Published online 18 January 2018

Nucleic Acids Research, 2018, Vol. 46, No. 7 ***e40***
doi: 10.1093/nar/gky016

Highly selective retrieval of accurate DNA utilizing a pool of *in situ*-replicated DNA from multiple next-generation sequencing platforms

Hyeonseob Lim[1], Namjin Cho[1], Jinwoo Ahn[1], Sangun Park[1], Hoon Jang[1], Hwangbeom Kim[1], Hyojun Han[1], Ji Hyun Lee[2,*] and Duhee Bang[1,*]

[1]Department of Chemistry, Yonsei University, Seoul 03722, Korea and [2]Department of Clinical Pharmacology and Therapeutics, College of Medicine, Kyung Hee University, Seoul 02447, Korea

Received August 01, 2017; Revised December 01, 2017; Editorial Decision January 05, 2018; Accepted January 11, 2018

ABSTRACT

Scalable and cost-effective production of error-free DNA is critical to meet the increased demand for such DNA in the field of biological science. Methods based on 'Dial-out PCR' have enabled the high-throughput error-free DNA synthesis from a microarray-synthesized DNA pool by labeling with retrieval PCR tags, and retrieving error-free DNA of which the sequence is identified via next generation sequencing (NGS). However, most of the retrieved products contain byproducts due to background amplification of redundantly labeled DNAs. Here, we present a highly selective retrieval method of desired DNA from a pool of millions of DNA clones from NGS platforms. Our strategy is based on replicating entire sequence-verified DNA molecules from NGS plates to obtain population-controlled DNA pool. Using the NGS-replica pool, we could perform improved and selective retrieval of desired DNA from the replicated DNA pool compared to other dial-out PCR based methods. To evaluate the method, we tested this strategy by using 454, Illumina, and Ion Torrent platforms for producing NGS-replica pool. As a result, we observed a highly selective retrieval yield of over 95%. We anticipate that applications based on this method will enable the preparation of high-fidelity sequenced DNA from heterogeneous collections of DNA molecules.

INTRODUCTION

Scalable production of error-free DNA may provide a variety of genetic material in diverse fields of biological science. Currently, steps related to controlled-pore glass (CPG)-based oligonucleotide synthesis (1), molecular cloning (2), and selection by Sanger sequencing (3) are standard protocols for *in vitro* production of high-fidelity oligonucleotides (Supplementary Figure S1A). Due to the prohibitive cost of traditional oligonucleotide synthesis for high-throughput biological applications, oligonucleotides cleaved from microarrays have been used for large-scale DNA preparation as a low-cost alternative (4–7). However, a large proportion of microarray-derived oligonucleotides contains synthesizing errors depending on the size of oligonucleotides and the synthesis method (8). Selection of error-free oligonucleotides has been previously inefficient due to laborious cloning procedures and costly Sanger sequencing. While some error-reduction methods have been developed (9–12), these methods still involve labor- and cost-intensive efforts.

Recently, Matzas *et al*. (13) made an advancement in accurate gene synthesis with respect to retrieval of sequence-verified DNA (Supplementary Figure S1B, upper). They prepared programmable oligonucleotides cleaved from a microarray, performed 454 sequencing (14), and utilized a robotic pick-in-place pipette to retrieve error-free sequence reads from the sequencing flow cell. In addition, Lee *et al.* reported 'Sniper Cloning' (15), which enables fast retrieval of targets on NGS platforms using laser-pulse technology (Supplementary Figure S1B, lower). However, these bead-based mega-clone strategies are currently limited to 454-based technologies and are not applicable in widely-utilized Illumina sequencing platforms (16).

Alternatively, a PCR-based DNA retrieval method (17–19) termed 'dial-out PCR', which can specifically amplify desired DNAs from a pool of complex DNA libraries, (Supplementary Figure S2A) was also reported. In this method, designed oligonucleotides are synthesized on a programmable microarray, to which ∼20-bp degenerate barcoded nucleotides ('dial-out tag') (20) are attached as flanking sequences, and resulting oligonucleotides are subsequently read by NGS. The dial-out tag serves as both DNA identification tags and PCR priming loci to selec-

*To whom correspondence should be addressed. Tel: +82 2 2123 2633; Email: duheebang@yonsei.ac.kr
Correspondence may also be addressed to Ji Hyun Lee. Email: hyunihyuni@khu.ac.kr

tively retrieve the desired sequences. The dial-out PCR is cost-effective and does not require any specialized equipment, such as robotic pipettes or laser platforms, to retrieve clones. One of the dial-out PCR based methods recently showed a useful strategy utilizing combinatorial barcode tags (CBT) termed 'static tag library' (19) of which the dial-out PCR primers can be reused continuously at the next other attempts. Using this CBT approach, the great expense of preparing a myriad of barcoded primer pairs could be saved.

However, the dial-out PCR-based retrieval of error-free DNA has a scalability limitation. Only a sub-population of prepared libraries is subjected to NGS, which results in a discrepancy between the pool used for dial-out PCR ('pre-NGS pool') and the pool sequenced on the NGS flow cell. Thus, it was not possible to identify the selected dial-out tag whether it was uniquely or redundantly labeled in pre-NGS pool (Supplementary Figure S2B).

If the tags of retrieval target were misidentified as unique, non-targeted products would also be retrieved and left in the background. The previous report utilizing 'static tag library' (i.e. CBT library) (19) is also in agreement with this trend. Almost 22% of retrieved product was observed as byproducts when ca. 5 million pairs of dial-out tags were used for labeling only 250 target DNAs. We expect that dial-out PCR-based retrieval may become less specific when applied to an increasingly complex DNA mixture. To prohibit misidentification of a unique pair of dial-out tags, a new method is needed to control the number of molecules to synchronize the population of DNA pool for dial-out PCR and NGS data.

Here, we present a method by replicating the library pool from a NGS plate or a flow cell whose entire clonal population is comprehensively sequenced. Using the method, we successfully reduced the pool's DNA population size to millions of clones ('NGS-replica pool') then NGS-replica pool was used for a dial-out PCR template instead of pre-NGS pool (Figure 1). This population size is significantly smaller than that of the pre-NGS pool, which leads to a lower probability of redundant labeling with CBT. To evaluate the method, 454-, Illumina- and Ion Proton-based sequencing platforms were tested to be replicated and used for retrieving microarray-derived DNA or sheared genomic DNA. Then, retrieval fidelity and selectivity were compared to general dial-out PCR method.

Moreover, we extended our method to a tag-directed assembly method which utilizes full- or sub-assembled fragments as a retrieval target whose length are usually longer than the common length of sequencing read (21). In the previous method, retrieval step is proceeded after a round of full- or sub-assembly (up to almost 500 bp), barcode tagging, shotgun sequencing and *de novo* assembly of NGS data. By performing these steps, this method can use the longer building blocks and reduce the number of retrieving fragments. However, this method is limited by the absence of appropriate population-control method. To overcome the limitation, serial dilution step is only used to minimize the population so that huge NGS data are utilized to cover all constructs in the pool. We expected that our method would be useful for controlling the population without the dilution step. Thus, we tested whether the fully-assembled error-free KRAS (570 bp) and GFP (810 bp) genes could be selectively retrieved from the assembled construct.

MATERIALS AND METHODS

Simulation for demonstrating specificity of the CBT-based labeling method

Prior to the experiment, we investigated specificity of CBT-based labeling method for pre-NGS pool (Supplementary Figure S3) and NGS-replica pool by counting the number of DNA molecules per CBT. For this simulation, we carried out the following procedures using Python scripts: (i) A virtual library (DNA pool) was composed of 100 million unique molecules (equivalent to ca. 0.03 ng NGS library containing ca. 300 bp fragments), and each DNA molecule was simplified as a unique integer. (ii) Each DNA molecule in the virtual library was randomly tagged with one of 2000 forward barcodes (f1, f2, f3, . . . , f2,000) and one of 2000 reverse barcodes (r1, r2, r3, . . . , r2000), which could generate 4×10^6 CBTs. The resulting sequences were considered as the 'pre-NGS pool'. (iii) One hundred thousand barcode-tagged DNA molecules (throughput of a 454 GS Junior sequencer) were randomly picked from the pre-NGS pool and considered this library as 'NGS-replica pool'. (iv) The number of DNA molecules per CBT was counted for all CBTs in the pre-NGS and NGS-replica pools. This value indicates the tagging specificity of the library; for example, a value of 1 denotes a unique CBT.

