nature communications

Article

https://doi.org/10.1038/s41467-025-56856-0

Oligonucleotide subsets selection by single nucleotide resolution barcode identification

Received: 26 July 2024

Accepted: 3 February 2025

Published online: 12 February 2025

Check for updates

Woojin Kim[1,6], Mingweon Chon[2,6], Yoonhae Koh[1,6], Hansol Choi[2,3,5,6], Eunjin Choi[1], Hyewon Park[4], Yushin Jung[4], Taehoon Ryu[4], Sunghoon Kwon[2,3] ✉ & Yeongjae Choi[1] ✉

Effective subset selection from complex oligonucleotide libraries is crucial for genomics, synthetic biology, and DNA data storage. The polymerase chain reaction, foundational for amplifying target subsets is limited by primer design and length for specificity, which constrains the scalability of oligo libraries and increases the synthesis burden for primers. We introduce an oligo subset selection methodology that utilizes sequence-specific cyclic nucleotide synthesis and blocking of the template oligos. This approach eliminates the need for primers for selective hybridization and enables the encoding and selection of hundreds of subsets with barcode lengths of fewer than five nucleotides. Moreover, cyclic selection enables a hierarchical data structure in the oligo library, enhancing the programmability. This advancement offers a scalable and cost-effective solution for handling complex oligo libraries.

Polymerase chain reaction (PCR) is one of the most widely used techniques for the selective amplification of target oligonucleotide molecules in complex library, identified by specific primer sequences[1,2]. PCR facilitates the selective amplification of both biological DNA and synthetic oligos from complex libraries containing millions of distinct sequences[3–6]. For effective PCR, each target subset must possess a unique selection region for primer pairs, comprising forward and reverse primers ~ 20 nucleotides (nt) in length, to ensure the specificity of primer hybridization[7]. Despite advancements, including the optimization of primer design and the incorporation of molecular probe-based mechanisms, the specificity of selecting oligos still largely depends on the length of the unique region for hybridization[8]. Also, users must identify the selection regions within the target and synthesize the corresponding primer pairs, with the synthesis burden increasing in proportion to the number of subsets[9]. In addition to PCR, other nucleic acid amplification, enrichment, and selection methods, such as hybridization-based capture and the CRISPR system, rely on hybridization and share this dependency on the length of the unique selection region[7,10–18]. Consequently, the scalability of current oligo-selection methods is limited. In addition, given that the current length limit for de novo oligonucleotide synthesis is 200 nt, dedicating 40 nt to the selection region regardless of the number of oligo subsets is inefficient[7,16,19–21]. To prevent library depletion during repeated subset selection, it is essential to enable oligo library amplification, which requires an additional 40 nt replication region where universal primers bind[7,13,22]. Although recent biotechnological advancements have expanded the scale by exploring diverse nucleic acid libraries, primer requirements have constrained these efforts. Furthermore, despite the introduction of multiplexed target selection methods, such as multiplexed PCR, which aim to select multiple oligo subsets in a single reaction, the requirement for primer synthesis proportionally increases with the number of target subsets, posing a bottleneck[9,22,23].

Here, we introduce a highly efficient method for selecting oligo subsets that can recognize barcodes in single-nucleotide resolutions, eliminating the need for individual primers for each subset. Through the implementation of synthesis and selection, which involves single-nucleotide resolution cyclic DNA synthesis and blocking of template

[1]School of Materials Science and Engineering, Gwangju Institute of Science and Technology (GIST), Gwangju, Republic of Korea. [2]Department of Electrical and Computer Engineering, Seoul National University, Seoul, Republic of Korea. [3]Bio-MAX Institute, Seoul National University, Seoul, Republic of Korea. [4]ATG Lifetech Inc., Seoul, Republic of Korea. [5]Present address: Department of Biological Chemistry and Molecular Pharmacology (BCMP), Harvard Medical School, Boston, MA, USA. [6]These authors contributed equally: Woojin Kim, Mingweon Chon, Yoonhae Koh, Hansol Choi. ✉e-mail: skwon@snu.ac.kr; yeongjae@gist.ac.kr

oligos, we successfully demonstrated oligo subset selection with only a few nucleotides and without primer synthesis for each oligo subset. This advancement enables the application of a hierarchical data structure to oligo subsets, similar to the management of digital data in computer science, where data are organized into files with directories for streamlined searching, thereby facilitating the efficient selection of multiple subsets.

The introduction of single-nucleotide resolution barcode identification through synthesis and selection enabled the development of a programmable and scalable directory system within a complex oligo library (Fig. 1). Theoretically, four to the power of N oligo subsets can be encoded by N nt, with four possible bases in each nucleotide. Thousands of oligo subsets can be identified using only 6 nt of selection region (4^6) – referred to as a "barcode" in our proposed method – reducing the complexity of primer design compared to PCR-based methods that require unique primer pairs for each subset[7] (Fig. 1a). Moreover, the proposed method allows the selection of oligo subsets in a hierarchical manner. Because the synthesis- and selection-based approach involves cyclic reactions, higher groups in the hierarchy can be selected with fewer cycles. For example, if oligo subsets are encoded with a 3 nt barcode, three cycles of selection with sequences A, T, and G permit the selection of a single subset. However, if the selection stops after the 2nd cycle with barcodes A and T, the four subsets within the ATN barcode can be simultaneously selected. The proposed hierarchical selection contrasts with PCR-based selection, where the amplification of a subset exhibits a one-to-one correspondence between the primer and oligo subset, resulting in a planar and non-hierarchical structure[22–24].

The proposed synthesis and selection-based oligo selection involves cyclic coupling of specific types of nucleotides with reversible terminators (such as 3′-O-azidomethyl deoxynucleotides) that match the target directory barcode, while others are coupled with nucleotides with irreversible terminators (such as dideoxynucleotides) (Fig. 1b, see Methods for detail). The original oligo library was immobilized onto solid substrates, such as magnetic beads, which facilitated the replacement of reagents during selection cycles. The selection process began with the hybridization of a primer complementary to the universal primer region. Following hybridization, two-step cycles continue until the desired barcode is reached. First, various combinations of nucleotides with reversible and irreversible terminators are introduced. For example, to select a single barcode A, a reversible terminator corresponding to the complementary T and irreversible terminators corresponding to A, C, and G were employed. Second, the blocker of the reversible terminator was removed using tris(2-carboxyethyl) phosphine hydrochloride treatments, enabling subsequent

