analysis was performed with 12 sets of whole-genome sequencing data prepared using the Nextera Mate-pair Library Construction Kit (25), which is an updated version of BLBEC that uses adaptors for circularization instead of blunt ends and is sequenced in paired-end mode with >100-bp reads. However, 24.4% of read-pairs were found to have < 1 kb of aligned distance in total for each sample, and 80.8% of these read-pairs were uniquely aligned from both ends (**Figure 5D**). Moreover, by searching with the Tn5 sequence (36) in all the sequencing reads, ~23.7% and ~23.1% of read 1 (or 1st read) and read 2 (or 2nd read) were found to have the full length of such

sequence classified as “adapter-contaminated” reads, causing a large reduction of read utility.

Next we evaluated ability of MpCP to detect SVs. Using PE50 sequencing for each sample, five apparent insertion translocations were identified with 24 linkage events in total. The affected bands were consistent with original karyotypes. To evaluate the feasibility of detecting SVs using shorter reads in MpCP, the read-pairs from each sample were also trimmed to 26 bp and subjected to the same analysis. Unsurprisingly, the detection results from this “new” dataset only show 22 clusters from these five samples (91.6%, 22/24, **Figure 6**). Comparing the alignment results of the read-pairs originally supporting the two “missed” clusters, all supporting read-pairs appeared to have at least one read become non-uniquely aligned when the read length was shortened. In addition, 183 out of 553 supporting read-pairs from the other 22 clusters also have at least one end become non-uniquely aligned with the shorter read length, thus reducing the average number of supporting read-pairs in each cluster from 25.1 to 16.8 (33.1% reduction).

Although the identified events using PE50 were apparently consistent with the cytogenetic results, the linkage events with one end of four segments remained undetected (**Figure 6C**). To specifically map the breakpoints and demonstrate the feasibility and advantage of obtaining a longer sequencing template with MpCP, we performed paired-end 100-bp (PE100) sequencing for each of these six cases. Interestingly, another two clusters/linkages supporting the “missed” four breakpoints were identified, providing 26 linkage events in total. Compared to the results identified with paired-end 100-bp sequencing, the diagnostic yields of PE50 and PE26 were only 92.3% (24/26) and 84.6% (22/26), respectively. This result indicates that the longer read

length attained in this study (100 bp) could enable the comprehensive detection of chromosomal structural rearrangements, particularly complex rearrangement events. Of note, in these five positive cases, compared with the original karyotypes, all events were classified as complex rearrangements (>3 breakpoints), which are beyond the resolution of G-banded chromosome analysis (**Table 1**).

Discussion

Mate-pair library construction is a commonly used protocol for detecting structural rearrangements/variants and delineating the associated breakpoints; its popularity can be attributed to its ability to accomplish the following: (1) it can identify unique sequences in the flanking regions of repetitive elements that potentially reveal the precise breakpoint(s); (2) it can increase physical coverage and thus decrease the cost per test. However, problems such as a high DNA input requirement and high cost limit its clinical application (**Table 3**) (5,14). In this study, we developed a cost-effective and high-efficiency mate-pair library construction method based on advanced controlled DNA polymerization during both nick translation and primer extension with a novel adapter ligation reaction recently developed by some of the current authors (23).

Compared to previously reported approaches (3,5,7,14,18,37), our method includes the following four improvements: (1) using an adapter with a 14-bp overhang increases circularization efficiency (from ~4% to ~32% by 8-fold), thus decreasing DNA input requirement; (2) no gel purification is required; (3) obtaining the DNA template directly from the primer extension reaction after nick translation instead of using biotin enrichment results in a much lower percentage of read-pairs with an aligned distance < 1 kb (0.7%, or a 39.3-fold reduction from 27.5% for BLBEC; 24.4% with the Nextera Mate-pair Library Construction Kit). In addition, ~23% of "adapter-contaminated" reads from Nextera were found due to random fragmentation after DNA circularization, which further decreases the read utility efficiency; (4) generating a longer DNA read length (100 bp or longer) by controlled polymerization instead of template selection with *EcoP*15I resulted in an 18.2% increased detection rate for structural rearrangement events (four

additional linkage events detected from 22). Furthermore, by providing a 3-fold excess of dATPs/dTTPs in the nick translation procedure, the percentage of autosomal exome with <60% of mean coverage was improved from 6.6% to 0.7%, whereas the coding region with all SNPs called increased from 95.9% to 98.1% (**Table 2**), thus demonstrating more uniformly distributed genome coverage. In addition, a significantly high concordance (r>0.98) of the GC percentage between the reference genome and the aligned reads was observed in each sample (**Figure 5**), suggesting that the GC bias generated by this method is minimal compared to the existing methods (21). With these improvements, an approximately 50% increased physical coverage was obtained, which reduces the minimum read requirements. Although MpCP can generate paired-end lengths longer than 150 bp (**Supplementary Figure S2**), we only performed PE100 to keep sequencing cost low and decrease turn-around-time while examining its feasibility in clinical applications.

In this study, we have demonstrated the effectiveness of controlling polymerization during nick translation and primer extension with 3'BL, a special ligation event. Although studies show that nick translation can be performed by adjusting reaction temperature and duration prior to exonuclease digestions, a gel purification step is required due to the large range of templates generated (14). In this study, we have demonstrated the feasibility and robustness of performing nick translation by adjusting the dNTP quantities with a fixed reaction temperature and duration. Furthermore, we demonstrated the feasibility of performing primer extension by either adjusting the dNTP amount (naCPE) or adjusting reaction temperature and duration (ttCPE). Although adjusting dNTP levels seems less laborious than ttCPE, it is difficult to precisely quantify

the ssDNA templates for naCPE after the 3'BL of the naCNT products. Therefore, for low-pass WGS, ttCPE was adopted instead of naCPE. With this combination, a narrow range of DNA inserts is obtained by eliminating the "inward" read-pairs and the small DNA fragments (<1 kb). This approach will not only be useful to increase the read utility for SV detection in low-pass WGS applications but would also be very useful for providing the true distance when building up scaffolds/contigs in *de novo* assembly (22,31), another critical application of mate-pair library construction. Last, but not least, the controlled primer extension reactions created by successfully combining nick translation with 3'BL can eliminate the use of exonuclease digestion and/or streptavidin beads, which is costly and inefficient.

For low-pass WGS, only a limited number of read-pairs (50-70 million) are required for analysis in each sample; a pooling step is suitable to increase the sample volume in each batch (7). However, in this study, due to the flexibility of sequencing orientation for ssDNA, pooling was performed for the DNA soon after adding adapters incorporating dUTPs and 10-bp barcode sequences instead of completing the whole process of library construction (**Figure 1** and **Supplementary Figure S1**). This change will reduce the labor cost for all subsequent procedures for each particular case by treating the pooled library as a single sample, greatly increasing the efficiency of library construction by handling more cases simultaneously. Furthermore, eliminating gel purification or biotin capture throughout the entire process makes MpCP a highly cost-effective and efficient approach for mate-pair library construction.

Due to all the improvements described above, the reagent cost for this approach is only ~1/10th that of biotin-labeled blunt-end circularization (US $40 vs $350, **Table 3**)

(5), which makes low-pass WGS more practical for routine clinical SV detection. Although these great advances enable the application of low-pass WGS, several limitations of our workflow remain: (1) PCR amplification with primers incorporating dUTPs is used, and this will become inefficient for DNA templates >10 kb, thereby limiting the utility of this method for DNA libraries with longer fragment sizes; and (2) DNA samples used for this study were fragmented physically (Hydroshear); alternative methods such as fragmentation by Transposases (i.e., Tn5) (38) would reduce labor cost and turn-around-time. Further investigation is warranted for these types of improvements.

In conclusion, we provide a cost-effective and high-efficiency method of mate-pair library construction that provides higher utility of read-pairs, better genome coverage uniformity, and improved detection of SVs by providing longer read length.