CBT-labeled DNA library preparation for 454 sequencing-based experiment

For the preparation of DNA substrates, human genomic DNA (NA 12878) was utilized as a model. We sheared 1 μg genomic DNA into 180-bp fragments using a M220 focused ultrasonicator™ (Covaris, Woburn, MA, USA), repaired both ends of the DNA fragments, dA tailed on the 3′ end of the DNA fragments, ligated NEBNext Adaptors (New England BioLabs, Ipswich, MA, USA) for flanking CBT library, and cleaved ideoxyU bases on the adaptor sequence. End-repair, dA-tailing, and ligation reactions were performed according to standard protocols of the SPARK™ DNA Sample Prep Kit (Enzymatics, Beverly, MA, USA). After the ligation, product was enriched by PCR using common primers. 14 μl ligated-product, 2.5 μl CBT_flk_fwd primer, 2.5 μl CBT_flk_rev primer, 6 μl dH_2O, and 25 μl KAPA HiFi polymerase (KAPA Biosystems, Wilmington, MA, USA) were mixed, and the reaction was performed under the following conditions: 5 min at 95°C; 6 cycles of 30 s at 95°C, 30 s at 65°C, 30 s at 72°C; and 10 min at 72°C. The amplified-products were termed as the 'sheared gDNA library'.

For preparation of CBT sequences, 2133 forward and 2133 reverse barcode primer sequences were designed, and barcode libraries were synthesized using a programmable microarray (CustomArray, Inc., Bothell, WA, USA) (Figure 1B, Supplementary Table S1). We designed CBT sequences with little similarity and more specificity by following these design principles: (i) melting temperature (T_m) close to 60°C, (ii) designing barcodes with three or more

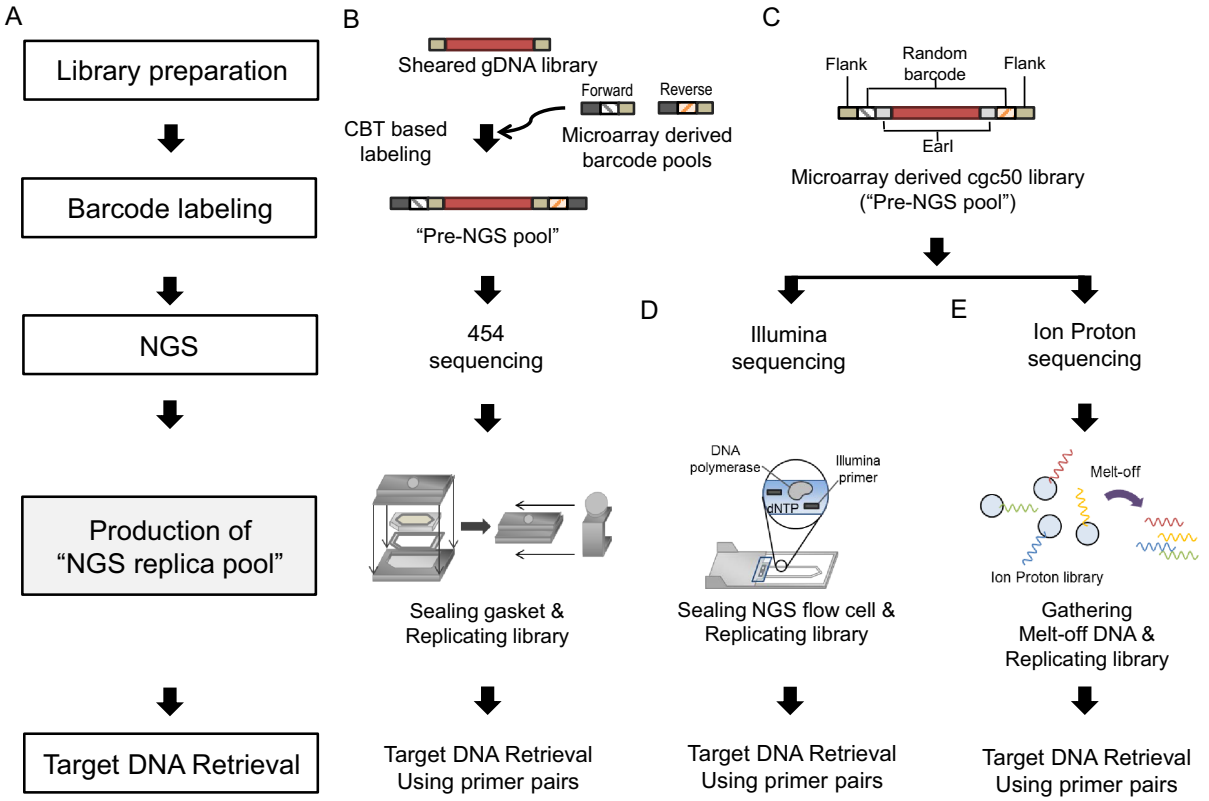

Figure 1. Schematic of *in situ* replication of DNA molecules from next-generation sequencing (NGS) platforms and subsequent PCR-based retrieval of target sequences. (**A**) Process flow chart for PCR-based methods for the retrieval of error-free DNA targets from an NGS-replica pool. (**B**) Preparation strategy of 454 GS Junior sequencing-based retrieval. Combinatorial barcode-tagged (CBT) pools were processed from microarray-synthesized oligonucleotides and subsequently ligated to the sheared genomic DNA as flanking sequences. The library was replicated in a sealed NGS plate. (**C**) Preparation strategy of a pre-NGS pool (MiSeq and Ion Proton). The barcoded library (cgc50 pool) was directly synthesized on a microarray. (**D**) Schematic of library replication in a MiSeq flow cell. (**E**) Schematic of library replication using melt-off DNA in the Ion Proton system. This process could be automatically performed using an Ion OneTouch™ ES system.

base differences from one another at each nucleotide position and (iii) excluding barcodes with three or more repeated bases (e.g. 'AAA') to avoid homopolymer sequencing errors. For amplification of the forward barcoded oligonucleotide pool, 0.5 μl tagged oligonucleotide pool DNA, 2.5 μl 454_fwd primer, 2.5 μl flk_fwd primer, 25 μl KAPA HiFi polymerase, and 19.5 μl dH_2O were mixed and placed in a thermal cycler. Polymerase chain reactions were performed under the following conditions: 5 min at 95°C; 20 cycles of 30 s at 95°C, 30 s at 60°C, and 30 s at 72°C; and 10 min at 72°C. Amplicons were electrophoresed on a 2% agarose gel and were purified with a MinElute Gel Extraction Kit (Qiagen, Valencia, CA, USA). The reverse barcode pool was amplified under the same conditions but with 454_rev and flk_rev primers instead of its forward counterparts. After amplification, the forward and reverse barcode pool DNA were flanked to both ends of the sheared human genomic DNA fragments by assembly PCR under the following conditions: 5 min at 95°C; 20 cycles of 30 s at 95°C, 30 s at 60°C, and 30 s at 72°C; and 10 min at 72 °C. The libraries were then subjected to 454 sequencing with a GS Junior sequencer (454 Life Sciences, Branford, CT, USA).

NGS analysis of 454 sequencing data

Raw FASTA format data were converted to FASTQ format. To align the sequence data to the human reference genome, barcode and adaptor sequences were trimmed, and the barcode information was saved in a new file using an in-house program. The trimmed sequences were aligned to hg19 (UCSC Genome Browser) using Novoalign (V2.07.18; http://www.novocraft.com), then the barcode information was re-attached to the aligned data. Improperly barcoded reads were removed, and duplicate reads with the same barcode, loci, and size information were counted. If multiple DNA substrates were tagged with the same barcode, that barcode was removed from the target retrieval list. Then, remaining sequences were considered candidates for retrieval.