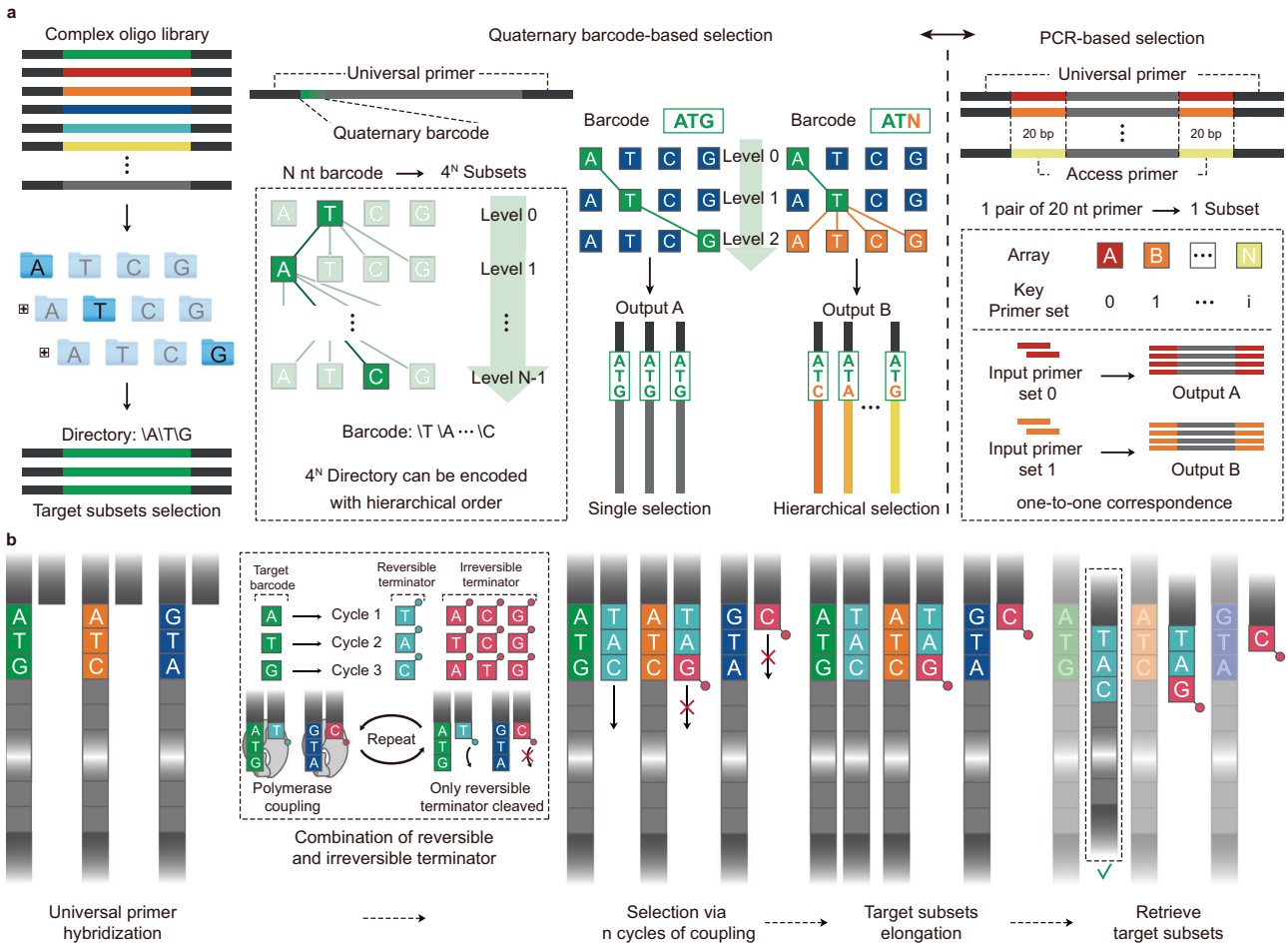

Fig. 1 | Synthesis and selection-based oligo selection enables 4^N subsets access with N nucleotide region and a hierarchical structure in complex oligo library. **a** Oligo library subset selection employing a hierarchical structure. Accessing a higher level of the hierarchy enables reaching all the oligo subsets below it. The DNA sequence-based barcoding system utilizes four base types and employs a quaternary barcode-based method for representing directories. By contrast, traditional PCR-based oligo selection requires a one-to-one mapping of specific primers to oligo subsets. The hierarchical structure within the oligo library demonstrates scalability through reduced nucleotide length requirements and improved programmability in oligo selection. **b** Details of the oligo selection include single-base resolution synthesis and selection cycles on the complementary template oligo strands. During synthesis and selection, nucleotides matching the target directory are introduced with reversible terminators, such as 3′-O-azidomethyl deoxynucleotides, while others are coupled with irreversible terminators, such as dideoxynucleotides. By repeating the cycles after deprotecting the reversible terminators, only the oligos with target barcodes will be retrieved.

reactions. As a result, oligos that do not align with the target barcode are coupled with irreversible terminators and become incapable of further reactions. After identification of all barcodes, target oligo subsets were retrieved by denaturation from the original template, following the elongation with dNTPs to synthesize full-length oligos (See Methods for detail).

Results

Demonstration of multiple modes of oligo subset selection using five distinct oligos

To validate the proposed method, five oligos with distinct barcodes and varying lengths (54, 64, 74, 84, and 94 bp) were designed to facilitate differentiation via electrophoresis (Fig. 2). Each oligo contained an identical universal primer sequence of 20 nt at one end, followed by a barcode region (Supplementary Table 1). Theoretically, a 2 nt barcode is sufficient to differentiate and encode the five subsets. However, to demonstrate the capability of hierarchical and simultaneous selection of multiple barcodes, we assigned a 4 nt barcode (Fig. 2a). Oligos were mixed in equal molar proportions and selected using a combination of reversible and irreversible terminators corresponding to the target barcode. For demonstrations, single, hierarchical, and multiple subset selections are conducted. For the selection of single subsets, the synthesis and selection processes were continued until the barcodes of each subset differed from one another (Fig. 2b and Supplementary Fig. 1). The oligos that were denatured and analyzed by polyacrylamide gel electrophoresis demonstrated varying band lengths depending on the selection, in contrast to the control, where all five oligo bands were visible. The capability of hierarchical selection of all "files" within a "folder" simultaneously by targeting a top-level barcode sequence was then examined (Fig. 2c). Selecting the first barcode, T, resulted in bands of 74, 64, and 54 bp oligos. Extending selection to the second barcode A isolated files under the folder labeled A, bands for the 64 and 54 bp oligos were acquired. In addition, this method was applied to select subsets with multiple barcodes simultaneously (Fig. 2d). For example, by employing reversible terminators for G and T and irreversible terminators for A and C, simultaneous selection to the barcodes G and T was achieved. Subsequent selection of the second barcodes T and C enabled specific selection of the 94 bp and 74 bp oligos in the sub-layer of T. The simultaneous selection of multiple barcodes was feasible regardless of the barcode sequence or hierarchical position. Thus, the proposed synthesis and selection method can be applied for oligo subset selection in a highly programmable manner, thereby enabling multiple modes of subset selection.

