doi: 10.1093/nar/gky112

High-throughput construction of multiple *cas9* gene variants *via* assembly of high-depth tiled and sequence-verified oligonucleotides

Namjin Cho[1,†], Han Na Seo[1,†], Taehoon Ryu[2,†], Euijin Kwon[2], Sunghoon Huh[1], Jinsung Noh[3], Huiran Yeom[3], Byungjin Hwang[1], Heejeong Ha[1], Ji Hyun Lee[4,5], Sunghoon Kwon[3,6,7,*] and Duhee Bang[1,*]

[1]Department of Chemistry, Yonsei University, Seoul 03722, Republic of Korea, [2]Celemics Inc., 371-17, Gasan-dong, Geumcheongu, Seoul 153-718, Republic of Korea, [3]Department of Electrical and Computer Engineering, Seoul National University, 1, Gwanak-ro, Gwanak-gu, Seoul, 08826, Republic of Korea., [4]Department of Clinical Pharmacology and Therapeutics, College of Medicine, Kyung Hee University, Seoul 02447, Republic of Korea, [5]Kyung Hee Medical Science Research Institute, Kyung Hee University, Seoul 02447, Republic of Korea, [6] Institute of Entrepreneurial Bio Convergence, Seoul National University, 1, Gwanak-ro, Gwanak-gu, Seoul, 08826, Republic of Korea. and [7]Seoul National University Hospital Biomedical Research Institute, Seoul National University Hospital, 101, Daehak-ro Jongno-gu, Seoul, 03080, Republic of Korea.

Received October 08, 2017; Revised January 14, 2018; Editorial Decision February 06, 2018; Accepted February 09, 2018

ABSTRACT

Selective retrieval of sequence-verified oligonucleotides (oligos) from next-generation sequencing (NGS) flow cells, termed megacloning, promises accurate and reliable gene synthesis. However, gene assembly requires a complete collection of overlapping sense and nonsense oligos, and megacloning does not typically guarantee the complete production of sequence-verified oligos. Therefore, missing oligos must be provided *via* repetitive rounds of megacloning, which introduces a bottleneck for scaled-up efforts at gene assembly. Here, we introduce the concept of high-depth tiled oligo design to successfully utilize megacloned oligos for gene synthesis. Using acquired oligos from a single round of the megacloning process, we assembled 72 of 81 target Cas9-coding gene variants. We further validated 62 of these *cas9* constructs, and deposited the plasmids to Addgene for subsequent functional characterization by the scientific community. This study demonstrates the utility of using sequence-verified oligos for DNA assembly and provides a practical and reliable optimized method for high-throughput gene synthesis.

INTRODUCTION

Gene-sized synthetic DNA can be constructed using a variety of oligo assembly methods, such as assembly PCR (1),ligase chain reaction (LCR) (2) and Gibson assembly (3). The oligos utilized for gene synthesis are typically designed and synthesized to overlap each other by up to 50% of their length. DNA microarrays, which contain an assortment of thousands to millions of synthesized oligos, enable high-throughput oligo synthesis in a cost-effective manner and have thus been employed for large-scale gene synthesis (4–7). However, because microarray-derived oligos often contain higher synthetic error rates than those synthesized by conventional, column-based synthesis (8), microarray-derived oligos are more likely to lead to imperfect target gene sequence. Consequently, an efficient method for sequence verification of synthesized gene products is needed to obtain error-free products, and this introduces a bottleneck for high-throughput gene synthesis. To address this issue, several error correction methods based on enzymatic mismatch correction have been reported (9–11). These procedures all involve two steps: (i) generation of a heteroduplex containing a mismatch through heating and re-annealing and (ii) correction of the error using an enzyme that detects mismatch. Critically, however, mismatch-based methods were found to be more suitable for error correction of single targets than for use on complex gene library pools. Additionally, their error correction efficiency decreases as the target gene length increases.

*To whom correspondence should be addressed. Tel: +82 2 2123 2633: Email: duheebang@yonsei.ac.kr
Correspondence may also be addressed to Sunghoon Kown. Email: skwon@snu.ac.kr.
†These authors contributed equally to this work as first authors.

To circumvent these problems, megacloning, a method whereby error-free, microarray-derived oligos are selectively retrieved based on next-generation sequencing (NGS), has been adopted in the field of DNA writing. Several techniques for the retrieval of sequence-verified, error-free oligos, which provide a source for accurate and effective target gene synthesis, have been reported. For example, in dial-out PCR, a degenerate sequence at both ends of the target DNA allows tag-directed retrieval from a complex oligo library, after sequence verification by PCR with pairs of tag-specific retrieval primers (7,12). This method is advantageous because it requires only target-specific retrieval primer sets and involves instrument-independent retrieval methods. Conversely, PCR-based retrieval becomes laborious and prohibitive when applied to high-throughput and large-scale gene-sized DNA synthesis. To address this, other retrieval systems that can extract target DNA directly from a sequencer flow-cell after NGS have also been developed. In the pick-and-place approach, targets are retrieved using robotic micropipettes (13), whereas they are obtained using an optomechanical apparatus in a process known as 'Sniper Cloning' (14). The Sniper Cloning technique, in particular, allows for precise DNA retrieval in a high-throughput and automated manner. Moreover, it can be used for different kinds of retrieval, such as for individual beads or for the transfer of multiple beads from a sequencer plate to a single tube. Such features suggest the possibility of using this method for high-throughput and large-scale gene synthesis.

A main drawback associated with the amplification of microarray-derived oligo pools that must be addressed when using these for gene synthesis is that this technique generates a stochastic PCR bias (15,16), and depending on the oligos used, this may lead to an irregular abundance of oligos in the amplified population (16). Consequently, the resulting imbalance between different oligos can generate NGS results in which particular populations were missing (Figure 1A). Therefore, in order to effectively mediate gene synthesis, the missing populations must be recovered, and additional sequence verification procedures are required.

In this study, we present a strategy for high-throughput and efficient gene synthesis with sequence-verified microarray oligo pools. As PCR bias is an inevitable, and uncontrollable, factor in amplification, which impedes retrieval of all oligos corresponding to each target gene, we introduce a high-depth tiling design, where each oligo is highly overlapped with several DNA fragments, depending on the degree of depth tiling (Figure 1B and Supplementary Figure S1A). To determine the optimal depth tiling design, we performed computer simulations with various depth tiling conditions and parameters, such as oligo error rate and sequencing throughput. Based on these simulations, we adopted 10× depth tiling in microarray synthesis and performed NGS and error-free DNA retrieval by an optomechanical retrieval system to obtain sequence-verified DNA for gene synthesis. Using these techniques, through one round of microarray chip synthesis and 454 GS Junior sequencing, we successfully synthesized *cas9* gene variants from 72 different species, encoding a total length of ∼276 kilobase pairs (kb) of DNA.

MATERIALS AND METHODS

Optimized oligo tiling depth simulation

Identification of the most effective depth tiling condition for gene synthesis was performed by generating virtual NGS data for 100-nt oligos with 70 000 reads (average sequenced read number of a 454 Junior sequencing system). This process was performed using various depth tiling (2×, 4×, 5× and 10×) strategies with different error rates (from one synthetic error/52 bp to one synthetic error/120 bp), and the simulation was executed 500 times for each condition.