Replication of the 454 DNA library in a sealed plate

The sequenced picotiter plate was removed from the 454 GS Junior sequencer before the final bleaching step, sealed using an in-house prepared gasket (Figure 1B), and filled with a PCR mixture of 60 μl 454_fwd primer, 60 μl 454_rev primer, 804 μl dH_2O, 24 μl dNTPs, 12 μl Phusion DNA polymerase, and 240 μl 5X Phusion HF Buffer (New England BioLabs). The sequenced 454 DNA library plate was then replicated in an isothermal incubator via five cycles

of 5 min at 95°C and 5 min at 70°C. The replicated pool (approximately 1 ml) was subsequently collected, purified with Agencourt AMPure XP beads (Beckman-Coulter, Indianapolis, IN, USA) and eluted with 50 μl dH_2O. For enrichment of the NGS-replica pool, 10 additional cycles of PCR amplification were performed. Three microliters of the NGS-replica pool, 7 μl dH_2O, 1 μl 454_fwd primer, 1 μl 454_rev primer and 10 μl KAPA HiFi polymerase were mixed per reaction ($n = 8$), and PCR was carried out under the following conditions: 5 min at 95°C; 10 cycles of 30 s at 95°C, 30 s at 60°C, and 30 s at 72°C; and 10 min at 72°C. Products were then purified with Agencourt AMPureXP beads.

Random-barcode labeled DNA library preparation for testing platform compatibility

Seventy-bp-long building block DNA flanked with restriction enzyme sites (EarI), 19-bp degenerate nucleotide-based barcode sequences (5′-NNNNANNNNTNNNNANNNN-3′ at both ends with $4^{16} \fallingdotseq 4 \times 10^9$ complexity), and 20-bp sequences were designed for the synthesis of 50 cancer-associated genes as target sequences. (Figure 1C and Supplementary Table S2). The 'cgc50 pool' (3742 unique oligonucleotides) was designed, synthesized and cleaved from the microarray (CustomArray). One microliter of the cgc50 pool, 1 μl illu_flk_fwd primer, 1 μl illu_flk_rev primer, 7 μl dH_2O and 10 μl KAPA HiFi polymerase were mixed, and the PCR amplification was performed as followed: 5 min at 95°C; 20 cycles of 30 s at 95°C, 30 s at 60°C and 30 s at 72°C; and 10 min at 72°C. Next, either Illumina adaptor or Proton adaptor was attached to the product. Illumina adaptor was ligated using standard protocols of the SPARK™ DNA Sample Prep Kit, and Proton adaptor was attached using PCR. 1 μl amplified product, 1 μl proton_fwd primer, 1 μl proton_rev primer, 7 ul dH_2O and 10 μl KAPA HiFi polymerase were mixed and amplification was carried out as follows: 5 min at 95°C; 20 cycles of 30 s at 95°C, 30 s at 60°C and 30 s at 72°C; and 10 min at 72°C. Each library was sequenced using the Illumina MiSeq instrument (Illumina, San Diego, CA, USA) and Ion Proton instrument, respectively.

NGS analysis and barcode verification of Illumina & Ion Proton sequencing data

NGS data were analyzed by the following procedure. (i) Content sequence was obtained from sequences located between 'CTCTTC' and 'GAAGAG' sequence (i.e. EarI) in a raw FASTQ file. (ii) 19-bp left barcode located between 'CTCTTC' and left flanking sequence (i.e. 'GACTCAGT GAGCGGAACGAT'), and 19-bp right barcode located between 'GAAGAG' and right flanking sequence (i.e. 'AT CACCGACTGCCCATAGAG') were obtained. (iii) Error-introduced contents were filtered (iv) Redundant pair of barcodes labeling more than two different contents were removed. Then, duplicates were counted.

Replication of the Illumina DNA library in a flow cell

PCR mixture was injected into the inlet, flow cell was sealed using sealing film (BioRad, Hercules, CA, USA) (Figure 1D), and replication reaction was carried out. The PCR mixture consisted of 5 μl illu_fwd primer, 5 μl illu_rev primer, 40 μl dH_2O and 50 μl KAPA HiFi polymerase under the same conditions used for the picotiter plate-based replication. For enrichment of the NGS-replica pool, PCR was performed for 10 additional cycles. Each reaction ($n = 10$) consisted of 5 μl NGS-replica pool DNA, 3 μl dH_2O, 1 μl illu_fwd primer, 1 μl illu_rev primer and 10 μl KAPA HiFi polymerase, which were mixed and amplified under the following conditions: 5 min at 95°C; 10 cycles of 30 s at 95°C, 30 s at 60°C, and 30 s at 72°C; and 10 min at 72°C. The products were then purified with Agencourt AMPureXP beads.

Replication of Ion Proton DNA library using a melt-off library

Instead of the Ion PI™ Chip, a melt-off waste of which the population is identical to the sequenced population was used as a template of NGS-replica pool (Figure 1E). Melt-off DNA was automatically collected from accessory instrument 'Ion OneTouch™ ES', and PCR purification using the MinElute PCR purification kit (Qiagen) was performed for neutralization. The product was amplified in reactions of 1 μl DNA, 1 μl proton_fwd primer, 1 μl proton_rev primer, 7 μl dH_2O and 10 μl KAPA HiFi polymerase under the following conditions: 5 min at 95°C; 25 cycles of 30 s at 95°C, 30 s at 60°C and 30 s at 72°C; and 10 min at 72°C. The products were electrophoresed on a 2% agarose gel and purified with a MinElute Gel Extraction Kit (Qiagen).

CBT-labeled DNA library preparation for Illumina sequencing-based experiment

To use the CBT-based labeling method on the Illumina sequencer, another microarray oligonucleotides, consisting of cgc50 pool and CBT library, were redesigned and synthesized to be compatible with the Illumina platform (Supplementary Table S3). Then, cgc50 pool was labeled with CBT library using the same protocol of 454-based experiment with primers listed in Supplementary Tables S3 and S4. With Illumina MiSeq instrument, the product was sequenced and analyzed, and NGS-replica pool was obtained from flow-cell.

Retrieval and validation of target DNA fragments for 454-, Illumina-, Ion Proton-based experiment

Primers were prepared by solid-phase oligonucleotide synthesis (Macrogen, Seoul). For retrieval from the random barcode labeled library, while considering the low T_m of the 19-bp barcode sequence (Supplementary Figure S4 and Supplementary Tables S5–S8), 3-bp common sequences (*e.g.* 'CTC') were added to the retrieval primers. Then, each primer pair was mixed with 1 μl template (0.1–1 ng per retrieval), 1 μl forward tag primer, 1 μl reverse tag primer, 7 μl dH_2O, and 10 μl KAPA HiFi polymerase. The retrieval reaction was performed under the following conditions: 5 min at 95°C; 30 cycles of 30 s at 95°C, 30 s at 60°C and 30 s at 72°C; and 10 min at 72°C. The products were electrophoresed on a 2% agarose gel and correct bands were

size-selected and purified with a MinElute Gel Extraction Kit (Qiagen). If the band of the product was not clear in gel running data, five additional cycles were tested to sharpen the PCR band, knowing the non-target amplicon could also be sharpened. The retrieval products were validated by Sanger sequencing (Macrogen).

Partial PCR products were purified using Ampure XP bead without size-selection, mixed, and subjected to NGS to evaluate target selectivity of dial-out retrieval. Then, primer and flanking sequences were trimmed, and data were aligned to the targeted sequence using Novoalign software. Aligned reads with mapping quality score of <20 were trimmed and the reads that did not result from the retrieval primer (not containing primer sequence) were filtered out. Ratio of reads, which did not contain a retrieved target, was calculated from cleaned reads.