Subset selection in hierarchically encoded complex oligo libraries

To validate the scalability of the subset selection via synthesis and selection, we synthesized a complex oligo library that encodes digital data and selected various target subsets (Fig. 3). The data of four classical music pieces were encoded into a library composed of 12,000 oligos of 200 nt, each differentiated by unique 4 nt barcodes (Fig. 3a, Supplementary Data 1, Supplementary Note 1, and Supplementary Fig. 2). We designed a hierarchical barcode structure: the first sequence of barcodes identified the musical piece, the second specified the instrument, and the third and fourth sequences denoted sections of the music. For instance, the first sequence 'A' represents Pachelbel's Canon in D Major, and the second sequence 'T' denotes the viola part, while subsequent sequences denote the sections. The 4 nt barcodes used in the experiments accounted for 128 subsets, with each subset assigned between 60 and 130 oligos. Based on the designed barcode structure, we selected various hierarchical levels of subsets from the complex oligo library (Fig. 3b, c). First, 2 nt level subsets were selected from a group of 14 subsets. We targeted the trombone part of Mozart's Requiem in D minor and analyzed the enrichment results by normalizing and plotting NGS read counts before and after selection (Fig. 3b). The read count of the oligo pool before selection indicated that the values for each barcode were aligned with the designed ratio. The CG and CC barcodes were intentionally excluded, resulting in a two-fold prevalence of the CA and CT barcodes (Supplementary Fig. 3). After the TG barcode was selected, all barcodes except TG fell below the original ratio, while TG soared to more than 10-fold from 6.25% to 73.25%. The efficiency of target enrichment was also maintained in

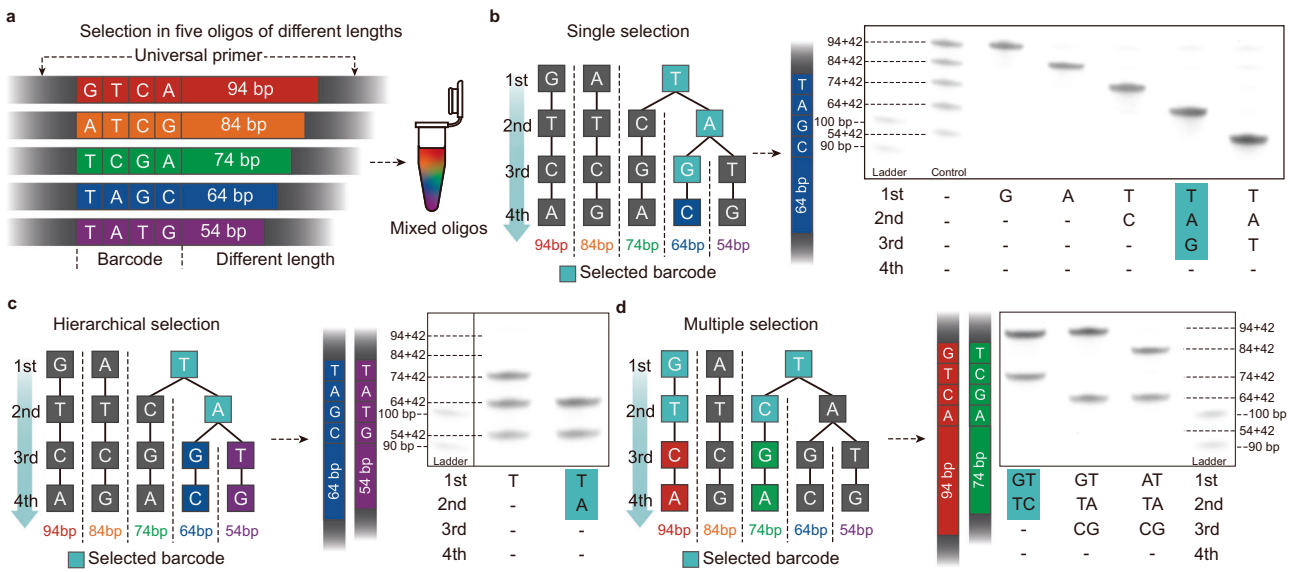

Fig. 2 | Multiple modes of oligo subset selection were enabled by synthesis and selection along with a hierarchical barcode system. a Five oligos of varying lengths with identical universal primers were designed and mixed at equal molar concentrations before selection. A 4 nt barcode was assigned to demonstrate various selection methods. **b** Single, (**c**) hierarchical, and (**d**) multiple selections of oligo subsets were demonstrated. The hierarchical structure of the barcodes for the five oligos is depicted on the left, where selecting a barcode at a higher level enables simultaneous selection of all connected lower-level barcodes, akin to accessing files within a folder. Barcodes highlighted in cyan are examples used for selection, with the corresponding selected oligos illustrated next to them. The results of polyacrylamide gel electrophoresis after selection are displayed, listing the selected barcodes in sequence below the image. These experiments (**b**, **c**, **d**) were repeated three times independently with similar results.

https://doi.org/10.1038/s41467-025-56856-0

Fig. 3 | Scalable, programmable oligo subset selection within a hierarchically structured complex oligo library. a The complex oligo library is designed to accommodate the hierarchical organization of barcodes. The library encodes data from four classical music pieces totaling 96.88 kB, spread across 12,000 oligos. Each oligo is 200 bp long, with 4 nt for the barcode. Barcodes are hierarchical; the first two nucleotides (N1, N2) differentiate pieces and instruments, while the next two (N3, N4) address specific sections. **b** Two nt and (**c**) 4 nt hierarchical barcode-based selection were demonstrated and analyzed using NGS. Read counts for each designed oligo were normalized and plotted in bar graphs. The left graph shows the distribution before selection; the right shows the results after. The red dashed line indicates the normalized read count, averaged at one. Pie charts show the proportion of total read counts, with blue for target barcodes before selection, green for target barcodes after, and gray for non-targets. The 2 nt barcode reveals the library's hierarchical structure. The barcode 'TG' targets the trombone part of Mozart's Requiem in D minor. Selecting a 4 nt barcode decodes specific musical segments. The barcode 'ATAC' targets a segment of Pachelbel's Canon in D Major Viola part. The sheet music is the result of converting the original MIDI file of the music into an SVG file (https://musescore.com/user/94946395/scores/23202934?share=copy_link) using the MuseScore app (https://musescore.org/ko). The MIDI file is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 3.0 license. **d** Multiple selection scenarios within the oligo library, with graphs representing selection to 1, 2, and 3 nt barcodes from the left, and the lower graph for a 4 nt barcode. Two barcodes at each length were selected simultaneously, marked in cyan on the tree above each graph.