To generate a virtual data set with the PCR bias pattern, we generated the PCR bias profile to reflect the relative portion of sequenced read numbers within the microarray oligo pool. The simulated distribution was based on the two distributions of the sequenced diverse oligo pool (Supplementary Table S1). First, we normalized the sequenced read number for each oligo by dividing the number to the average read count of a distribution. The normalized relative sequenced read numbers of each oligo pool were integrated to create a reference distribution in order to generate the PCR bias profile. The relative sequenced read numbers were randomly assigned from this distribution to simulate the oligo pool of a given tiling design. The total sequenced reads were set to 70 000, reflecting the various error rates.

In order to proceed with a simulation that reflects the actual error distribution pattern, we used an equation to calculate the error rate for each position of the oligo. We defined the equation for the frequency of error that increases from the 5′ to 3′ position, as follows:

$$\text{error rate} = 0.0001 \times (\text{base position of oligos}) + \text{C}$$

The increase in the synthetic error rate according to the oligo base position is based on the empirical distribution in Supplementary Figure S2, and the constant term (C) was added as a variable to fit this equation of simulation process. From this simulation, we concluded that target genes were 'synthesizable' when the overlap length between neighboring oligos was at least 20 bp.

Target *cas9* gene and oligo sequence design

The *cas9* gene from 81 different species was chosen as our synthetic target based on previous classification of Type II clustered regularly interspaced palindromic repeat (CRISPR)–Cas systems (17). Each Cas9 protein sequence was retrieved from UniProt and NCBI protein databases and reverse-translated to DNA sequence with human codon usage, utilizing the codon usage ratios of synonymous codon sets. Some codons, including those for leucine (TTA, CTA), serine (TCG), and arginine (CGT), were discarded, as their coding frequency in synonymous codons is less than 0.1. After discarding these codons, we randomly sampled a codon for each amino acid from a pool generated with respect to its codon usage ratio in its synonymous codon set. Once the reverse translation was performed, we changed a nucleotide in a region to allow for the generation of the forward and reverse complements of the *Bsa*I Type IIS endonuclease recognition site, without altering the protein sequence.

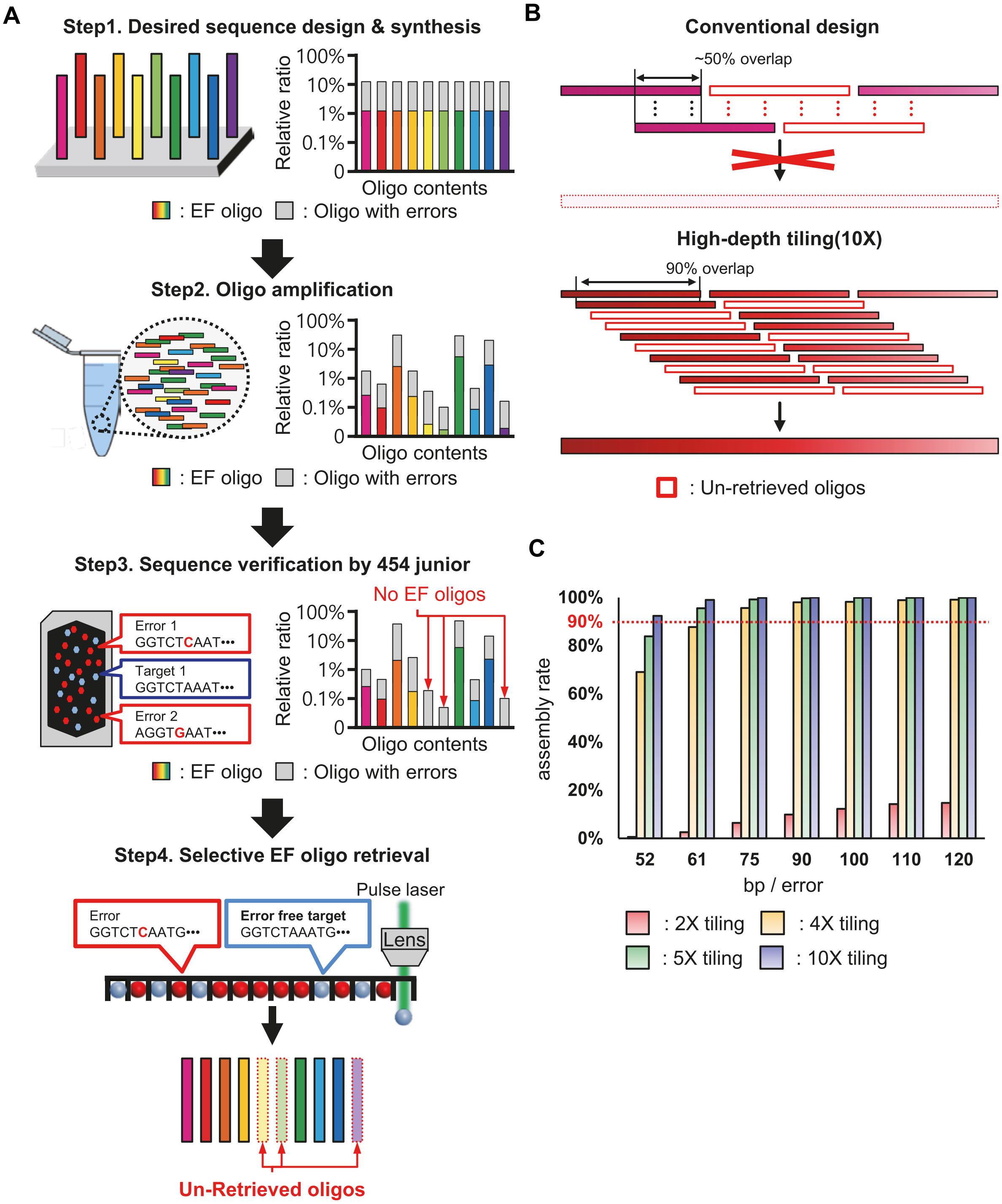

Figure 1. Effectiveness of high-depth oligo tiling. (**A**) Schematic visualization of PCR bias during each step of the megacloning process. In Step 1, designed oligos are synthesized on a DNA microarray. Although in this step, every oligo is synthesized with the same ratio on the microarray, amplification of microarray-derived oligos in Step 2 alters the relative ratio of each oligo population due to PCR-based bias. Consequently, certain populations with a relatively low amplification rate can be lost during sequence verification in Step 3. In Step 4, error-free (EF) oligos are selectively retrieved, although some populations that were not sequenced during Step 3 due to a low relative ratio are unable to be retrieved. (**B**) Under the conventional 2X tiling design, even single un-retrieved populations delay the assembly process. In contrast, because high-depth tiling is designed to fill the missing regions created by un-retrieved oligos, target DNA can be assembled in a stable manner. (**C**) Simulation predicting the assembly rate of a target gene library comprised of 20 genes, by manipulating tiling depth and the error rate of synthesized oligos.

We then attached the following flanking sequences to the ends of the reverse-translated DNA sequence: left, TGGTTATTGTGCTGTCTCATCATTTTGGCAAAGAATTCGCGGCCGCCACC and right, AGCAGGGCTGACCCCAAGAAGAAGAGGAAGGTGAGGTCCGGCGGCGGAGA. These were designed to be 50 bp in length to ensure the required size of the overlapping region for DNA synthesis. If the size of a sequence did not match to a multiple of 50, we elongated the sequence by adding the following flanking sequences: left, TTCCTACAGCTCCTGGGCAACGTGC and right, GGGCAGAGGAAGTCTTCTAACATG, without the BsaI site, to both ends alternatively until it reached the criterion level, resulting in a total sequence length of 312,250 bp. DNA sequences were targeted at 10X depth for the oligo tiling method, and oligos were designed by sliding 10 bp to the next oligo, resulting in 10 oligos for each final synthesized region. All designed oligos were tagged with universal flanking sequences, including the BsaI endonuclease recognition site on both ends, and synthesized by a CustomArray B3™ Synthesizer, using the 12K array chip.

