Low bias DNA for sustainable and efficient data storage

Yanmin Gao[1,2], Xin Chen[3], Jianye Hao[4], Chengwei Zhang[5], Hongyan Qiao[1,2]and Hao Qi[1,2*]

[1] School of Chemical Engineering and Technology, Tianjin University, Tianjin, China.

[2] Key Laboratory of Systems Bioengineering (Ministry of Education), Tianjin University, Tianjin, China.

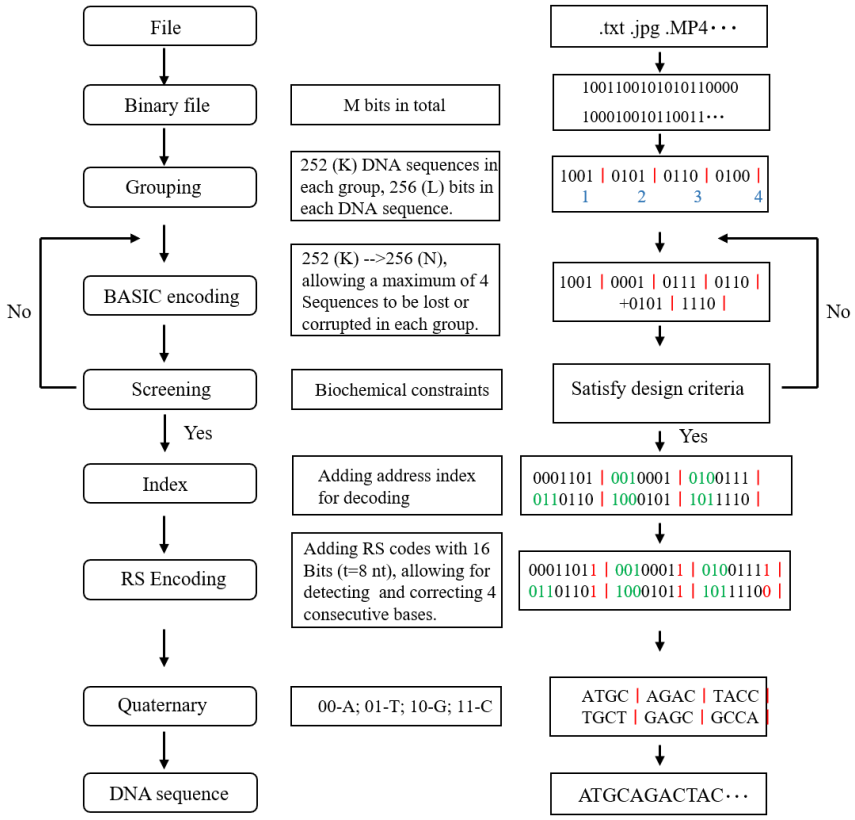
[3] Center for Applied Mathematics, Tianjin University, Tianjin, China.

[4] College of intelligence and computing, Tianjin University, Tianjin, China.

[5] College of Information Science and Technology, Dalian Maritime University, Dalian, China.

* Correspondence should be addressed to H.Q. (haoq@tju.edu.cn)

