Tagmentation on Microbeads: Restore Long-Range DNA Sequence Information Using Next Generation Sequencing with Library Prepared by Surface-Immoblized Transposomes

He Chen,† Jiacheng Yao,‡ Yusi Fu,† Yuhong Pang,† Jianbin Wang,*‡ and Yanyi Huang*†§

†Beijing Advanced Innovation Center for Genomics (ICG), Biodynamic Optical Imaging Center (BIOPIC), School of Life Sciences, and Peking-Tsinghua Center for Life Sciences, Peking University, Beijing 100871, China
‡School of Life Sciences, and Tsinghua-Peking Center for Life Sciences, Tsinghua University, Beijing 100084, China
§College of Engineering, and College of Chemistry and Molecular Engineering, Peking University, Beijing 100871, China

Supporting Information

ABSTRACT: The next generation sequencing (NGS) technologies have been rapidly evolved and applied to various research fields, but they often suffer from losing long-range information due to short library size and read length. Here, we develop a simple, cost-efficient, and versatile NGS library preparation method, called tagmentation on microbeads (TOM). This method is capable of recovering long-range information through tagmentation mediated by microbead-immobilized transposomes. Using transposomes with DNA barcodes to identically label adjacent sequences during tagmentation, we can restore inter-read connection of each fragment from original DNA molecule by fragment-barcode linkage after sequencing. In our proof-of-principle experiment, more than 4.5% of the reads are linked with their adjacent reads, and the longest linkage is over 1112 bp. We demonstrate TOM with eight barcodes, but the number of barcodes can be scaled up by an ultrahigh complexity construction. We also show this method has low amplification bias and effectively fits the applications to identify copy number variations.

KEYWORDS: microbeads, surface, tagmentation, sequencing, long-range information

INTRODUCTION

Next generation sequencing (NGS) technologies have been rapidly developed in the past decade and have revolutionized current research and applications in life sciences and medicine.1,2 For most NGS methods, especially those prevalent in high-throughput approaches, it is challenging to provide long-range DNA information because of the limitation in library size and read length. Such characteristics may hinder various applications in which the long-distance relationship of the DNA sequences is important—repeat-region sequencing,3 genome assembly,4 haplotype phasing,5 and alternative splicing analysis to name a few.6 Although single-molecule sequencing technologies are promising on long read-length,7,8 their sequencing accuracy and throughput are still insufficient.9,10 A practical approach to filling such genomic information gaps is to encode additional information in each short NGS read to retain the long-range relationship between reads. Matured and widely used examples include paired-end,11 mate-pair,12 HiC,13,14 synthetic long read,15,16 and several microfluidic-based technologies.10,17−19 In general, however, these experiments are labor-intensive and cost-inefficient. Therefore, there is a strong need for a high throughput, cost-effective, and versatile method to recapture such inter-read information in NGS.

Recently, hyperactive Tn5 transposase-mediated library preparation methods have shown great potential in simplifying the experimental process by performing both DNA fragmentation (cleaving) and adapter ligation (tagging) simultaneously.20,21 Two Tn5 transposases with each binds to a double-strand oligonucleotide containing the 19-bp mosaic end (ME) could be in vitro assembled into an active dimeric Tn5 transposome.22 The constructed Tn5 transposome can randomly fragmentize template DNA and ligate the oligonucleotides onto the ends of fragmented DNA.23 This tagmentation process is proven to be highly efficient and rapid, with low starting material requirement.20 Various
applications have been developed on the basis of Tn5 tagmentation and sequencing. For example, ATAC-seq maps the transposase-accessible chromatin region and provides genome-wide open chromatin profiles even for single cells.\textsuperscript{24} With the capability to label template DNA with specific oligonucleotides, the tagmentation process is an ideal editing tool to attach barcodes to each fragment thus leaving the information on the template DNA. Amini et al. developed CPT-seq which uses the contiguity preserving Tn5 tagmentation and combinatorial indexing to capture the haplotype information.\textsuperscript{25} However, this method requires sufficient resources (e.g., abundant Tn5 transposase, custom sequencing recipe and primers) and it is difficult to be adapted to other sequencing applications.

In this work, we report a tagmentation on microbeads (TOM) strategy to restore long-range DNA sequence information using next generation sequencing with library prepared by surface-immobilized transposomes. We use microbeads-attached identical-barcoded oligonucleotides and encode the information to template DNA through tagmentation. With each unique barcode representing the original DNA molecule, the long-range DNA information is retained. We did a proof-of-concept experiment with eight barcodes and verified the existence of this long-range linkage in the sequencing data. TOM successfully restored the linkage information and could be used on both 5′- and 3′-end tethered beads with high efficiency and low bias. It can also be scaled up to construct ultrahigh complexity barcodes for further applications.