References

1. Sudmant, P.H., Rausch, T., Gardner, E.J., Handsaker, R.E., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Hsi-Yang Fritz, M. *et al.* (2015) An integrated map of structural variation in 2,504 human genomes. *Nature*, **526**, 75-81.
2. Carvalho, C.M. and Lupski, J.R. (2016) Mechanisms underlying structural variant formation in genomic disorders. *Nat Rev Genet*, **17**, 224-238.
3. Redin, C., Brand, H., Collins, R.L., Kammin, T., Mitchell, E., Hodge, J.C., Hanscom, C., Pillalamarri, V., Seabra, C.M., Abbott, M.A. *et al.* (2017) The genomic landscape of balanced cytogenetic abnormalities associated with human congenital anomalies. *Nat Genet*, **49**, 36-45.
4. Latysheva, N.S. and Babu, M.M. (2016) Discovering and understanding oncogenic gene fusions through data intensive computational approaches. *Nucleic Acids Res*, **44**, 4487-4503.
5. Talkowski, M.E., Ernst, C., Heilbut, A., Chiang, C., Hanscom, C., Lindgren, A., Kirby, A., Liu, S., Muddukrishna, B., Ohsumi, T.K. *et al.* (2011) Next-generation sequencing strategies enable routine detection of balanced chromosome rearrangements for clinical diagnostics and genetic research. *Am J Hum Genet*, **88**, 469-481.
6. Chen, W., Kalscheuer, V., Tzschach, A., Menzel, C., Ullmann, R., Schulz, M.H., Erdogan, F., Li, N., Kijas, Z., Arkesteijn, G. *et al.* (2008) Mapping translocation breakpoints by next-generation sequencing. *Genome Res*, **18**, 1143-1149.
7. Dong, Z., Jiang, L., Yang, C., Hu, H., Wang, X., Chen, H., Choy, K.W., Hu, H., Dong, Y., Hu, B. *et al.* (2014) A robust approach for blind detection of balanced chromosomal rearrangements with whole-genome low-coverage sequencing. *Hum Mutat*, **35**, 625-636.
8. Dong, Z., Wang, H., Chen, H., Jiang, H., Yuan, J., Yang, Z., Wang, W.J., Xu, F., Guo, X., Cao, Y. *et al.* (2017) Identification of balanced chromosomal rearrangements previously unknown among participants in the 1000 Genomes Project: implications for interpretation of structural variation in genomes and the future of clinical cytogenetics. *Genet Med*.
9. Goodwin, S., McPherson, J.D. and McCombie, W.R. (2016) Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet*, **17**, 333-351.
10. Collins, R.L., Brand, H., Redin, C.E., Hanscom, C., Antolik, C., Stone, M.R., Glessner, J.T., Mason, T., Pregno, G., Dorrani, N. *et al.* (2017) Defining the diverse spectrum of inversions, complex structural variation, and chromothripsis in the morbid human genome. *Genome Biol*, **18**, 36.
11. Hampton, O.A., English, A.C., Wang, M., Salerno, W.J., Liu, Y., Muzny, D.M., Han, Y., Wheeler, D.A., Worley, K.C., Lupski, J.R. *et al.* (2017) SVachra: a tool to identify genomic structural variation in mate pair sequencing data containing inward and outward facing reads. *BMC Genomics*, **18**, 691.
12. Hanscom, C. and Talkowski, M. (2014) Design of large-insert jumping libraries for structural variant detection using Illumina sequencing. *Curr Protoc Hum Genet*, **80**, 7 22 21-29.
13. Li, W. and Freudenberg, J. (2014) Mappability and read length. *Front Genet*, **5**, 381.
14. van Heesch, S., Kloosterman, W.P., Lansu, N., Ruzius, F.P., Levandowsky, E., Lee, C.C., Zhou, S., Goldstein, S., Schwartz, D.C., Harkins, T.T. *et al.* (2013) Improving mammalian genome scaffolding using large insert mate-pair next-generation sequencing. *BMC Genomics*, **14**, 257.

15. Vergult, S., Van Binsbergen, E., Sante, T., Nowak, S., Vanakker, O., Claes, K., Poppe, B., Van der Aa, N., van Roosmalen, M.J., Duran, K. *et al.* (2014) Mate pair sequencing for the detection of chromosomal aberrations in patients with intellectual disability and congenital malformations. *Eur J Hum Genet*, **22**, 652-659.
16. Heavens, D., Accinelli, G.G., Clavijo, B. and Clark, M.D. (2015) A method to simultaneously construct up to 12 differently sized Illumina Nextera long mate pair libraries with reduced DNA input, time, and cost. *Biotechniques*, **59**, 42-45.
17. Murphy, R.R., O'Connell, J., Cox, A.J. and Schulz-Trieglaff, O. (2015) NxRepair: error correction in de novo sequence assembly using Nextera mate pairs. *PeerJ*, **3**, e996.
18. Nilsson, D., Pettersson, M., Gustavsson, P., Forster, A., Hofmeister, W., Wincent, J., Zachariadis, V., Anderlid, B.M., Nordgren, A., Makitie, O. *et al.* (2017) Whole-genome sequencing of cytogenetically balanced chromosome translocations identifies potentially pathological gene disruptions and highlights the importance of microhomology in the mechanism of formation. *Hum Mutat*, **38**, 180-192.
19. Weckselblatt, B., Hermetz, K.E. and Rudd, M.K. (2015) Unbalanced translocations arise from diverse mutational mechanisms including chromothripsis. *Genome Res*, **25**, 937-947.
20. Leggett, R.M., Clavijo, B.J., Clissold, L., Clark, M.D. and Caccamo, M. (2014) NextClip: an analysis and read preparation tool for Nextera Long Mate Pair libraries. *Bioinformatics*, **30**, 566-568.
21. Marine, R., Polson, S.W., Ravel, J., Hatfull, G., Russell, D., Sullivan, M., Syed, F., Dumas, M. and Wommack, K.E. (2011) Evaluation of a transposase protocol for rapid generation of shotgun high-throughput sequencing libraries from nanogram quantities of DNA. *Appl Environ Microbiol*, **77**, 8071-8079.
22. Chapman, J.A., Ho, I., Sunkara, S., Luo, S., Schroth, G.P. and Rokhsar, D.S. (2011) Meraculous: de novo genome assembly with short paired-end reads. *PLoS One*, **6**, e23501.
23. Wang, L., Xi, Y., Zhang, W., Wang, W., Shen, H., Wang, X., Zhao, X., Alexeev, A., Peters, B., Albe, A. *et al.* 3′ Branch Ligation: A Novel Method to Ligate Non-Complementary DNA to Recessed or Internal 3′ OH Ends in DNA or RNA.
24. Peters, B.A., Kermani, B.G., Sparks, A.B., Alferov, O., Hong, P., Alexeev, A., Jiang, Y., Dahl, F., Tang, Y.T., Haas, J. *et al.* (2012) Accurate whole-genome sequencing and haplotyping from 10 to 20 human cells. *Nature*, **487**, 190-195.
25. Zook, J.M., Catoe, D., McDaniel, J., Vang, L., Spies, N., Sidow, A., Weng, Z., Liu, Y., Mason, C.E., Alexander, N. *et al.* (2016) Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Sci Data*, **3**, 160025.
26. Drmanac, R., Sparks, A.B., Callow, M.J., Halpern, A.L., Burns, N.L., Kermani, B.G., Carnevali, P., Nazarenko, I., Nilsen, G.B., Yeung, G. *et al.* (2010) Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science*, **327**, 78-81.
27. Dong, Z., Ye, L., Yang, Z., Chen, H., Yuan, J., Wang, H., Guo, X., Li, Y., Wang, J., Chen, F. *et al.* (2018) Balanced Chromosomal Rearrangement Detection by Low-Pass Whole-Genome Sequencing. *Curr Protoc Hum Genet*, **96**, 8 18 11-18 18 16.
28. Carnevali, P., Baccash, J., Halpern, A.L., Nazarenko, I., Nilsen, G.B., Pant, K.P., Ebert, J.C., Brownley, A., Morenzoni, M., Karpinchyk, V. *et al.* (2012) Computational techniques for human genome resequencing using mated gapped reads. *J Comput Biol*, **19**, 279-292.
29. Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754-1760.