longer barcode-based selection. Next, we identified a larger number of 128 subsets using a 4 nt barcode (Fig. 3c and Supplementary Fig. 4). We selected the ATAC barcode to access a file comprising the third section of Pachelbel's Canon in D Major Viola. The ATAC barcode presence in the library was 0.83% before selection, but it increased to 31.04% after selection, increasing by 37.4 times. As a result, synthesis and selection-based selection could enrich hierarchically structured oligo subsets 10- to 30-fold. Also, most of the non-target sequences aligned with those containing unintended barcodes and showed no bias, suggesting that non-targets are likely due to non-specific binding.

https://doi.org/10.1038/s41467-025-56856-0

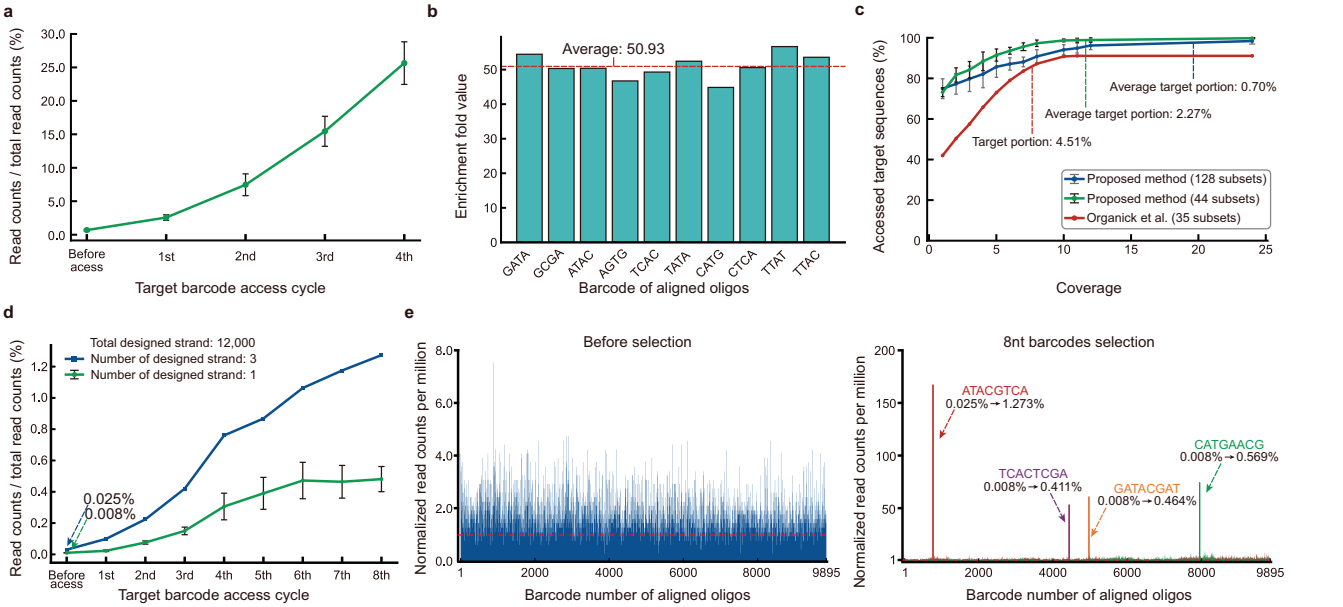

Fig. 4 | Selection efficiency of multiple cycles of selection and rare population of oligo subset. a The read count ratio for every synthesis and selection cycle of 4 nt barcodes was analyzed using 10 distinct barcodes. The bars represent one standard deviation from the average in this figure (n = 10 sequencing reads). **b** The Enrichment fold (EF) values of the target oligos were calculated after aligning the barcode and inserting the sequence to the reference sequence. Each barcode of aligned sequences showed no significant bias within EF value; the dotted line represents the average of each EF. **c** The ratio of selected target sequences by sequencing coverage (n = 10 sequencing reads). Results are compared with PCR-based selection. **d** Demonstration of subset selection within the rare population of an oligo library. The read count ratios for each cycle of four distinct 8 nt barcode selections were analyzed. The blue graph illustrates the ratio for one barcode assigned to three oligo designs, and the green graph illustrates the ratio for three barcodes each assigned to one designed strand, out of a diversity of 12,000. **e** The distribution of read counts per million of all aligned sequences were analyzed. The left graph shows the barcode number distribution before selection, and the right graph shows the distribution after selection. The ratio of target oligos were increased by an average of 57.86 times following the selection process.

In addition, we demonstrated multiplexed selection in a complex oligo library (Fig. 3d). In every synthesis and selection cycle, two barcodes were simultaneously selected. The selection of target bases generally resulted in enrichment levels more than 34 times those of non-target barcodes, demonstrating consistent enrichment across all barcode lengths without significant bias. The slight differences in percentages were influenced by the designed size differences between each subset. The advantages of synthesis- and selection-based selection demonstrate numerous selection modes in complex oligo libraries with reliable efficiency.

Selection efficiency of multiple cycles of selection and rare population of oligo subset

For an in-depth analysis of oligo subset selection from a complex oligo library, the selection efficiency of the 4 nt barcode was investigated at each step of cyclic DNA synthesis (Fig. 4). The oligo subset of 4 nt barcodes consists of 60, 80, and 100 distinct oligo designs from the oligo library, with a diversity of 12,000, corresponding to theoretical ratios of 0.5%, 0.67%, and 0.83%, respectively. Before selection, the average ratio of the subset was 0.68%, which increased with each cycle as the subset was continuously specified throughout the selection process. After four cycles of selection, the ratio reached 25.6%, which was a 37.6-fold increase due to selection. (Fig. 4a). We also measured the enrichment fold (EF) for each oligo subset selected using 10 distinct 4 nt barcodes. The EF can be calculated using the mathematical equation of the read fraction after selection (RFA) and the read fraction before selection (RFB)[25]:

$$\mathrm{EF} = \frac{\mathrm{RFA}(1-\mathrm{RFB})}{\mathrm{RFB}(1-\mathrm{RFA})}$$

The read fraction was calculated by aligning the barcode and inserting sequences into the target reference simultaneously. The average EF value was 50.93 (Fig. 4b), and because there was no significant bias for various barcodes, we believe that complexity could be increased by utilizing barcodes up to the theoretical level (4^N with an N-nt region). In addition, the recovery rate of the oligo subset of the selected barcode according to sequencing coverage was measured and compared with that of the PCR-based method, confirming that the proposed method is less prone to molecule loss after selection (Fig. 4c). For comparison, the selection of a single subset with a 4 nt barcode, and multiple subsets with a 3 nt barcode were analyzed (Supplementary Note 3). As a result, the PCR-based selection loses approximately 9.1% of the subset no matter how much the sequencing coverage; however, our method only causes a loss of less than 2% for single subset selection and 1% for multiple subset selection[23]. Although the original average target portion in the oligo library was smaller than that in the PCR-based selection, it demonstrated a high recovery rate of the target DNA subset and successful data recovery.

