Comparison of double-stranded and single-stranded cfDNA library construction methods for targeted genome and methylation sequencing

Jianchao Zheng [1,2], Zhilong Li[1,2], Xiuqing Zhang[1,3], Hongyun Zhang[2], Shida Zhu[2,4], Jianlong Sun[2†], and Yuying Wang[2†]

[1]College of Life Sciences, University of Chinese Academy of Sciences, Beijing 100049, China.

[2]BGI Genomics, BGI-Shenzhen, Shenzhen 518083, China.

[3]BGI-Shenzhen, Shenzhen 518083, China.

[4]Shenzhen Engineering Laboratory for Innovative Molecular Diagnostics, BGI-Shenzhen, Shenzhen, 518120, China

[†]Correspondence

Dr. Jianlong Sun and Dr. Yuying Wang, BGI Genomics, BGI-Shenzhen, Shenzhen 518083, China. E-mail: sunjianlong@genomics.cn and wangyuying@genomics.cn

Grants

This research was supported by Shenzhen Engineering Laboratory for Innovative Molecular Diagnostics (DRC-SZ[2016]884).

Conflict of Interest Statement

JZ, ZL, HZ, SZ, JS, and YW are employees of BGI Genomics, BGI-Shenzhen. XZ is an employee of BGI-Shenzhen.

Abstract

Cell-free DNA (cfDNA) profiling by deep sequencing (i.e., by next generation sequencing (NGS)) has wide applications in cancer diagnosis, prognosis, and therapy response monitoring. One key step of cfDNA deep sequencing workflow is NGS library construction, whose efficiency significantly affects the utilization efficiency of cfDNA molecules, and eventually determines effective sequencing depth and sequencing accuracy. In this study, we compared two different types of cfDNA library construction methods, namely double-stranded library (dsLib, the conventional method which captures dsDNA molecules) and single-stranded library (ssLib) preparation, which captures ssDNA molecules, for the applications of mutation detection and methylation profiling, respectively. Our results suggest that the dsLib method was suitable for mutation detection while the ssLib method proved more efficient for methylation analysis. Our findings could help researchers choose the more appropriate library construction method for corresponding downstream applications of cfDNA sequencing.

Keywords: cfDNA, NGS, library construction methods, target sequencing, methylation.

Introduction

Cell-free DNA (cfDNA), primarily derived from cell apoptosis, has been shown to be an important biomarker of many physiological and pathological conditions such as autoimmunity, infection, pregnancy, exercise, transplantation, and cancer [1-3]. It is detectable in almost all body fluids including plasma, serum, and urine, with a peak size at approximately 166 bp [4, 5]. In cancer patients, there is a subset of cfDNA known as circulating tumor DNA (ctDNA), which originates from tumor cells and carries genetic and epigenetic characteristics of the tumor [6, 7]. cfDNA has a short half-life in circulation (reported to be between 16 and 150 minutes), can be repeatedly sampled, and may potentially overcome intratumor heterogeneity compared to tissue biopsies [8]. These unique characteristics make cfDNA-based liquid biopsy an ideal approach for cancer diagnosis, prognosis, and therapy response monitoring [9, 10]. However, due to low concentration of cfDNA in plasma (~3 ng/ml in healthy individuals) and very small fraction of ctDNA among abundant cfDNA that derived from blood cells and normal tissues, accurate detection of ctDNA remains a challenging task [11, 12].

Massively parallel sequencing (MPS), also known as next generation sequencing (NGS), has been widely applied in both research and diagnostic fields [13, 14]. Millions of DNA fragments can be simultaneously sequenced and analyzed by NGS

[15]. Furthermore, targeted capture sequencing allows for deeper sequencing for target regions of interest at a lower cost [16]. Remarkably, efficient library construction before targeted capture sequencing determines effective sequencing depth and remains indispensable to successful sequencing of the target regions, and is particularly critical for sequencing of limited amount of cfDNA, and for identification of variants with lower allele fractions [9, 17]. During library construction, platform-specific adapters, which contain sample barcode sequence(s) and common primer binding sites for subsequent amplification and sequencing, are ligated to both ends of the original DNA fragments [18]. Various library construction methods have been introduced, aiming to improve DNA conversion efficiency (defined as the fraction of original DNA molecules that are successfully converted to the final library) [19-23]. Conventional double-stranded library (dsLib) construction workflow (such as what is used in the KAPA Hyper Prep kit) consists of following steps: (i) end repair and dA-tailing; (ii) adapter ligation to double-stranded DNA (dsDNA) templates; (iii) library amplification and purification [24]. On the other hand, single-stranded library (ssLib) construction was usually initialized by adaptor ligation to single-stranded DNA (ssDNA) templates and followed by library amplification and purification [19, 22, 25]. The ssLib construction method was originally invented to recover ancient and/or degraded DNA fragments which are usually poorly captured by the dsLib method [26, 27]. Although researchers have compared their performance in applications such as non-invasive prenatal testing, which is based on shallow whole-genome sequencing (WGS) and suggests that no significant difference of fetal

DNA enrichment was observed between ssLib and dsLib construction methods [28], there hasn't been systematic study to compare the performance of these two methods for cfDNA sequencing for cancer-related applications.

In this study, we compared the dsLib workflow and ssLib workflow for targeted deep sequencing (for variant detection) and methylation (for determining 5-methylcytosine (5-mC) status) sequencing of limited cfDNA inputs, two applications important for cancer diagnosis. We found that, for targeted deep sequencing, the dsLib method achieved better performance with a lower limitation of detection (LOD). For methylation sequencing, we compared the dsLib and ssLib workflow coupled with cytosine conversion using either bisulfite or enzyme-based method, and found that ssLib coupled with bisulfite conversion showed notably better performance.

Materials and Methods

Ethical Compliance

This study was approved by the institutional review board of BGI (NO. BGI-IRB: 19077).

Sample collection and cfDNA isolation

After obtaining informed consent, blood samples were collected from 37 healthy volunteers and 2 lung cancer patients in 10 mL K2 EDTA BD Vacutainer tubes. Blood was separated immediately by an initial centrifugation at 1,600 × g for 10 min and then

by a second centrifugation at 16,000 × g for 10 min. Plasma was pooled and splitted into 4ml per reaction for cfDNA isolation using MagPure Circulating DNA Maxi Kit (Magen, China) according to the manufacturer's instruction. Extracted cfDNA was pooled together to obtain sufficient homogeneous material for this study. cfDNA quantity was determined by Qubit dsDNA High Sensitivity Kit (Thermo Fisher Scientific, USA). The 1% Multiplex I cfDNA reference standards HD778 (Horizon Discovery, UK) were spiked in healthy donor cfDNA to obtain 0.1%, 0.25%, and 0.5% multiplex cfDNA.