30. Dong, Z., Zhang, J., Hu, P., Chen, H., Xu, J., Tian, Q., Meng, L., Ye, Y., Wang, J., Zhang, M. *et al.* (2016) Low-pass whole-genome sequencing in clinical cytogenetics: a validated approach. *Genet Med*, **18**, 940-948.
31. Cao, H., Wu, H., Luo, R., Huang, S., Sun, Y., Tong, X., Xie, Y., Liu, B., Yang, H., Zheng, H. *et al.* (2015) De novo assembly of a haplotype-resolved human genome. *Nat Biotechnol*, **33**, 617-622.
32. Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, J., Brown, C.G., Hall, K.P., Evers, D.J., Barnes, C.L., Bignell, H.R. *et al.* (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, **456**, 53-59.
33. Rothberg, J.M., Hinz, W., Rearick, T.M., Schultz, J., Mileski, W., Davey, M., Leamon, J.H., Johnson, K., Milgrew, M.J., Edwards, M. *et al.* (2011) An integrated semiconductor device enabling non-optical genome sequencing. *Nature*, **475**, 348-352.
34. Fehlmann, T., Reinheimer, S., Geng, C., Su, X., Drmanac, S., Alexeev, A., Zhang, C., Backes, C., Ludwig, N., Hart, M. *et al.* (2016) cPAS-based sequencing on the BGISEQ-500 to explore small non-coding RNAs. *Clin Epigenetics*, **8**, 123.
35. Valouev, A., Ichikawa, J., Tonthat, T., Stuart, J., Ranade, S., Peckham, H., Zeng, K., Malek, J.A., Costa, G., McKernan, K. *et al.* (2008) A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning. *Genome Res*, **18**, 1051-1063.
36. Van Nieuwerburgh, F., Thompson, R.C., Ledesma, J., Deforce, D., Gaasterland, T., Ordoukhanian, P. and Head, S.R. (2012) Illumina mate-paired DNA sequencing-library preparation using Cre-Lox recombination. *Nucleic Acids Res*, **40**, e24.
37. Talkowski, M.E., Ordulu, Z., Pillalamarri, V., Benson, C.B., Blumenthal, I., Connolly, S., Hanscom, C., Hussain, N., Pereira, S., Picker, J. *et al.* (2012) Clinical diagnosis by whole-genome sequencing of a prenatal sample. *N Engl J Med*, **367**, 2226-2232.
38. Picelli, S., Bjorklund, A.K., Reinius, B., Sagasser, S., Winberg, G. and Sandberg, R. (2014) Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. *Genome Res*, **24**, 2033-2040.

Conflicts of interest:

Employees of BGI and Complete Genomics have stock holdings in BGI. The other authors declare no conflicts of interest.

Acknowledgements

We would like to acknowledge the ongoing contributions and support of all Complete Genomics and BGI-Shenzhen employees, in particular the many highly skilled individuals that work in the libraries, reagents, and sequencing groups that make it possible to generate high-quality whole genome data. We thank Dr. Ya Gao for helpful discussions surrounding manuscript preparation. This work was supported in part by the National Key Research and Development Program of China (2017YFC0906500), the Shenzhen Peacock Plan (No.KQTD20150330171505310), and the Shenzhen Municipal Government of China (No. GJHS23017014152802146).

Table 1. Sample or data sequenced with paired-end 50-bp or trimmed 26-bp reads in this study

Sample/ data	Karyotype	Sequencing results	Library construction method and sequencing platform[2]	DNA input (μg)	Chimeric rate (%)[4]	Duplication rate (%)	Aligned distance <1 kb (%)	Aligned distance <1 kb (% uniquely aligned and non-duplicated)	Fraction of genome physically covered (%)
BCA01[1]	46,XX,t(2;20)(p13;p13)	46,XX,t(2;20)(p15;p13)	BLBEC+HS	3	9.04	15.7	27.6	5.2	98.6|91.1[3]
BCA02[1]	46,XX,t(2;3)(p21;p25)	46,XX,t(2;3)(p22.1;p26.1)	BLBEC+HS	3	10.92	15.9	22.3	3.9	98.5|93.7[3]
BCA11[1]	46,XX,t(4;5;14)(q31.3;q33;q13)	46,XX,t(4;5;14)(q32.3;q34;q21.3)	BLBEC+HS	3	6.70	12.5	29.7	6.0	96.7|91.9[3]
BCA16[1]	46,XX,t(14;18)(q24;q23)	46,XX,t(14;18)(q13.2;q22.1)	BLBEC+HS	3	8.04	12.0	30.5	8.0	96.7|91.2[3]
Sample01	46,XX,ins(10;13)(q11.2;q31q33)	46,XX,ins(10;13)(q21.3;q21.2q31.3)	MpCP+BS	1	2.84	16.8	0.6	0.1	97.7|96.6[3]
Sample02	46,XY,ins(6;2)(q23;p13p22)	46,XY,ins(6;2)der(6)(6pter->6q13::6q13->6q21.3::2p16.1<-2p16.1::6q21->6q22.31::2p16.1->2p11.2::2p22.2->2p16.1::6q22.31->6q26::6q13<-6q13::6q26->6qter);der(2)(2pter->2p22.2::2p11.2->2qter) seq[hg19] 6p21.31p22.1(2915033_33078330)x3	MpCP+BS	1	2.82	17.8	0.8	0.2	96.2|94.9[3]

[1] low-pass WGS with BLBEC was published in our pilot study.

[2] BLBEC and MpCP indicate library construction based on biotin-labeled blunt-end circularization and controlled polymerization during nick translation and primer extension, respectively; HS and BS refer to HiSeq 2000 (Illumina) and BGISeq 500 (BGI-Wuhan), respectively.

[3] The numbers before and after each vertical line reflect the fraction of the genome physically coverage rate by the data with paired-end 50-bp and trimmed 26-bp reads, respectively.

[4] Chimeric read-pairs were defined as the read-pairs aligned to different chromosomes or to the same chromosomes but with a distance larger than 10 kb.

Continued Table 1. Sample or data sequenced with paired-end 50-bp or trimmed 26-bp reads in this study

Sample/ data	Karyotype	Sequencing results	Library construction method and sequencing platform[2]	DNA input (μg)	Chimeric rate (%)[4]	Duplication rate (%)	Aligned distance <1 kb (%)	Aligned distance <1 kb (% uniquely aligned and non-duplicated)	Fraction of genome physically covered (%)
Sample03	46,XX,ins(2;18)(q31;q21.1q23)	46,XX,ins(2;18)(q32.2;q21.1q22.1);inv(18)(q21.1)	MpCP+BS	1	3.52	17.5	0.5	0.1	98.0|97.0[3]
Sample04	46,XY,der(4)ins(4)(q21.1;q31.1q31.3)inv(4)(p12q13.3)	46,XY,der(4)(pter->p12::q13.3<-p12::q31.1->q31.23::q13.3->q31.1::q31.23->qter)	MpCP+BS	1	2.74	19.0	0.9	0.4	96.3|95.0[3]
Sample05	46,XY,ins(6;3)(q13;q21q24)	46,XY,ins(6;2)der(6)(6pter->6q13::6q13->6q21.3::2p16.1<-2p16.1::6q21->6q22.31::2p16.1->2p11.2::2p22.2->2p16.1::6q22.31->6q26::6q13<-6q13::6q26->6qter);der(2)(2pter->2p22.2::2p11.2->2qter)	MpCP+BS	1	2.87	18.9	0.7	0.2	96.1|95.0[3]
Sample06	46,XX	46,XX	MpCP+BS	1	2.81	18.9	0.7	0.2	98.1|97.0[3]

[1] low-pass WGS with BLBEC was published in our pilot study.

[2] BLBEC and MpCP indicate library construction based on biotin-labeled blunt-end circularization and controlled polymerization during nick translation and primer extension, respectively; HS and BS refer to HiSeq 2000 (Illumina) and BGISEQ 500 (BGI-Wuhan), respectively.

[3] The numbers before and after each vertical line reflect the fraction of the genome physically coverage rate by the data with paired-end 50-bp and trimmed 26-bp reads, respectively.

[4] Chimeric read-pairs are defined as the read-pairs aligned to different chromosomes or to the same chromosomes but with a distance larger than 10 kb.

Table 2. Evaluation of different parameters/conditions for mate-pair library construction

Library	Nick translation	Primer extension[1]	% autosomal exome by GC with < 60% of mean coverage[2]	% fully called genome fraction[3]	% fully called coding sequence fraction[3]
#1	PolI	ttCPE by PfuCx	6.6	95.0	95.9
#2	PolI+3xAT	ttCPE by PfuCx	1.8	96.2	97.8
#3	PolI+3xAT	naCPE by Taq	0.7	96.7	98.1
#4	PolI/Taq+3xAT	naCPE by Taq	1.4	96.1	97.8

[1] ttCPE and naCPE refer to controlled primer extension by adjusting of reaction temperature and duration time or limiting the nucleotide input only

[2] % autosomal exome by GC with <60% of mean coverage measures the fraction of the autosomal exome regions where the normalized coverage of exome by cumulative base GC percentage is less than 60% of the mean coverage. Higher percentage indicates higher GC bias in the autosomal exome regions.

[3] fully called genome and coding sequence fractions; both alleles can be confidently called by comparing to our reported study (24).