Amplification of microarray-derived oligos

Lyophilized DNA microarray oligos were resuspended in 80 μl of Tris-EDTA (TE) buffer. For sequence verification by Roche 454 Junior sequencing and adapter tagging, 1 μl of microarray oligos was mixed with 25 μl of 2X KAPA HiFi HotStart ReadyMix, 2 μl of each 10 μM forward and reverse 454-tagged flanking primers, and 20 μl of distilled water (DW). The PCR reaction was performed as follows: (i) 95°C for 3 min, (ii) 95°C for 30 s, (iii) 60°C for 30 s, (iv) 72°C for 30 s, with repetition of steps (ii) to (iv) for 18 cycles, (v) 72°C for 5 min, and (vi) 4°C storage.

Sequence verification of adaptor-tagged oligos through 454 Junior sequencing

The adaptor-tagged oligo library was sequence-verified on a 454 Junior sequencer based on the standard protocol provided by GS Junior from Roche 454 Life Sciences. The main steps in 454 Junior sequencing, including library preparation, emulsion PCR, and sequencing, all followed standard protocols. The sequencing reaction was aborted just before the final wash step to allow further use of the 454 sequencer picotiter plate directly for target DNA retrieval.

Analysis of sequence-verified oligos

Sequencing results were aligned to 112 bp sequences, including the *Bsa*I endonuclease recognition site at both ends of the designed oligo sequence, using the Burrows-Wheeler Aligner (18). We analyzed the possibility of synthesizing target genes by selecting oligos having a 100% match to the reference from the aligned results. When more than one error-free oligo was present, the read that provided the best sequencing quality was selected as the high-throughput error-free oligo retrieval target.

Analysis of target gene synthesizability

When the length of overlap between all adjacent error-free oligos was greater than 20 bp, the gene was classified as synthesizable. However, we note that some of the 454 Junior sequencing results may have been inaccurate (19) due to the presence of homopolymeric regions in the non-overlapped regions. To identify actual errors in these homopolymeric regions, we individually retrieved oligos targeting homopolymeric regions and verified these using the Sanger sequencing method. Oligos that were identified as error-free by Sanger sequencing were further subjected to gene synthesis.

Retrieval of sequence-verified DNA by Sniper Cloning

Selected sequence-verified DNAs were retrieved as previously described (14). Briefly, a 454 Junior sequencing plate was scanned to identify coordinates of sequencing beads on which clonal DNA was immobilized, while charge-coupled device (CCD) pixel information for each sequence was extracted from the 454 Junior sequencing results. Recognized bead coordinates were cross-correlated with CCD pixel information by a diffusion-like local mapping algorithm. After a search of selected bead locations, target beads were retrieved by a Q-switched Nd:Yag laser system (Minilite, Continuum, 12 mJ at 532 nm). To reduce the number of oligo amplification reactions, error-free oligos for the same target gene were recovered by dividing them into 10 tubes, depending on the position of each retrieved target gene oligo, such that there would be no overlap. For example, the 1st, 11th and 21st oligos would be retrieved in the same PCR tube, as would the 3rd, 13th and 23rd oligos.

Amplification of retrieved oligos for the preparation of gene assembly

The error-free oligos recovered into 10 PCR tubes were subjected to a two-step amplification process. For the first amplification, 25 μl of 2X KAPA HiFi HotStart ReadyMix, 15 μl of DW, and 5 μl of each 10 μM forward and reverse primer were mixed with the retrieved beads. The amplification reaction was carried out after allowing the oligos to contact the PCR mix by vortexing, and the following parameters were used: (i) 98°C for 3 min, (ii) 98°C for 30 s, (iii) 60°C for 30 s, (iv) 72°C for 5 min, with steps (ii) to (iv) repeated for 14 cycles, (v) 72°C for 5 min and (vi) 4°C for storage. For the second amplification, 1.5 μl of the amplified product was mixed with 7.5 μl of 2X KAPA HiFi HotStart ReadyMix, 3 μl of DW, 1.5 μl of each 10 μM forward and reverse primer and amplified as follows: (i) 98°C for 3 min, (ii) 98°C for 30 s, (iii) 60°C for 30 s, (iv) 72°C for 30 s, with steps (ii) to (iv) repeated for 9 cycles, (v) 72°C for 5 min and (vi) 4°C storage.

GC content calculation

GC content calculation was performed using a simple in-house Python program. This program imported the gene sequence and calculated the GC content of 50 bp sliding windows, shifting by a single nucleotide position. It was designed such that, if the target gene had a GC content over

0.7 or under 0.35 across 5% of the target gene sequence, the program designated it as GC rich, respectively.

Gene construction using error-free oligos

Prior to digestion of the universal flanking sequences from the amplified oligos, amplified samples from the 10 individual tubes were pooled into one tube. To avoid sample loss and reduce the number of experimental steps, 6 μl of BsaI enzyme was added directly to the amplified PCR mix and incubated overnight at 37°C. Flanking sequence-cleaved products were then size-selected from 3% agarose gel electrophoresis. Various methods can be used for DNA recovery, including commercial gel purification kits, but to reduce labor, we used the 'Crush and Soak' method (Purification of Nucleic Acids from Miscellaneous Sources; section 23.1.2.3, Diffusion of DNA). To recover the target DNA, the sliced agarose gel is finely crushed, distilled water added to cover the gel, and incubated overnight.

Recovered error-free oligos were mixed into a single PCR tube, such that an average of 1 ng error-free oligo was contained in a total volume of 50 μl, including 25 μl of 2X KAPA HiFi HotStart ReadyMix. Assembly PCR was performed without amplification primers as follows: (i) 98°C for 3 min, (ii) 98°C for 30 s, (iii) ramp 0.1°C/s to 50°C and hold at 50°C for 30 s, (iv) 72°C for 30 s, repeating steps (ii) to (iv) for 10 cycles, (v) 72°C for 5 min and (vi) 4°C for storage. The temperature of step (iii) was modified depending on the sequence of the target gene (40°C for AT-rich target genes and 55°C for GC-rich target genes, using a PCR buffer for high GC templates).

Assembled products were then amplified to obtain high concentrations of each target gene; 5 μl of assembled product was mixed with 10 μl of 2X KAPA HiFi HotStart ReadyMix and 2.5 μl of each 10 μM forward and reverse primer and amplified as follows: (i) 98°C for 3 min, (ii) 98°C for 30 s, (iii) 65°C for 30 s, (iv) 72°C for 1 min/kb, repeating steps (ii) to (iv) for 15 cycles, (5) 72°C for 10 min and (vi) 4°C storage. The temperature of step (iii) was modified depending on the sequence of the target gene (60°C for GC-rich target genes, using a PCR buffer for high GC templates).

Cloning of synthetic *cas9* genes into the pUC19 vector

Synthetic *cas9* genes were cloned into the pUC19 vector using a one-step isothermal reaction (3). This reaction is normally performed by mixing 15 μl of Gibson enzyme mix and 5 μl of DNA mix (plasmid backbone and insert DNA). However, when the concentration of the sample DNA is low, an insufficient amount of this can be added to the reaction, due to the limit of the DNA mix volume. Therefore, to increase the allowable volume for the DNA mix, we used a 2-fold concentrated Gibson enzyme mix. The original Gibson enzyme mix is prepared by mixing 320 μl of 5X isothermal buffer, 1.2 μl of T5 Exonuclease, 20 μl of Phusion DNA polymerase, 160 μl of Taq Ligase, and 700 μl of DW, resulting in a final volume of ~1200 μl. We therefore generated a 2-fold concentrated enzyme mix by reducing the amount of DW to 100 μl, resulting in a final volume of ~600 μl.