Double-stranded cfDNA library construction

Duplex unique molecular identifier (UMI) adapter for MGISEQ-2000 sequencer was designed according to Newman et al [21] with the modification that 3-bp UMIs were chosen instead of 2-bp UMIs in order to accommodate a higher library complexity. To avoid potential issues during sequencing caused by low complexity at the T-A ligation position (constant base), 32 pairs of UMI adapters were incorporated with an additional base (G or C) before the T-A ligation position. Long oligonucleotides UMIxxL (5'-Phosphorylation-[C/G/-]-NNNAAGTCGGAGGCCAAGCGGTCTTAGGAAGACAA-3') and short oligonucleotides UMIxxS (5'-GACATGGCTACGATCCGACTNNN-[G/C/-]-T-3') were synthesized by BGI tech solutions (Beijing Liuhe co.limited). Each oligo was dissolved to 100 μM using TE buffer (pH=8.0). For each pair of adapters, 5 μL UMIxxL and 5 μL UMIxxS

oligos (100 μM) were combined in a 20 μL reaction volume with a final concentration of 20 mM Tris-HCl, 0.2 mM EDTA and 10 mM NaCl. Oligos were annealed for more than 30 min at room temperature. 64 UMI adapters (25 μM) were mixed and diluted to 5 μM using TE buffer, marked as UMI64M.

Double-stranded cfDNA libraries were prepared by KAPA Hyper Prep kit (Kapa Biosystems, cat. No. KK8504) according to the manufacturer's instruction or our self-developed library construction protocol. Briefly, 1-10 ng cfDNA in 42.5 μL was mixed with 7.5 μL end-repair master mix consisting of 1.5 μL T4 DNA polymerase (Enzymatics, cat. No. P7080L), 0.5 μL T4 polynucleotide kinase (Enzymatics, cat. No. Y9040L), 0.15 μL rtaq DNA polymerase (MGI, cat. No. 01E012MM), 0.35 μL dNTP (5mM each), and 5 μL 10× T4 DNA ligase buffer, and kept at 20□ for 30 min followed by 65□ for 30 min. Then 1 μL UMI64M adapter (5 μM) was added to the end-repair reaction product and mixed by pipetting, followed by adding 29 μL ligation master mix consisting of 4 μL golden T4 DNA ligase (MGI, cat. No. 02E004MM), 3 μL 10× T4 DNA ligase buffer, and 22 μL PEG6000 (Sigma Aldrich, 50%). The ligation reaction was incubated at 16□ for 60 min. After that, 50 μL nuclease-free water and 50 μL Agencourt AMPure XP beads were added to the 80 μL ligation product to purify adapter ligated cfDNA. The DNA was eluted in 30 μL nuclease-free water. Next, index PCR was performed in 60 μL reaction using 27.5 μL adapter ligated cfDNA template, 30 μL 2×Golden High-fidelity Readymix (MGI, cat. No. 01K00902MM), and 2.5 μL index PCR primer pair (20 μM each). Libraries were

amplified using following conditions: 98°C for 45 s, followed by a number of cycles of 98°C for 15 s, 60°C for 30 s, and 72°C for 30 s, and followed by 72°C for 1 min. Each library was given a specific number of PCR cycles based on cfDNA input amount. 90 μL Agencourt AMPure XP beads were added to the 60 μL PCR products to purify the library and the final library was eluted in 40 μL nuclease-free water. The concentration was determined by Qubit dsDNA High Sensitivity Kit.

Double-stranded cfDNA methylation sequencing libraries were prepared using a self-developed protocol with following modifications: (i) 0.05 ng fragmented lambda DNA was spiked in the 10 ng cfDNA to monitor bisulfite conversion rate; (ii) all cytosines of adapter were methylated; (iii) after purification of the ligation product, bisulfite conversion was performed using EZ DNA Methylation Gold kit (Zymo Research, cat. No. D5006) or EM-seq Conversion Module (NEB, cat. No. E7125); (iv) index PCR was performed by 2×Golden U+ High-fidelity Readymix (MGI, cat. No. 01K01701MM).

Single-stranded cfDNA library construction

Single-stranded cfDNA libraries were prepared according to the following protocol. Firstly, cfDNA was dephosphorylated using FastAP (Thermo Fisher Scientific) at 37□ for 20 min, denatured at 95□ for 3 min, and immediately placed on ice. Next, ss-Ad1 adapter was added to 3’ end of cfDNA by golden T4 DNA ligase (MGI, cat. No. 02E004MM) at 37□ for 15 min, followed by 95□ for 3 min and immediately placed

on ice. The ligation product was purified by MyOne C1 beads (ThermoFisher Scientific), mixed with extension reaction buffer and ss_Ext primer (100μM), and incubated by at 65□ for 3 min followed by immediately on ice for 5 min. Then extension was performed by Klenow Fragment (MGI, cat. No. 01E013MS) and incubated at 25□ for 30 min and 75□ for 10 min. After that, following procedures were performed repeatedly with distinctions in the third step. (1) Briefly spun the tube and placed it into a magnetic rack. (2) Discard the supernatant. (3) Resuspended beads respectively in: (i) A-tailing reaction mix consisting of Klenow (3'-5' exo-) (Enzymatics, cat. No. P7010-LC-L), 10× blue buffer, and dATP, and incubated at 37□ for 30 min; (ii) Stringency wash buffer and incubated at 45□ for 5 min; (iii) wash buffer B; (iv) ligation-2 master mix (golden T4 DNA ligase (MGI, cat. No. 02E004MM), 10× T4 DNA ligation buffer, 15% PEG6000 (Sigma Aldrich, 50%), 0.1% Tween-20, ss-Ad2 (1 μM), and nuclease-free water) and incubated at 20□ for 20 min; (v) Stringency wash buffer and incubated at 45□ for 5 min; (vi) wash buffer B; (vii) PCR reaction mix (nuclease-free water, 2×Golden High-fidelity Readymix (MGI, cat. No. 01K00902MM), and PCR primer pair (10 μM each)). Finally, libraries were amplified by a specific number of PCR cycles based on cfDNA input mass. NGS libraries were purified by Agencourt AMPure XP beads and eluted in nuclease-free water.

Single-stranded cfDNA methylation sequencing libraries were prepared according to the above protocol, while the bisulfite conversion of cfDNA was performed using

either the EZ DNA Methylation Gold kit (Zymo Research, cat. No. D5006) or the EM-seq Conversion Module (NEB, cat. No. E7125) before dephosphorylation reaction. To monitor the conversion rate, 0.05 ng fragmented lambda DNA was spiked into the 10 ng cfDNA.

Target capture and sequencing

A custom capture panel that spans 170 kb and covers 139 cancer associated genes was designed and then synthesized by IDT technologies. Targeted genome capture was performed using xGen® Lockdown® Reagents (IDT technologies) and BGI adapter-specific blocker (BGI). 400ng library per sample (6 or 8 multiplexing) was captured according to the manufacturer's instruction.

Targeted methylation capture was performed using TargetCap methylation capture probe (panel size, 198 kb) and reagents (BoKe Bioscience China, cat. No. MP121CD) and BGI adapter-specific blocker (BGI). 400ng library per sample (4 multiplexing) was captured according to the manufacturer's instruction.

The above captured cfDNA genome or methylation libraries were amplified by a 50 μL PCR mix consisting of 25 μL 2× KAPA HiFi Hot Start Ready Mix, 5 μL PCR primer pair (10 μM), and 20 μL beads suspensions. The PCR program was as follows: 98°C for 45 s, followed by a number of cycles of 98°C for 15 s, 60°C for 30 s, and 72°C for 30 s, and followed by 72°C for 1 min. Targeted methylation library and

targeted genome library were amplified with 11 and 13 cycles respectively. 75 μL AMPure XP beads were added to purify amplified libraries. The libraries were eluted in 30 μL nuclease-free water and the concentration was determined by Qubit dsDNA High Sensitivity Kit.

Target captured libraries were sequenced on MGISEQ-2000 sequencer (MGI, China) using the 2 × 100 paired-end sequencing method.