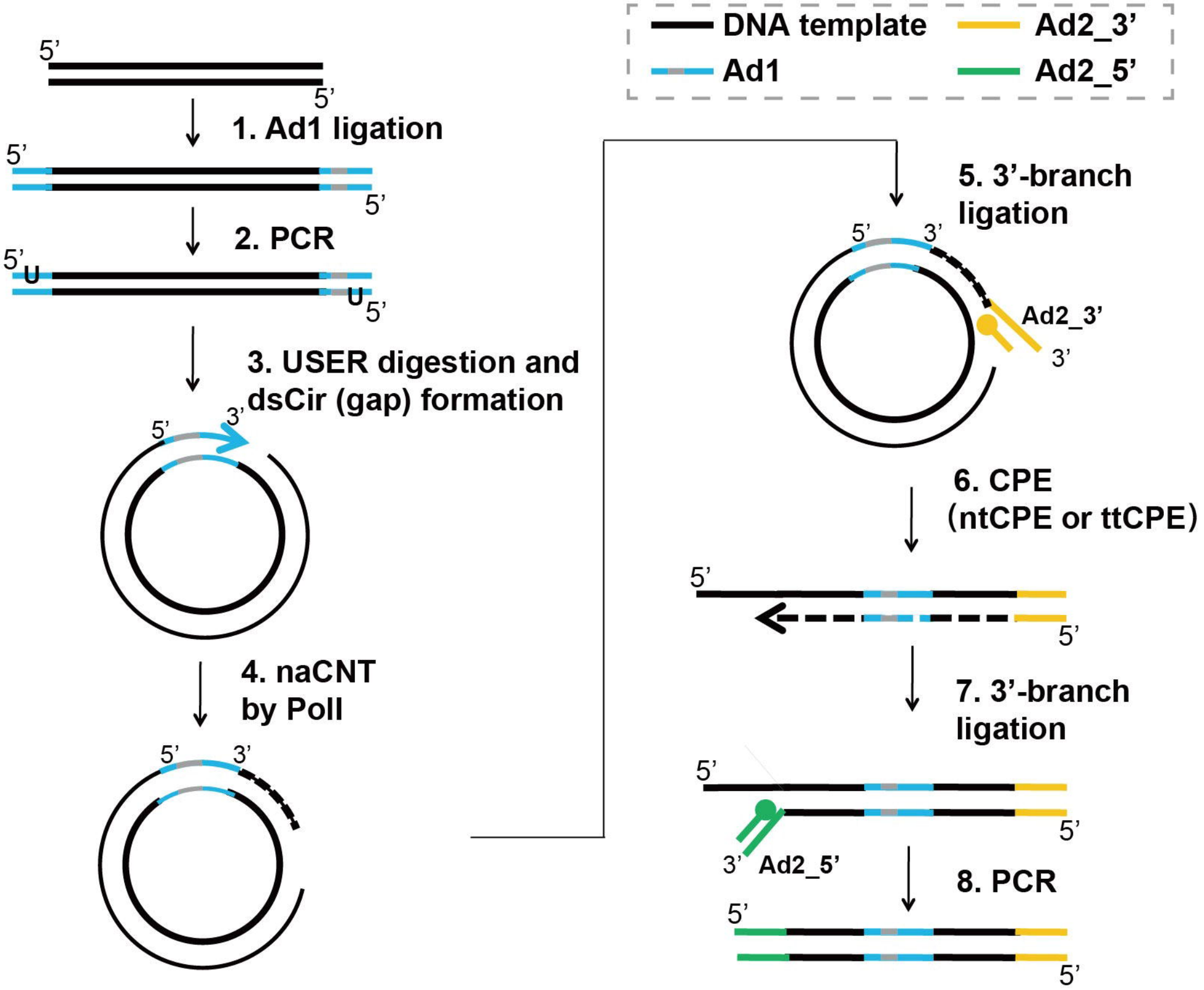
Table 3. Method Comparison between Reported Approaches and Our Workflow

Method[1]	Reagent cost (USD)[2]	Turn-around-time (days)	DNA input amount (μg)	Chimeric rate (%)	Longest read-length (bp)
BLBEC	350	3	3|20[3]	~8.7	>100
Custom Jumping Library	188	3	5~20	-	~26
ttCNT	~4,000/12 samples	2	1~5	24.0-58.4	60
Nextera	4,233/48 gel-free samples or 12 gel-plus samples	1.5~2	1|4[3]	~3.0	>100
MpCP	~40	3	1	~2.9	>100

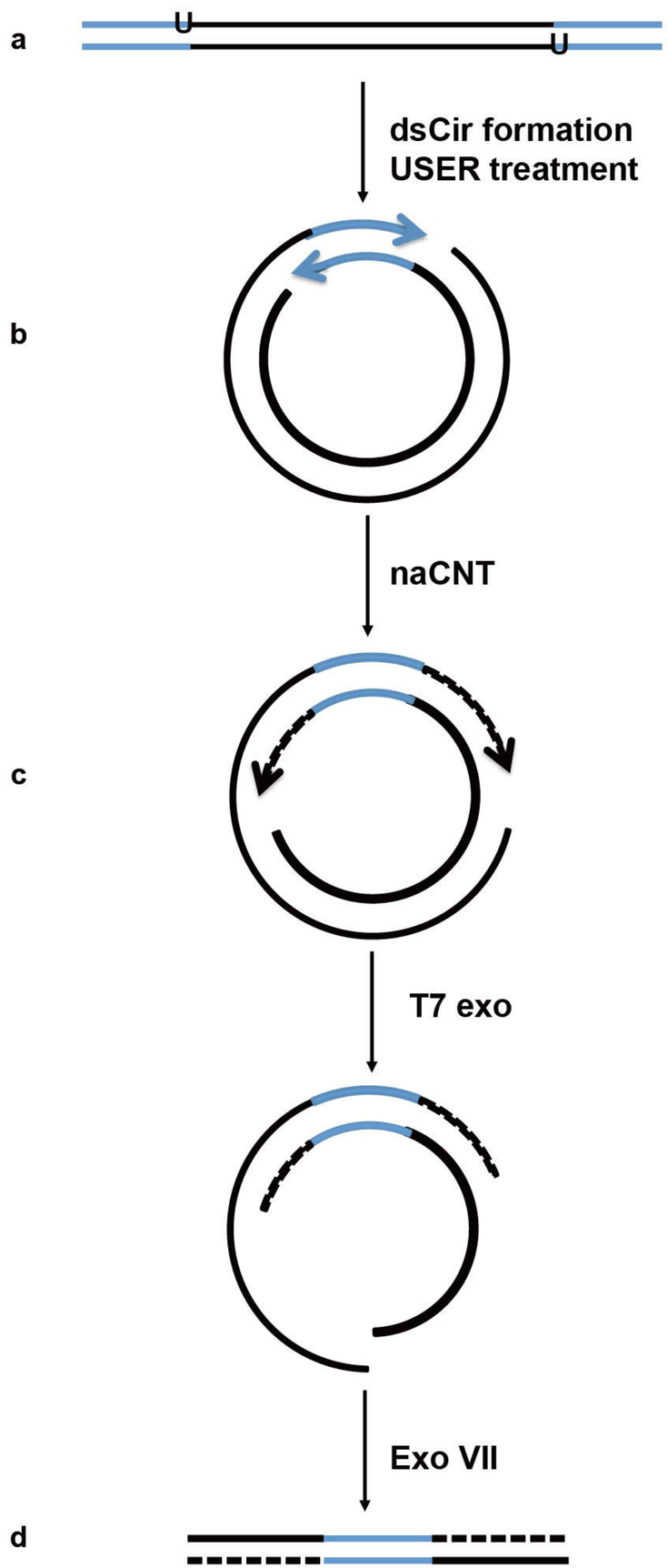
[1]Library construction method: BLBEC [biotin-labeled blunt-end circularization (7)]; custom jumping library (5,12); time and temperature controlled nick translation [ttCNT (14)]; Nextera mate-pair library construction (25); MpCP (mate-pair library construction by controlled polymerization during nick translocation and primer extension) used in this study.

[2]Reagent cost was obtained from a published study (5) and supplier websites.