Constructing of synthetic gene libraries, and generating of NGS-replica pool

KRAS and *GFP* were selected for fully-assembly target, and considered as sub-assembled product in Hiatt *et al.* (21). Each gene was designed into 60-bp-long building blocks, synthesized by solid-phase oligonucleotide synthesis (Macrogen, Seoul, all sequences related to this experiment were listed in Supplementary Table S9) and pooled respectively. 10 μl each oligo pool (0.1 μM per each oligo) and 10 μl KAPA HiFi polymerase were mixed, and assembly PCR was performed under the following conditions: 5 min at 95°C; 20 cycles of 30 s at 95°C, 30 s at 60°C and 30 s at 72°C; and 10 min at 72°C. The products were size-selected and amplified under 20 cycles of PCR. The assembled products were purified by Agencourt AMPure XP beads. Subsequently, 24 nt-degenerate barcodes ('NNNNANNNNTNNNNANNNNTNNNN') were flanked on both ends of each gene by PCR. 1 μl assembled product, 1 μl NNN_fwd for each gene, 1 μl NNN_rev for each gene, 7 μl dH_2O, and 10 μl KAPA HiFi polymerase were mixed and amplified under following conditions: 5 min at 95°C; 8 cycles of 30 s at 95°C, 30 s at 60°C and 30 s at 72°C; and 10 min at 72°C. The products were purified by Agencourt AMPure XP beads, prepared to Illumina NGS library using the standard protocols of the SPARK™ DNA Sample Prep Kit, subjected to Illumina MiSeq instrument, and sequenced. Then, NGS-replica pool of each gene was obtained and barcode pair information was extracted and saved as an index file.

Preparation of sheared sequencing library using NGS-replica pool for tag-directed assembly

200 ng of NGS-replica pool was sheared into 150–450 bp fragments using a M220 focused ultrasonicator™. The product was prepared as Illumina NGS library using standard protocols of the SPARK™ DNA Sample Prep Kit, and was sequenced using Illumina HiSeq platform.

Sequence verification of synthetic gene libraries using tag-directed assembly

From the NGS data, barcode and flanking sequences were trimmed, aligned to reference of each gene using Novoalign software and optionally downsampled using Picard (version 1.128). The barcodes were re-attached to the aligned data. To assemble the whole consensus sequence from the shot-gun data, the barcodes were mated based on index file. Contigs were made using breakpoint read (Figure 4A), and were merged into a consensus of which a sequence was decided by selecting major base on each position from the contigs. If indels existed in the contig, 'I' and 'D' symbols were used instead of base. Among these, error-free consensus sequences were selected and retrieved using dial-out PCR and validated by Sanger sequencing. To assess the accuracy, error-containing consensus were also tested at the same time.

RESULTS

General experimental scheme

We formulated the above-described method based on the following procedures: (i) utilizing a pool of DNA molecules containing sheared human genomic DNA or oligonucleotides cleaved from a microarray, (ii) performing NGS of the DNA pool, (iii) generating the NGS-replica pool via *in situ* replication of sequenced DNA from the NGS platform and (iv) retrieving the desired DNA from the NGS-replica pool via PCR amplification (Figure 1). We applied this method to 454 GS Junior, Illumina MiSeq, and Ion Proton sequencing. Based on the sequencing capacities of each flow cell, CBT and degenerate barcode based labeling methods were used to cover the entire population of NGS-replica pool.

Simulation for predicting success of 454 sequencing-based target DNA retrieval

Before the experiment, we performed a simulation to test if CBT-based labeling is successful when approximately 2000 × 2000 barcodes pairs are used to tag substrates. Based on the Monte Carlo method, we randomly labeled a pre-NGS pool composed of 100 million unique molecules, or an NGS-replica pool containing ∼100 000 DNA clones from 454 Junior sequencing with 2000 × 2000 barcodes pairs (Supplementary Figure S3A). The unique barcode combination ensures that a single barcode pair combination labels one substrate exclusively. According to the simulation results, no unique barcode combinations existed in the pre-NGS pool (Supplementary Figure S3B). At least four substrates were redundantly labeled with one CBT, and, in most cases, 10–40 substrates shared one CBT. This trend could demonstrate why our previous dial-out PCR method using CBT primers to amplify specific DNA molecules from a pre-NGS pool was unsuccessful. In contrast, 98.9% of barcode combinations were identified as unique barcodes in the NGS-replica pool simulation.

454 sequencing-based target DNA retrieval

Based on the simulation results, we prepared 2133 × 2133 CBT pairs (see Materials and Methods) for labeling the target DNA library. A human genomic DNA library was prepared as a target and labeled with CBT pairs. We chose sheared genomic DNAs as a target library model, because

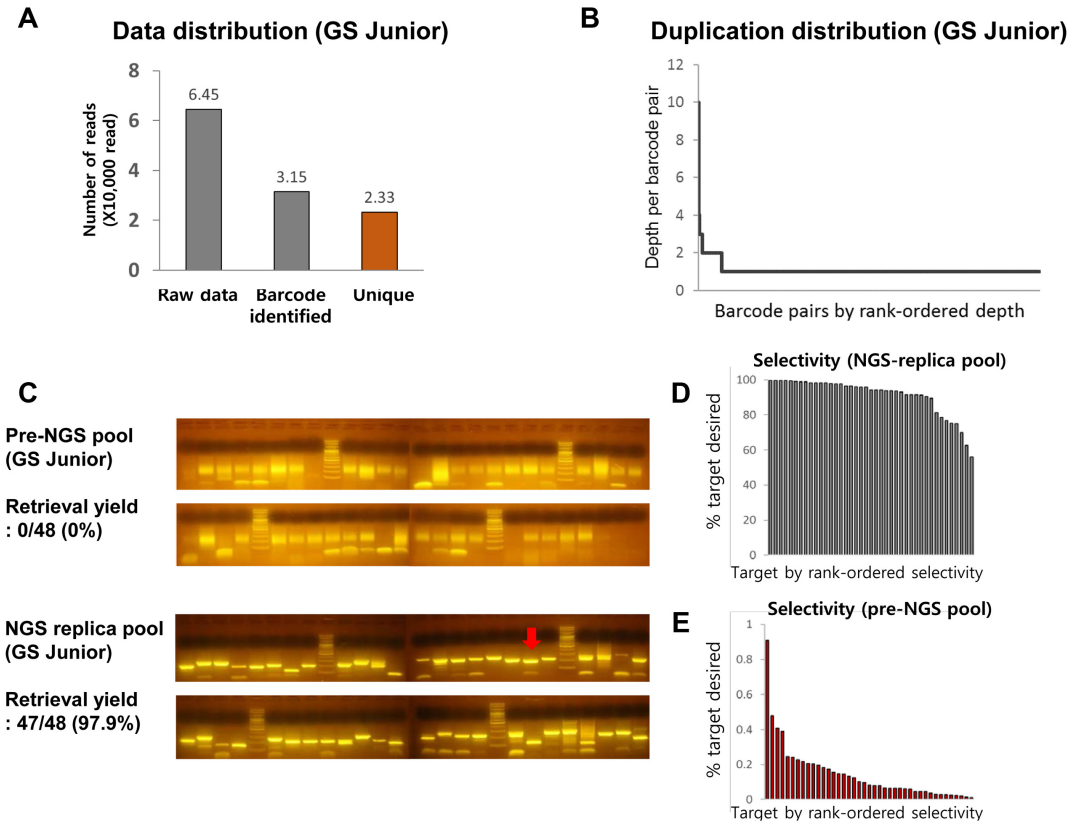

Figure 2. 454 sequencing-based retrieval of target sequences. (**A**) Distribution of reads, (**B**) duplicate distribution, and (**C**) comparison of retrieval yields between pre-NGS and NGS-replica pools using the 454 GS Junior are shown (Red arrows denote non-specific products). (**D**, **E**) Plot for selectivity of each retrieved target from NGS-replica pool (D) and pre-NGS pool (E).

the model is a highly complex pool of DNA molecules that could help us evaluate the utility of our retrieval strategy. The library (pre-NGS pool) was subjected to sequence verification with a 454 GS Junior sequencer. After aligning the reads to CBT pair sequences and to the reference human genome, 23 308 reads were identified as candidates for retrieval using unique CBT pairs (Figure 2A). All the sequence-verified fragments were labeled uniformly with a low duplication bias (Figure 2B). In this case, we chose 48 retrieval targets from the retrieval candidates containing no more than 3-bp homopolymers to avoid false discovery by homopolymer sequencing errors (22) in 454 sequencing result. However, we note that homopolymer sequencing errors would not be considered when microarray-synthesized oligo is used for target. We will only select the target exactly matched with the desired sequence.