Data analysis

Adapter trimming and quality control of sequencing data were performed using fastp (v0.19.7) [29]. Paired-end reads of TGS (Targeted Genome Sequencing) and TMS (Targeted Methylation Sequencing) were aligned to the hg19 reference human genome using bwa (v0.7.17) and BitMapperBS (v1.0.0.8), respectively [30, 31]. Duplications were marked and deduped using sambamba (v0.6.8) [32]. Removal of sequencing errors using duplex UMIs and variant calling were performed using custom python scripts. Methylation rate of cytosine was called as #C/(#C+#T) for each CpG site with at least 4x coverage, and M-bias of sequence reads was analyzed using MethylDackel (v0.3.0) (https://github.com/dpryan79/MethylDackel). The conversion rate was calculated using the methylation ratio of the spiked-in lambda DNA. GC-bias metrics were analyzed using Picard Tools (v 2.10.10) (http://broadinstitute.github.io/picard). Insert size distribution, base distribution of reads, on-target rate, and depth were analyzed using custom Perl scripts.

Data Access

The data that support the findings of this study have been deposited in the CNSA (https://db.cngb.org/cnsa/) of CNGBdb with accession number CNP0001331.

Results

Comparison of dsLib *vs.* ssLib method for targeted deep sequencing of cfDNA

Firstly, we compared double-stranded library (dsLib) preparation and single-stranded library (ssLib) preparation methods for cfDNA mutation detection using deep sequencing (Figure 1A, see Methods for more details). KAPA Hyper Prep kit, a widely used NGS library construction kit, was also included as a reference to evaluate performance of our self-developed workflow. Both dsLib preparation method and KAPA workflow used a unique molecular identifier (UMI) strategy to reduce noises that may derive from PCR and/or sequencing errors [21] (see Methods for more details). Since in clinical practice the amount of extracted cfDNA was often limited and highly variable [33], we used 1 ng, 5 ng, and 10 ng cfDNA as inputs for library construction respectively (Supplementary Table 2). Prepared libraries underwent hybridization-based target enrichment procedure and captured libraries were sequenced to > 20000x raw average depth (see Methods for more details). Results showed that dsLib and ssLib workflow produced similar amount of libraries (Supplementary Figure 1A). Notably, the ssLib workflow achieved similar deduplicated depths as the dsLib workflow with comparable amount of sequencing

clean reads (Figure 1B and Supplementary Table 2). Moreover, the ssLib workflow was more complicated and time-consuming than the dsLib method (8h vs 3.5h, see Methods for more details). Remarkably, our self-developed dsLib protocol showed significantly better performance than the commercial KAPA kit (Figure 1B and Supplementary Figure 1).

ctDNA has been proven to be shorter than cfDNA originated from normal cells [34]. Theoretically, the ssLib method preferentially enriches short DNA molecules and therefore may enrich ctDNA [28]. Copy number variation (CNV) is a hallmark of cancer and could be used as a biomarker for ctDNA [35]. Here, we compared CNV detectability of plasma cfDNA from lung cancer patients using either dsLib or ssLib preparation to test the hypothesis that ssLib may enrich for shorter ctDNA and improve its detection. Results showed that there is no significant difference in copy number variation detection by the ssLib workflow *vs.* dsLib (Figure 1C), consistent with previous study which showed that ssLib workflow does not enrich for fetal DNA for the application of non-invasive prenatal testing (NIPT), despite the finding that ssLib does enrich for shorter cfDNA fragments [36]. Taken together, these results indicated that our self-developed dsLib was suitable for cfDNA mutation detection.

To further validate its ability to detect low abundance mutations in cfDNA and confirm the limitation of detection (LOD), we applied the dsLib method on 10 ng cfDNA samples spiked-in with cfDNA reference standards (see Methods for more

details). The measured allele frequencies correlated well to the expected variant allele frequencies (VAFs) (Figure 1D). All mutations were detected in cfDNA samples expected to carry 1% or 0.5% VAFs, and 75% (6/8) mutations were detected in cfDNA samples with 0.25% expected VAFs (Figure 1D and Supplementary Table 3). The results indicated that our self-developed dsLib workflow provides satisfactory sensitivity for low abundance mutations in cfDNA.

Comparison of dsLib *vs.* ssLib for cfDNA methylation sequencing

Bisulfite sequencing has been a widely used sequencing technology for methylation profiling, where 5mC status could be determined with single-nucleotide resolution. This technology leverages the fact that methylated cytosine remains unaffected under bisulfite treatment, whereas unmethylated cytosine is converted to uracil [37].

To compare performance of the single-stranded bisulfite sequencing library construction (ssmLib) and the double-stranded bisulfite sequencing library construction (dsmLib) on cfDNA methylation profiling (Figure 2A), we applied these two workflows on 1 ng, 5 ng, and 10 ng cfDNA as inputs for library construction respectively and captured with the Boke 198 kb TargetCap methylation capture panel and sequenced (Supplementary Table 4). Sequencing results showed that the ssmLib produced significantly higher library yield and deduplicated depth than the dsmLib method; the on-target rates were also slightly higher in ssmLib libraries than dsmLib libraries (Figure 2 B-C and Supplementary Figure 2). Notably, libraries produced by the ssmLib method had more short insert fragments than those produced by the

dsmLib method (Figure 2D). These results can be attributed to the DNA degradation caused by the bisulfite conversion process, which involves high temperature and low pH conditions [38]: during ssmLib workflow, the resulted short cfDNA fragments can still be captured by the ssDNA-based adaptor ligation; on the other hand, during dsmLib workflow, since bisulfite was applied to the adaptor-ligated dsDNA, excessive damage of the templates will cause the libraries to be un-amplifiable and lost during subsequent process, resulting in much lower library yield and effective sequencing depth. For the measured CpG methylation rates, the correlation of technical replicates was favorable for both methods (Supplementary Figure 3) and there was also a high correlation between the two methods different DNA inputs (Figure 2E).

Methylation bias (M-bias) is the term describing measured methylation levels that deviate from true values, often observed near the 3' end of sequenced fragments due to unmethylated cytosines introduced by the end-repair step during dsDNA ligation based library preparation workflow [39, 40]. Theoretically, libraries produced by the ssmLib method may show less to no M-bias since there is no end-repair step involved (Figure 2A). Indeed, we observed severe M-bias in Read 2 of dsmLib libraries, but not in ssmLib libraries (Figure 2F). Taken together, these results suggest that ssmLib method is more suited for the application of cfDNA methylation sequencing.

Finally, since recently several enzyme-based cytosine conversion methods had been

developed as gentler substitutes for bisulfite conversion [41, 42], we also compared performance of an enzyme-based workflow (using the NEB EM-seq Conversion Module) with the conventional bisulfite conversion workflow (using the widely used ZYMO EZ DNA Methylation Gold kit) (Figure 3A). Briefly, 10 ng cfDNA was used as the input of each reaction and the comparison was done for both ssmLib and dsmLib workflow (Figure 2A). As expected, the ssmLib produced by the ZYMO bisulfite conversion kit had more short insert fragments than by the NEB enzyme-based conversion kit due to DNA degradation caused by bisulfite; notably, these fragments would be expected to be ligatable with the ssmLib instead of the dsmLib workflow (Figure 3B). Remarkably, the ZYMO kit generated significantly higher library yield and deduplicated depth than NEB kit; similar cytosine conversion efficiency was observed for the two kits (Figure 3C-E and Supplementary Table 4). For the dsmLib method, however, the NEB kit showed higher conversion efficiency while there was no significant difference in library yield and deduplicated depth between these two kits (Supplementary Figure 4 and Supplementary Table 4). The difference in CT conversion rate may be due to the fact that bisulfite converted DNA tends to degrade more leaving unconverted DNA for sequencing thus reducing the conversion rate for the dsmLib method. The NEB enzyme-based conversion kit is more time-consuming (8h vs 3.5h) than the ZYMO bisulfite conversion kit. Taken together, our results favor bisulfite conversion coupled with ssLib workflow for cfDNA methylation profiling.