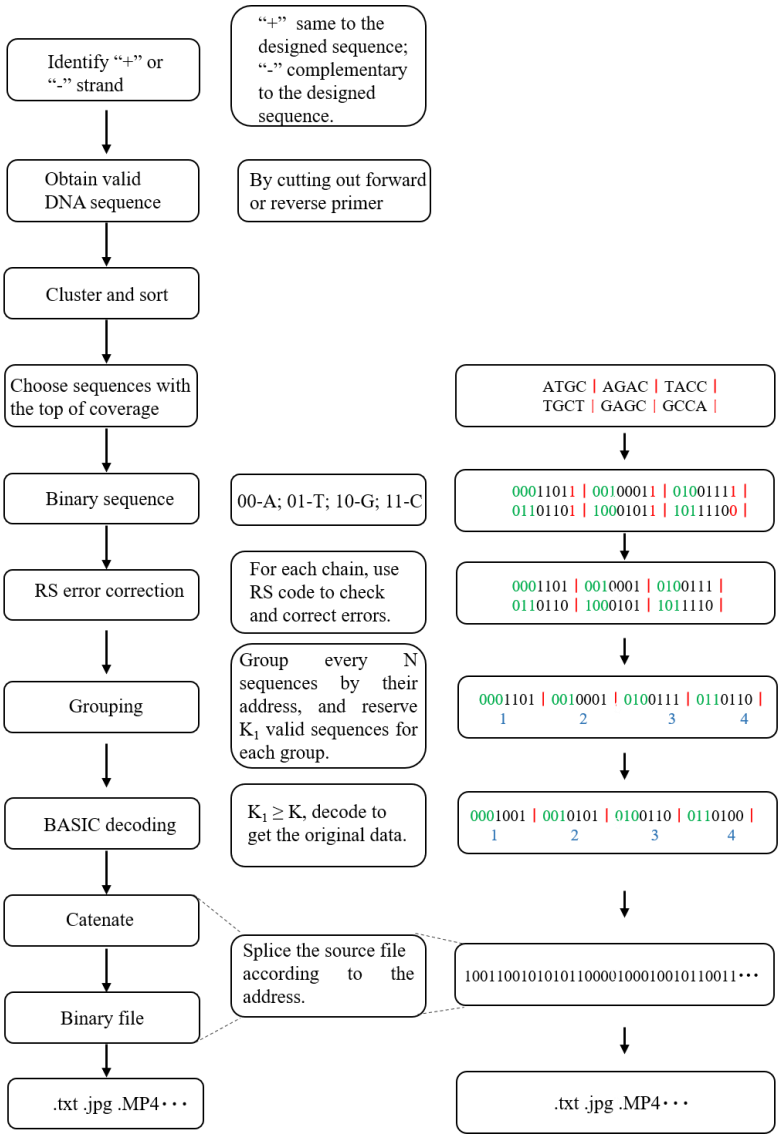
SUPPLEMENTARY INFORMATION



Supplementary Figure 1

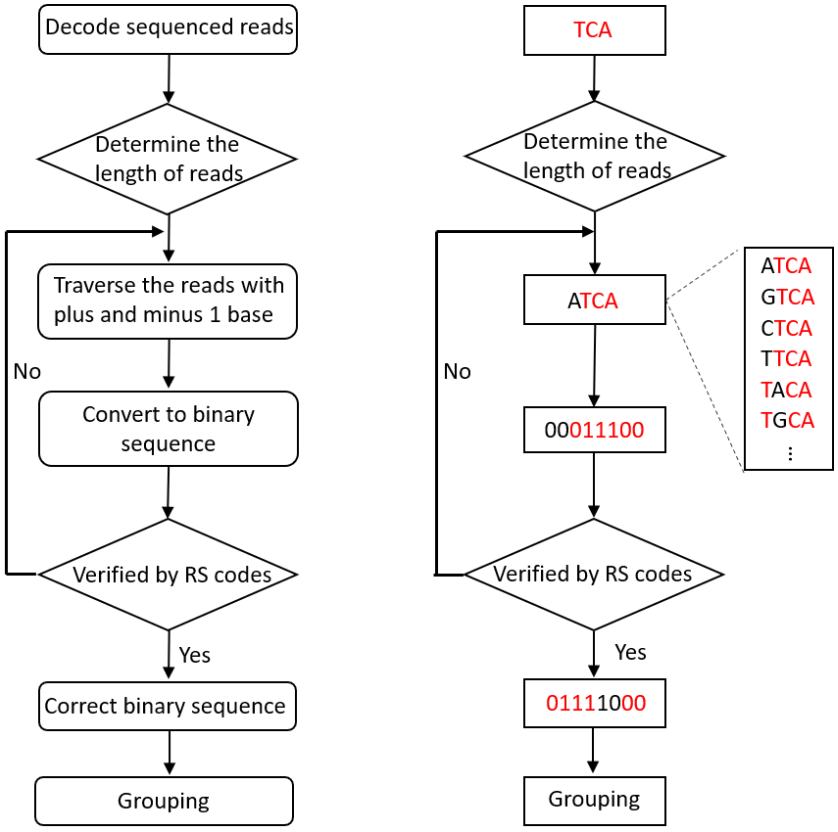
DNA BASIC encoding**.**

(Left) The process of encoding in detail. K and L are arbitrarily defined. Here, K is the number of DNA sequence in each group; L is the length of binary data in each DNA sequence; K codewords of length L are encoded into N codewords by polynomial matrix operations on the polynomial cycle domain $\mathcal{C}_p$, where N>K, N-K is the maximum number of corrupted or lost sequences per group. Otherwise, we would not be perfect decoding. RS codes can detect and correct error within t/2 bases in each DNA sequence (Here, N=256, K=252, L=256, and t=8 in the design of pool 1 and pool 2; N=256, K=252, L=192, and t=8 in the design of pool 3). **(Right)** Example file.



Supplementary Figure 2.

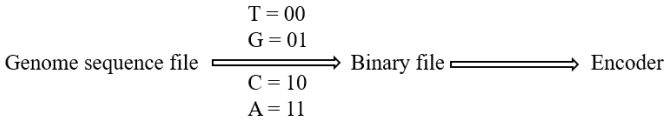
DNA BASIC decoding.



Supplementary Figure 3

RS error correction.

(Left) The detail process of decoding. **(Right)** Example of repairing the DNA sequence with one missing base. Each sequence contains L/2 nt (L bits). Repair the DNA sequence with plus and minus one base via verification of RS codes, which proceed $4^{L/2}$ times at most.



Supplementary Figure 4

The rule that genome sequences were converted to binary information.

Human mitochondrial genome sequences (Chinese, Italy, Native American, and South African) and genome sequences of one artificial bacteria cell (JCVI-syn.1.0) were converted to binary data by the simple rule.

	Number	E/N logical density (bits/nt)	Redundancy	Error correction
Church et al.	0.055 million	0.83/0.6	0	-
Goldman et al.	0.153 million	0.29/0.19	300%	+
Grass et al.	0.005 million	1.16/0.86	0	Reed-Solomon
Erlich et al.	0.072 million	1.55/1.18	7%	Reed-Solomon
Microsoft	13.4 million	1.1/0.8	15%	Reed-Solomon
This work	0.110 million	1.65/1.25	1.56%	Reed-Solomon

Supplementary Figure 5

A comparison of encoding system features to prior reported systems.

By BASIC code, relative high information density 1.65 bits/nt was achieved with a 1.56% of coding redundancy.

Data	File Size	Number of DNA strands	Pool
Central dogma (.jpg)	35 KB	990	Pool 1/2
DNA helix (.gif)	81 KB	2292	Pool 1/2
China Classical literature (.txt)	164 KB	4641	Pool 1/2
A Brief History of Element (.txt)	34 KB	962	Pool 1/2
Panda burn incense (.rar)	66 KB	1867	Pool 1/2
Human Mitochondrial	65 KB	768	Pool 1/2
Bitcoin (.txt)	165 KB	7019	Pool 3
Dictionary of idioms (.txt)	70 KB	2978	Pool 3
Black hole (.jpg)	86 KB	3659	Pool 3
Chaplin (.MP4)	659 KB	28034	Pool 3
JCVI-syn.0.1	1054 KB	44838	Pool 3
Total	**2924 KB**	**109,568**	
-*	-	256	Pool 1/2 – OPN 1.0
**	-	1024	Pool 4 – OPN 2.0

Supplementary Figure 6

Files encoded within these 2.85 MB of data.

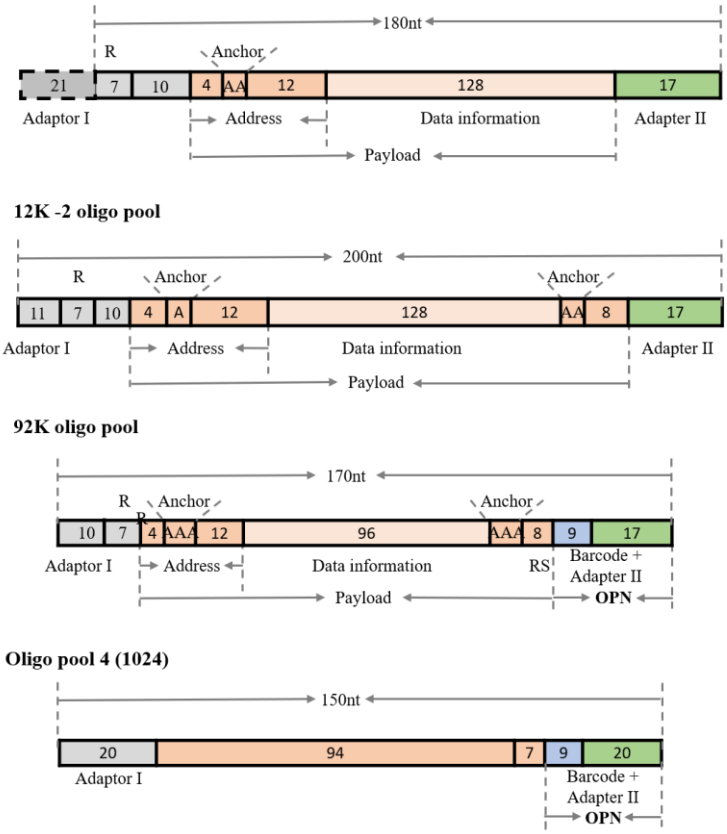
Both pool 1 and pool 2 contain 11,520 DNA sequences. Pool 3 contains 86,528 DNA sequences.

Pool	Sample	Perfect Decoding (Noisy reads)	Coverage (Dropout=0)	Coverage (Dropout=1.56 %)
Pool 1	Free-iDR	Yes	50x	9x
	PCR	Yes	80x	12.5x
	ss	Yse	350x	18x
	#1 PCR	Yes	-**	17x
	#5 PCR	Yes	-	160x
	#10 PCR	No*	-	-***
	#1 iDR	Yes	70x	11x
	#5 iDR	Yes	285x	12x
	#10 iDR	Yes	400x	12x
Pool 2	PCR	Yes	70x	6x
	iDR	Yes	40x	5x
Pool 3	PCR	Yes	-	65x
	iDR	Yes	-	49x
Pool 1/2-OPN 1.0	PCR	Yes	65x	27x
	iDR	Yes	40x	13x
	OPN-iDR	Yes	20x	8x
Pool 4-OPN 2.0	PCR	Yes	223x	--
	OPN-iDR	Yes	87x	--

Supplementary Figure 7

Samples with PCR or iDR amplified were whether could be successfully decoded under the condition of total sequenced reads (noisy reads) used. And the coverage was showed when the dropout rate was 0% and 1.56%.