By performing barcode selection beyond 4 nt barcode selection up to 8 nt, we also checked whether it was possible to recover the rare oligo subset within an oligo library consisting of three distinct sequences out of a diversity of 12,000, which represents a theoretical ratio of 0.025% (Fig. 4d). In addition, to verify the possibility of more severe cases, we selected a library with a theoretical ratio of 0.008%, that consists of one sequence out of a diversity of 12,000. For a total of four 8 nt barcodes, the selection efficiency was investigated at each step of cyclic DNA synthesis. An increasing trend in efficiency was observed for each cycle as the subset was continuously specified throughout the selection process. After eight cycles of selection, the ratios reached 1.27% and 0.48% for three distinct oligo designs and one oligo design, respectively. This represents a 42.4-fold and 48.1-fold increase in counts, respectively, due to the selection process. When the

normalized read count per million before and after the selection of an 8 bp barcode was measured, we confirmed that the ratio of the target barcode improved after selection compared with that before selection (Fig. 4e). This value is significantly higher than that of non-target barcodes, and it is possible to effectively access the data encoded in the corresponding DNA sequences.

Targeted subset replacement in a hierarchically structured oligo library

Moreover, we verified the possibility of replacing target subsets without affecting the original library by negative synthesis and selection followed by new subsets addition (Fig. 5). The aim of this experiment is to replace a contrabass file with a violin file. For file replacement, we selected up to the third directory containing the target file and then blocked the target file while allowing the selection of the others (Fig. 5a). The subset-replacement process begins with the synthesis and selection of the target subset. The subset to be replaced was the fifth subset of the contrabass part of Pachelbel's Canon in D Major. Since the barcode of the subset is ACTG, synthesis and selection proceed up to the ACT sequence, allowing selection within the hierarchical structure down to the "folder" containing the fourth through sixth subsets of contrabass. At the final barcode, negative synthesis and selection were applied, where an irreversible terminator was coupled to the target subset and a reversible terminator was coupled to the non-target subsets. This approach blocks the fifth subset of the contrabass and, after elongation, enables the enrichment of all contrabass subsets, except for the fifth subset. Next, a newly synthesized oligo subset is introduced, which is designed with the same barcode as the original fifth subset of contrabass, but instead encodes violin data. This allowed us to add a new subset in the exact same position within the hierarchical structure, retaining the original barcode while replacing the content with different data (Fig. 5b and Supplementary Fig. 5). To verify whether file replacement was successful, we plotted the NGS results of the read counts using barcodes (Fig. 5c). If the negative synthesis and selection processes worked properly, only the reads for the ACTG barcode were reduced from the ACT selection results. Indeed, the reads for ACTG decreased, whereas those for ACTA and ACTC increased. To determine whether there was a difference in efficiency between the original selection approach and negative synthesis and selection, we compared the selection results for the ACTA, ACTG, and ACTC barcodes. The reads for unselected barcodes appeared to be comparable across both methods. After adding a newly synthesized oligo library following negative synthesis and selection, we observed reads from the new library containing the ACTG barcode, confirming that the file replacement was successful.

Discussion

In this study, we propose a synthesis and selection-based oligo subset selection method that distinguishes target molecules from a complex oligo library by single-nucleotide resolution with high efficiency and programmability. To the best of our knowledge, this is the first attempt at selecting oligo subsets from a complex library that does not rely on selective hybridization. Conventional methods, such as PCR and hybridization-based capture, require individually designed primers for each subset. Moreover, selection efficiency is low if primer orthogonality is not ensured (Supplementary Fig. 6). In contrast, our proposed method can encode N distinct oligo subsets with only approximately [$\log_4$N] nt barcode regions, thereby eliminating the need for subset-specific primers and enhancing selection efficiency[7,8]. For instance, 14 to 128 types of subsets were encoded with only 2 to 4 nt barcode regions, which is less than 2% of the total oligo length. This is a substantial improvement, considering that previous studies required approximately 15–25% of the total oligo length for the selection region. Furthermore, there are additional restrictions in primer sequence design to minimize secondary structure and crosstalk between distinct barcodes, and approximately 6000 subsets were designed with 40 nt barcodes (Table 1)[23,26]. By contrast, the proposed method allows programmable barcode design in lengths that can be adjusted based on the number of subsets, and 47,088 subsets were encoded with 8 nt barcodes—approximately 39.2 times more barcode per nt than that of PCR-based methods. Furthermore, 415 billion subsets can theoretically be encoded using a 20 nt barcode.

We have enriched the target oligo subsets with two synthesis and selection cycles from 6.25% to 73.25%, whereas that of other subsets