Conclusion

The double-stranded library preparation method has strength for cfDNA mutation detection thanks to the high data quality and easy workflow. On the other hand, the single-stranded library preparation method coupled with bisulfite conversion showed overall better performance for cfDNA methylation sequencing. Our results suggest that when performing high-throughput sequencing for cfDNA, depending on the downstream applications, these two library preparation methods should be chosen accordingly.

Discussion

In recent decades, thanks to the development of NGS technology, the cost of high-throughput DNA sequencing had dropped dramatically, making it affordable for researchers worldwide [43]. Library construction is a key step for successful NGS. In this study, we compared dsLib and ssLib methods for cfDNA deep sequencing (i.e., for ctDNA variant detection) and methylation profiling. Results showed that dsLib was more suitable for cfDNA variant detection, whereas the single-stranded ssLib coupled with bisulfite conversion would be a better choice for cfDNA methylation profiling. These findings could help researchers maximize the efficiency of NGS library preparation and produce better quality sequencing data.

Shorter insert fragments were observed in libraries prepared by ssLib workflow than those by dsLib workflow for both direct sequencing and bisulfite sequencing (Figure

2D, Supplementary Figure 1B, and Supplementary Table 2 and 4), especially for the latter, where a clear difference in fragment length distribution was observed between ssLib and dsLib libraries. This observation can be explained by the fact that cfDNA molecules harboring single-strand breaks (also called nicks) could be utilized by the ssLib method but would not be ligatable when using the dsLib method (Figure 1A and 2B). Naturally nicked cfDNA molecules may only be a very small fraction hence this difference would be expected to be small and may not cause significant impact on the library conversion efficiency. Indeed, we observed similar deduplicated depth for cfDNA libraries generated using ssLib or dsLib workflow (Figure 1B); in fact, deduplicated depth of ssLib libraries were even slightly inferior than dsLib, possibly due to the fact that ssLib workflow is lengthier and requires more beads purification and may cause template loss. Moreover, application of dsLib workflow further allows utilization of duplex UMIs, which make it possible to recover original dsDNA fragments following paired-end sequencing and utilize the information from complementary strands of DNA molecules to correct possible sequencing and/or PCR errors, achieving an extra low base error rate [21]. On the contrary, a clear advantage was observed for ssLib-based libraries for bisulfite sequencing (Figure 2B-C). This is because sequencing libraries were constructed before bisulfite conversion during the dsmLib workflow (Figure 2A), and the nicked DNA resulting from the bisulfite conversion won’t be sequenced due to the lack of paired adapters. For the ssmLib workflow, however, cfDNA ligation happens after the bisulfite treatment and the nicked DNA molecules can still be ligated as ssDNA and sequenced, therefore

preserving more DNA templates (Figure 2A and 2D).

Previous studies have shown that ctDNA is shorter than cfDNA originated from normal cells [34, 44], and the ssLib method could enrich for short DNA fragments [19, 25]. Therefore theoretically, the ssLib method is likely to preferentially enrich ctDNA. However, using detected CNV level as an indicator of ctDNA fraction, our results indicate that there was no significant enrichment of ctDNA by ssLib workflow compared to dsLib workflow (Figure 1C), consist with previous research conducted in the setting of NIPT which showed that ssLib workflow does not enrich for shorter fetal DNA [28, 36]. It was suggested that intrinsic biological differences between fetal DNA and maternal DNA molecules might account for the failure of ssDNA workflow to enrich for fetal DNA [28, 36], and similar mechanism may also explain our results for ctDNA. Further study is needed to deepen our understanding of cfDNA/ctDNA generation processes and/or to develop novel library construction methods for ctDNA enrichment.

For cfDNA methylation profiling, methylation bias (M-bias) was proposed to be an important library preparation quality metric since its existence could cause significant bias in measurements of methylation level [39, 40]. M-bias is caused by the end-repair step in the dsmLib workflow which typically recruits unmethylated cytosines instead of methylated cytosines during the fill-in reaction (Figure 2A). The filled-in cytosines were then converted to uracils regardless of the original cytosine

methylation states in the genome, resulting in incorrect methylation level being assigned to the 3' end of the sequenced reads [39, 40]. The ssmLib method could perfectly overcome this problem since it does not involve an end-repair step and is a post–bisulfite conversion library construction method (Figure 2A and 2F).

We also compared the ZYMO EZ DNA Methylation Gold kit, a widely used commercial kit for bisulfite conversion, with the NEB EM-seq Conversion Module, a novel enzyme-based 5-mC conversion kit, in this study. The enzyme-based conversion method was developed in order to minimize DNA degradation caused by the chemical conversion and was suggested to be more suitable for cfDNA with limited input amount. Our results, however, still favored the bisulfite conversion method due to the higher library yield as well as higher deduplicated depth, suggesting that there may be excessive loss of templates during the enzyme-conversion workflow (Figure 3C-E and Supplementary Figure 4). Although the bisulfite conversion would cause loss of input DNA templates by degradation, the two rounds of bead purification in the enzyme-based conversion workflow may explain the excessive template loss and the lower effective depth of the libraries. Also, the NEB EM-seq Conversion Module workflow is more labor- and time-consuming. More effective enzyme-based cytosine conversion methods may require improvements in template recovery and simplified workflow.