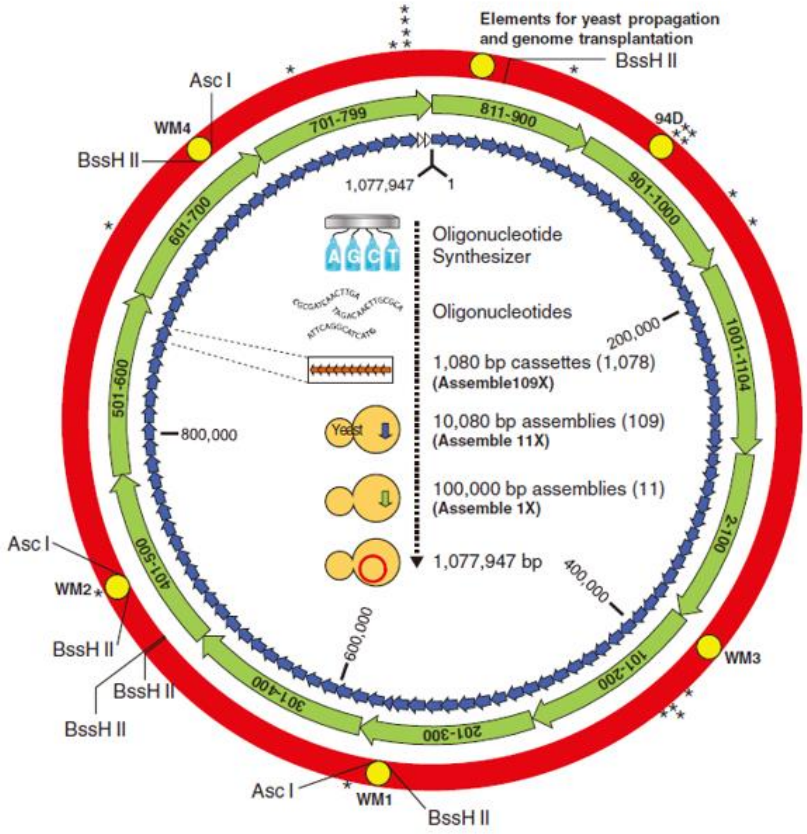
Note: '*' Total noisy reads with missing 14.1% of given sequences (11,520) could not successfully retrieve original file. '**' The dropout rate was beyond 0% even when total noisy reads were used. '***' The dropout rate was beyond 1.56% even when total noisy reads were used.



Supplementary Figure 8

The structure of oligos.

11,520 oligos (Pool 1, Twist Bioscience) without RS code were encoded and synthesized (Adaptor I was added by PCR). And 11,520 oligos (Pool 2, Twist Bioscience) with RS code were encoded and synthesized. 86,528 oligos (Pool 3, CustomArray) were encoded and synthesized. (R: recognition sequence of nickase).



Supplementary Figure 9

The genetic map of JCVI-syn1.0 reported by Gibson, D.G. et al. (Gibson, D.G. et al. Creation of a bacterial cell controlled by a chemically synthesized genome. *Science* **329**, 52-56 (2010).)

Nickase	Recognition site
Nt.BstNBI	5’ GAGTCNNNNN 3’ 3’ CTCAGNNNNN 5’
Nt.BspQI	5’ GCTCTTCN 3’ 3’ CGAGAAGN 5’
Nt.BbvCI	5’ CCTCAGC 3’ 3’ GGAGTCG 5’

Supplementary Figure 10

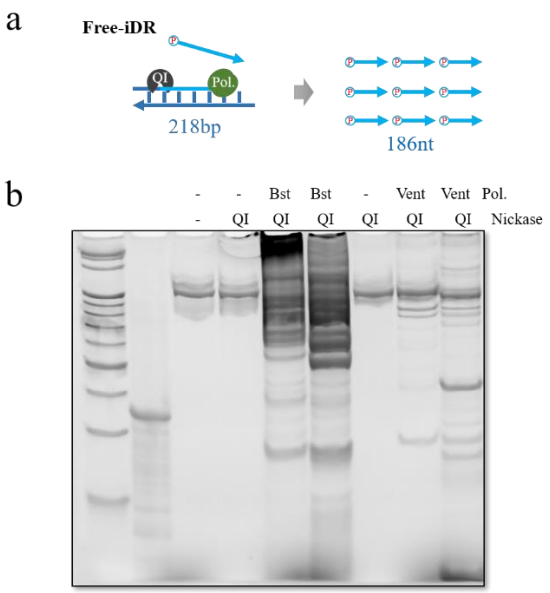
Recognition site of nickase including Nt.BstNBI, Nt.BspQI and Nt.BbvCI we used. Nickase cleaves only one strand of DNA on a double-stranded DNA substrate.

Polymerase	Fidelity
Q5® High-Fidelity DNA Polymerase	1.02 x 10^{-6}
Bst 2.0 WarmStart® DNA Polymerase	62 (±5) x 10^{-6}
Vent® DNA Polymerase	57 x 10^{-6}
Bsu DNA Polymerase, Large Fragment	-
Klenow Fragment (3′→5′ exo-)	100 x 10^{-6}

Supplementary Figure 11

The fidelity of polymerases we used.

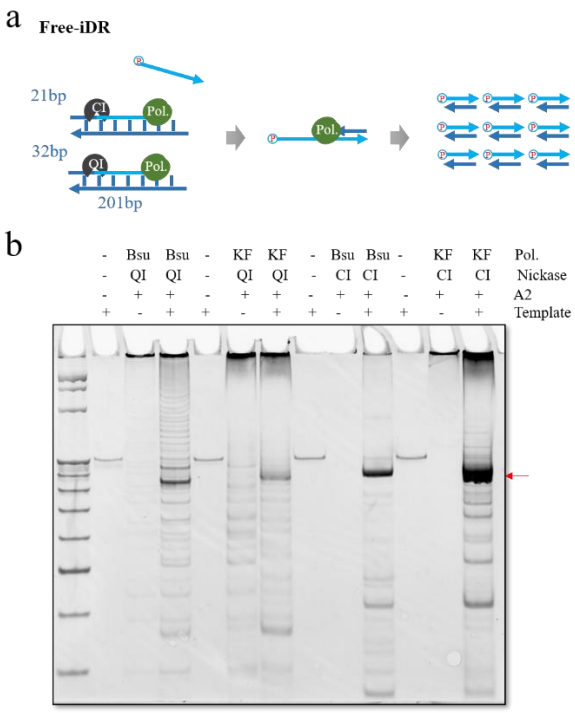
The fidelity of Q5 High-Fidelity DNA polymerase is two orders of magnitude higher than Klenow fragment (3'→5' exo-).



Supplementary Figure 12

iDR reaction under 55℃.

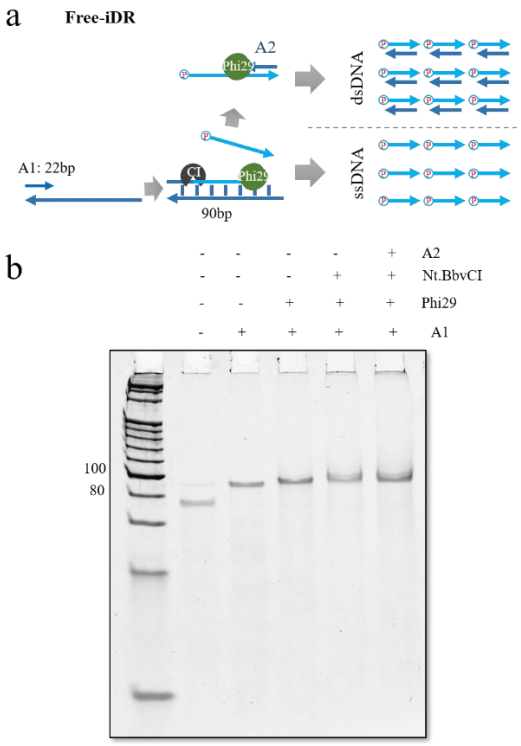
(a) Workflow of iDR reaction. **(b)** 10% Native PAGE results. The higher yield of by-product was occurred using Nt.BspQI and Bst for iDR. Meanwhile, the yield of iDR product using Nt.BspQI and Vent was little and the by-product was also occurred. QI: Nt.BspQI; Bst: Bst2.0 DNA polymerase, large fragement (3'→5 exo$^-$); Vent: Vent DNA polymerase.