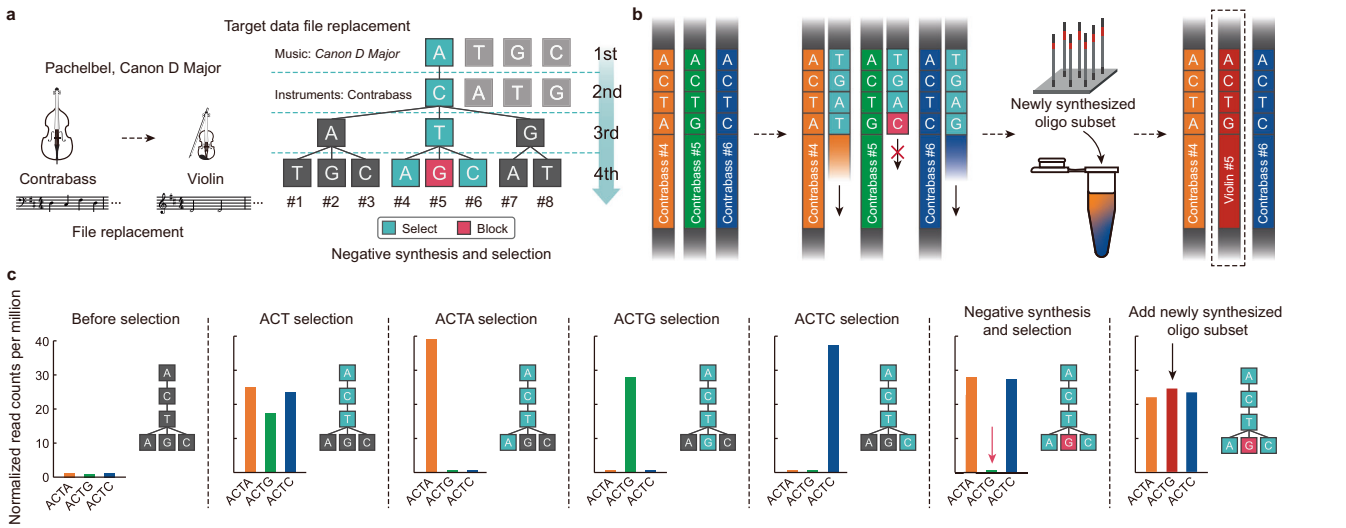

Fig. 5 | Demonstrating subset replacement through negative synthesis and selection in a hierarchically structured oligo library. a Each layer of the barcode sequences corresponds to different levels of data, from musical pieces to specific instrument sections. The barcode 'ACTG' targets a segment of Pachelbel's Canon in D Major. The process represents the replacement from the original data (contrabass #5) with new data (violin #5) through negative synthesis and selection. The sheets of music are the results of converting the original MIDI file of the music into an SVG file (contrabass #5: https://musescore.com/user/94946395/scores/23072350?share=copy_link, violin #5: https://musescore.com/user/94946395/scores/23072362?share=copy_link) using the MuseScore app (https://musescore.org/ko). The MIDI file is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 3.0 license. **b** The negative synthesis and selection process followed by file replacement involves blocking the target barcode while allowing the selection of the others. Subsequently, a newly synthesized oligo subset is added to replace the file. **c** To verify the target file replacement, we plotted the normalized read counts from various barcode selection scenarios in a bar graph.

https://doi.org/10.1038/s41467-025-56856-0

Table 1 | Comparison between the synthesis and selection-based methods and previously reported methods of oligo subset selection

Access method	PCR	Hybridization	Proposed method 2 barocde (TC access)	Proposed method 3 barocde (TCA access)	Proposed method 4 barocde (TCAC access)
Maximum number of designable subsets	5,625	–	~ 4^{11}*	~ 4^{11}*	~ 4^{11}*
Designed subsets	35	3	14	44	128
Index / data region length	40 bp / 110 bp (26.7% of 150 bp)	20 bp / 117 bp (14.6% of 137 bp)	2 bp / 158 bp (1.25% of 160 bp)	3 bp / 157 bp (1.86% of 160 bp)	4 bp / 156 bp (2.5% of 160 bp)
Target enrichment before access / after access (100% match)	3.86% / 45.23% (11.7X)	3.85% / ~ 62.5% (16.2X)	4.83% / 63.77% (13.2X)	2.00% / 40.81% (20.4X)	0.68% / 22.92% (33.7X)
Target enrichment before access / after access (98% match)	3.86% / 61.13% (15.8X)	–	5.85% / 78.31% (13.4X)	2.39% / 47.90% (20.0X)	0.84% / 29.19% (34.7X)
Ratio of missing sequence (100% match)	1.94%	–	0.26%	1.03%	1.00%
Ratio of missing sequence (98% match)	1.80%	–	0.26%	1.03%	1.00%

For PCR and hybridization, the number of required primers or capture probes increases proportionally with the number of subsets. In contrast, the proposed method requires only universal primers and leverages a hierarchical data structure.
*n = barcode length; the barcode length is determined by the size of the library. Homopolymers longer than 3 nucleotides are excluded.

was decreased to 1.96% or 37.4-fold. The increased target subset ratio enabled the decoding of all target subsets within the oligo library with reduced sequencing depth. A possible reason for limited enrichment is non-specific binding and the nucleotide coupling efficiency of the polymerase. We believe that enrichment can be improved by implementing harsher washing conditions and engineering the performance of the polymerase[19,27–30]. Although a synthesis and selection require universal primers, these can be attached through blunt-end ligation, which along with reduced barcode regions, can lower both synthesis and sequencing costs. Finally, our approach significantly enhances the utility of complex oligo libraries, which are crucial for applications in gene synthesis, perturbation screening, and especially DNA data storage, and can be further applied to the identification of various targets of interest with high sequence similarity from complex biological pools.

Methods

Process of synthesis and selection-based selection

The synthesis and selection cycle involved introducing 3′-O-azidomethyl-dNTPs (Jena Bioscience, cat. no. NU-937, 938, 939, and 940) complementary to the barcode, while ddNTPs (Jena Bioscience, cat. no. NU-1015, 1016, 1017, and 1018) excluding the complementary base were added. Tris(2-carboxyethyl) phosphine (TCEP) (Sigma-Aldrich, cat. no. C4706) was employed to cleave the azidomethyl groups. After each coupling and cleavage step, the beads were washed three times with 1x ThermoPol® Reaction Buffer (New England Biolabs, cat. no. B9004S).

For the coupling step, a mixture containing 1 µL of 1 mM 3′-O-azidomethyl-dNTP, 1 µL each of 2 mM ddNTP, 5 µL of ThermoPol® Reaction Buffer, 1 µL of Therminator™ III DNA Polymerase (New England Biolabs, cat. no. M0333), and 40 µL of nuclease-free water was incubated at 65 °C for 30 s. Cleavage was executed by treating the beads with 50 µL of 100 mM pH 9.0 TCEP at 65 °C for 1 min.

Following the final cycle, the beads were treated with Bst DNA Polymerase (New England Biolabs, cat. no. M0275S) to elongate the dNTPs and washed three times with 1x ThermoPol® Reaction Buffer. The beads were then treated with 50 µL of 8 mM urea and incubated at 70 °C for 3 min to ensure denaturation. The supernatant separated from the beads was purified using Monarch® PCR & DNA Cleanup Kit (New England Biolabs, cat. no. T1030) to complete the subset selection process.