Reference

1. Lui, Y.Y.N., et al., *Predominant Hematopoietic Origin of Cell-free DNA in Plasma and Serum after Sex-mismatched Bone Marrow Transplantation.* Clinical Chemistry, 2002. **48**(3): p. 421-427.
2. Snyder, Matthew W., et al., *Cell-free DNA Comprises an In Vivo Nucleosome Footprint that Informs Its Tissues-Of-Origin.* Cell, 2016. **164**(1): p. 57-68.
3. Aravanis, A.M., M. Lee, and R.D. Klausner, *Next-Generation Sequencing of Circulating Tumor DNA for Early Cancer Detection.* Cell, 2017. **168**(4): p. 571-574.
4. Lo, Y.M.D., et al., *Maternal Plasma DNA Sequencing Reveals the Genome-Wide Genetic and Mutational Profile of the Fetus.* Science Translational Medicine, 2010. **2**(61): p. 61ra91.
5. Chan, A.K., R.W. Chiu, and Y.M. Lo, *Cell-free nucleic acids in plasma, serum and urine: a new tool in molecular diagnosis.* Ann Clin Biochem, 2003. **40**(Pt 2): p. 122-30.
6. Schwarzenbach, H., D.S.B. Hoon, and K. Pantel, *Cell-free nucleic acids as biomarkers in cancer patients.* Nature Reviews Cancer, 2011. **11**(6): p. 426-437.
7. van der Pol, Y. and F. Mouliere, *Toward the Early Detection of Cancer by Decoding the Epigenetic and Environmental Fingerprints of Cell-Free DNA.* Cancer Cell, 2019. **36**(4): p. 350-368.
8. Bronkhorst, A.J., V. Ungerer, and S. Holdenrieder, *The emerging role of cell-free DNA as a molecular marker for cancer management.* Biomolecular detection and quantification, 2019. **17**: p. 100087-100087.
9. Volik, S., et al., *Cell-free DNA (cfDNA): Clinical Significance and Utility in Cancer Shaped By Emerging Technologies.* Molecular Cancer Research, 2016. **14**(10): p. 898.
10. Corcoran, R.B. and B.A. Chabner, *Application of Cell-free DNA Analysis to Cancer Treatment.* New England Journal of Medicine, 2018. **379**(18): p. 1754-1765.
11. Gormally, E., et al., *Circulating free DNA in plasma or serum as biomarker of carcinogenesis: Practical aspects and biological significance.* Mutation Research/Reviews in Mutation Research, 2007. **635**(2): p. 105-117.
12. Chabon, J.J., et al., *Integrating genomic features for non-invasive early lung cancer detection.* Nature, 2020.
13. Schuster, S.C., *Next-generation sequencing transforms today's biology.* Nature Methods, 2008. **5**(1): p. 16-18.
14. Shendure, J., et al., *DNA sequencing at 40: past, present and future.* Nature, 2017. **550**(7676): p. 345-353.
15. Tucker, T., M. Marra, and J.M. Friedman, *Massively parallel sequencing: the next big thing in genetic medicine.* American journal of human genetics, 2009. **85**(2): p. 142-154.
16. Green, E.D., E.M. Rubin, and M.V. Olson, *The future of DNA sequencing.* Nature, 2017. **550**(7675): p. 179-181.
17. Mardis, E.R., *DNA sequencing technologies: 2006–2016.* Nature Protocols, 2017. **12**(2): p. 213-218.
18. Lundberg, D.S., et al., *Practical innovations for high-throughput amplicon sequencing.* Nature Methods, 2013. **10**(10): p. 999-1002.
19. Gansauge, M.T., et al., *Single-stranded DNA library preparation from highly degraded DNA using T4 DNA ligase.* Nucleic Acids Res, 2017. **45**(10): p. e79.

20. Newman, A.M., et al., *An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage.* Nat Med, 2014. **20**(5): p. 548-54.
21. Newman, A.M., et al., *Integrated digital error suppression for improved detection of circulating tumor DNA.* Nat Biotechnol, 2016. **34**(5): p. 547-555.
22. Raine, A., et al., *SPlinted Ligation Adapter Tagging (SPLAT), a novel library preparation method for whole genome bisulphite sequencing.* Nucleic Acids Res, 2017. **45**(6): p. e36.
23. Olova, N., et al., *Comparison of whole-genome bisulfite sequencing library preparation strategies identifies sources of biases affecting DNA methylation data.* Genome Biol, 2018. **19**(1): p. 33.
24. Keats, J.J., et al., *Whole Genome Library Construction for Next Generation Sequencing,* in *Disease Gene Identification: Methods and Protocols*, J.K. DiStefano, Editor. 2018, Springer New York: New York, NY. p. 151-161.
25. Gansauge, M.T. and M. Meyer, *Single-stranded DNA library preparation for the sequencing of ancient or damaged DNA.* Nat Protoc, 2013. **8**(4): p. 737-48.
26. Glocke, I. and M. Meyer, *Extending the spectrum of DNA sequences retrieved from ancient bones and teeth.* Genome Res, 2017. **27**(7): p. 1230-1237.
27. Sanchez, C., et al., *Circulating nuclear DNA structural features, origins, and complete size profile revealed by fragmentomics.* JCI insight, 2021. **6**(7): p. e144561.
28. Vong, J.S., et al., *Single-stranded DNA library preparation preferentially enriches short maternal DNA in maternal plasma.* J Clinical chemistry, 2017. **63**(5): p. 1031-1037.
29. Chen, S., et al., *fastp: an ultra-fast all-in-one FASTQ preprocessor.* Bioinformatics, 2018. **34**(17): p. i884-i890.
30. Li, H. and R. Durbin, *Fast and accurate short read alignment with Burrows–Wheeler transform.* Bioinformatics, 2009. **25**(14): p. 1754-1760.
31. Cheng, H. and Y. Xu, *BitMapperBS: a fast and accurate read aligner for whole-genome bisulfite sequencing.* bioRxiv, 2018: p. 442798.
32. Tarasov, A., et al., *Sambamba: fast processing of NGS alignment formats.* Bioinformatics, 2015. **31**(12): p. 2032-2034.
33. Bronkhorst, A.J., J. Aucamp, and P.J. Pretorius, *Cell-free DNA: Preanalytical variables.* Clinica Chimica Acta, 2015. **450**: p. 243-253.
34. Underhill, H.R., et al., *Fragment length of circulating tumor DNA.* J PLoS genetics, 2016. **12**(7): p. e1006162.
35. Li, J., et al., *Cell-free DNA copy number variations in plasma from colorectal cancer patients.* Molecular oncology, 2017. **11**(8): p. 1099-1111.
36. Moser, T., et al., *Single-Stranded DNA Library Preparation Does Not Preferentially Enrich Circulating Tumor DNA.* Clinical Chemistry, 2017. **63**(10): p. 1656-1659.
37. Hayatsu, H., *Discovery of bisulfite-mediated cytosine conversion to uracil, the key reaction for DNA methylation analysis--a personal account.* Proceedings of the Japan Academy. Series B, Physical and biological sciences, 2008. **84**(8): p. 321-330.
38. Grunau, C., S.J. Clark, and A. Rosenthal, *Bisulfite genomic sequencing: systematic investigation of critical experimental parameters.* Nucleic Acids Res, 2001. **29**(13): p. E65-5.
39. Hansen, K.D., B. Langmead, and R.A. Irizarry, *BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions.* Genome Biology, 2012. **13**(10): p. R83.
40. Ryan, D.P. and D. Ehninger, *Bison: bisulfite alignment on nodes of a cluster.* BMC

Bioinformatics, 2014. **15**(1): p. 337.

41. Vaisvila, R., et al., *EM-seq: Detection of DNA Methylation at Single Base Resolution from Picograms of DNA.* bioRxiv, 2019: p. 2019.12.20.884692.
42. Liu, Y., et al., *Bisulfite-free direct detection of 5-methylcytosine and 5-hydroxymethylcytosine at base resolution.* Nat Biotechnol, 2019. **37**(4): p. 424-429.
43. Kchouk, M., et al., *Generations of sequencing technologies: from first to next generation.* 2017. **9**(3).
44. Jiang, P., et al., *Lengthening and shortening of plasma DNA in hepatocellular carcinoma patients.* Proceedings of the National Academy of Sciences of the United States of America, 2015. **112**(11): p. E1317-E1325.

Figure 1: Comparison of dsLib and ssLib workflow for cfDNA mutation detection by targeted deep sequencing. (A) Schematic view of our self-developed dsLib and ssLib workflow. See Methods for more details. **(B)** Deduplicated depths of libraries constructed by dsLib, ssLib, and the KAPA kit. Two replicates were performed for each experimental condition. Data are presented as mean ± SD. N.S, p>=0.05; *, 0.01<=p<0.05; **, 0.001<=p<0.01; ****, p<0.0001, as calculated by Student's t-test. **(C)** CNVs detected by dsLib and ssLib methods respectively, in plasma cfDNA samples from lung cancer patients P1 and P2. X-axis, chromosome. Y-axis, CNV adjusted by GC content and map-ability. **(D)** Detection of low AF mutations by dsLib workflow in simulated cfDNA samples. Each shape represents one type of spiked-in mutation. Pearson correlation was shown for measured AFs vs. expected AFs.