To construct an NGS-replica pool from the 454 sequencing flow cell, the picotiter plate was assembled with an in-house gasket (Figure 1B) and filled with PCR mixture without leakage before the replication reaction was performed. We obtained the NGS-replica pool from the picotiter plate. Next, the retrieval process was carried out targeting 48 loci of the genome that was selected randomly. As a result, 48 targets from the NGS-replica pool were retrieved whereas none of the targets were selectively amplified from the pre-NGS pool based on agarose gel imaging (Figure 2C; Supplementary Table S5). All bands of retrieval products from the pre-NGS pool were shown smeary. Although some off-target bands were observed in products from NGS-replica pool, target bands were sharper than off-target bands except for the primer dimer.

To assess the selectivity of retrieval reactions, products, except seven short or low-yielded targets, were mixed and validated using NGS. Although no difference was observed between 7 excluded targets and other targets in Sanger sequencing, we avoided NGS quality-drop by short NGS library and lower input concentration. We expected that the loss of the seven targets would not affect the general trends. According to the NGS data, 91% of contents were confirmed as desired targets retrieved from NGS-replica pool (Figure 2D), whereas only 0.15% of contents were observed as the targets from pre-NGS pool (Figure 2E). NGS results were also consistent with the Sanger sequencing and gel image results. Abundance, and other properties of the primer, did not affect the selectivity. The off-targets had proper flanking sequences on both sides. However, their contents aligned to another locus of the genome. We assumed that these off-targets were redundantly tagged contents or PCR errors (e.g. template switching). From this result, we demonstrated that the use of NGS-replica pool is helpful to reduce byproducts of dial-out PCR reaction.

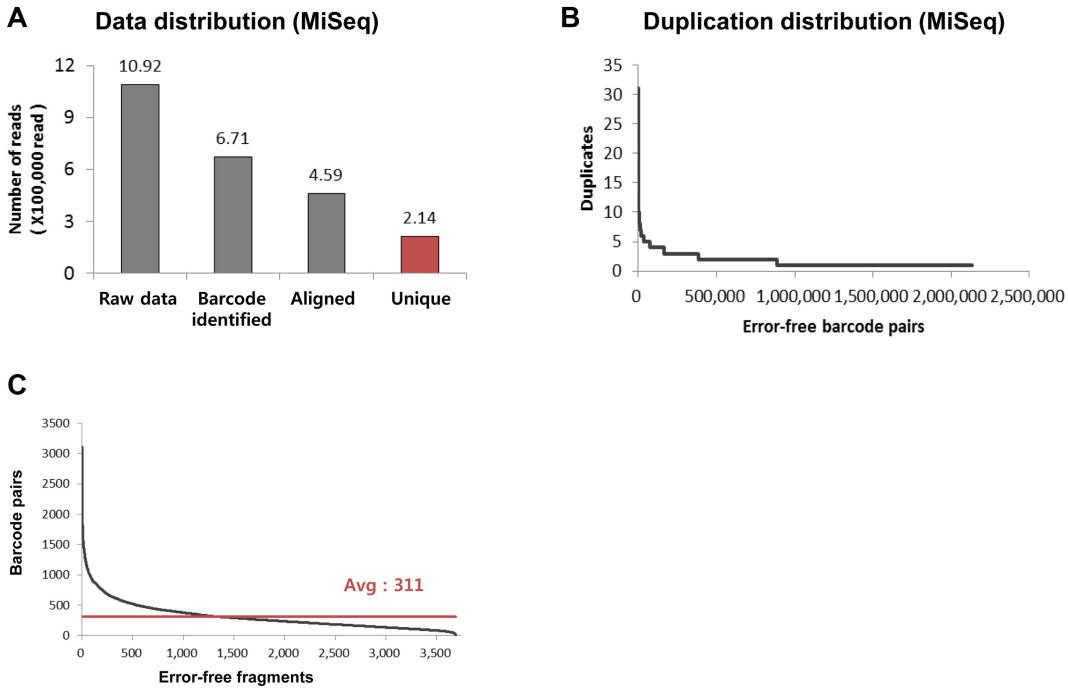

Figure 3. Illumina sequencing-based retrieval of target sequences. (**A**) Distribution of reads, and (**B**) duplicate distribution are shown. (**C**) Plot of the number of barcode pairs for each error-free fragment sorted in descending order. Nearly all designed error-free oligonucleotides were covered in the MiSeq run.

Illumina- and Ion Proton- sequencing based target DNA retrieval

To apply our method based on Illumina MiSeq and Ion Proton platforms, different target, labeling strategy, and replicating method for obtaining NGS replica pool were used. Target library was designed to be comprised of 3742 oligonucleotides for synthesis of 50 cancer-associated genes reported in the catalogue of the Cancer Gene Census and termed as 'cgc50 pool'. Twenty-bp degenerate sequences were used for labeling target library for almost unlimited dial-out PCR combinations, because 2133 × 2133 CBT pairs (4.5 million pair) could not cover the throughput of the Illumina MiSeq (15 million read) or Ion Proton platforms (60 million read). We estimated that preparation of a larger number of CBT pairs would be required for dealing with a NGS-replica pool from Illumina MiSeq sequencer. Based on the simulation result, 15 000 × 15 000 CBT pairs of CBTs would compose of 94.0% of barcode combinations as unique CBT pairs in NGS-replica pool containing 15 million DNA molecules.

Then, 3742 labeled oligonucleotides were synthesized, cleaved from the microarray (Figure 1C), and amplified using PCR. Sequencing adaptors were attached to the library (pre-NGS pool) and Illumina MiSeq and Ion Proton sequencing were carried out. According to the results, low biased read count was observed and melting temperatures of primers and GC contents of target DNA were independent with read count (Figure 3 and Supplementary Figure S4). After analyzing the data, 48 error-free DNA fragments were chosen for each experiment.

For *in situ* replication of the Illumina-sequenced pool, we injected the PCR mixture through the inlet of the MiSeq flow cell, sealed the inlet and outlet holes with an adhesive sealing film, and carried out replication of the entire DNA pool (Figure 1D). In case of Ion Proton based experiment, we collected the melted single-stranded DNA from the sequencing library, purified the DNA, and recovered it as a double-stranded DNA via PCR (Figure 1E). Then, NGS-replica pool of each platform was obtained, and the targets were retrieved from MiSeq-, Ion Proton-replica pool and pre-NGS pool of MiSeq. The pre-NGS pool of each platform was assumed to be almost the same, and the only difference was inclusion of a step for NGS adaptor attachment. Therefore,pre-NGS pool of the MiSeq was not examined. We observed 47 targets were retrieved from MiSeq-replica pool and all of the targets were retrieved from Ion Proton-replica pool (Supplementary Figure S5A, B and Supplementary Tables S6 and S7). However, in contrast to the retrieval using pre-454 NGS pool (retrieval yield of 0.15%), we noticed that ∼80% of targets (41 targets) were also retrieved from pre-NGS library for MiSeq platform. We assumed that the all pairs of twenty-bp degenerate tag, which could have 4^{40} possible combinations (septillion scale), labeled the molecules uniquely.

Comparison of target DNA retrieval efficiency among three sequencing methods

Specifications and retrieval performance of the three tested platforms are shown in Table 1. Based on information provided by the manufacturers, the three platforms exhibit differences in capacity and possible read length. In terms of retrieval performance, retrieval yields were over 98% with all three sequencing approaches when the NGS- replica pool

Table 1. Specifications and retrieval performance of the three tested sequencing platforms

Sequencing platform Retrieval barcode tagging method	GS Junior CBT	MiSeq Degenerate	Ion Proton Degenerate
Sequencing throughput (M reads)	0.1	15	60
Possible read length (bp)	400	300 × 2[a]	200
Barcoding capacity (M reads)	4.5	4300	4300
Barcode identified fragments from NGS analysis (%)	35	62	10
Error-free fragments (%)	N/A	42	7
Error-free coverage (%)	N/A	98	93
Retrieval yield (pre-NGS pool) (%)	0	85.4	-
Retrieval yield (NGS-replica pool) (%)	98	98	100
Error-free validated proportion (%)	100	100	100

[a]Paired end-sequencing strategy could be used in the Illumina MiSeq platform.