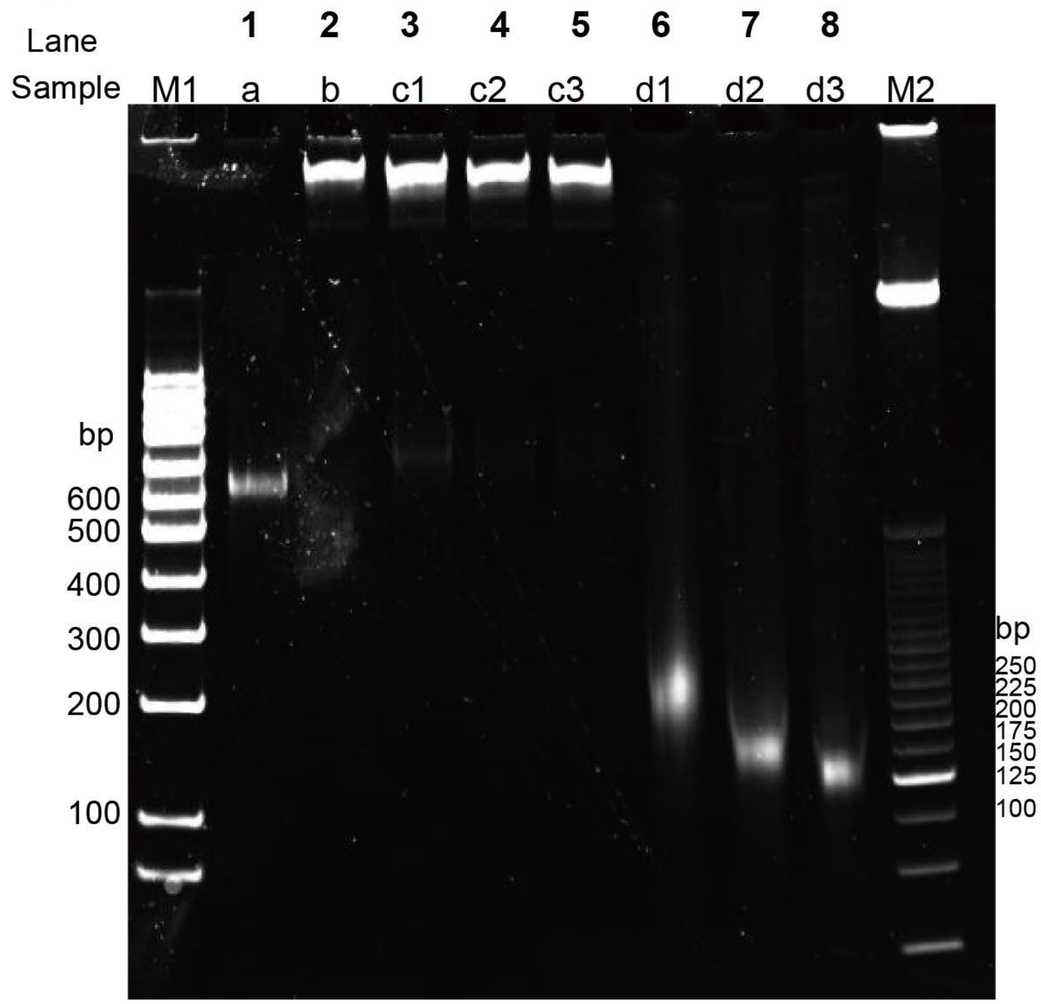
[3]The numbers before and after the vertical line reflect DNA input for non-size-selected and size-selected library construction, respectively.


DNA template
Ad1
Ad2_3'
Ad2_5'
5'
5'
1. Ad1 ligation
5'
5'
2. PCR
5'
U
U
5'
3. USER digestion and
dsCir (gap) formation
5'
3'
4. naCNT
by PolI
5'
3'
5. 3'-branch
ligation
5'
3'
Ad2_3'
3'
6. CPE
(ntCPE or ttCPE)
5'
5'
7. 3'-branch
ligation
5'
5'
3'
Ad2_5'
8. PCR
5'
5'