Figure 2: Comparison of dsmLib and ssmLib workflow for cfDNA methylation profiling by bisulfite sequencing. (A) Schematic view of our dsmLib and ssmLib procedures. See Methods for more details. **(B)** On-target rates and **(C)** deduplicated depths of libraries prepared by dsmLib and ssmLib workflow. Two replicates were performed for each experimental condition. Data are presented as mean ± SD. N.S, p>=0.05; *, 0.01<=p<0.05; **, 0.001<=p<0.01; ****, p<0.0001, as calculated by Student's t-test. **(D)** Size distribution of library insert fragments. **(E)** Pearson correlation of methylation rates between ssmLib (x-axis) and dsmLib (y-axis) at different DNA inputs. **(F)** M-bias plots of libraries prepared by dsmLib and ssmLib workflow. For each row from left to right: Read 1 ++ strand, Read 1 -+strand, Read 2

+- strand, and Read 2 --strand. X-axis, position in read (bp). Y-axis, methylation level (%).

Figure 3: Comparison of the ZYMO bisulfite conversion kit and the NEB enzyme-based conversion kit cfDNA methylation analysis. (A) Technical principles of ZYMO and NEB kits. 5caC, 5-carboxylcytosine. T, thymine. **(B)** Size distribution of library insert fragments. X-axis, fragment size (bp). Y-axis, frequency count. **(C)** CT conversion rates, **(D)** library yields, and **(E)** deduplicated depths of ssmLib libraries using ZYMO and NEB kits for 5mC conversion. Two replicates were performed for each experimental condition. Data are presented as mean ± SD. N.S, $p>=0.05$; *, $0.01<=p<0.05$, as calculated by Student's t-test.

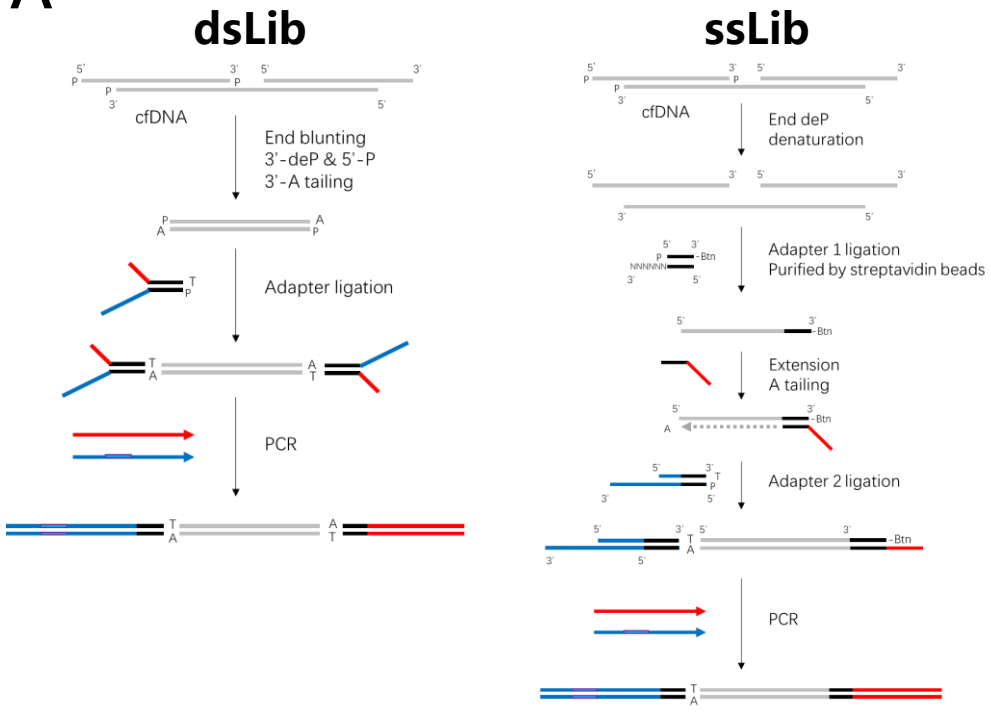
Supplementary Figure 1: (A) Library yield and (B) mean fragment length of sequenced libraries constructed by dsLib, ssLib, and the KAPA kit. Two replicates were performed for each experimental condition. Data are presented as mean ± SD. N.S, p>=0.05; *, 0.01<=p<0.05; **, 0.001<=p<0.01; ****, p<0.0001, as calculated by Student's t-test.

Supplementary Figure 2: Library yield of libraries constructed by dsmLib and ssmLib methods. Two replicates were performed for each experimental condition. Data are presented as mean ± SD. N.S, p>=0.05; *, 0.01<=p<0.05; **, 0.001<=p<0.01, as calculated by Student's t-test.

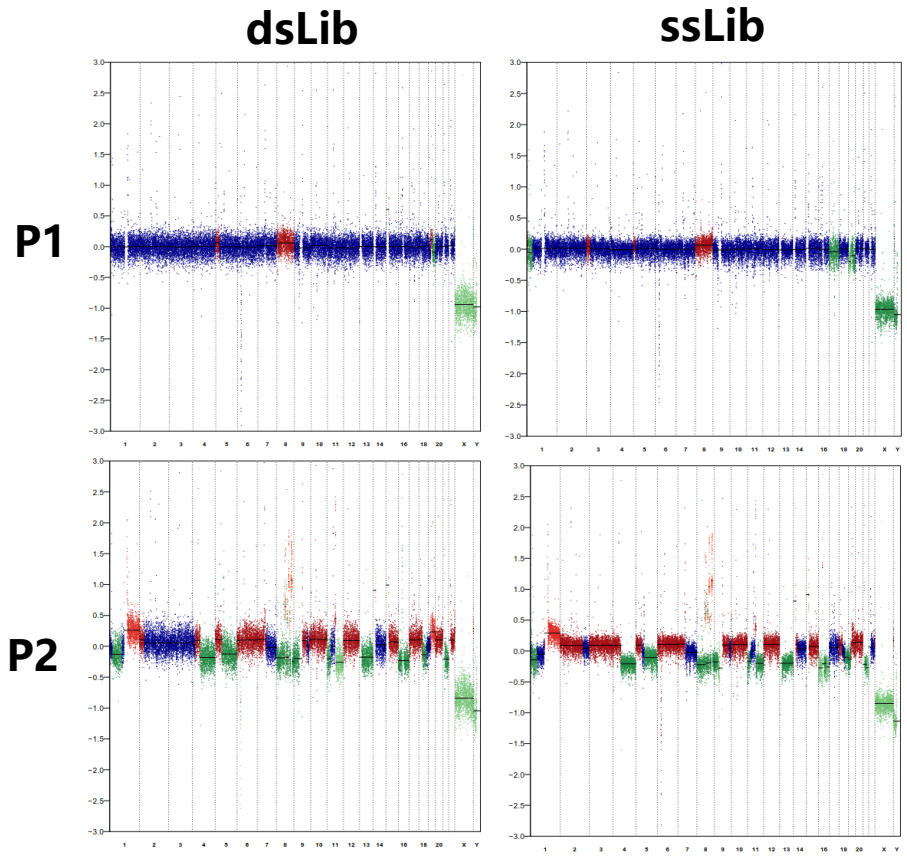
Supplementary Figure 3: Pearson correlation of methylation rates between (A) dsmLib or (B) ssmLib technical replicates at different DNA inputs.