Information regarding the throughput and possible read length is cited from the platforms' manufacturers. Barcoding capacity represents the number of possible combinations of barcodes used in each experiment. The values of barcode identified fragments (%), error-free fragments (%) and error-free coverage (%) were determined from NGS data. The values of retrieval yield (pre-NGS pool) (%), retrieval yield (NGS-replica pool) (%) and error-free validated proportion (%) were determined from Sanger sequencing data. In the case of GS Junior sequencing, sheared genomic DNA was used as a substrate; therefore, error-free contents and error-free coverage were not evaluated. In the case of Ion Proton sequencing, the retrieval experiment from the pre-NGS pool was not performed, so its retrieval yield (pre-NGS pool) was not evaluated. Differences observed among retrieval yields (pre-NGS pool) and barcoding capacity of the sequencing platforms were due to differences in barcoding strategies. CBT, combinatorial barcode tag.

was used. Although the same substrate was used, the proportion of error-free fragments sequenced by MiSeq and Ion Proton platforms differed; 42.1% error-free reads were identified in the MiSeq reads, whereas only 6.64% of fragments were evaluated as error-free from Ion Proton sequencing. We presumably accounted for this discrepancy due to the characteristic weak point of Ion Proton platform that additional sequencing errors (e.g. homopolymeric insertion or deletion errors) (23) could be introducible during the process of sequencing repeated base. However, sufficient throughput and in-house programming for strict selection of error-free fragments could resolve this problem of sequencing error. In summary, we demonstrated that our method can be selectively applied to three major sequencing platforms.

Adjusting population-control by mixing control library

Although sequencing platform is selectable, unique labeling is still difficult when barcoding capacity of CBT is as low as the previous case that 4.5 million pairs of CBT could not cover 15 million clusters in MiSeq flow-cells. We expected that capacity limit could be solved by adjusting the population of the MiSeq replica pool to be less than barcoding capacity by tuning the ratio of a control library such as *PhiX* or other differently indexed sequencing library. As the ratio of a control library increased, population ratio of the target library decreased as we intended. To test this approach, pre-NGS library of cgc50 library labeled with 4.5 million pairs of CBTs was prepared, mixed with the control library to account for 20% of the total sequencing throughput, and applied to MiSeq-sequencing based protocol. As a result, we obtained NGS replica pool with the population of 3 million clones (20% of 15 million). From the NGS replica pool, a total of 43 targets was retrieved, and verified by NGS. Selectivity was observed as 19.9% from NGS-replica pool (Supplementary Figure S7). Although the selectivity is lower than the 454-based experiment, retrieval selectivity from the NGS-replica pool was much higher than that from the pre-NGS pool, which was 0.07%. We presumed that this decreased selectivity is caused by some escaped redundant CBT during the data filtering procedure and misidentified as a unique CBT. This presumption can also explain the tendency of lowered selectivity of MiSeq compared to 454-sequencing. Additional CBT escapees could have happened in MiSeq because extra redundant CBTs exist in 3 million clones from MiSeq compared to 0.07 million clones from 454 based experiment. As shown in the simulation (Supplementary Figure S8), 66% of the combinations were identified as unique barcodes in MiSeq based simulation whereas 98.9% were unique in 454 based simulation. We expected that the use of additional CBT or adjusting the ratio to be less than 20% can improve the selectivity of retrieval. To investigate the selectivity at lower ratios, we first performed a simulation using various MiSeq throughputs (0.1%, 1%, 2.5%, 5%, 7.5%, 10%, 12.5%, 15%, 17.5% and 20%). We found there was a decrease in the unique CBT ratio as the sequencing population increased (Supplementary Figure S9A). Although a simulation result is not an exact match with a selectivity result, we did find a negative correlation between the sequencing population and the ratio of unique CBTs. Based on the simulation results, 5% of the sequencing population was examined. Sixteen targets were then retrieved and sequenced using NextSeq. We found 35% selectivity (on average) using the NGS-replica pool; 0.14% selectivity was observed using the pre-NGS pool (Supplementary Figure S9B and C). The selectivity was still lower than that of the 454-based experiment. However, we found that there was a 15% increase in the selectivity. Use of a greater number of CBTs would also improve the selectivity.

Tag-directed assembly using NGS-replica library

We expected that our method could be comprehensively utilized for other synthetic methods that require a population-size control. Tag-directed assembly (21) is one of the examples. Contrary to dial-out PCR, sub-assembled DNA product whose length is longer than the usual sequencing read is used as a building block of total assembly. To enable evaluation of the sub-assembled product, barcoding, shearing, NGS, and *de novo* assembly were performed in the tag-directed assembly method. During the process, serial dilu-

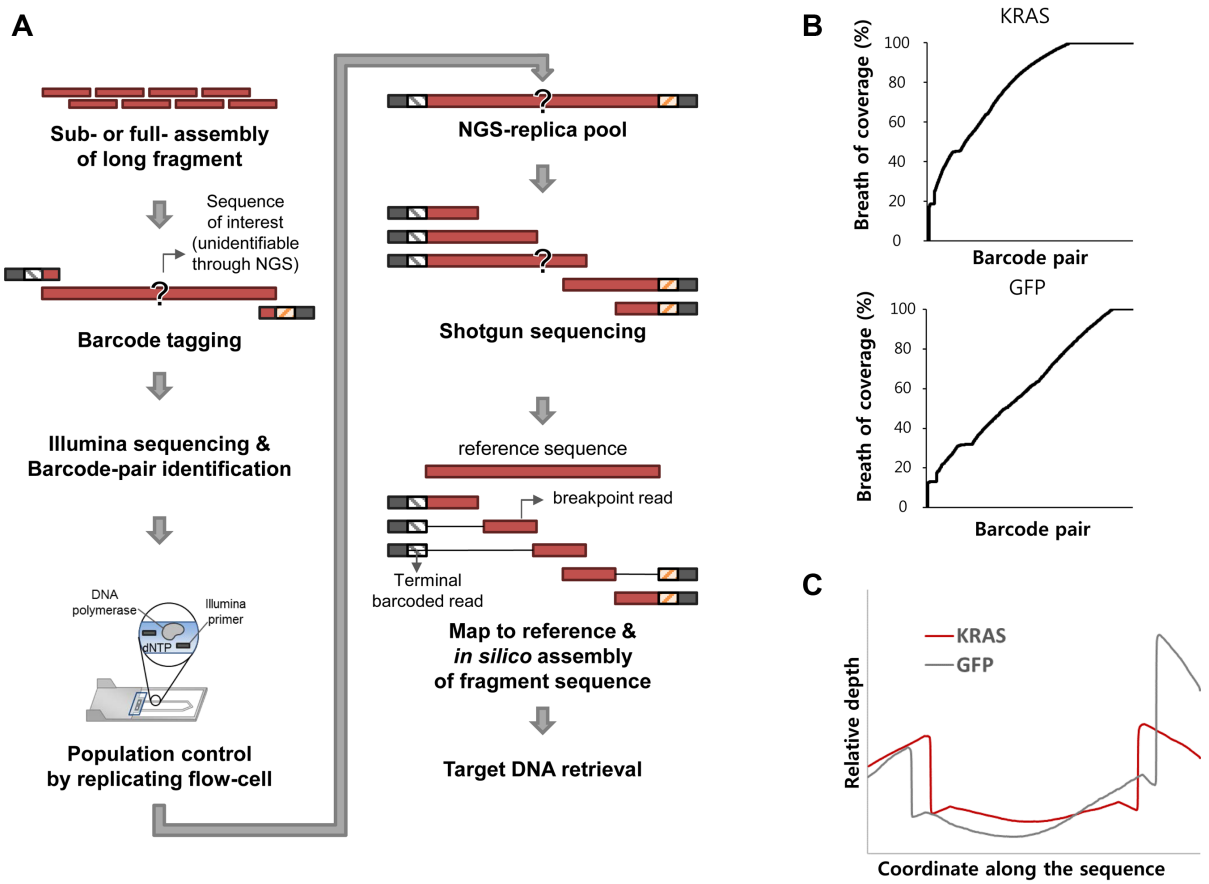

Figure 4. Application to tag-directed assembly. (**A**) Schematic flow of population controlled tag-directed assembly method. (**B**) Distribution of coverage of each assembly (upper: *KRAS* and lower: *GFP*). (**C**) Distribution of relative depth according to their coordinate along the sequence.

tion was proceeded for population control (21). Instead of using serial dilution, however, we adapted our method to control the DNA population, and NGS-replica pool was prepared to a shot-gun sequencing library (Figure 4A).