\section*{RESULTS AND DISCUSSION}

In conventional NGS experiment, DNA fragments are sequenced independently. To restore the long-range information, a connection between reads has to be built. We took advantage of the barcoding ability of Tn5 transposome. During the tagmentation process, Tn5 transposome cuts the template DNA and adds the oligonucleotides with barcodes to both ends of the fragmented DNA, which would be presented in different fragment ends (reads) during sequencing. This feature provides the possibility to link two “neighbor” fragments that cut by one Tn5 transposome, because they share the two barcodes from that single Tn5 transposome. However, barcodes with ultrahigh complexity were required to distinguish every fragment, which in theory should be higher than the number of fragmented molecules (∼10\textsuperscript{10} for 5 ng genomic DNA), making this approach hard to implement and extremely expensive. To reduce the barcode complexity requirement, we took advantage of another intrinsic feature of Tn5 transposition, a 9-bp target sequence duplication after tagmentation and repairing.\textsuperscript{26} Such 9-bp end-duplications are present at both ends of molecules after in vitro gap-filling and can be treated as natural extensions of barcodes. Because the Tn5 cut-sites are randomly distributed along the template DNA,\textsuperscript{20} it is unlikely that two molecules share the same cut-site by coincidence. Thus, fragments with the same barcode but different 9-bp end-duplications should not have been cut by the same transposome. Theoretically, combining the barcode and the 9-bp end-duplication sequence as a unique fragment identifier (UFI) can accurately identify the neighbor fragments with barcode complexity 5 orders of magnitude (4\textsuperscript{9} = 262 144) lower than the theoretical requirement. Ideally, with 100% reaction efficiency and recapture rate of all fragments, we could use UFIs to concatenate every fragment to reconstruct the original full-length template DNA molecules, only requiring the barcode complexity to be larger than the count of the most abundant fragment. The one essential requirement for this approach is that the two barcodes from one transposome should be either the same or with
certain linkage. However, such requirement is difficult to achieve in in-solution reaction but feasible when barcodes are immobilized on solid-phase.

We designed a Tn5 tagmentation strategy with solid-phase immobilized barcoded oligonucleotides to recover long-range information and reduce the requirement for high complexity barcodes. The overall experimental process of our TOM method (Figure 1A) contains two major steps: (1) use solid-phase immobilized barcoded oligonucleotides with Tn5 transposase to tagment the DNA fragments; (2) amplify the DNA fragments with PCR, followed by standard NGS procedures. The first step is critical, requiring careful design of the surface immobilized oligonucleotides (Figure 1B). Each bead is linked, through biotin–streptavidin binding, with $10^5$–$10^6$ copies of DNA oligonucleotides with identical sequence. The oligonucleotide contains several parts: a 5′-biotin modification, a linker to reduce steric effect, a PCR handler to amplify the fragment with barcodes after tagmentation, a 12-bp barcode, and a 19-bp hyperactive Tn5 mosaic end sequence (Tn5ME) that is necessary for transposome construction. To ensure each DNA molecule was tagmented by only one bead, which will result in identical barcode representing one molecule, beads with a comparable number of input DNA molecules were added to the reaction ($>10^6$ in each reaction). If more than one DNA molecules were tagmented by one bead, these DNA molecules were unlikely to be from the same genomic region, and they could be distinguished by their genomic position when aligned to the reference genome. Also, considering the input DNA length ($>20$ kb or $<6.8$ μm), we used beads with comparable circumference (2 μm in diameter). The small diameter of the beads can help keeping beads in suspension to reduce the probability of tagmentation occurring on multiple beads. After tagmentation and gap-filling, the neighbor fragments will have the same barcode and 9-bp end-duplication that provide reliable linkage between fragments (Figure 1C and Figure S1).

As a proof-of-principle test, we started with 5 ng of genomic DNA from a human cell line HEK293 as input to verify whether the barcodes combined with the end-duplication can be used as the UFIs to link the neighbor fragments in a low-complexity-barcode scenario. We first synthesized oligonucleotides with eight different barcodes and immobilized them on beads separately. With $10^5$–$10^6$ copies of identical barcoded oligonucleotides on each bead, we merged equal portions of eight beads to form a pool (identical-on-bead). By adding Tn5 transposase to the beads pool, Tn5 transposome were assembled. Using such bead-immobilized Tn5 transposome for tagmentation, each fragment from one template DNA was linked on the same bead with the same barcode through phosphodiester bond. The tagmentation reaction was conducted for 60 min to maximize the tagmentation efficiency. Because of the magnetic beads employed, the reaction can be terminated by changing buffer after magnet separation without requirement for purification. Then 12–18 cycles of PCR were performed to enrich the fragments with barcodes at both ends, followed by a standard NGS library preparation with slight modification.