Supplementary Figure 13

iDR reaction under 37℃.

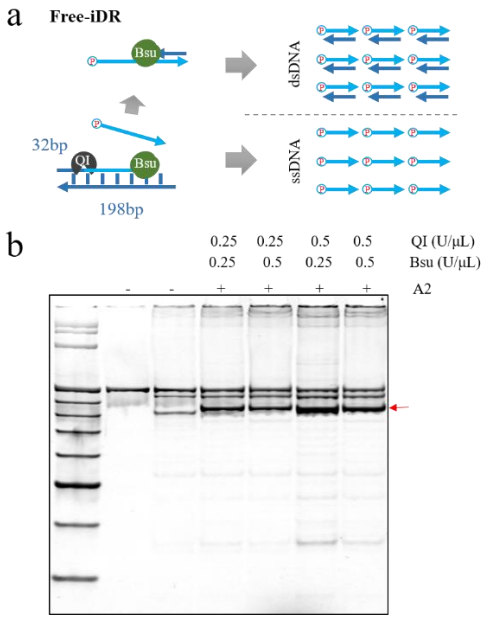
(a) Workflow of iDR reaction. **(b)** 10% Native PAGE results. Red arrow indicates the product of iDR. The combination of nickases and polymerases were Nt.BspQI/Nt.BbvCI and Bsu DNA polymerase/Klenow fragment (3'→5' exo$^-$). QI: Nt.BspQI; CI: Nt.BbvCI; Bsu: Bsu DNA polymerase, large fragement (3'→5' exo$^-$); Klenow fragment (3'→5' exo-). A2: Adaptor 2.



Supplementary Figure 14

iDR reaction under 30℃.

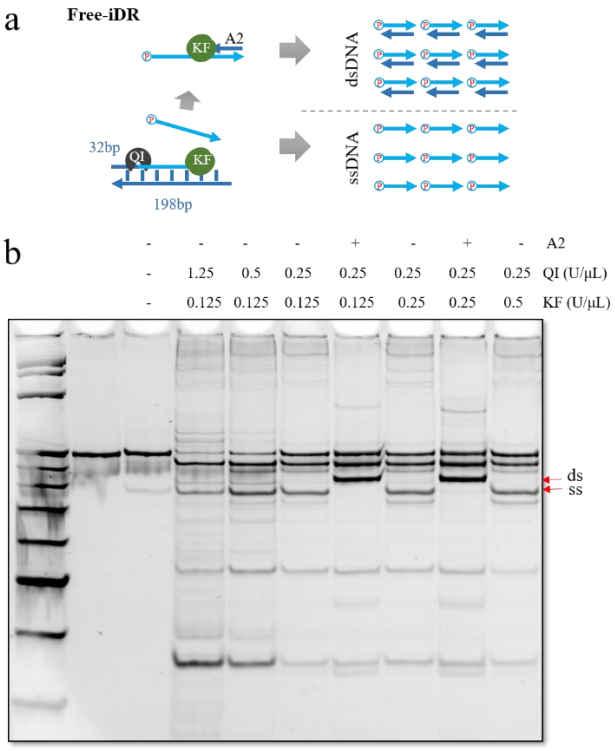
(a) Workflow of iDR reaction. **(b)** 12% Native PAGE results. No amplification was observed on the gel. Phi29: Phi29 DNA polymerase; A1: Adaptor 1; A2: Adaptor 2.



Supplementary Figure 15

Effect of the concentration ratio of Nt.BspQI and Bsu on iDR.

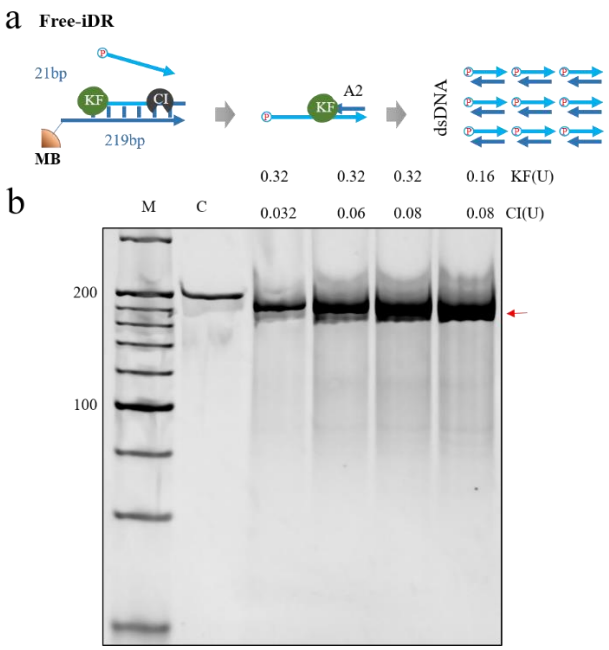
(a) Workflow of iDR reaction. **(b)** 10% Native PAGE results. Red arrow indicates the product of iDR. The yield of iDR product was increased with increasing amount of QI while the by-product was also increased. But the iDR product was slightly decreased as Bsu DNA polymerase was increased. We considered that the optimal concentration of Nt.BspQI and Bsu DNA Polymerase were 0.5 U/μL and 0.25 U/μL separately. Red arrow represents the product of iDR. QI: Nt.BspQI; Bsu: Bsu DNA polymerase, large fragment (3'→5' exo-); A2: Adaptor 2.



Supplementary Figure 16

Effect of the concentration ratio of Nt.BspQI and KF(exo$^-$) on iDR.

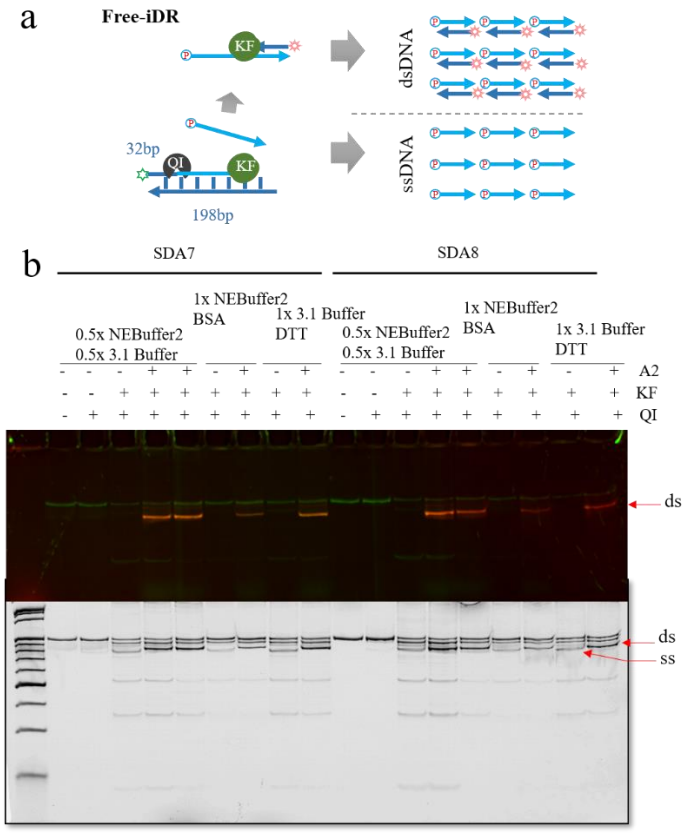
(a) Workflow of iDR reaction. **(b)** 10% Native PAGE results. ss: single-strand DNA product of iDR; ds: double-strand DNA product of iDR. The yield of iDR product was increased with increasing amount of QI rather than KF. Here, the highest yield product of iDR was observed at 0.25 U/μL Nt.BspQI and 0.25 U/μL Klenow fragment (3'→5' exo-). Red arrow represents the product of iDR. QI: Nt.BspQI; KF: klenow fragment (3'→5' exo-); A2: Adaptor 2.