A



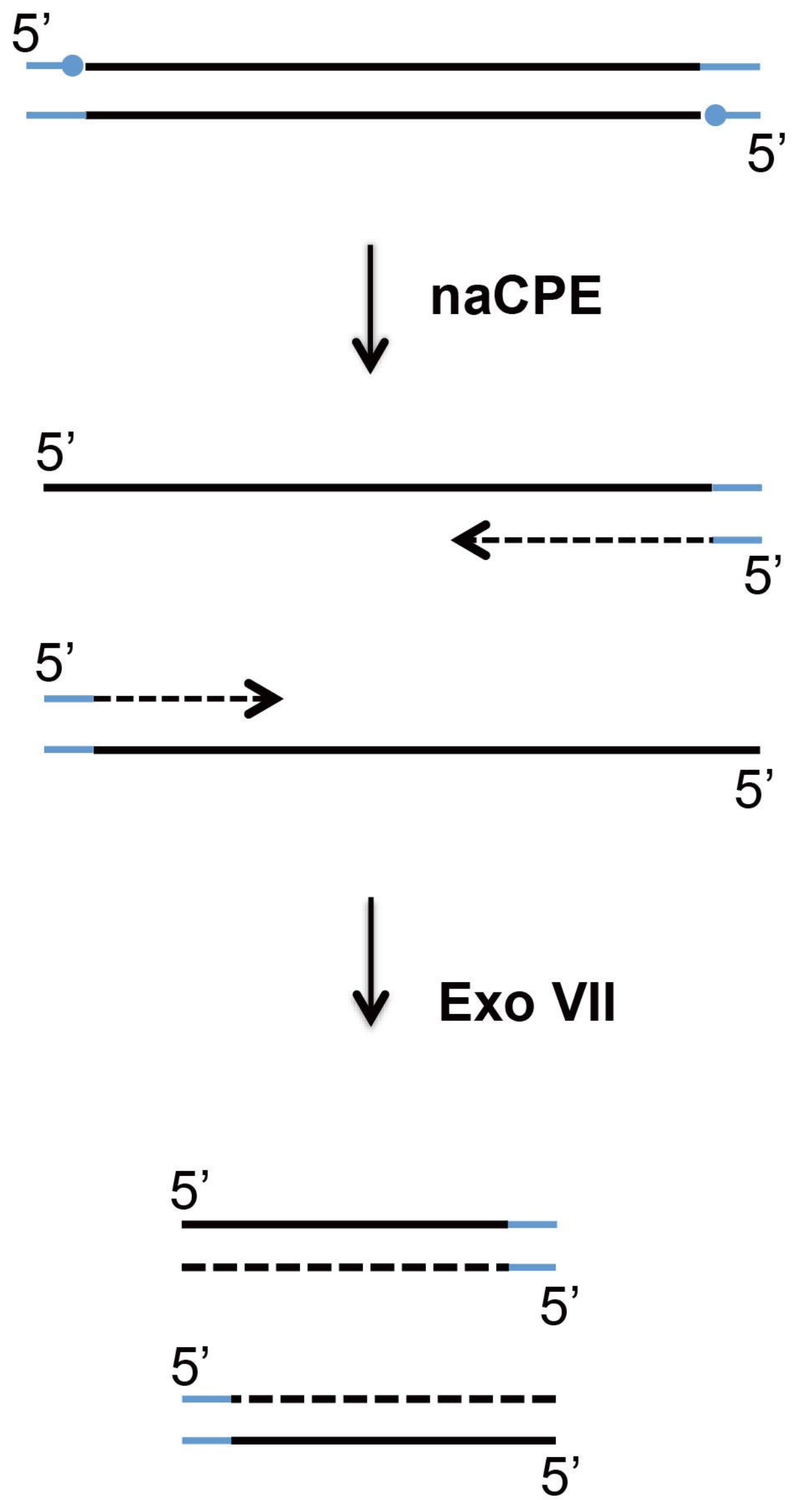
B



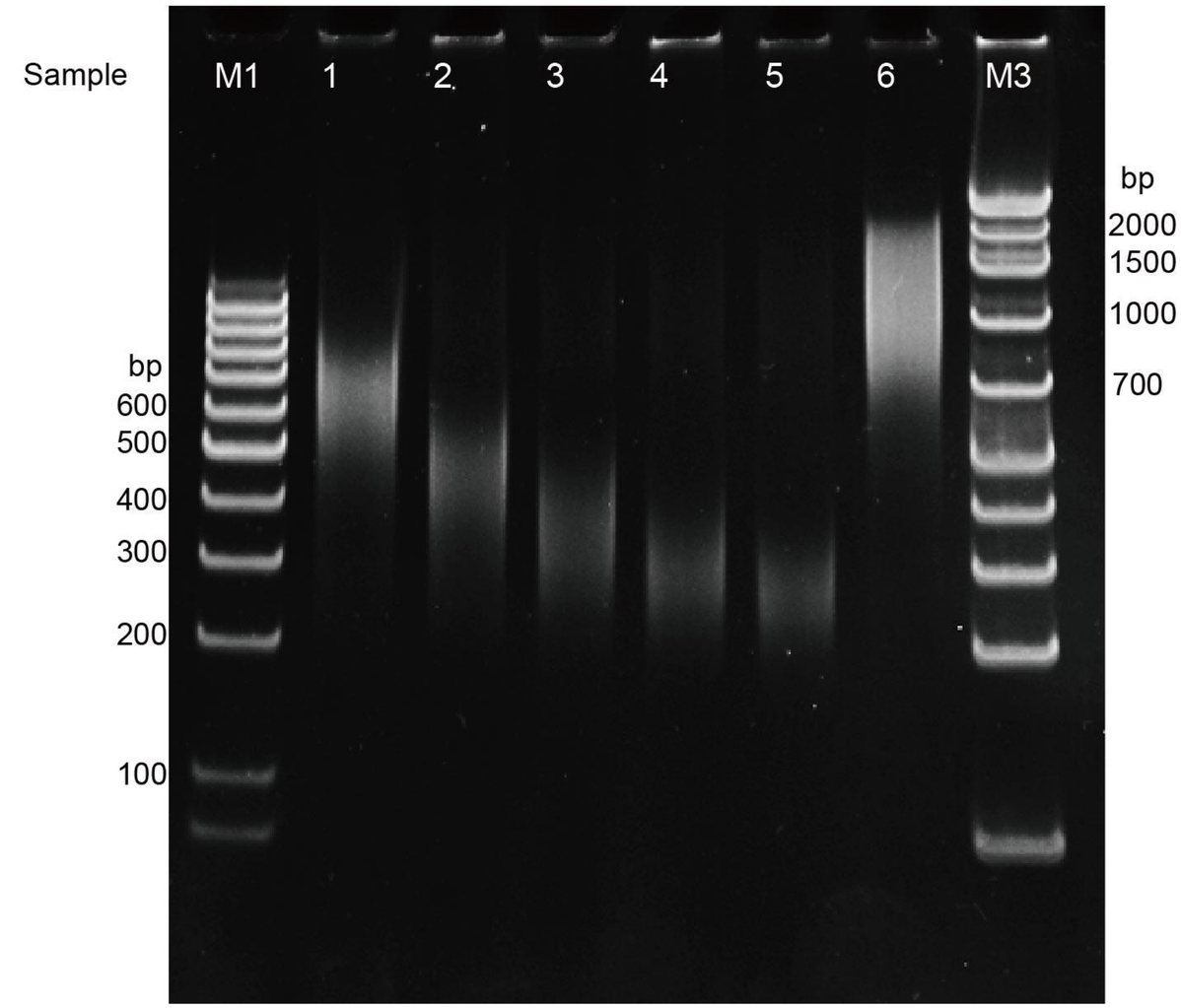
C

Sample	dsDNA circles (pmol)	IS[1] (pmol)	dNTP (pmol)	naCNT reaction conditons	Ratio dNTPs::IS	Predicted polymerization length per arm[2] (bp)	Observed length range (bp)	Median length (bp)	CNT extension length per arm[3] (bp)
d1	0.23	0.46	8.0	PolI, 10ºC, 20min	17.4	70	175-250	213	62
d2	0.23	0.46	4.0	PolI, 10ºC, 20min	8.7	35	125-175	150	30
d3	0.23	0.46	2.6	PolI, 10ºC, 20min	5.7	23	125-150	138	24