Paired-end sequencing was performed to obtain barcodes at both ends of insert DNA. The barcodes could be easily recognized from the sequencing reads through their characteristic positions and sequence context. We then examined the barcodes of each pair of reads to check whether each template DNA molecule was tagmented by one bead. If both reads in a pair had identical barcodes, this fragment was either tagmented by a single bead, or tagmented by two beads with the same barcode which was a rare event through increasing the variety of barcodes. In the sequencing data, we observed that majority of the paired-end reads (>85%) had the same barcode, which indicates each template DNA was tagmented by one bead. As a control experiment, we immobilized eight different barcodes on one bead pool (mixed-on-bead). In this way, the barcodes in the paired-end reads were randomly arranged after tagmentation and sequencing. As expected, a small portion of the paired-end reads (<13%) had the same barcode, which was the result from random assignment (Figure 2A). These observations proved that tagmentation for each template DNA molecule indeed occurred on a single bead. Therefore, using TOM, we can encode barcodes into DNA molecules as unique molecule identifiers (UMIs) to label and quantify molecules.

The application of UMI is limited if the number of barcodes is smaller than the number of identical molecules. Low barcode complexity is not sufficient to determine the origin of fragments because fragments from different template DNA molecules could share the same barcode. Thus, it still requires high barcode complexity. Recently, Zhang et al. developed a method call CPTv2-seq using the same principle to resolve haplotype
phasing by a beads library with ~150k barcodes,\textsuperscript{28} which is not feasible for most researchers. Conventionally, the genomic position (Tn5 cut-site in our case) can also be considered as a barcode. So, in our low-barcode-complexity scenario, when the barcode is combined with the 9-bp end-duplication at the cut-site for each fragment end, some linkage between the neighbor fragments might be restored. We first confirmed the 9-bp end-duplications exist in the NGS sequencing data by comparing the fragments prepared by tagmentation in solution-phase Tn5 transposome with those by sonication.\textsuperscript{3} The signature of 9-bp the fragments prepared by tagmentation in solution-phase Tn5 duplications exist in the NGS sequencing data by comparing barcode. So, in our low-barcode-complexity scenario, when the feasible for most researchers. Conventionally, the genomic on-bead DNA synthesis strategy can be employed.\textsuperscript{31} Conventionally, oligonucleotides are synthesized using the 3′-to-5′ phosphoramidite chemistry,\textsuperscript{32,33} which produces oligonucleotides with their 3′ end attached to the surface. Because the Tn5 transposase ligates the 3′ terminus of each transposon strand to target DNA by 3′−OH groups mediated nucleophilic attacks and strand transfer,\textsuperscript{34,35} oligonucleotides with the 3′ end attached to the beads rely on their opposite strand to be linked to the target DNA. We hence employed a different immobilization scheme by tethering 3′ end of the oligonucleotides to the microbeads to verify whether the orientation of the immobilized oligonucleotides and the additional synthesis step for the opposite strand would affect tagmentation. To generate these functional sequences, we performed high-fidelity 5′-3′ elongation initiated by a forward elongation (FE) primer (Figure 3A,B). We measured the reaction efficiency using a FAM labeled FE primer and a TAMRA-labeled probe. Replacement of the TAMRA signal by the FAM signal due to elongation-induced strand displacement indicates the gener-

**Figure 3.** Workflow for 3′-end tethered beads construction and flow cytometry analysis. (A) Workflow for construction of barcoded beads. Each barcoded oligonucleotide was tethered on beads independently. (B) Workflow for Tn5 transposome on beads construction. Two fluorescent-labeled primers were annealed to the on-bead oligonucleotides. 5′-3′ elongation was performed to generate the forward-strand barcode and Tn5ME sequence, followed by on-bead Tn5 transposome assembly. (C) Kernel density estimate plots of flow cytometry results in log scale. Beads: beads without any primer. N/A: beads with two primers but not elongated by polymerase.
Tagmentation resulted in a broad distribution of fragment size, and after library construction and size selection, the mean size is ∼450 bp, which is compatible with current NGS platforms (Figure 4A). Because of the high-fidelity elongation, more than 80% of the barcodes at both ends of paired-end reads were recognized (Figure 4B). Our demonstration showed that oligonucleotides tethered to beads by 3′-end can be used to create high-quality double-strand DNA for barcoding tagmentation. The compatibility of different beads indicated the potential of applying ultrahigh barcode complexity DNA synthesis.