Supplementary Figure 17

Effect of the concentration ratio of Nt.BbvCI and KF (3'→5' exo$^-$) on iDR.

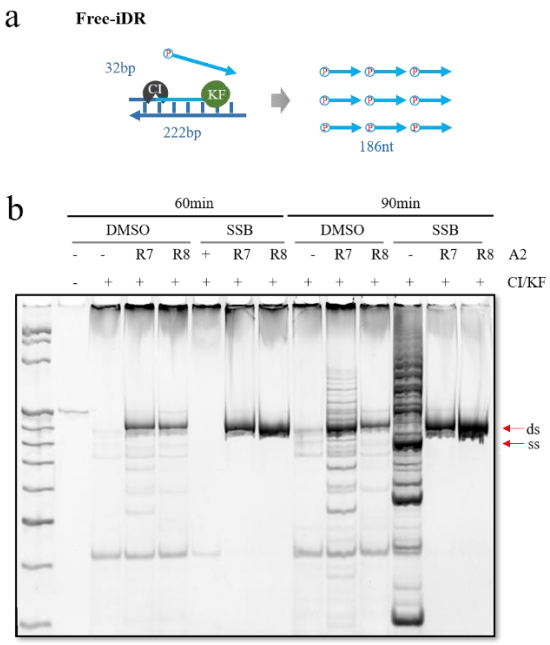
(a) Workflow of iDR reaction. **(b)** 10% Native PAGE results. M: 20 bp DNA Ladder; C: the template of iDR. The yield of iDR product was increased with increasing amount of CI. But the iDR product was kept constant as Klenow fragment DNA polymerase was decreased. Here, the highest yield product of iDR was observed at 0.08 U/μL Nt.BbvCI and 0.16 U/μL Klenow fragment (3'→5' exo-). The concentration of Nt.BbvCI and KF was applied in subsequent assays. Red arrow represents the product of iDR. CI: Nt.BbvCI; KF: Klenow fragemen(3'→5' exo-).



Supplementary Figure 18

Effect of reaction buffer on iDR.

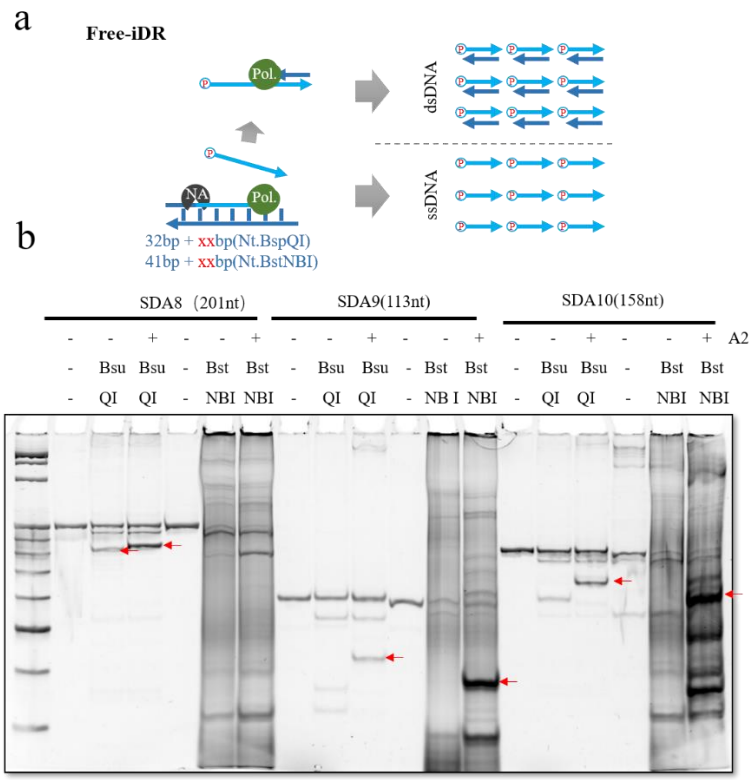
(a) Workflow of iDR reaction. **(b)** 10% Native PAGE results. The DNA with 5'-FAM was used as a template in the iDR reaction. The adaptor 2 with Cy5 was used to generate dsDNA product with red fluorescence of iDR. Red arrow indicates the product of iDR. QI: Nt.BspQI; KF: Klenow fragment (3'→5' exo-); A2: Adaptor 2.



Supplementary Figure 19

Effects of DMSO and SSB on iDR.

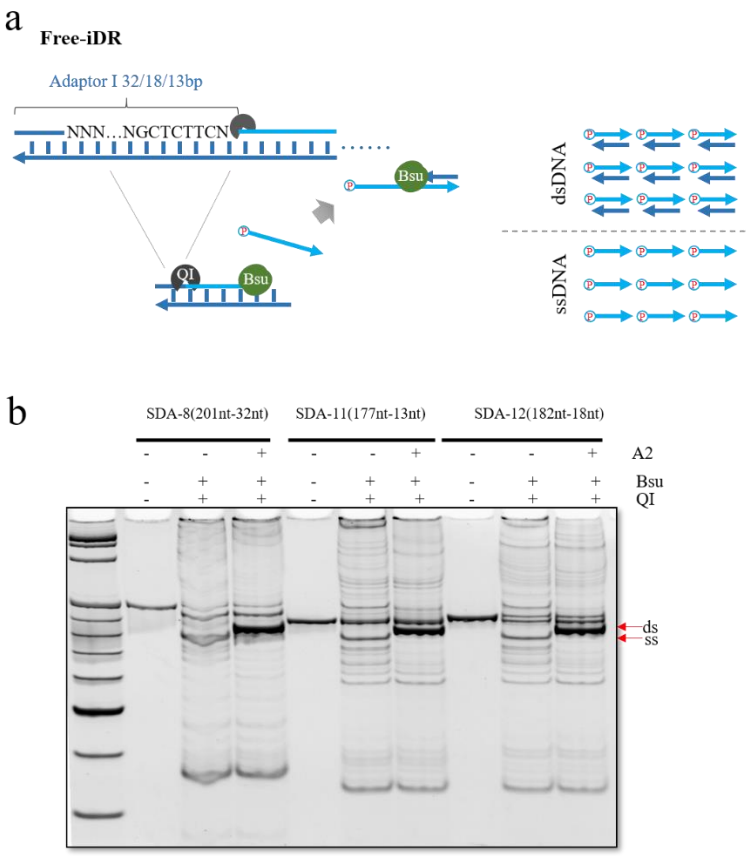
(a) Workflow of iDR reaction. **(b)** 10% Native PAGE results. The yield of iDR product was increased while the by-product was decreased with addition of T4 Gene 32 protein. Red arrow represents the product of iDR. ss: single-strand DNA product of iDR; ds: double-strand DNA product of iDR. CI: Nt.BbvCI; KF: Klenow fragment(3'→5' exo$^-$); A2: Adaptor 2.