The pUC19 plasmid backbone was linearized with *Xma1* and then amplified by PCR using a flanking sequence-tagged primer. For this reaction, 10 μl of 2X KAPA HiFi HotStart ReadyMix, 2.5 μl of each 10 μM forward and reverse primer, 10 ng of linearized template, and DW to a total volume of 20 μl were added, and amplification was performed as follows: (i) 98°C for 3 min, (ii) 98°C for 30 s, (iii) 60°C for 30 s, (iv) 72°C for 3 min, repeating steps (ii) to (iv) for 20 cycles, (v) 72°C for 10 min and (vi) 4°C for storage.

For cloning, we mixed 50 ng of flanking sequence-tagged linear pUC19 backbone with an equimolar ratio of amplified *cas9* gene product. If the volume of the sample mix was less than 12.5 μl, we added DW to a final volume of 12.5 μl, and then added 7.5 μl of the 2-fold concentrated assembly master mixture described above. Reactions were incubated at 50°C for 1 h, and 2.5 μl of cloned samples were transformed into competent *Escherichia coli* cells (C2566, NEB, USA).

Plasmid extraction using magnetic beads

To sequence the synthesized and cloned *cas9* genes, plasmid extraction was performed using Sera-Mag SpeedBeads (6515-2105-050350, Thermo Scientific, USA). *Escherichia coli* colonies containing the synthetic *cas9* plasmids were individually cultured in 1 ml of Luria-Bertani media with ampicillin at 37°C for 16 h. Bacterial cells were pelleted by centrifugation, and after the media was decanted, 50 μl of S1 solution from the Exprep™ Plasmid mini kit (101–102, GeneAll, Seoul, Korea) was added to cell pellet and pipetted for resuspension. We then added 100 μl of S2 cell lysis solution, and samples were shaken for 5 min to ensure complete lysis. Samples were neutralized by adding 100 μl of S3 solution, followed by centrifugation to pellet the debris. We then dispensed 110 μl of the clear lysate into clean PCR tubes, and added 10 μl of the magnetic bead solution and 80 μl of 100% isopropanol. The PCR tubes were placed on a magnet for 15 min to pellet the beads. The supernatant was then discarded, and 200 μl of 80% ethanol was added twice to wash the magnetic beads. Lastly, the plate was dried at room temperature or at 37°C until all ethanol had evaporated, and DNA was eluted by dispensing 30 μl of DW into each well and incubating for 5 min.

Sequence verification by Tn5 tagmentation and next-generation sequencing

The pTXB1 vector, which encodes the Tn5 protein, was kindly provided by Rickard Sandberg's group (20), and Tn5 enzyme was produced as previously described. For efficient tagmented pool indexing, we added eight forward and eight reverse barcode sequences to Mosaic End Double-Stranded (MEDS) oligos to distinguish eight samples in a single Illumina index.

For plasmid tagmentation, 2 μl of Tn5 mix with MEDS solution was added to 100 ng of plasmid with 4 μl of 5X TAPS-DMF and DW up to 20 μl. The tagmentation mixture was incubated for 7 min at 55°, then 5 μl of 0.2% of sodium dodecyl sulfate (SDS) solution was added, and it was further incubated for 7 min at 55° for Tn5 inactivation.

After Tn5 enzyme inactivation, 5 μl of the tagmented pool was added to 10 μl of 2X KAPA HiFi HotStart ReadyMix and 2.5 μl of each 10 μM forward and reverse Illumina index primer, and pools were amplified as follows:

(i) 72°C 3 min, (ii) 98°C for 3 min, (iii) 98°C for 30 s, (iv) 60°C for 30 s, (v) 72°C for 30 s, repeating steps (iii) to (v) for 12 cycles, (vi) 72°C for 5 min and (vii) 4°C for storage. Amplified target pools were cleaned using a 1.2 volume of Sera-Mag magnetic beads, and the tagged samples were sequenced with the HiSeq 4000 platform.

TnClone software-based sequence verification of synthesized genes

Sequencing data analysis was performed using in-house-built software (manuscript submitted). Briefly, the analysis pipeline contained three major steps: (i) quality assessment of NGS reads, (ii) *de novo* assembly, (iii) and analysis of assembled results (contigs). For each NGS read, the sequencing adaptor sequence was trimmed, and sequencing reads with low base quality were removed. The reads were then entered into the *de novo* assembly module, and plausible contigs generated by this method underwent contig analysis to select error-free contigs. During analysis, contigs were aligned to reference sequence, and variants were reported. The software then provided a summary file containing the number of error-free clones and DNA/protein error-free clone information.

When TnClone analysis revealed that two or more sequences were present, we concluded that multiple clones were analyzed in one reaction, and they were excluded from the analysis. Clones with a deletion of 100 bp or longer at a single poinwere deemed to be misassembled and excluded. These clones were analyzed to see what errors they contained, but thewere excluded from the calculation of the gene synthesis error rate (Supplementary Table S4).

Error analysis of oligos retrieved from 454 picotiter plates

Error analysis for the oligos recovered from the 454 picotiter plate was performed. Amplified oligos were further processed for sequence verification through the Illumina sequencing platform, using the SPARK DNA Sample Prep Kit (Enzymatics, Beverly, MA, USA). In brief, samples were subjected to end-repair, dA-tailing, and Illumina adaptor ligation procedures. Sequence verification was performed through Illumina HiSeq4000. Sequencing results were aligned to the reference sequence using the Burrows-Wheel Aligner (BWA), and error analysis was performed using SAMtools (version 1.1, htslib 1.1) (21) mpileup command.

Virtual gene sequence generation through a computational method

To generate a virtual gene sequence for the desired gene sample, the target gene sequence and the analyzed error spectrum were used. We randomly selected the presence or absence of errors at all base positions of the target gene, and the probability that an error base was generated reflected the overall error rate and error spectrum. The generated virtual gene sequence was analyzed for errors in DNA sequence and protein sequence, and all random functions were based on Python 2.7.5 random module and choice function.

Error-free target gene region amplification

For each gene, we selected two sequence-identified clones that could be used to generate an error-free clone by assembling DNA error-free or amino acid error-free regions from each one (Supplementary Table S4). For example, one clone could have error-free region at the front part of the gene and the other clone could have error-free region at the rear part of the gene. Error-free construct could then be assembled gluing these parts together. Primers used to amplify the error-free region in each selected clone were designed to allow the amplified error-free fragment to have an overlap of 25–35 bp with the neighboring amplified fragments. To amplify the error-free fragments, 1 μl of sequence-verified clone plasmid containing the error-free region, 10 μl of KAPA HiFi HotStart ReadyMix, and 2.5 μl of each 10 μM forward and reverse primer were added to the PCR mix. DW was added to a total volume of 20 μl, and reactions were performed as follows: (i) 98°C for 3 min, (ii) 98°C for 30 s, (iii) 65°C for 30 s, (iv) 72°C for 1 min/kb, repeating steps (ii) to (iv) for 20 cycles, (v) 72°C for 10 min, and (vi) 4°C storage.