To link the fragments and reconstruct the sequence of original template DNA, a high-reaction-efficiency procedure with low bias and low tolerance for fragment loss is required, which highly depends on barcode design. In our previous design, both ends of the fragments have the same PCR handles, barcodes, and Tn5ME sequences. These long (54 bp) sequences may cause intramolecule hairpin formation during PCR (Figure SSA), and significantly lower the amplification efficiency and introduce bias. Capillary electrophoresis analysis showed that when using beads with mixed barcodes, the size distribution of the fragments was shifted downward compared to those using identical barcoded beads (Figure 5A). This shift suggests an amplification bias toward longer molecules when using identical barcoded beads. This bias might result from the intramolecule hairpin for short fragments. To eliminate this bias, we designed a “wobble” version of a barcode by introducing random nucleotides into the barcode sequence at specific positions to prevent hairpin formation (Figure S5A). With optimized annealing temperature, after this optimization, the fragment size distribution was similar between experiments with mixed and identical barcoded beads. Because the random nucleotides were only located in certain positions in the barcode, we can still accurately identify more than 80% of the barcodes based on their context (Figure S5B). Furthermore, we tested whether different barcodes would introduce sequence-dependent amplification bias by using three different sets of barcodes. We found even distributions of these barcodes, which

Figure 4. TOM performance using different microbeads. (A) Fragment-size distribution after tagmentation (blue) and library generation (yellow) using different microbeads. (B) Percentage of the fragments that have recognized barcodes in both ends while using 3′-end tethered beads and 5′-3′ elongation.

Figure 5. Bias analysis. (A) Boxplot of the fragments size distribution when using two types of beads (Mixed and Identical) and two types of barcodes (Fixed and Wobble). (B) Percentage of each barcode presented in fragments when using different numbers of barcodes (labeled as B). Dashed line represents the expectation. (C) Whole genome copy number profiles using TOM.
indicates no amplification bias was introduced (Figure 5B). We also verified TOM has no sequence dependent bias at whole genome level by performing CNV analysis in a 1M-bin resolution (Figure 5C). When compared with solution-phase Tn5 tagmentation, our method shows highly reproducible and low noise at the whole genome level.

CONCLUSIONS

We developed a method, TOM, to restore long-range information by a sequencing library that was created through tagmentation and barcoding with microbead-immobilized Tn5 transposome. We showed that TOM has acceptable efficiencies for both 5′- and 3′-ends tethered oligonucleotides on microbeads. Tethering oligonucleotides with 5′-ends can construct mid/high complexity barcodes, whereas the 3′-end can be more suitable for ultrahigh-complexity barcodes that required split-and-pool DNA synthesis. We demonstrated that this method could reconstruct template DNA molecule by linking adjacent reads through UFLs, which combine uniquely designed barcodes with the 9-bp duplications at the cut-sites. By using TOM, 4.5% of the sequenced fragments could be linked with their neighbor, and the longest linkage is over 1112 bp.

We are the first to show that the Tn5 transposition generated 9-bp duplications at cut-sites can be used to reduce the barcode complexity requirement for stitching adjacent reads. TOM can be widely used in various sequencing applications, such as haplotype phasing, structure variations detection, alternative splicing analysis, and only requires low complexity of barcodes. Our method could be further improved by increasing the template DNA capture efficiency with the Tn5 transposome on beads and by employing a barcode system with higher complexity. In general, there was a difference requirement trade-off between barcode complexity and capture efficiency, which should be taken into consideration for different applications.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acsami.8b01560.

Detailed experimental process, bioinformatics analysis, and Figures S1−S5 (PDF)

AUTHOR INFORMATION

Corresponding Authors
*E-mail for J.W.: jianbinwang@mail.tsinghua.edu.cn.
*E-mail for Y.H.: yanyi@pku.edu.cn.

ORCID

Yanyi Huang: 0000-0002-7297-1266

Author Contributions

H.C., J.W., and Y.H. conceived the project. J.Y. performed the protein purification. H.C. performed the experiments and analyzed the data. H.C., Y.F., Y.H., and J.W. wrote the manuscript. All authors contributed to the discussion and reviewed the manuscript.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors thank Dr. Shuang Geng for discussions and comments at primary stage of the project. We also thank Bo Wu for help with the protein purification. This work was supported by the Ministry of Science and Technology of China (2016YFC0900100 for J.W. and Y.P.), and the National Natural Science Foundation of China (21327808, 21525521 to Y.H. and 21675098 for J.W.).

REFERENCES