Supplementary Figure 20

Effects of fragment length on iDR.

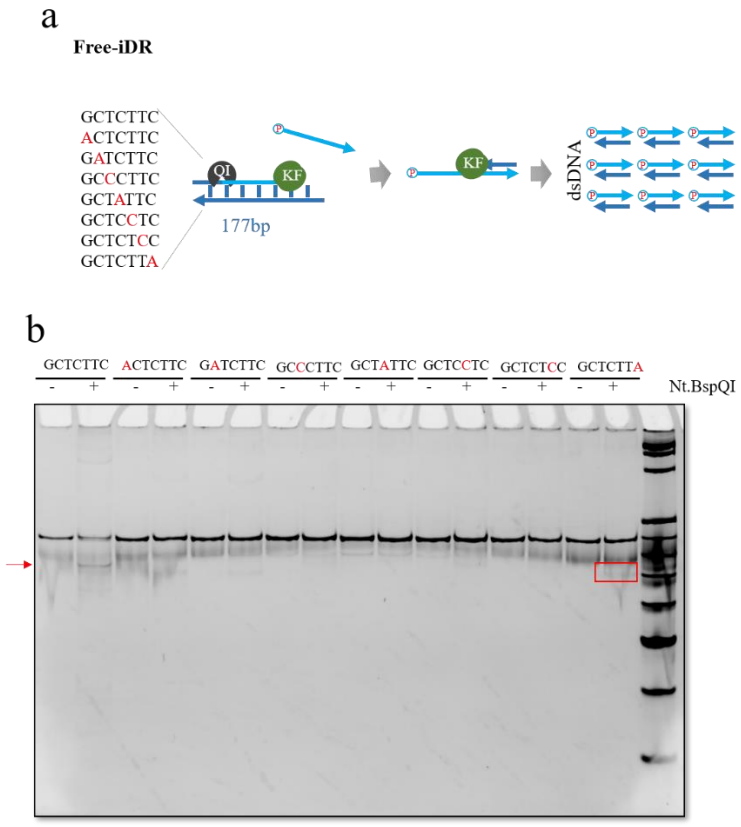
(a) Workflow of iDR reaction. **(b)** 10% Native PAGE results. The iDR reaction using Nt.BspQI and Bsu DNA polymerase is not related to the length of DNA fragment. However, the iDR reaction using Nt.BstNBI and Bst DNA polymerase, large fragment has a preference for the DNA with less than 200bp. Red arrow represents the product of iDR. QI: Nt.BspQI; NBI: Nt.BstNBI; KF: Klenow fragment (3'→5' exo$^-$); Bsu: Bsu DNA polymerase, large fragment (3'→5' exo$^-$); A2: Adaptor 2.



Supplementary Figure 21

Effects the length of adaptor I on iDR.

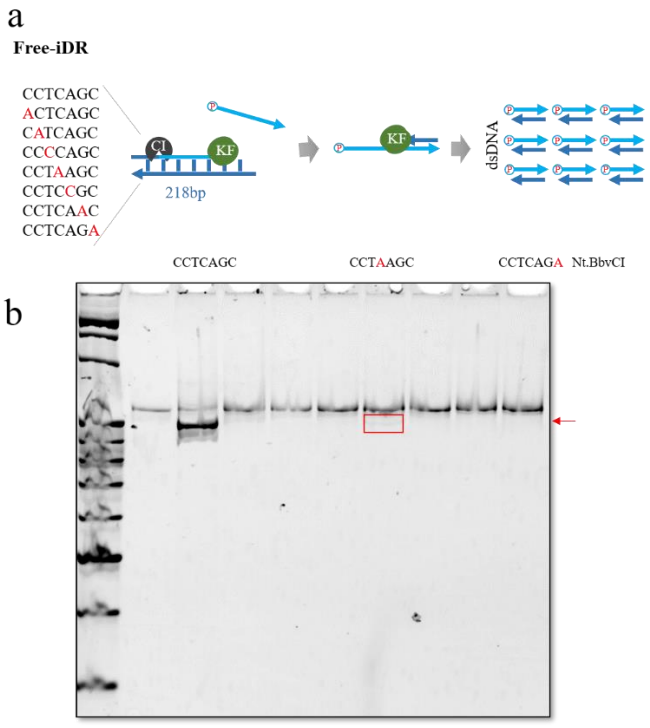
(a) Workflow of iDR reaction. **(b)** 10% Native PAGE results. The minimum length of Adaptor I was 13 nt which was enough stable for avoiding the adaptor I dissociating. Red arrow represents the product of iDR; ss: single-strand DNA product of iDR; ds: double-strand DNA product of iDR. QI: Nt.BspQI; Bsu: Bsu DNA polymerase, large fragment (3'→5' exo$^-$); A2: Adaptor 2.



Supplementary Figure 22

Effect of variations in recognition region on the nicking efficiency of Nt.BspQI.

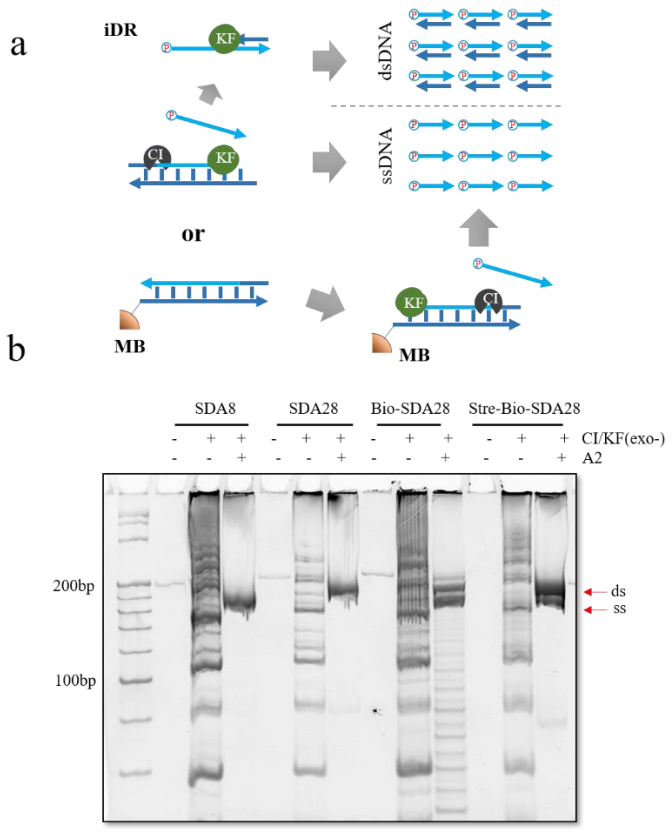
(a) Workflow of iDR reaction. **(b)** 10% Native PAGE results. The iDR product was observed when GCTCTTC, recognition sequence of Nt.BspQI, was converted into GCTCTTA. It demonstrated that Nt.BspQI was unspecific to recognize GCTCTTC. Thus, GCTCTTN was evaded in the sequence we encoded. Red arrow represents the product of iDR.



Supplementary Figure 23

Effect of variations in recognition region on the nicking efficiency of Nt.BbvCI.