RESULTS

Oligo tiling method optimization *via* computer simulation

To effectively address the inevitable PCR bias associated with oligo pool amplification, computer simulations were conducted in order to determine which tiling depth method is most stable and effective for gene library synthesis. For these simulations, the target gene library was set to 20 genes of 4 kb length, and the length of coding region contained within each oligo was set to 100 nt. We then ran simulations with oligos designed using 2×, 4×, 5× and 10× tiling methods, in which neighboring oligos overlap by 50%, 75%, 80% and 90%, respectively (Supplementary Figure S1A). Virtual microarray oligo sequence verification results were then generated based on these designed oligos, resulting in up to 70 000 reads, an amount chosen to resemble the normal data throughput of a 454 Junior sequencing system (22). The number of synthesizable genes was analyzed under the condition that 'gene synthesis is possible only when there is a minimum of 20 nt overlap between the nearest error-free oligo', and the process was repeated 500 times. Because of the inconsistency inherent in the synthesis of a DNA microarray oligos, our simulation was also carried out with various synthetic errors, including the insertion of a single error between 52 and 120 bp, on average (Figure 1c).

When we analyzed the results of our megacloning simulation using oligos designed by the 2X tiling method, we found that the number of predicted synthesizable genes was quite low. Specifically, only 15% (3 genes) of the total 20 gene library was synthesizable, even with the highest quality oligos (one synthetic error per 120 bp, on average). However, using oligos designed with the 4× tiling method, >90% (18 genes) of the gene library was synthesized with oligos containing a single synthetic error per 75 bp or fewer. In simulations with 5× and 10× tiling, the same results were obtained with oligos containing a single synthetic error per 60 bp and a single synthetic error per 52 bp, respectively. This indicated that with increased tiling depth, the number of synthesizable genes can be increased, even with lower quality oligos. Based on these data, we proceeded with the 10× tiling method for the subsequent studies.

Synthesis of 10 target genes as a pilot experiment

We next performed a pilot experiment, in which we constructed 10 target *cas9* genes. For this procedure, we designed 10X-tiled oligos, which were then synthesized on a 12K DNA microchip (termed Chip 0). These microarray oligos were subsequently sequence-verified using the 454 Junior platform, and we retrieved error-free oligos *via* an automated Sniper Cloning method. For this process, we divided the oligos into 10 tubes to reduce potential interference during the amplification of retrieved oligos (Figure 2A). Critically, oligos were divided such that none shared overlapping sequence with any other oligo in the same tube to prevent non-specific assembly and elongation prior to full-length target gene synthesis (see Materials and Methods for details). Using this method, all requisite error-free oligos for target gene synthesis were recovered.

We PCR-amplified the retrieved oligos in a pooled manner, and flanking sequences were digested prior to the assembly process (Supplementary Figure S3). Assembly of the 3–5 kb genes was then performed using a two-step procedure. The first step entailed assembly without primer, and this was followed by amplification of the full-length gene using primers with the assembled product as a template. To confirm the optimal amount of template DNA for gene synthesis, gene assembly was performed under various template concentration conditions (Supplementary Figure S4). With this procedure, we successfully assembled all 10 target genes (Figure 2B and Materials and Methods).

Synthesis of the remaining 71 target genes and method optimization

Synthesis of the 10 *cas9* gene variants using the 10× depth tiling design in our pilot experiment demonstrated that our method effectively allows for the construction of a target gene library through a single round of megacloning. Therefore, we decided to synthesize the remaining 71 *cas9* genes with the 10× depth tiling method. For these target genes, we designed 27 016 oligos using the same method as in the pilot experiment. Oligos were synthesized on three 12K DNA microchips, designated Chips 1, 2 and 3, which target 25, 23 and 23 genes, respectively (Supplementary Table S2). These were verified using 454 sequencing, as in our 10-gene pilot experiment; based on this analysis, total error rates were 10.9×10^{-3}, 10.2×10^{-3} and 10.3×10^{-3} errors/bp, respectively, and error-free read rates were 17.8%, 23.3% and 23.5%, respectively. From optomechanical oligo retrieval procedures, we obtained 7474, 6720 and 6453 error-free desired oligos from Chips 1, 2 and 3, respectively, which represent 81.4%, 74.6% and 72.8%, respectively, of the total designed oligos (Supplementary Table S3).

Based on these results, we were unable to retrieve sufficient overlapping oligos to assemble three, two, and five target genes from microarray oligo pools 1, 2 and 3, respectively. The region that lacked error-free oligos was subject to sequence analysis, and homopolymeric regions longer than 7 bp were observed to be present in seven out of 11 regions (Supplementary Table S2). In addition, one of the remaining four regions had an error at the position where the different short homopolymeric sequences were continuously connected. It has been known that the 454 Junior sequencing results for homopolymeric regions may be inaccurate (19). Thus, we concluded that sequencing errors from homopolymeric regions might cause this lack of error-free oligos for gene assembly. Therefore, we validated this observation using Sanger sequencing after retrieving oligos with errors at homopolymeric regions (Supplementary Figure S5 and Supplementary Table S2). Additionally, in one of the remaining three regions, we observed early termination of the 454 Junior sequencing results at a specific position in the reference gene, leading to short reads. We also retrieved target oligos and sequence-validated them by Sanger sequencing. Based on the combined 454 Junior sequencing and Sanger sequencing data, we were able to retrieve sufficient error-free oligos for the assembly of 24/25, 23/23, and 22/23 target genes from Chips 1, 2 and 3, respectively. In the case of the C0FXH5 gene from Chip 1 and Q9CLF2 from Chip 3, we analyzed the regions where error-free oligos were not recovered but were unable to identify any clear problems. Target error-free oligos were extracted by Sniper Cloning in 10 PCR tubes for each target gene. We then utilized optimized assembly protocols (see Materials and Methods) for the successful synthesis of 20, 21 and 21 target genes, from Chips 1, 2 and 3, respectively (Supplementary Figure S6). Combined with the 10 synthetic genes from Chip 0, in total we obtained sequence validation for 72 *cas9* gene variants.

Sequence verification of 72 synthesized gene products

To verify correct target gene synthesis, assembled gene products were cloned into the pUC19 plasmid and transformed into *Escherichia coli*. Sequence analysis was performed on up to 12 colonies per target gene. To effectively analyze target gene sequences of 3–5 kb, random fragmentation of the plasmids using Tn5 transposase (20) was followed by sequencing with Illumina HiSeq. The clones for sequence verification were subjected to plasmid extraction, random fragmentation by Tn5 transposase, and indexing by Illumina index primers, respectively. Additionally, during the Tn5-based fragmentation process, gene-specific labeling was performed using Mosaic End Double-Stranded (MEDS) with custom barcodes, and each clone samples from same gene were identified using various combinations of Illumina index primers. The sequence verification results for cloned target genes were then analyzed using an inhouse *de novo* plasmid assembly-based clonal analysis platform named 'TnClone' (manuscript submitted).