To simplify the application, two constructs with coding sequences of KRAS and GFP genes (570 and 810 bp, respectively) were selected for targets. We assembled these genes by using multiple overlapping oligos. The products were subsequently labeled with a degenerate barcode containing 20 random bases, and were sequenced in MiSeq instrument to obtain NGS-replica pool. The barcode pair information obtained from paired-end sequencing data was used as indexes for shotgun data assembly for the secondary MiSeq. We randomly sheared NGS-replica pools, shot-gun sequenced in HiSeq, and *in silico* assembly was performed for each barcode pair.

As a result, 69.8% of tag pairs identified at MiSeq were found in shotgun sequencing (65.7% for *KRAS* and 73.8% for *GFP*). Among them, 30.5% of *KRAS* pairs and 9.8% of *GFP* pairs were fully reconstructed using *in silico* assembly (Figure 4B and Supplementary Table S10). 17.3% of *KRAS* assemblies, and 2.46% of *GFP* assemblies were identified as error-free. We assumed that the difference between yields was accounted by additional needs of contigs to cover insufficient center region of longer construct. We observed that 8 and 20 contigs were minimally needed to assemble the full construct of *KRAS* and *GFP*, respectively, and the most of contigs were distributed on both side of the gene (Figure 4C). Among the consensus, 20 error-free and 15 error-containing targets for *KRAS*, and 1 error-free and 1 error-containing targets for *GFP*, whose retrieval tags have the appropriate melting temperature ($55°C < T_m < 65°C$), were retrieved and validated by Sanger sequencing. Most retrieval products (34 of 37 products) were exactly matched with each analyzed sequence including indel and substitution error (Supplementary Table S11). However, unexpected heterozygous substitution errors were observed in two cases while large deletions were observed in two cases (in one case, both of errors were simultaneously observed, Supplementary Figure S10). We assumed that heterozygous substitution error might be introduced during a PCR amplification because the probability that two different molecules with the same molecular tag pair is very low (ca. 10^{-24}). On the other hand, large deletion errors may be accounted by misalignment of breakpoint read. We expected that these unusual errors could be minimized by increasing the sequencing depth.

DISCUSSION

In this study, we developed a method for *in situ* replication of DNA from various NGS platforms that achieves highly efficient target DNA retrieval. We also overcame the major

limitation of previous dial-out PCR methods: unavailability of cost-effective CBT-based labeling due to the discrepancy between DNA in the pre-NGS library and sequence information generated from the NGS.

In our 454 GS Junior-based experiment, we introduced the CBT-based labeling method and showed the increase in selectivity of the dial-out PCR reaction when using the NGS-replica pool. This observation indicated that we could reduce the primer-preparing cost almost to a square root. Also, if we stored forward and reverse CBT primers in premade plates (usually 10 nmol per synthesis), primers could be reused for 1000 times (10 pmol per retrieval) meaning that CBT based labeling method would be less expensive by 1000-fold (∼0.002 USD per primer). However, in order to use CBTs more effectively, specificity should be enhanced for a large-scale retrieval. Although our result was encouraging, background amplification cannot be ignored at some retrieval reactions performed in Illumina-CBT based experiment. We expected that it can be analytically avoided by stringent filtering of CBT pairs. For example, rechecking unique primers using short-read aligner while considering hamming distances could be helpful for preventing omitted redundant CBT pairs. Also, we could improve the selectivity by designing additional barcodes as illustrated above (i.e. 15 000 × 15 000 CBT pairs for 94.0% of barcode combinations as unique CBT pairs in a NGS-replica pool containing 15 million DNA molecules). Despite this expandability, the cost of synthesizing 15 000 × 15 000 CBT primers remains high. However, we propose this obstacle could be overcome by introducing additional tags adjacent to the two flanking CBT sequences (Supplementary Figure S11).

We showed our method is compatible to Illumina- and Ion Proton sequencing. Although degenerate barcodes were used as tags instead of CBT at first attempt due to the larger capacity of NGS platforms, we found another advantage of our method even in these cases. Increased efficiency of the retrieval yield from the MiSeq NGS-replica pool was observed compared to the yield from its corresponding pre-NGS pool. However, some drop-outs were observed, which means the targets were analyzed but not retrieved. We tried to find the reason of these drop-outs by investigating secondary structure or potential interactions between primers. However, we could not clearly explain the reason of the drop-outs. We also showed that CBT based labeling could be used in MiSeq based experiment by adjusting ratio of the library. Retrieval was performed from 3 million clones of which the number is equivalent to 30 times of GS Junior and we observed enhanced selectivity compared to pre-NGS pool. However, selectivity was relatively lower than 454 based experiment because mixing ratio was not optimized. Further optimization of experimental procedure should be studied to reduce the drop-outs and background amplifications. Although drop-outs and polymerase errors were rarely observed in this study, only a small proportion of the pool was examined. Investigation of the replica pool using repeated experiments and NGS will be helpful to understand the effects of drop-outs and background amplification. Estimation of appropriate CBT complexity, sequencing a population using simulation, or calculation of a general equation (see Supplementary Note 1) are also useful approaches. Exact prediction of selectivity is impossible because of the effects of systematic errors (e.g., PCR bias and template switching error). However, we found that the selectivity increased when we adjusted the population based on the simulation. Sequencing errors could also affect the uniqueness during analysis. However, we think that effect of sequencing error is negligible because we removed the error-containing target, regardless of sequencing or synthesis error. The error-containing fragment could be misidentified as error-free by sequencing error, but the probability is very low. Considering synthesis error rate (usually one error per 200 bases), sequencing error rate (0.1% for Illumina platform and 1% for Ion Torrent and 454 platforms), and probability of these errors occurring on the same position, we assumed that the probability would be <0.01%.

Despite these improvements, the retrieval reaction is still a laborious process. As the target number is increased over the hundreds of thousands, the amount of replica pool will not be sufficient for retrieving all targets. Additional PCR can amplify the template amount, but additional errors will be introduced. Therefore, developing a high-throughput retrieval procedure is still desirable, and some target capture methods utilizing hybridization probes (24) or molecular inversion probes (25) could be the solution for the limitation. In contrast to PCR-based methods, target capture strategies are capable of accurate target enrichment from a heterogeneous pool via a simple reaction. Although these methods sometimes exhibit off-target capturing, they can be utilized when a pool of desired targets is required for retrieval by a high-throughput approach. Additionally, an instrument-based DNA retrieval method, such as Sniper Cloning, that enables retrieval of thousands of targets in a few hours could be a labor-saving solution by modifying the optical laser to operate on MiSeq or HiSeq platforms. However, it is not possible to efficiently adapt this approach to Illumina system yet. We also showed that our system could also be applied to a tag-directed assembly for controlling the population-size. As we mentioned, the ability to reduce the size of population was improved with our method compared with using the serial dilution. However, some cautions will be needed for designing a target. First, assembly or retrieval of a high-GC target such as *CDKN2A* gene of which the GC-content is 70% is difficult by a secondary structure (Supplementary Figure S12). Even if an alternative protocol using high-GC buffer could be helpful for the assembly PCR, more PCR errors would be introduced. Second, the length of assembled product is restricted by cluster formation limit in NGS flow cell (almost 1.5 kb is limit in Illumina device) for generating NGS-replica pool. Although the length of product is limited, among 51 710 CDSs of gene reported in RefSeq database, 19 281 CDSs were shorter than 1000 nt and 30 951 CDSs were shorter than 1500 nt. It demonstrates that almost 60% of gene could be synthesized using our method. We expect that this method could be used for retrieval of assembled gene for high-throughput synthesis of gene library.