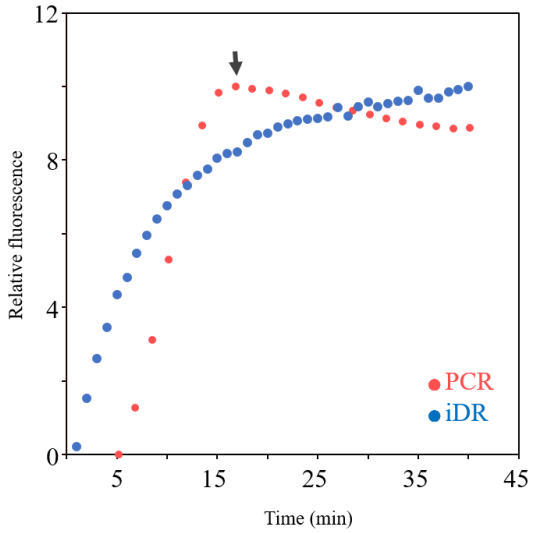
(a) Workflow of iDR reaction. **(b)** 10% Native PAGE results. The iDR product was observed when CCTCAGC, recognition sequence of Nt.BspQI, was converted into CCTAAGC. It demonstrated that Nt.BspQI was unspecific to recognize CCTCAGC. Thus, CCTNAGC was evaded in the sequence we encoded. Red arrow represents the product of iDR.



Supplementary Figure 24

Effect of templates attached to the magnetic beads on iDR.

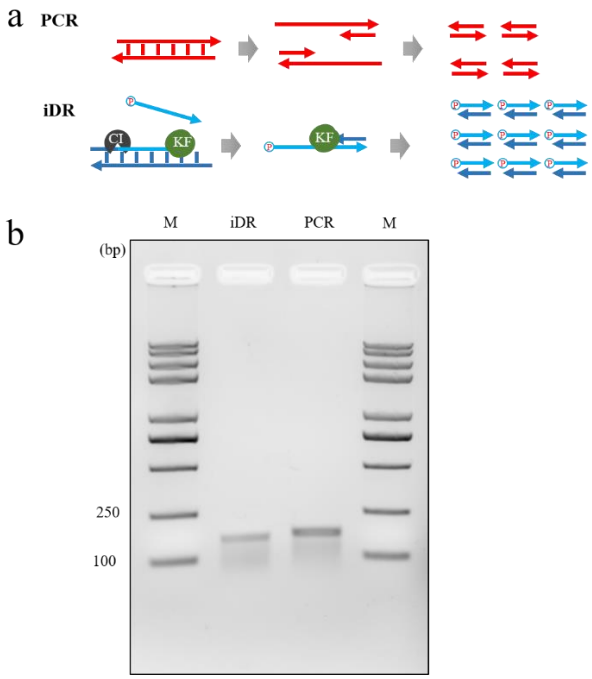
(a) Workflow of iDR reaction. **(b)** 10% Native PAGE results. DNA templates modified with biotin and attached to the magnetic beads had no influence on iDR reaction. Red arrow represents the product of iDR. ss: single-strand DNA product of iDR; ds: double-strand DNA product of iDR. CI: Nt.BbvCI; KF: Klenow fragment(3'→5' exo-); Bio-SDA28: the DNA (SDA28) modified with biotin; Stre-Bio-SDA28: the DNA template (SDA28) was attached to the streptavidin magnetic beads by the interaction of streptavidin and biotin; A2: Adaptor 2.



Supplementary Figure 25

Real-time PCR and iDR. Black arrow represents the tenth cycle of PCR.

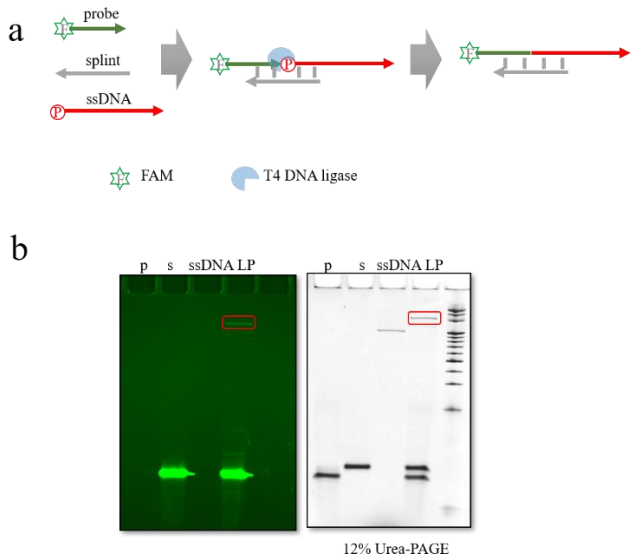
The iDR reaction mixtures contained 1 µL of the template (10 ng/µL), 0.25 mM dNTPs, 2.5 µL 10x NEBuffer 2, 0.08 U/µL Nt.BbvCI, 0.16 U/µL KF polymerase (exo$^-$), 4 µM SSB, 0.2 mg/mL BSA, 0.5 µM adaptor 2 (For production of ssDNA, adaptor was not added.) and 0.5 µL 3.75xSYBR Green I. The mixtures were incubated at 37℃ for 30 min and detected every 30s. PCR was performed using Q5 High-Fidelity DNA Polymerase and forward primer/adaptor 2 (10ng DNA master pool, 2 µL of forward primer (100 µM); 2 µL of adaptor 2 (100 µM)), 1 µL 3.75x SYBR Green I, and 10 µL 5x Q5 reaction buffer in a 50 µL reaction. Thermocycling conditions were as follows: 5 min at 98℃; 10 cycles of: 30 s at 98, 30 s at 58℃, 10 s at 72℃ (detection), followed by extension at 72℃ for 5 min.



Supplementary Figure 26

iDR product and PCR product.

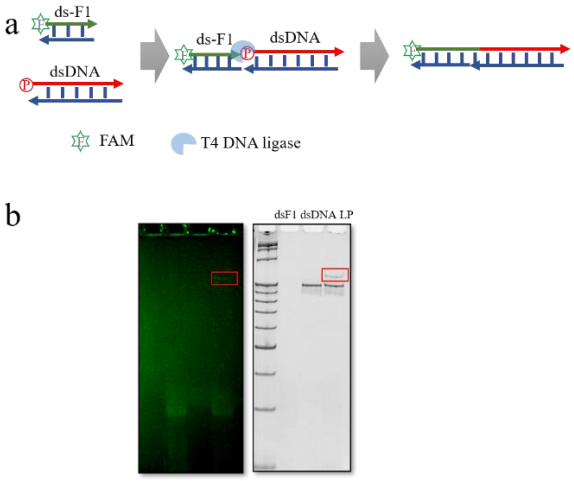
(a) PCR and iDR reaction. **(b)** 2% agarose gel results. Lane M: Trans2K Plus II DNA Maker; lane iDR: double-strand DNA product of iDR; lane PCR: PCR product.



Supplementary Figure 27

5’-phosphate group of iDR ssDNA product confirmed through ligation assay.