Immobilization of oligo library on magnetic beads

To amplify the oligonucleotides, a forward primer (ACACTCTTTCCC TACACGACGCTCTTCCGATCT) and a reverse primer (GTGACTG GAGTTCAGACGTGTGCTCTTCCGATCT) were used. The reaction mixture, comprising 2 µL of each 10 µM primer, 0.2 µL of AccuPrime™ Taq DNA Polymerase (Thermo Scientific™, cat. no. 12339016), 5 µL of AccuPrime™ PCR Buffer I, 2 µL of template DNA (1.17 ng/µL), and 38.8 µL of nuclease-free water, was incubated with the following protocol: (1) initial denaturation at 94 °C for 15 s, (2) denaturation at 94 °C for 15 s, (3) annealing at 58 °C for 15 s, (4) extension at 68 °C for 30 s, with steps 2–4 repeated for 11 cycles for five oligos and 24 cycles for oligo library. The amplicons were stored at − 20 °C before use.

The amine-modified reverse primer was immobilized onto magnetic beads coated with N-hydroxysuccinimide (NHS) ester reactive groups (Thermo Scientific™, cat. no. 88826) (Supplementary Note 4). To anneal the amplified oligonucleotide to the primer on the bead, the mixture underwent denaturation at 95 °C for 30 s, followed by a gradual cooling from 95 °C to 65 °C at a rate of 1 °C per 30 s. Following annealing, The beads were washed three times with 1x ThermoPol® Reaction Buffer. the extension phase involved adding 1 µL of Bst DNA Polymerase, 3 µL of 100 mM Magnesium Sulfate ($MgSO_4$) (New England Biolabs, cat. no. B1003S) Solution, 5 µL of ThermoPol® Reaction Buffer, 1 µL of 10 mM dNTP, and 40 µL of nuclease-free water. The

mixture was then incubated at 65 °C for 1 min. The beads were washed three times with 1x ThermoPol® Reaction Buffer.

To prepare for single-stranded DNA selection, 50 µL of 8 mM urea was added to the beads with double-stranded DNA. This mixture was denatured at 70 °C for 3 min and washed three times with 1x ThermoPol® Reaction Buffer to retain only the single-stranded DNA complementary to the amplified oligo on the bead. Then, 1 µL of 10 µM forward primer was added, and the annealing process was repeated as initially described (Supplementary Fig. 7).

Polyacrylamide gel electrophoresis (PAGE) analysis

To verify the bands of oligos obtained from the synthesis and selection process, PCR was conducted before PAGE analysis. Before amplification, the cycle threshold was measured using Luna® Universal qPCR Master Mix (New England Biolabs, cat. no. M3003S) and the CFX Connect Real-Time PCR Detection System (Bio-Rad) with the following protocol: (1) initial denaturation at 95 °C for 1 min, (2) denaturation at 95 °C for 15 sec, (3) extension at 60 °C for 30 s and plate read, with steps 2–3 repeated for 35 cycles. Amplification was carried out through a saturation cycle using AccuPrime™ Taq DNA Polymerase (Supplementary Fig. 8). The amplicon was electrophoresed on an 8% polyacrylamide denaturing gel containing 7 M urea at 200 V for 30 min. The gel was stained with SYBR Gold (Thermo Scientific™, cat. no. S11494) and imaged using the Invitrogen iBright FL1500 Imaging System (Thermo Scientific™, cat. no. A44115) to confirm the presence of bands. Full scans of the PAGE are provided in the Source Data file.

NGS library preparation

Following synthesis and selection, qPCR (CFX Connect Real-Time PCR Detection System, Bio-Rad) was performed to quantify the selected oligos. The optimal number of PCR cycles was determined based on the qPCR saturation point. PCR amplification was conducted with a reaction mixture containing 2 µL each of 10 µM Nextera R2 primer and universal reverse primer, 0.2 µL AccuPrime™ Taq DNA Polymerase, 5 µL AccuPrime™ PCR Buffer I, 2 µL of selected oligos, and 38.8 µL nuclease-free water, in a total volume of 50 µL. PCR products were purified using the Monarch® PCR & DNA Cleanup Kit according to the manufacturer's instructions. Library preparation was performed at ATG Lifetech (Seoul, Republic of Korea) using an overlap PCR approach with Illumina index primers to attach adapters to the amplified products. The library was purified via PCR clean-up prior to sequencing on the iSeq platform.

Data encoding and decoding process

A total of 96.88KB of Musical Instrument Digital Interface (MIDI) files were encoded into a DNA sequence diversity of 12,000 using DNA fountain code[31] and synthesized by Twist Bioscience. For a targeted subset replacement, 766 bytes of the MIDI file were encoded into 80 DNA sequences, also synthesized by Twist Bioscience. The data were encoded within 156 nt of the 200 nt synthesized oligos. Following selection, the decoding process was performed from the raw data obtained through sequencing. Error correction was applied during the decoding back to MIDI files for sequences that did not fully align with the reference due to sequencing errors. This error correction was facilitated by a Reed-Solomon (RS) code of 2–10 nt incorporated into the 156 nt during the encoding process.

NGS data analysis

Raw FASTQ files were obtained from NGS on an Illumina iSeq platform with 150 bp paired-end sequencing. Paired-end reads of 160 nt in length were merged using FLASH[32] for further analysis. Merged reads were filtered using FASTP[33] to ensure a quality score above 30. Barcode and insert sequences were treated as single sequences and aligned with the reference sequences using BWA[34] at once. The SAM files of aligned reads were converted into BAM files utilizing SAMtools (http://www.htslib.org/doc/samtools.html), and a text file containing sequences and their respective read counts is obtained based on the BAM files.

NGS read counts normalization

For each barcode, only perfectly matched reads from the reference-aligned BAM files were counted. The read count for each barcode was calculated as read count per million (RPM). The RPM was further normalized by setting the average value for the barcodes within each hierarchical level (based on the barcode length) to one. For example, we normalized the total read counts so that the sum of the counts for each level reflected the number of possible barcodes at that level: four for 1-nt barcodes (four possible sequences), 14 for 2-nt barcodes, 44 for 3-nt barcodes, and 128 for 4-nt barcodes.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All sequencing data that support the findings of this study are available on Sequence Read Archive (SRA) under accession numbers PRJNA1202662. Source data are provided in this paper.

Code availability

The custom codes are available at https://doi.org/10.5281/zenodo.14048650 and at https://github.com/mgchon/OSS[35].