In summary, we have developed a method for reducing complex libraries and efficiently retrieving desired DNA sequences. Notably, we used the NGS flow cell and melt-off DNA, which is generally discarded after sequencing, as a source of the NGS-replica pool and introduced improvements in the target retrieval yield. Moreover, we demon-

strated that CBT-based labeling is suitable for our method and provides a better cost-effective PCR-based alternative, as pre-designed primers can be used for more rapid retrieval of target sequences, even from a complex DNA pool.

DATA AVAILABILITY

Our sequencing data are available at the NCBI Sequence Read Archive (accession number SRP124419).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR online.

ACKNOWLEDGEMENTS

We thank members of the Duhee Bang and Ji Hyun Lee laboratories for their critical comments.

FUNDING

Pioneer Research Center Program [NRF-2012-0009557]; Mid-career Researcher Program [2015R1A2A1A10055972]; Bio & Medical Technology Development Program [NRF-2016M3A9B6948494], Basic Science Research Program [NRF-2015R1A2A2A03006577] through National Research Foundation of Korea funded by the Ministry of Science, ICT & Future Planning. Funding for open access charge: Mid-career Researcher Program [2015R1A2A1A10055972] through National Research Foundation of Korea funded by the Ministry of Science, ICT & Future Planning.

Conflict of interest statement. D.B., H.L., N.C., S.P., H.K. and H.H. are authors of a patent application for the method described in this paper (METHODS FOR RETRIEVING SEQUENCE-VERIFIED NUCLEIC ACID FRAGMENTS AND APPARATUSES FOR AMPLIFYING SEQUENCE VERIFIED NUCLEIC ACID FRAGMENTS (14/975873), Method of collecting nucleic acid fragments separated from the sequencing process (10-1648252) and Method of collecting sequence-verified nucleic acid fragments and the equipment for amplifying sequence-verified nucleic acid fragments (10-1576709). The remaining authors declare no competing financial interest.

REFERENCES

1. Caruthers,M.H. (1985) Gene synthesis machines: DNA chemistry and its uses. *Science*, **230**, 281–285.
2. Holton,T.A. and Graham,M.W. (1991) A simple and efficient method for direct cloning of PCR products using ddT-tailed vectors. *Nucleic Acids Res.*, **19**, 1156.
3. Sanger,F., Nicklen,S. and Coulson,A.R. (1977) DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U.S.A.*, **74**, 5463–5467.
4. Borovkov,A.Y., Loskutov,A.V., Robida,M.D., Day,K.M., Cano,J.A., Le Olson,T., Patel,H., Brown,K., Hunter,P.D. and Sykes,K.F. (2010) High-quality gene assembly directly from unpurified mixtures of microarray-synthesized oligonucleotides. *Nucleic Acids Res.*, **38**, e180.
5. Kim,H., Jeong,J. and Bang,D. (2011) Hierarchical gene synthesis using DNA microchip oligonucleotides. *J. Biotechnol.*, **151**, 319–324.
6. Kosuri,S., Eroshenko,N., LeProust,E.M., Super,M., Way,J., Li,J.B. and Church,G.M. (2010) Scalable gene synthesis by selective amplification of DNA pools from high-fidelity microchips. *Nat. Biotechnol.*, **28**, 1295–1299.
7. Quan,J., Saaem,I., Tang,N., Ma,S., Negre,N., Gong,H., White,K.P. and Tian,J. (2011) Parallel on-chip gene synthesis and application to optimization of protein expression. *Nat. Biotechnol.*, **29**, 449–452.
8. Baker,M. (2011) Microarrays, megasynthesis. *Nat. Methods*, **8**, 457–460.
9. Carr,P.A., Park,J.S., Lee,Y.J., Yu,T., Zhang,S. and Jacobson,J.M. (2004) Protein-mediated error correction for de novo DNA synthesis. *Nucleic Acids Res.*, **32**, e162.
10. Kim,H., Han,H., Shin,D. and Bang,D. (2010) A fluorescence selection method for accurate large-gene synthesis. *ChemBioChem*, **11**, 2448–2452.
11. Linshiz,G., Yehezkel,T.B., Kaplan,S., Gronau,I., Ravid,S., Adar,R. and Shapiro,E. (2008) Recursive construction of perfect DNA molecules from imperfect oligonucleotides. *Mol. Syst. Biol.*, **4**, 191.
12. Saaem,I., Ma,S., Quan,J. and Tian,J. (2012) Error correction of microchip synthesized genes using Surveyor nuclease. *Nucleic Acids Res.*, **40**, e23.
13. Matzas,M., Stahler,P.F., Kefer,N., Siebelt,N., Boisguerin,V., Leonard,J.T., Keller,A., Stahler,C.F., Haberle,P., Gharizadeh,B. *et al.* (2010) High-fidelity gene synthesis by retrieval of sequence-verified DNA identified using high-throughput pyrosequencing. *Nat. Biotechnol.*, **28**, 1291–1294.
14. Margulies,M., Egholm,M., Altman,W.E., Attiya,S., Bader,J.S., Bemben,L.A., Berka,J., Braverman,M.S., Chen,Y.J., Chen,Z. *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376–380.
15. Lee,H., Kim,H., Kim,S., Ryu,T., Kim,H., Bang,D. and Kwon,S. (2015) A high-throughput optomechanical retrieval method for sequence-verified clonal DNA from the NGS platform. *Nat. Commun.*, **6**, 6073.
16. 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.
17. Kim,H., Han,H., Ahn,J., Lee,J., Cho,N., Jang,H., Kim,H., Kwon,S. and Bang,D. (2012) 'Shotgun DNA synthesis' for the high-throughput construction of large DNA molecules. *Nucleic Acids Res.*, **40**, e140.
18. Schwartz,J.J., Lee,C. and Shendure,J. (2012) Accurate gene synthesis with tag-directed retrieval of sequence-verified DNA molecules. *Nat. Methods*, **9**, 913–915.
19. Klein,J.C., Lajoie,M.J., Schwartz,J.J., Strauch,E.M., Nelson,J., Baker,D. and Shendure,J. (2016) Multiplex pairwise assembly of array-derived DNA oligonucleotides. *Nucleic Acids Res.*, **44**, e43.
20. Shoemaker,D.D., Lashkari,D.A., Morris,D., Mittmann,M. and Davis,R.W. (1996) Quantitative phenotypic analysis of yeast deletion mutants using a highly parallel molecular bar-coding strategy. *Nat. Genet.*, **14**, 450–456.
21. Hiatt,J.B., Patwardhan,R.P., Turner,E.H., Lee,C. and Shendure,J. (2010) Parallel, tag-directed assembly of locally derived short sequence reads. *Nat Methods*, **7**, 119–122.
22. Luo,C., Tsementzi,D., Kyrpides,N., Read,T. and Konstantinidis,K.T. (2012) Direct comparisons of Illumina vs. Roche 454 sequencing technologies on the same microbial community DNA sample. *PLoS One*, **7**, e30087.
23. Salipante,S.J., Kawashima,T., Rosenthal,C., Hoogestraat,D.R., Cummings,L.A., Sengupta,D.J., Harkins,T.T., Cookson,B.T. and Hoffman,N.G. (2014) Performance comparison of Illumina and ion torrent next-generation sequencing platforms for 16S rRNA-based bacterial community profiling. *Appl. Environ. Microbiol.*, **80**, 7583–7591.
24. Gnirke,A., Melnikov,A., Maguire,J., Rogov,P., LeProust,E.M., Brockman,W., Fennell,T., Giannoukos,G., Fisher,S., Russ,C. *et al.* (2009) Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat. Biotechnol.*, **27**, 182–189.
25. Yoon,J.K., Ahn,J., Kim,H.S., Han,S.M., Jang,H., Lee,M.G., Lee,J.H. and Bang,D. (2015) microDuMIP: target-enrichment technique for microarray-based duplex molecular inversion probes. *Nucleic. Acids Res.*, **43**, e28.