A

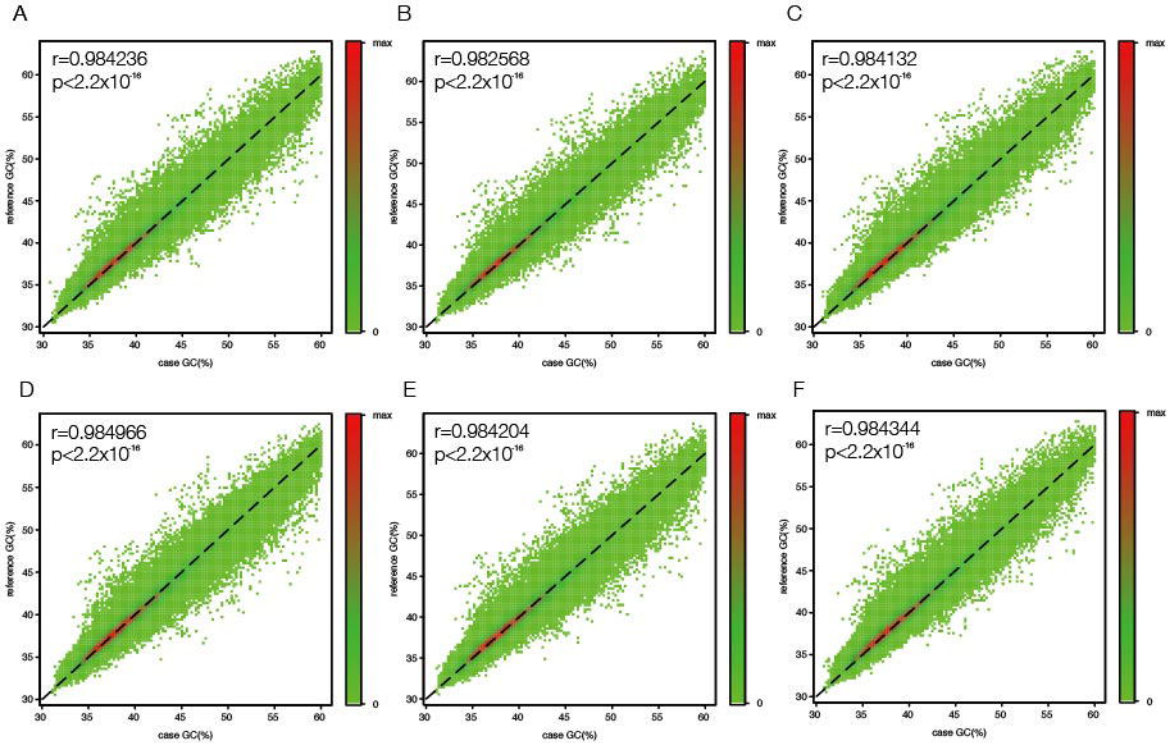


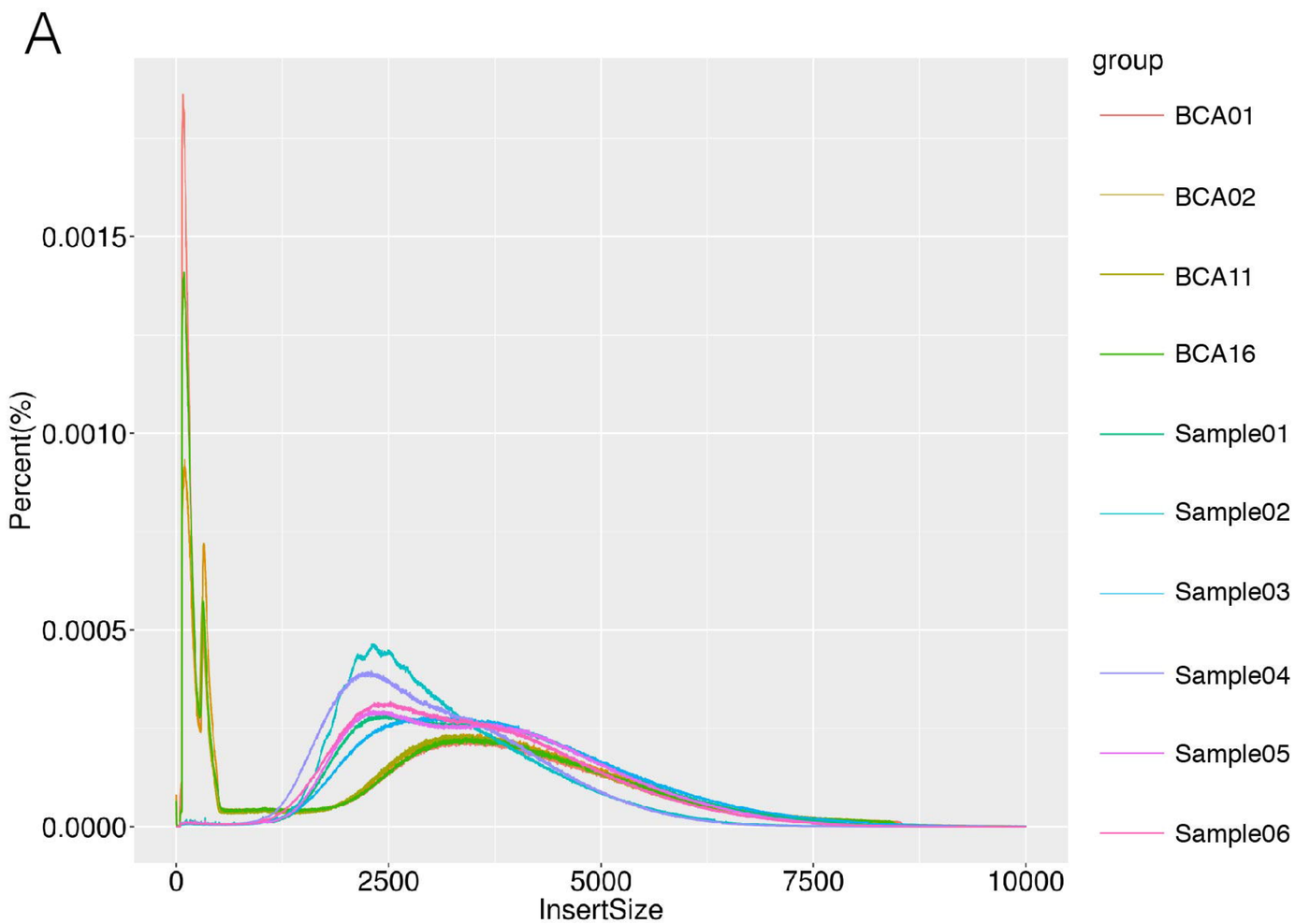
B

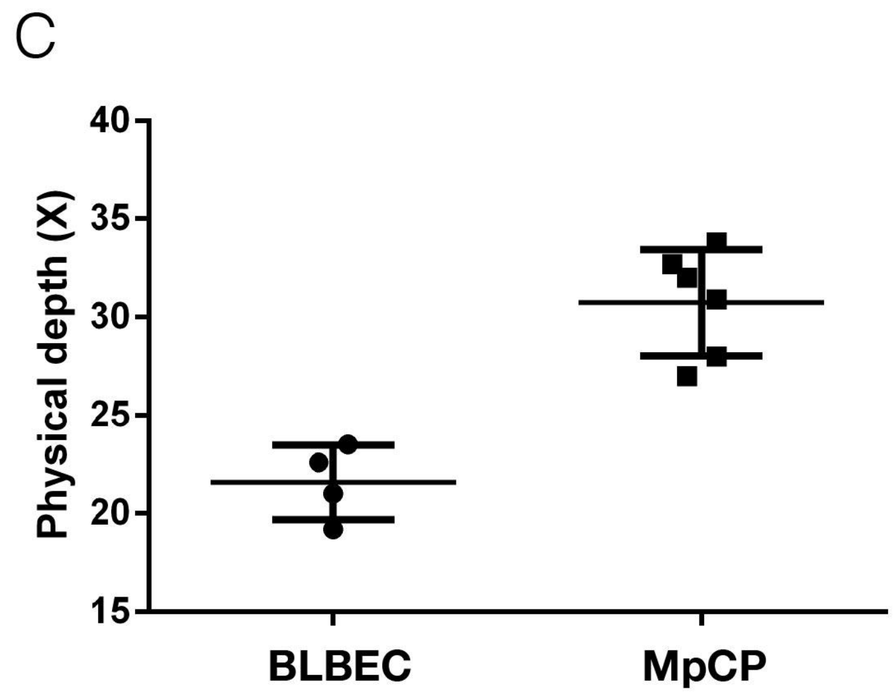


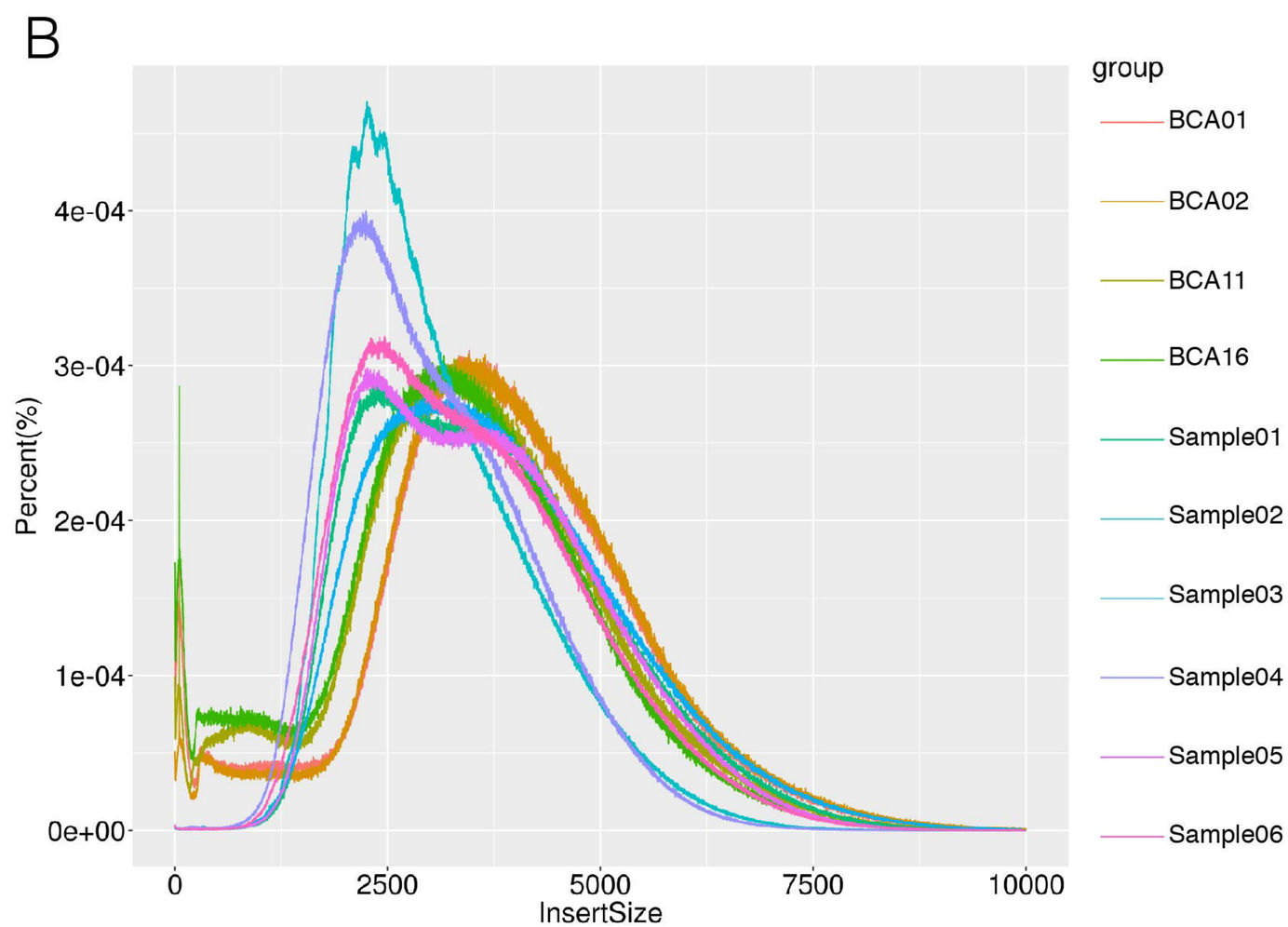
C

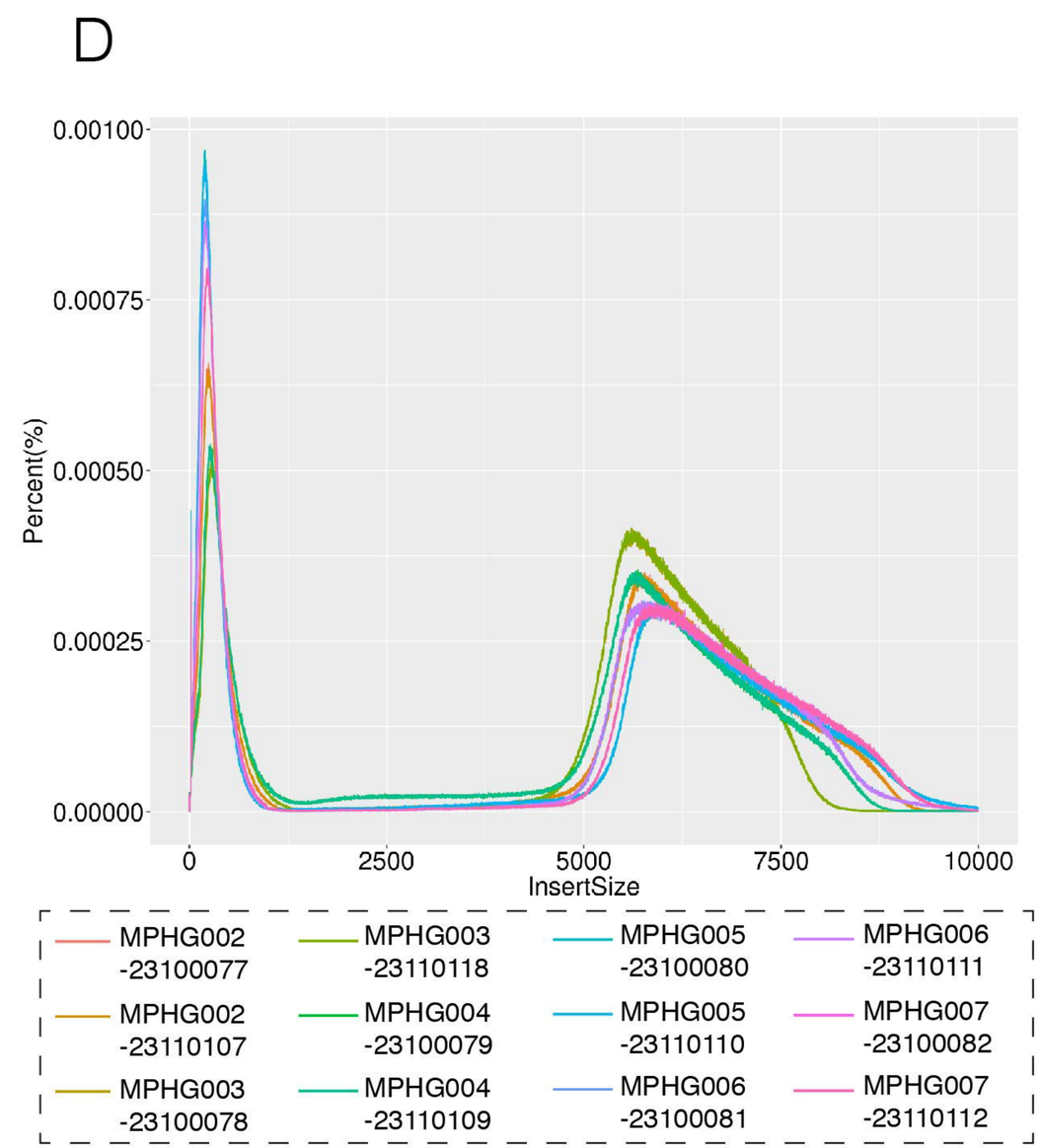
Sample	DNA (pmol)	IS (pmol)	dNTP (pmol)	Ratio dNTPs::IS	Observed length range (bp)	PE length range (bp)
1	0.16	0.32	33.3	104.1	500-700	460-670
2	0.16	0.32	25	78.1	380-510	340-470
3	0.16	0.32	20	62.5	270-400	230-360
4	0.16	0.32	16.7	52.2	220-320	180-280
5	0.16	0.32	14.3	44.7	210-300	170-260
6	0.16	0.32	5,000	15,625	700-3,000	700-3,000

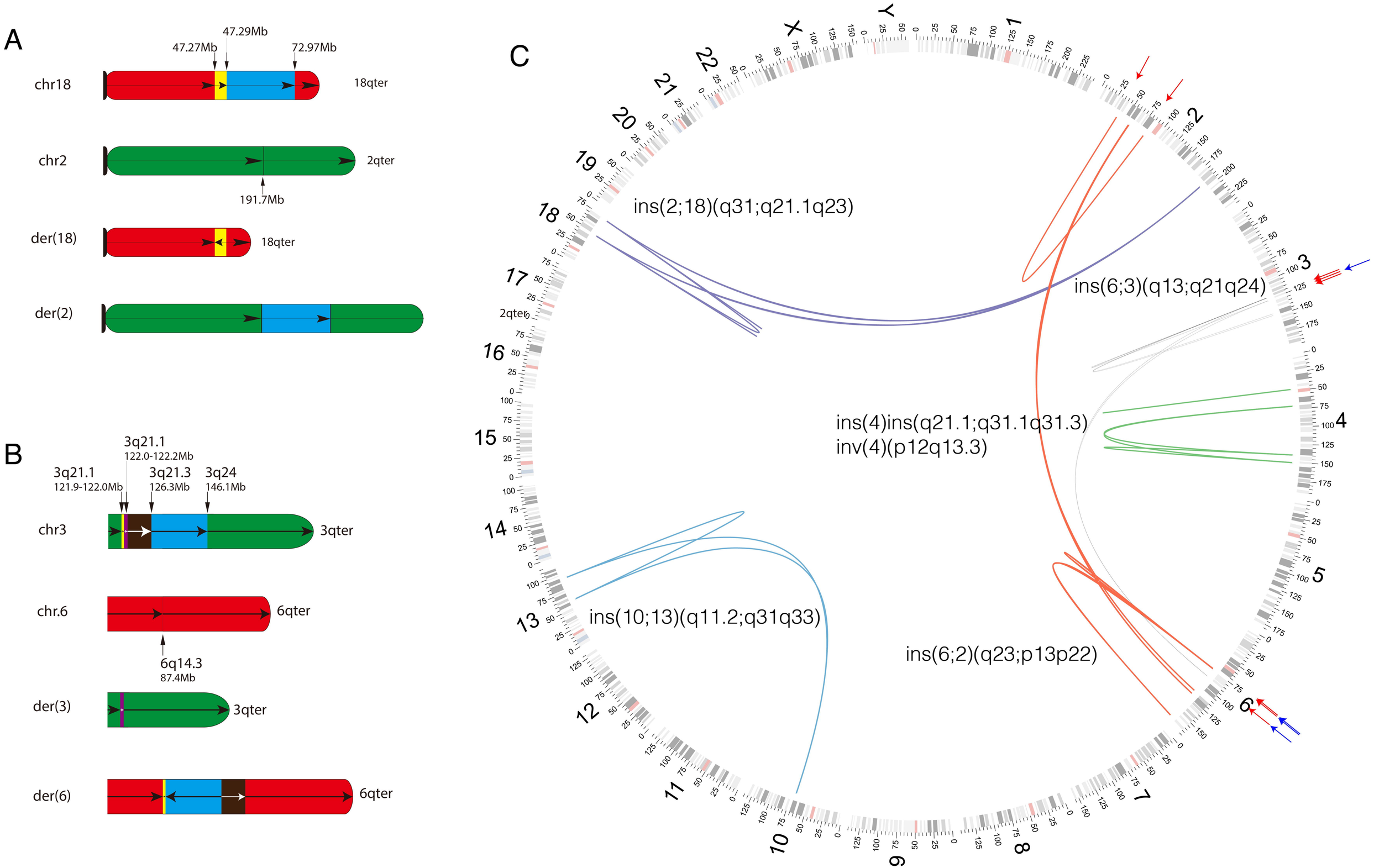

A
r=0.984236
p<2.2x10-16
B
r=0.982568
p<2.2x10-16
C
r=0.984132
p<2.2x10-16
D
r=0.984966
p<2.2x10-16
E
r=0.984204
p<2.2x10-16
F
r=0.984344
p<2.2x10-16
reference GC(%)
case GC(%)
max
0


A
group
BCA01
BCA02
BCA11
BCA16
Sample01
Sample02
Sample03
Sample04
Sample05
Sample06
Percent(%)
0.0015
0.0010
0.0005
0.0000
0