References

1. Mullis, K. B. & Faloona, F. A. Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction. *Methods Enzymol.* **155**, 335–350 (1987).
2. Kosuri, S. et al. Scalable gene synthesis by selective amplification of DNA pools from high-fidelity microchips. *Nat. Biotechnol.* **28**, 1295–1299 (2010).
3. Saiki, R. K. et al. Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science* **239**, 487–491 (1988).
4. Wang, D. G. et al. Large-scale identification, mapping, and genotyping of single-nucleotide Polymorphisms in the Human Genome. *Science* **280**, 1077–1082 (1998).
5. Gibson, D. G. et al. Creation of a bacterial cell controlled by a chemically synthesized genome. *Science* **329**, 52–56 (2010).
6. Klein, J. C. et al. Multiplex pairwise assembly of array-derived DNA oligonucleotides. *Nucleic Acids Res.* **44**, e43 (2015).
7. Chevet, E., Lema^e, G. & Katinka, M D.Low concentrations of tetramethylammonium chloride increase yield and specificity of PCR. *Nucleic Acids Res.* **23**, 3343–3344 (1995).
8. Breslauer, K. J., Frank, R., Blöcker, H. & Marky, L. A. Predicting DNA duplex stability from the base sequence. *Proc. Natl. Acad. Sci. USA* **83**, 3746–3750 (1986).
9. Xie, N. G. et al. Designing highly multiplex PCR primer sets with Simulated Annealing Design using Dimer Likelihood Estimation (SADDLE). *Nat. Commun.* **13**, 1881 (2022).
10. Singh, R. R. Target enrichment approaches for next-generation sequencing applications in oncology. *Diagnostics* **12**, 1539 (2022).
11. Ng, S. B. et al. Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **461**, 272–276 (2009).
12. Plesa, C., Sidore, A. M., Lubock, N. B., Zhang, D. & Kosuri, S. Multiplexed gene synthesis in emulsions for exploring protein functional landscapes. *Science* **359**, 343–347 (2018).
13. Tian, J. et al. Accurate multiplex gene synthesis from programmable DNA microchips. *Nature* **432**, 1050–1054 (2004).
14. Wang, J. S., Yan, Y. H. & Zhang, D. Y. Modular probes for enriching and detecting complex nucleic acid sequences. *Nat. Chem.* **9**, 1222–1228 (2017).

15. Lee, J. et al. CRISPR-Cap: multiplexed double-stranded DNA enrichment based on the CRISPR system. *Nucleic Acids Res.* **47**, e1 (2019).
16. Liu, L., Huang, Y. & Wang, H. H. Fast and efficient template-mediated synthesis of genetic variants. *Nat. Methods* **20**, 841–848 (2023).
17. Tong, X. et al. Fast and sensitive CRISPR detection by minimized interference of target amplification. *Nat. Chem. Biol.* **20**, 885–893 (2024).
18. Zhang, J., Hou, C. & Liu, C. CRISPR-powered quantitative keyword search engine in DNA data storage. *Nat. Commun.* **15**, 2376 (2024).
19. Choi, H. et al. Purification of multiplex oligonucleotide libraries by synthesis and selection. *Nat. Biotechnol.* **40**, 47–53 (2022).
20. Yeom, H. et al. Barcode-free next-generation sequencing error validation for ultra-rare variant detection. *Nat. Commun.* **10**, 977 (2019).
21. Hoose, A., Vellacott, R., Storch, M., Freemont, P. S. & Ryadnov, M. G. DNA synthesis technologies to close the gene writing gap. *Nat. Rev. Chem.* **7**, 144–161 (2023).
22. Schwartz, J. J., Lee, C. & Shendure, J. Accurate gene synthesis with tag-directed retrieval of sequence-verified DNA molecules. *Nat. Methods* **9**, 913–915 (2012).
23. Organick, L. et al. Random access in large-scale DNA data storage. *Nat. Biotechnol.* **36**, 242–248 (2018).
24. Brownie, J. The elimination of primer-dimer accumulation in PCR. *Nucleic Acids Res.* **25**, 3235–3241 (1997).
25. Song, P. et al. Selective multiplexed enrichment for the detection and quantitation of low-fraction DNA variants via low-depth sequencing. *Nat. Biomed. Eng.* **5**, 690–701 (2021).
26. Lin, K. N., Volkel, K., Tuck, J. M. & Keung, A. J. Dynamic and scalable DNA-based information storage. *Nat. Commun.* **11**, 2981 (2020).
27. Guo, J. et al. Four-color DNA sequencing with 3′- *O* -modified nucleotide reversible terminators and chemically cleavable fluorescent dideoxynucleotides. *Proc. Natl. Acad. Sci. USA* **105**, 9145–9150 (2008).
28. Gardner, A. F. et al. Therminator DNA polymerase: Modified nucleotides and unnatural substrates. *Front. Mol. Biosci.* **6**, 28 (2019).
29. Pfeiffer, F. et al. Systematic evaluation of error rates and causes in short samples in next-generation sequencing. *Sci. Rep.* **8**, 10950 (2018).
30. Kircher, M., Stenzel, U. & Kelso, J. Improved base calling for the Illumina Genome Analyzer using machine learning strategies. *Genome Biol.* **10**, R83 (2009).
31. Erlich, Y. & Zielinski, D. DNA Fountain enables a robust and efficient storage architecture. *Science* **355**, 950–954 (2017).
32. Magoč, T. & Salzberg, S. L. FLASH: Fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27**, 2957–2963 (2011).
33. Chen, S., Zhou, Y., Chen, Y. & Gu, J. Fastp: An ultra-fast all-in-one FASTQ preprocessor. in *Bioinformatics* **34**, i884–i890 (2018).
34. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
35. Kim, W. et al. Oligonucleotide subsets selection by single nucleotide resolution barcode identification. *Zenodo* https://doi.org/10.5281/zenodo.14048650 (2024).

Acknowledgements

This research was supported by the Pioneer Research Center Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (NRF-2022M3C1A3081366) (Y.C.); the Bio&Medical Technology Development Program of the NRF funded by the Korean government (MSIT) (RS-2024-00440370) (Y.C.); and NRF grant funded by the Korea government (MSIT) (No. RS-2023-00302766) (S.K.).

Author contributions

W.K., H.C., and Y.C. initiated and designed the experiments. W.K., M.C., Y.K., H.C., E.C., S.K., and Y.C. wrote the manuscript. W.K., M.C., Y.K., H.P., Y.J., and T.R. conducted the research including DNA synthesis and analysis.

Competing interests

W.K., M.C., Y.K., H.C., S.K., and Y.C. are inventors of a patent application for the method described in this article. The remaining authors declare no competing interest.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-025-56856-0.

Correspondence and requests for materials should be addressed to Sunghoon Kwon or Yeongjae Choi.

Peer review information *Nature Communications* thanks Wei Chen and Seung Soo Oh for their contribution to the peer review of this work. A peer review file is available.