We performed sequence verification for 796 colonies; among these, 693 colonies had properly assembled gene clones, 56 colonies were misassembled clones with long (>100 bp) deletions, and 47 colonies had multiple plasmids. Based on these sequencing results, we found that the average error rate for each gene ranged from 0.59×10^{-3} to 2.1×10^{-3} errors/bp (Figure 3A and Supplementary Table S4). We obtained error-free DNA clones from 14 target genes and amino acid error-free clones (i.e., DNA sequence error with synonymous codon changes) from 15 ad-

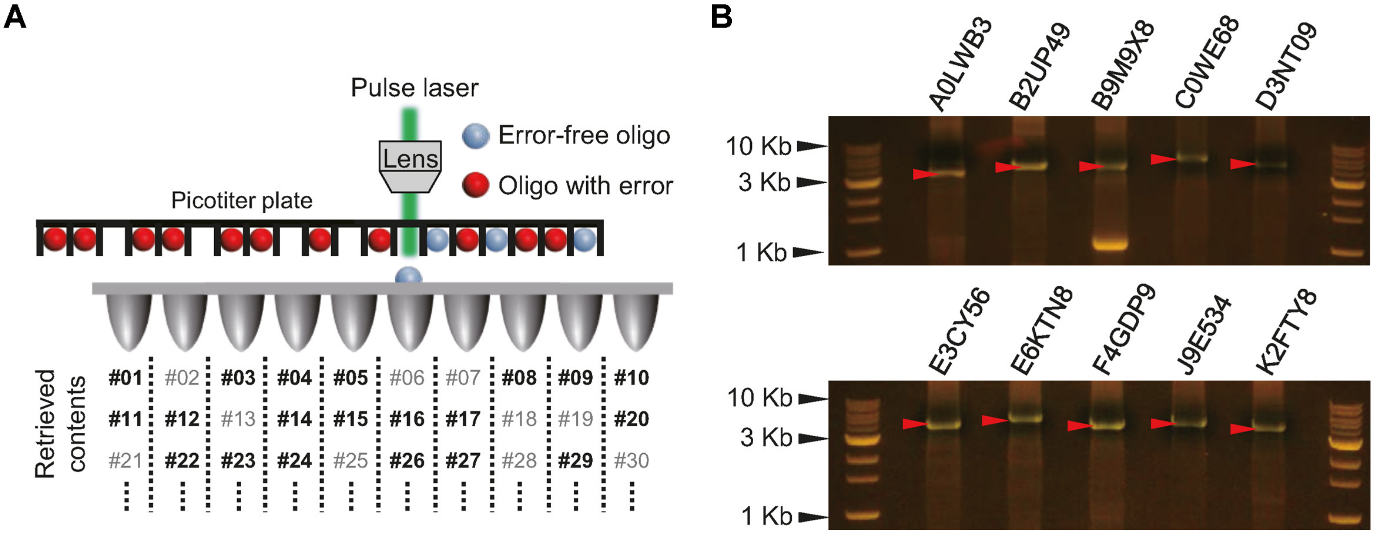

Figure 2. Error-free oligo retrieval in 10 PCR tubes and synthesis of 10 target genes. (**A**) Retrieved oligos for each gene were amplified in 10 separate tubes to reduce possible interference during the amplification process. None of the oligos in each tube share overlapping sequence with any other oligo in the same tube to prevent non-specific assembly and elongation prior to full-length target gene synthesis. The position of the tube into which each oligo was recovered was determined by the order of the designed oligo. The numbers in light gray letters indicate designed oligos that did not have error-free results as determined by 454 Junior sequencing. (**B**) Agarose gel electrophoresis of the assembled products of 10 target genes. Red triangles indicate the synthesized gene product.

ditional genes (Supplementary Table S4). We further found that >90% of errors in the synthesized gene clones were substitution errors (Supplementary Figure S7). To identify when these substitution errors were introduced into our targets, we performed sequence analysis of the oligos that were used to synthesize the first 10 target genes. Sequencing the retrieved oligos from megacloning of Chip 0 was performed for a total of 44,391,353 bases, revealing an error rate in retrieved oligos that ranged from 0.74×10^{-3} to 0.89×10^{-3} errors/bp (Supplementary Figure S8). From these results, we conclude that substitution errors in the recovered oligos were a major source of error in the synthesized genes.

Error-free clone acquisition probability simulation

Despite gene synthesis using sequence-verified oligos, only 29 error-free clones, from a total of 72 synthesized gene products were obtained. Therefore, we sought to determine a method for obtaining error-free clones for the remaining 43 genes. The simplest possibility would be to perform additional sequencing until an error-free clone is recovered. We estimated how many sequenced clones would be required to obtain an error-free clone. First, we designed a virtual gene sequence that contained a mutation pattern reflecting the error rate and error spectrum analyzed to this point (see Materials and Methods). To observe the changes in the production of DNA error-free clones according to gene length, this process was performed for 72 synthesized target genes (Figure 3B). These simulation results revealed that the percentage of DNA error-free clones ranged from 0.37% to 3.50%. Additionally, even when the ratio of protein sequence error-free clones was added, the percentage of error-free clones ranged only from 1.42% to 7.95%. These results indicate that the probability of obtaining error-free clones is not as high as we expected. We therefore concluded that screening additional colonies would be inefficient for application to the remaining genes.

As a workaround, we chose to generate error-free clones by amplifying the error-free region from clones that had already been sequenced. For this process, we screened already sequenced genes and error-free regions from two clones for an assembly of an error-free gene (Supplementary Table S4). The primer was designed to amplify the error-free region of each screened clone and to overlap with other error-free clones. The amplified samples were then cloned and sequenced by the same procedures as described above. With this method, we obtained an additional 30 DNA error-free clones and three protein sequence error-free clones. The 62 plasmids encoding these sequence-verified *cas9* genes are available from Addgene.

DISCUSSION

Here, we describe the development of an efficient method for high-throughput gene synthesis. To our knowledge, this is the first report demonstrating large-scale gene synthesis with sequence-verified oligos, following NGS and error-free DNA retrieval. Critically, the introduction of our high-depth tiling design provides a way in which to stably synthesize target genes through a single round of microarray oligo synthesis, sequence verification, and error-free oligo retrieval.

We expect that previously uncharacterized *cas9* gene variants from different species that were synthesized during this study will function as useful starting materials for studies by other researchers. The Cas9 nuclease is encoded within Type II CRISPR loci, which function in bacteria to prevent attack by invasive DNA elements, such as bacteriophages. This protein can recognize and mediate cleavage at its target site in a sequence-dependent manner, and it is directed