Supplementary Figure 4: (A) CT conversion rate, (B) library amount, and (C) deduplicated depth of dsmLib libraries using ZYMO and NEB kits. Two replicates were performed for each experimental condition. Data are presented as mean ± SD. N.S, p>=0.05; *, 0.01<=p<0.05, as calculated by Student's t-test.

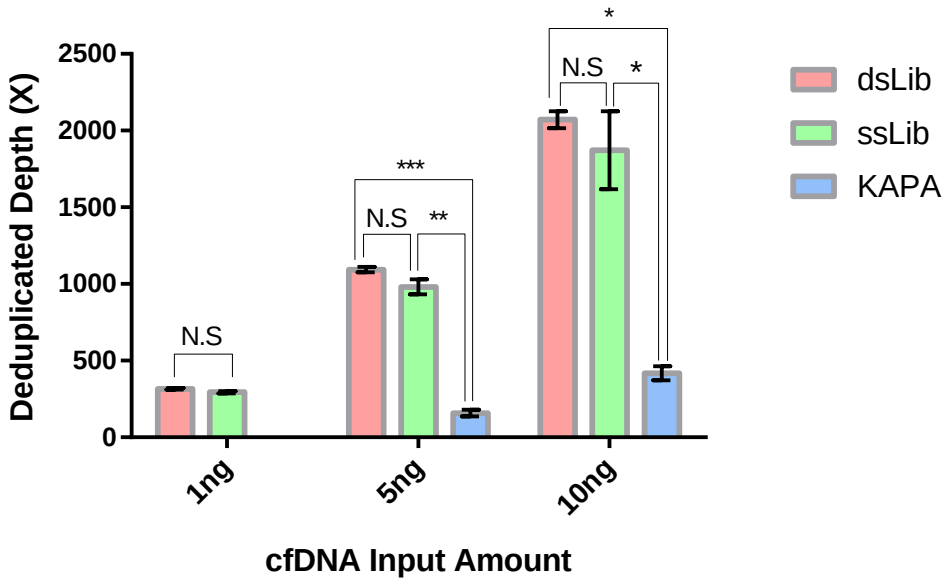
A


dsLib
ssLib
cfDNA
End blunting
3'-deP & 5'-P
3'-A tailing
Adapter ligation
PCR
cfDNA
End deP
denaturation
Adapter 1 ligation
Purified by streptavidin beads
Extension
A tailing
Adapter 2 ligation
PCR