2500
5000
7500
10000
InsertSize


C
Physical depth (X)
40
35
30
25
20
15
BLBEC
MpCP


B
group
BCA01
BCA02
BCA11
BCA16
Sample01
Sample02
Sample03
Sample04
Sample05
Sample06
Percent(%)
4e-04
3e-04
2e-04
1e-04
0e+00
0
2500
5000
7500
10000
InsertSize


D
Percent(%)
0.00100
0.00075
0.00050
0.00025
0.00000
0
2500
5000
7500
10000
InsertSize
MPHG002 -23100077
MPHG003 -23110118
MPHG005 -23100080
MPHG006 -23110111
MPHG002 -23110107
MPHG004 -23100079
MPHG005 -23110110
MPHG007 -23100082
MPHG003 -23100078
MPHG004 -23110109
MPHG006 -23100081
MPHG007 -23110112


A
47.27Mb
47.29Mb
72.97Mb
chr18
18qter
chr2
2qter
191.7Mb
der(18)
18qter
der(2)
B
3q21.1
122.0-122.2Mb
3q21.1
121.9-122.0Mb
3q21.3
126.3Mb
3q24
146.1Mb
chr3
3qter
chr.6
6qter
6q14.3
87.4Mb
der(3)
3qter
der(6)
6qter
C
ins(2;18)(q31;q21.1q23)
ins(6;3)(q13;q21q24)
ins(4)ins(q21.1;q31.1q31.3)
inv(4)(p12q13.3)
ins(10;13)(q11.2;q31q33)
ins(6;2)(q23;p13p22)
2qter