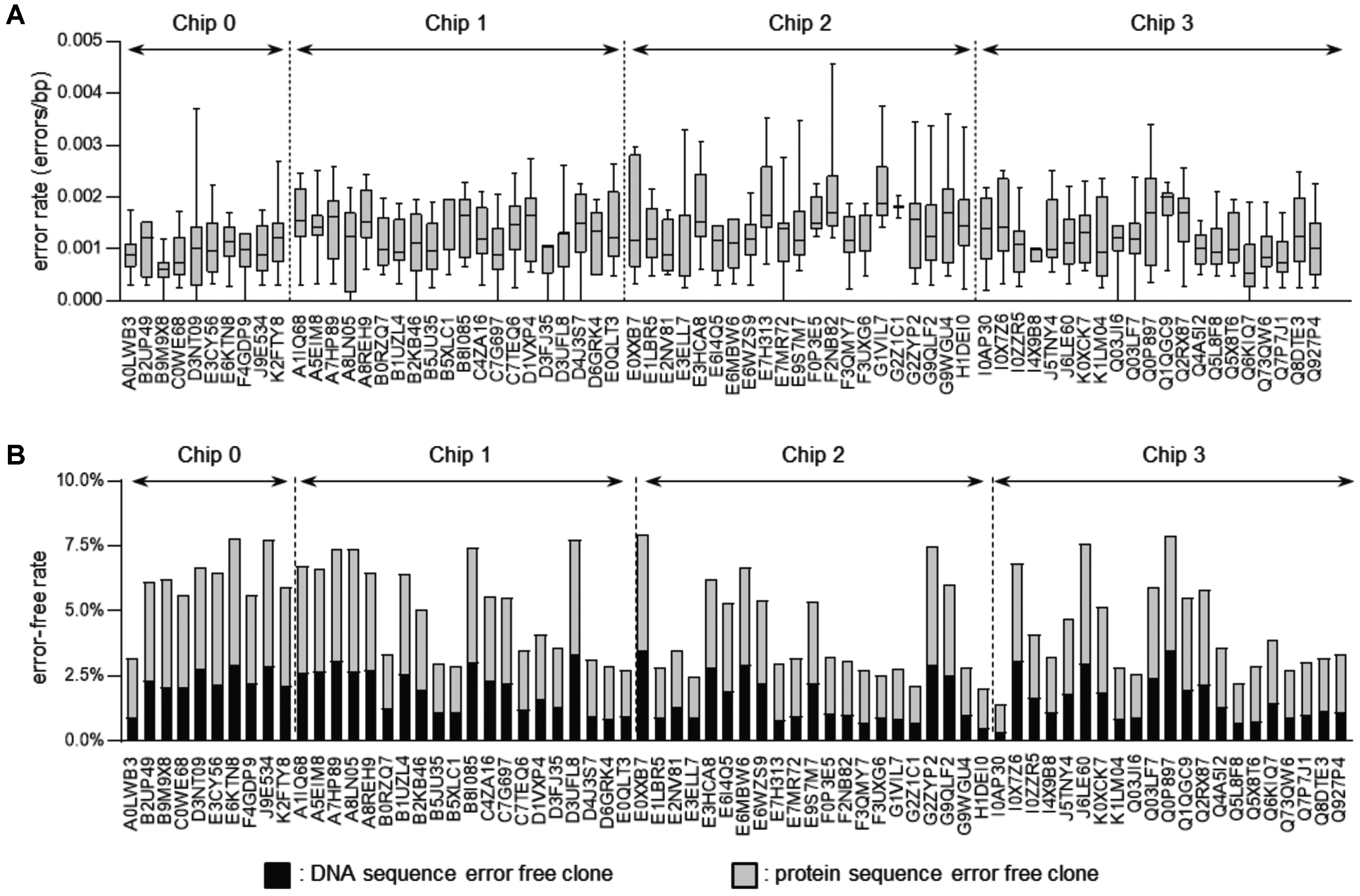

Figure 3. Error rate of sequence-verified target genes and error-free clone rates from computationally generated virtual gene sequences. (**A**) Error rates for the 72 synthesized target genes are shown. Sequence verification of each target gene was performed for up to 12 clones. Box plots shows the 5%-95% range of error rate for each target gene, and error bars indicate the standard deviation of each target gene's error rate. The first 10 target genes synthesized in our pilot experiment are labeled as Chip 0. In the case of G2Z1C1, sequence verification was performed for only two clones, so a box was unable to be generated, and the error rate is represented as a line. (**B**) Computer simulation results depicting the probability of recovering DNA error-free or protein sequence error-free clones for 72 synthesized target genes. The probability of recovering DNA error-free clones was found to be <4%. Even when the probability of recovering protein error-free clones is considered, this probability is <8%.

to these sites by two small RNAs, the CRISPR RNA (crRNA) and the trans-activating CRISPR RNA (tracrRNA). Critically, the *Streptococcus pyogenes* Cas9 has been widely exploited for various types of research, including genome editing (23,24). Several studies have also reported the function of Cas9 proteins from various other species, and those we have identified are summarized in Supplementary Table S5. For the Cas9 protein to function properly, species-specific guide RNA and protospacer adjacent motif (PAM) sequences must be identified and, to our knowledge, this information has only been reported for 18 species at present. Of the 62 *cas9* gene variants that we deposited to Addgene, seven were derived from species whose PAM and guide RNA sequences have been identified; however, we found no information for the *cas9* gene variants from the remaining 55 species.

In this study, we hypothesized that a sufficient amount of error-free oligos could not be recovered from a single mega-cloning process because of the PCR bias and the synthetic error of the oligo pool. Using the conventional 2X tiling strategy, we could not synthesize the full-length gene construct even when a single oligo contained an error. Thus, we sought to resolve this problem by utilizing high-depth tiling of the oligo design.

To verify this strategy, we performed simulation study. First, we generated the PCR bias profile based on the distribution of the two sequenced oligo pools. From this distribution, we sampled the sequenced read numbers to mimic the sequenced read number distribution of the 454 Junior sequencing data (average 70 000 reads). Next, we also analyzed these empirical pools to inspect the actual synthetic error distribution of oligo pools. We confirmed that the error rate in the 3′ direction was higher than that in the 5′ direction, and increased linearly. Since the direction of synthesis of the oligos proceeds from the 3′ to the 5′ direction, we hypothesize that this imbalance in the error rate is due to the accumulation of events such as depurination of the 3′ direction region of oligos (because of the longer exposure to the synthesis cycle) (8,25).

Taking all this into account, we performed a simulation to identify a depth tiling method that is most suitable for robust gene library synthesis. We found that a 2× tiling strategy is inefficient in synthesizing a full-length gene construct. However, a 10× tiling method could robustly synthesize gene constructs in the presence of any simulated error rate condition.

Because megacloning based on Sniper Cloning shows optimal performance on the 454 Junior platform, we sequence-verified our synthesized oligos on this device. However, we note that gene synthesis based on the 454 Junior platform has a number of limitations. For example, there is insufficient throughput in this platform to carry the entire population required to synthesize our target genes. Thus, to examine the correlation between the numbers of assembled genes and sequencing read number, we ran an additional computer simulation with variable amounts of sequencing data, ranging from 100,000 (i.e., typical sequencing reads from the 454 Junior platform) to 10 million reads (i.e., typical sequencing reads from Illumina MiSeq platform) with the 2× tiling design. The simulation was further performed on the condition that the oligos used for megacloning have, on average, a single synthetic error per 100 bp. From this simulation, 1.6 (8%) out of 20 genes could be assembled from 100,000 reads. However, with 10 million sequencing reads, which corresponds to an Illumina MiSeq level of data throughput, the number of assembled genes increased to 19.39 (96.97%) out of 20 target genes (Supplementary Figure S10).

Another limitation is that, when using the 454 Junior platform, our method is vulnerable to homopolymeric regions. This sequencing method uses a light signal and does not call bases directly, and as such, it cannot correctly read a series of homopolymeric nucleotides (19). We compared the differences in the error-free result ratio between oligos with and without long (>7 bp) homopolymeric regions, based on 454 sequencing data. These long homopolymeric regions were identified in 1,011 designed oligos. From the oligo sequence verification results, the average error-free ratios of oligos without long homopolymeric regions were 27%, 20%, 26% and 27% for chip 0, chip 1, chip 2 and chip 3, respectively (Supplementary Table S6). However, error-free ratios of oligos with long homopolymeric regions were 9%, 8%, 13% and 11% for chip 0, chip 1, chip 2 and chip 3, respectively. Given the synthetic errors of the oligos, the differences in error-free ratios indicate that the sequencing error occurred in the homopolymeric region. To avoid this issue, manipulation of homopolymeric regions *via* codon optimization should be considered.

The errors identified from sequence verification of our target genes were mainly substitution errors. We initially hypothesized that these mutations resulted from polymerase errors that accumulated during gene assembly. However, the spectrum of substitution errors identified in our synthesized gene clones did not fully match those associated with the KAPA Biosystems HIFI HotStart DNA polymerase, which was used for the amplification and assembly of the retrieved error-free oligos (26). This polymerase was chosen due to its high fidelity and the fact that it generates the lowest PCR bias when amplifying a DNA library pool (27). It belongs to the Family B DNA polymerase group, and substitution errors of these polymerases have been reported to occur mainly at G:C base pairs. Conversely, our sequence verification results for both retrieved error-free oligos and synthesized gene clones confirm that a significant amount of substitution errors occurred at A:T base pairs. These data indicate that the retrieved oligos were affected by the Platinum *Taq* DNA polymerase, which was used to perform sequencing *via* the 454 Junior platform (28). Taq polymerase belongs to the Family A DNA polymerase group, which causes substitutions mainly in A:T base pairs. This type of substitution error was confirmed in the NGS analysis of retrieved error-free oligos (Supplementary Figures S6 and S7). We propose that substitution errors at A:T base pairs can be overcome in future by using polymerases with a higher fidelity in the NGS process.