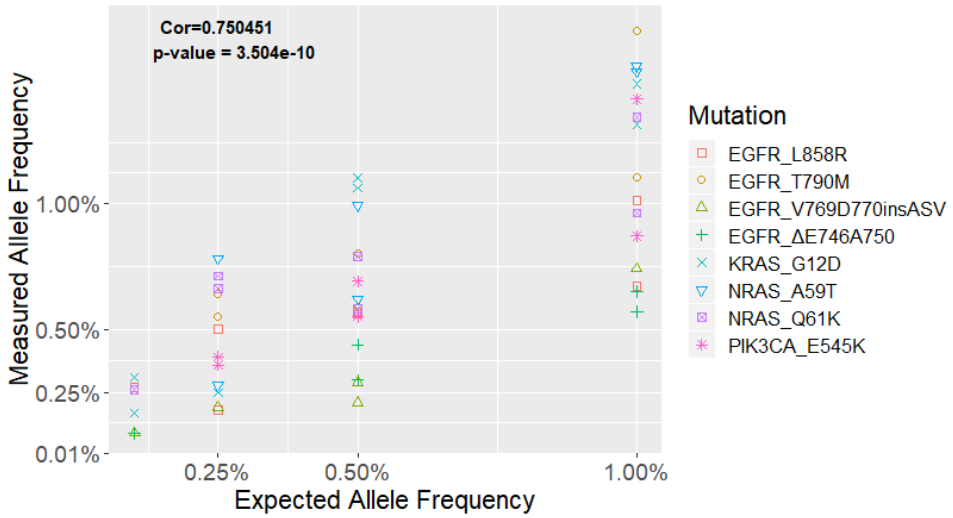
C

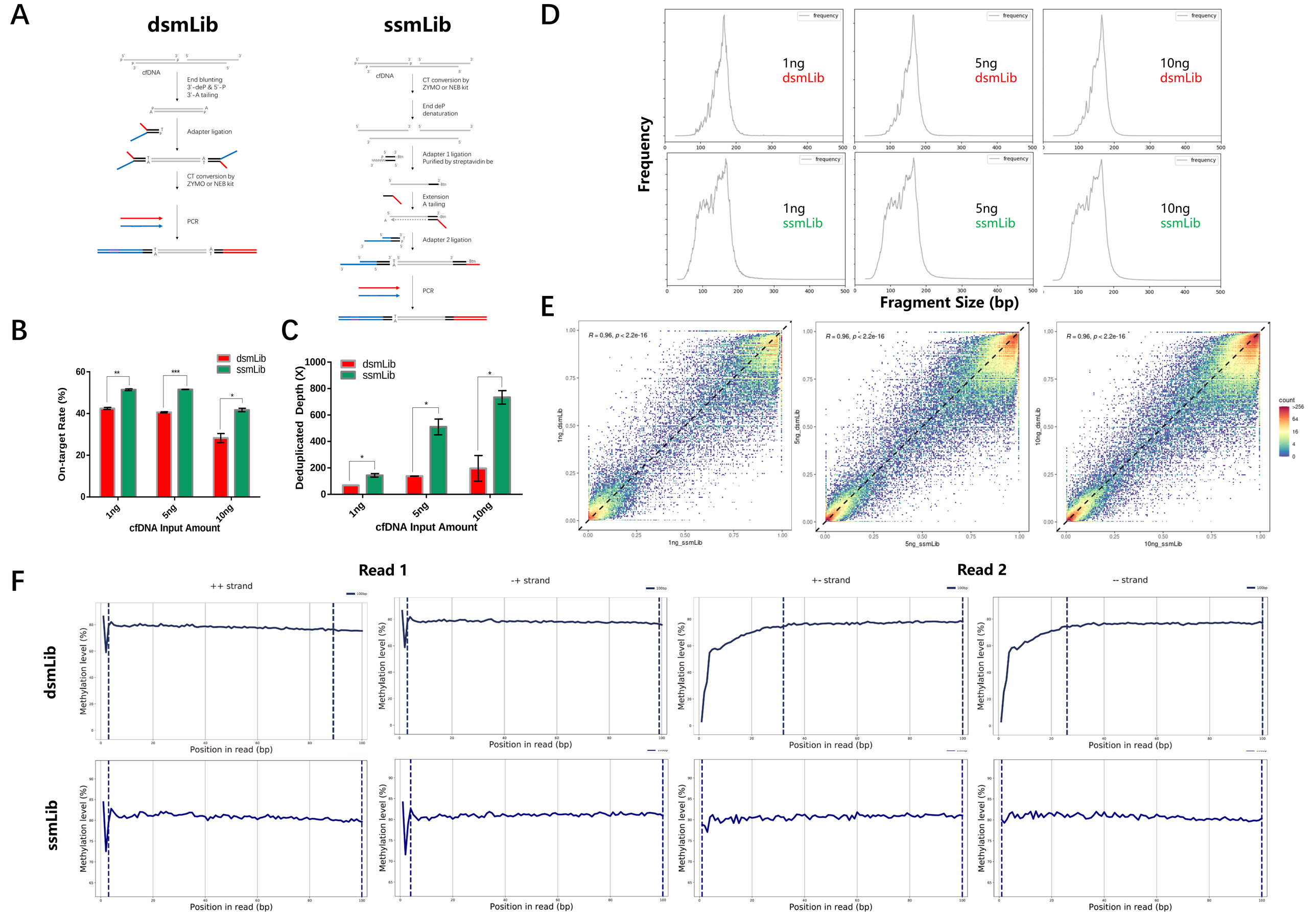

dsLib
ssLib
P1
P2

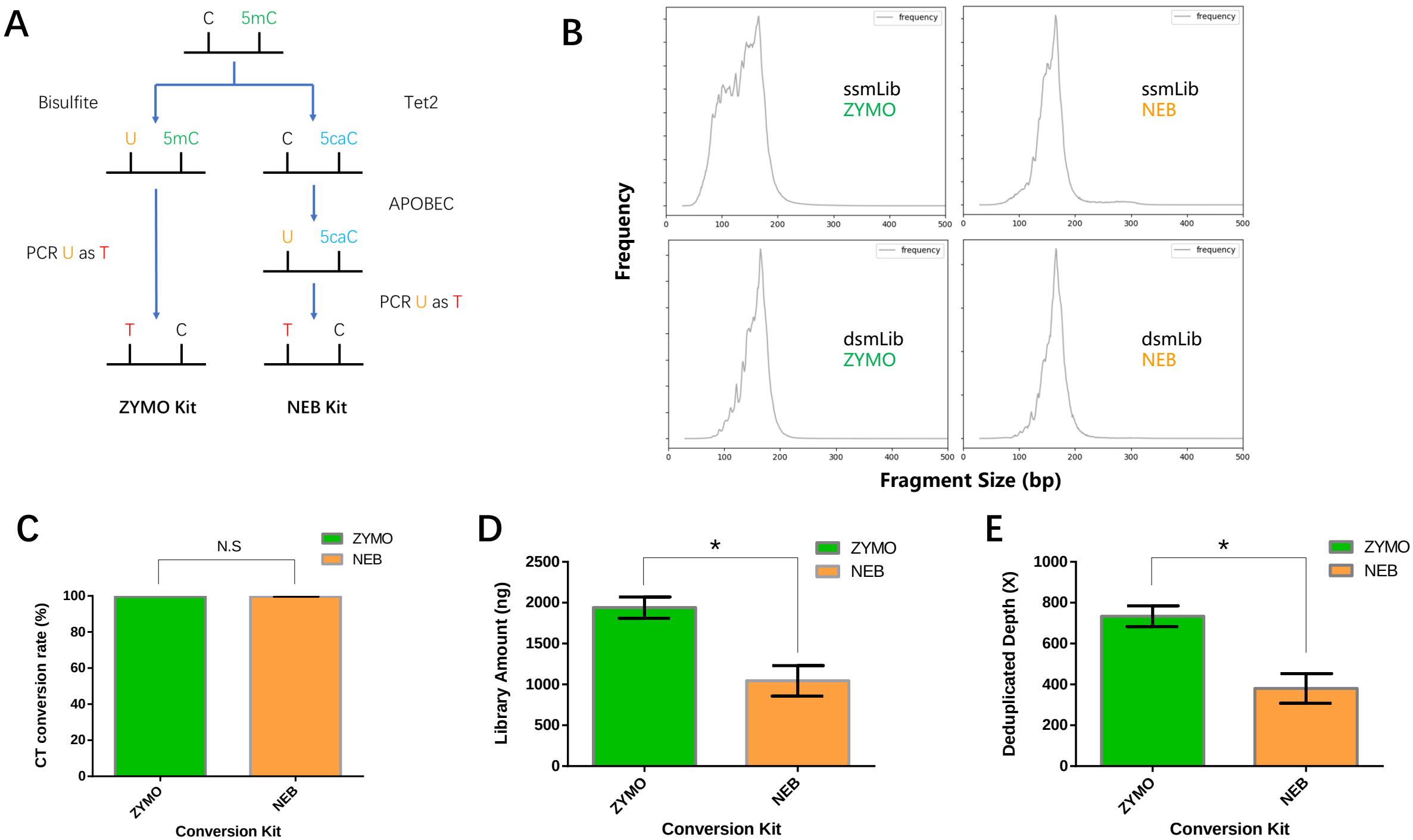
B


Deduplicated Depth (X)
2500
2000
1500
1000
500
0
N.S
***
N.S
**
*
N.S
*
1ng
5ng
10ng
cfDNA Input Amount
dsLib
ssLib
KAPA

D


Cor=0.750451
p-value = 3.504e-10
Measured Allele Frequency
1.00%
0.50%
0.25%
0.01%
0.25%
0.50%
1.00%
Expected Allele Frequency
Mutation
EGFR_L858R
EGFR_T790M
EGFR_V769D770insASV
EGFR_ΔE746A750
KRAS_G12D
NRAS_A59T
NRAS_Q61K
PIK3CA_E545K


A
dsmLib
ssmLib
cfDNA
End blunting
3'-deP & 5'-P
3'-A tailing
Adapter ligation
CT conversion by
ZYMO or NEB kit
PCR
CT conversion by
ZYMO or NEB kit
End deP
denaturation
Adapter 1 ligation
Purified by streptavidin be
Extension
A tailing
Adapter 2 ligation
PCR
B
dsmLib
ssmLib
On-target Rate (%)
cfDNA Input Amount
1ng
5ng
10ng
C
dsmLib
ssmLib
Deduplicated Depth (X)
cfDNA Input Amount
D
Frequency
Fragment Size (bp)
1ng dsmLib
5ng dsmLib
10ng dsmLib
1ng ssmLib
5ng ssmLib
10ng ssmLib
E
R = 0.96, p < 2.2e-16
1ng_dsmLib
1ng_ssmLib
5ng_dsmLib
5ng_ssmLib
10ng_dsmLib
10ng_ssmLib
count
F
Read 1
Read 2
++ strand
-+ strand
+- strand
-- strand
dsmLib
ssmLib
Methylation level (%)
Position in read (bp)


A
C
5mC
Bisulfite
Tet2
U
5mC
C
5caC
APOBEC
U
5caC
PCR U as T
PCR U as T
T
C
T
C
ZYMO Kit
NEB Kit
B
Frequency
frequency
ssmLib
ZYMO
ssmLib
NEB
dsmLib
ZYMO
dsmLib
NEB
0
100
200
300
400
500
Fragment Size (bp)
C
N.S
ZYMO
NEB
CT conversion rate (%)
0
20
40
60
80
100
ZYMO
NEB
Conversion Kit
D
*
ZYMO
NEB
Library Amount (ng)
0
500
1000
1500
2000
2500
ZYMO
NEB
Conversion Kit
E
*
ZYMO
NEB
Deduplicated Depth (X)
0
200
400
600
800
1000
ZYMO
NEB
Conversion Kit