Although the large deletion was not a major error type, we identified 56 clones containing this error. We noticed that this type of long deletions occurred frequently at specific positions of specific genes. Therefore, we compared the originally designed sequences and the misassembled sequences of clones with long deletions. As a result, from 47 out of 56 clones, misassembled events were derived from regions containing a short (7–14 bp) sequence. In the case of the remaining nine clones, the reason for misassembly was not clear (Supplementary Table S4). We expect that misassembly will be avoided through rational target sequence design.

Lastly, we ran a cost analysis for synthesizing 4-kb segments of a 25-target gene library using high-depth tiled and sequence-verified oligos. We first considered factors that account for major costs in the gene synthesis process, including DNA microarray synthesis, 454 Junior sequencing, Sniper Cloning, PCR primers, PCR reagents, Tn5 transposase, and Illumina HiSeq sequencing, for verification of the assembled gene sequence. Using this streamline, we estimate that 4-kb target genes can be produced for $36.64/kb (Supplementary Tables S7 and S8). If we include other factors that are difficult to price accurately, such as labor costs, gel isolation, and DNA concentration, the cost of synthesizing a 4-kb target gene will increase. However, the cost can be lowered through the automation of these processes (including retrieved oligo amplification, enzyme digestion, gene synthesis, and random fragmentation by Tn5 transposase).

CONCLUSIONS

In this study, we developed a method for high-throughput gene synthesis using sequence-verified oligos. In particular, introduction of a high-depth tiling oligo design and the extraction of multiple sequence-verified, error-free oligos in a single PCR tube, allowed for large-scale gene synthesis in a stable and robust manner. This technique further surmounted both the PCR-based bias that often results from the amplification of a large number of oligos derived from a microarray-pool and the limited sequencer throughput. We believe that the scalability of our methods will improve, not only DNA writing technology, but also protein engineering, genetic refactoring, and functional genomics in the synthetic biology field.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank members of Bang lab and Kwon lab for their critical comments during this work.

FUNDING

Pioneer Research Center Program [NRF-2012–0009557] through the National Research Foundation of Korea, funded by the Ministry of Science, ICT & Future Planning; Mid-career Researcher Program [NRF-2015R1A2A1A10055972] through the National Research Foundation of Korea, funded by the Ministry of Science, ICT & Future Planning; Bio & Medical Technology Development Program [NRF-2016M3A9B6948494] through the National Research Foundation of Korea, funded by the Ministry of Science, ICT & Future Planning; Center Program [NRF-2015K1A4A3047345] through the National Research Foundation of Korea, funded by the Ministry of Science, ICT & Future Planning; Nano•Material Technology Development Program [NRF-2012M3A7A9671610] hrough the National Research Foundation of Korea, funded by the Ministry of Science, ICT & Future Planning. Funding for open access charge: Mid-career Researcher Program [NRF-2015R1A2A1A10055972] through the National Research Foundation of Korea, funded by the Ministry of Science, ICT & Future Planning.

Conflict of interest statement. D.B., N.C., H.N.S. and E.K. are the authors of a patent application for the method described in this manuscript (Method for synthesizing gene using high-depth oligonucleotide tiling, US 15/184,805, 2016.06.16). The remaining authors declare they have no competing financial interest.

REFERENCES

1. Stemmer,W.P., Crameri,A., Ha,K.D., Brennan,T.M. and Heyneker,H.L. (1995) Single-step assembly of a gene and entire plasmid from large numbers of oligodeoxyribonucleotides. *Gene*, **164**, 49–53.
2. Wiedmann,M., Wilson,W.J., Czajka,J., Luo,J., Barany,F. and Batt,C.A. (1994) Ligase chain reaction (LCR)–overview and applications. *PCR Methods Appl.*, **3**, S51–S64.
3. Gibson,D.G., Young,L., Chuang,R.Y., Venter,J.C., Hutchison,C.A. 3rd and Smith,H.O. (2009) Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods*, **6**, 343–345.
4. Tian,J., Gong,H., Sheng,N., Zhou,X., Gulari,E., Gao,X. and Church,G. (2004) Accurate multiplex gene synthesis from programmable DNA microchips. *Nature*, **432**, 1050–1054.
5. 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.
6. 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.
7. 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.
8. Kosuri,S. and Church,G.M. (2014) Large-scale de novo DNA synthesis: technologies and applications. *Nat. Methods*, **11**, 499–507.
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. Young,L. and Dong,Q. (2004) Two-step total gene synthesis method. *Nucleic Acids Res.*, **32**, e59.
11. Saaem,I., Ma,S., Quan,J. and Tian,J. (2012) Error correction of microchip synthesized genes using Surveyor nuclease. *Nucleic Acids Res.*, **40**, e23.
12. 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.
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. 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.
15. Aird,D., Ross,M.G., Chen,W.S., Danielsson,M., Fennell,T., Russ,C., Jaffe,D.B., Nusbaum,C. and Gnirke,A. (2011) Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol.*, **12**, R18.
16. Kebschull,J.M. and Zador,A.M. (2015) Sources of PCR-induced distortions in high-throughput sequencing data sets. *Nucleic Acids Res.*, **43**, e143.
17. Chylinski,K., Makarova,K.S., Charpentier,E. and Koonin,E.V. (2014) Classification and evolution of type II CRISPR-Cas systems. *Nucleic Acids Res.*, **42**, 6091–6105.
18. Li,H. and Durbin,R. (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, **26**, 589–595.
19. 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.
20. Picelli,S., Bjorklund,A.K., Reinius,B., Sagasser,S., Winberg,G. and Sandberg,R. (2014) Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. *Genome Res.*, **24**, 2033–2040.
21. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G., Durbin,R. and Genome Project Data Processing, S. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
22. Loman,N.J., Misra,R.V., Dallman,T.J., Constantinidou,C., Gharbia,S.E., Wain,J. and Pallen,M.J. (2012) Performance comparison of benchtop high-throughput sequencing platforms. *Nat. Biotechnol.*, **30**, 434–439.
23. Barrangou,R. and Doudna,J.A. (2016) Applications of CRISPR technologies in research and beyond. *Nat. Biotechnol.*, **34**, 933–941.
24. Jiang,F. and Doudna,J.A. (2017) CRISPR-Cas9 structures and mechanisms. *Annu. Rev. Biophys.*, **46**, 505–529.
25. Ma,S., Saaem,I. and Tian,J. (2012) Error correction in gene synthesis technology. *Trends Biotechnol.*, **30**, 147–154.
26. Potapov,V. and Ong,J.L. (2017) Examining Sources of Error in PCR by Single-Molecule Sequencing. *PLoS One*, **12**, e0169774.
27. Quail,M.A., Otto,T.D., Gu,Y., Harris,S.R., Skelly,T.F., McQuillan,J.A., Swerdlow,H.P. and Oyola,S.O. (2011) Optimal enzymes for amplifying sequencing libraries. *Nat. Methods*, **9**, 10–11.
28. 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.