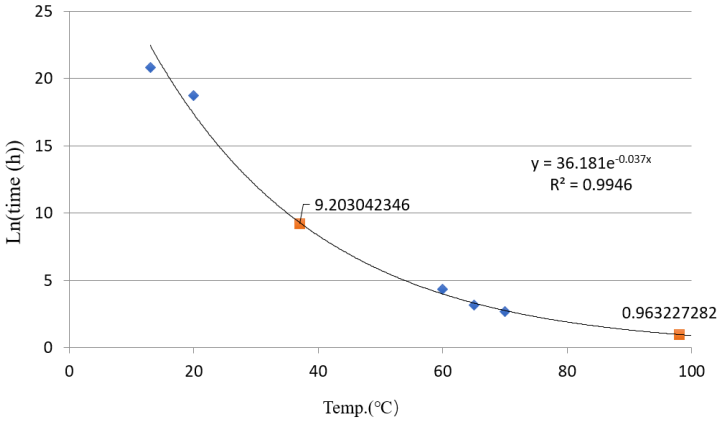
(a) Workflow of ligation. **(b)** Verification of 5’-phosphate group of iDR product, as shown on 12% TBE-urea gel. Lane p: probe; lane s: splint; Lane ssDNA: single-strand product of iDR; lane LP: Ligation product (LP) was generated in the red box. The ligation reaction was carried out in 20 μL reaction mixtures containing 0.5 μM of the probe with FAM at 5’ terminal, 0.5 μM of the splint, 4 μL of the ssDNA (9 ng/μL), 2 μL of 10xT4 DNA ligase reaction buffer, 25 U T4 DNA ligase. Before adding T4 DNA ligase, the mixture was annealed from 95℃ to 25℃ at a ramp of 0.1℃/s. After annealing, T4 DNA ligase was added to the mixture and incubated at 37℃ for 3 h. The reaction was terminated by incubation at 65℃ for 10 min. Image of the gel was recorded using a Cy3 channel of the C300 Imaging System. Then the gel was stained with SYBR Gold for 20 min at room temperature, followed by recorded using UV channel of the C300 Imaging System.



Supplementary Figure 28

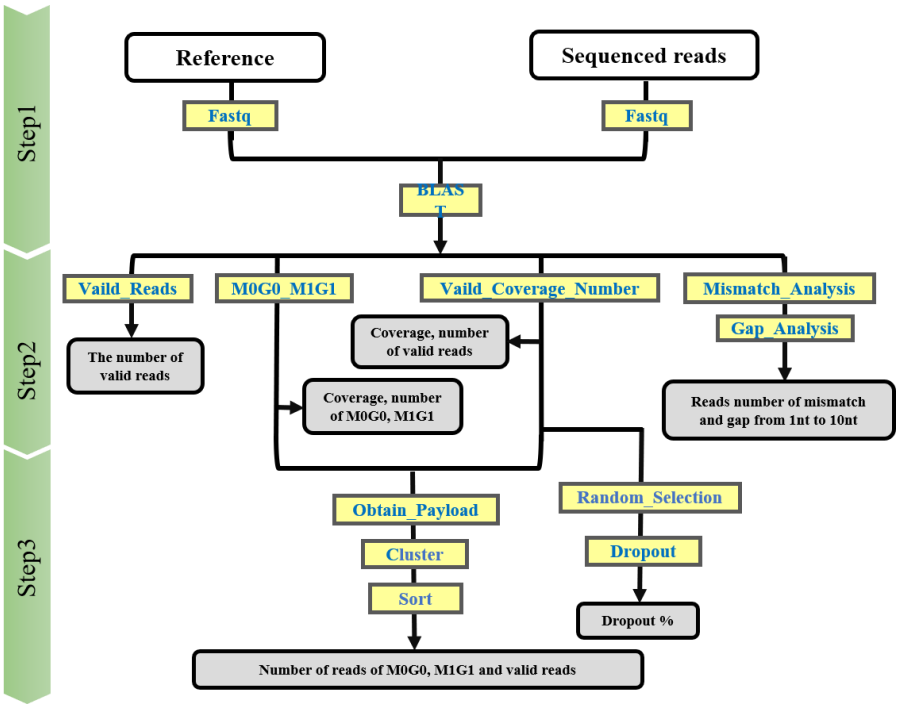
5'-phosphate group of iDR dsDNA product confirmed through ligation assay.

(a) Workflow of ligation. **(b)** Verification of 5'-phosphate group of iDR product, as shown on 12% native PAGE. Ligation product (LP) was generated in the red box. The ligation reaction was carried out in 20 μL reaction mixtures containing 0.5 μM of the probe (ds-probe) with FAM at 5' terminal, 4 μL of the dsDNA (20 ng/μL), 2 μL of 10xT4 DNA ligase reaction buffer, 25 U T4 DNA ligase. Before adding T4 DNA ligase, the mixture was annealed from 95℃ to 25℃ at a ramp of 0.1℃/s. After annealing, T4 DNA ligase was added to the mixture and incubated at 16℃ for overnight. The reaction was terminated by incubation at 65℃ for 10 min. Image of the gel was recorded using a Cy3 channel of the C300 Imaging System. Then the gel was stained with SYBR Gold for 20 min at room temperature, followed by recorded using UV channel of the C300 Imaging System.



Supplementary Figure 29

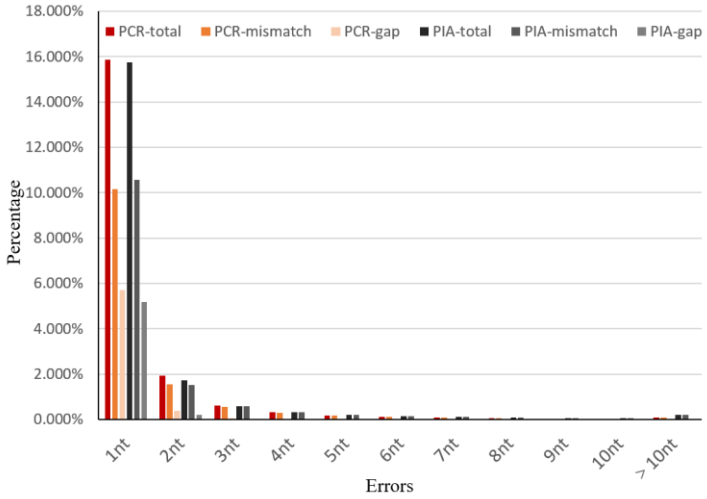
The temperature – logarithm of time curve of DNA decay (Supplementary Note 8)**.**



Supplementary Figure 30

The workflow of bioinformatic statistical analysis.

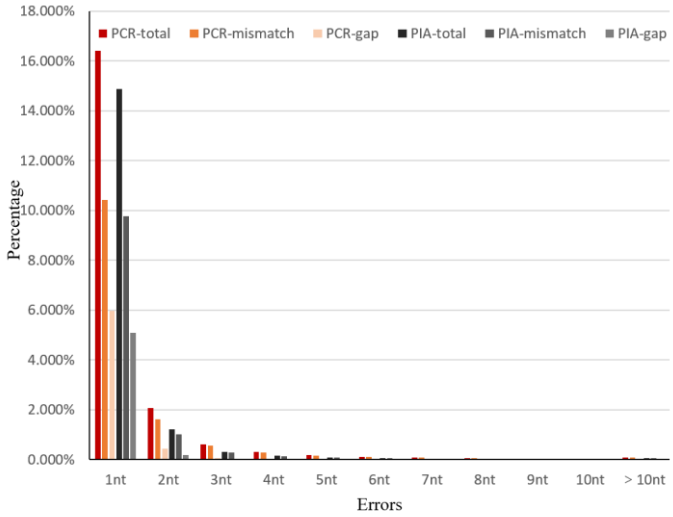
Bioinformatics analysis programs are in the yellow boxes. Results generated by bioinformatics analysis programs are gray boxes.



Supplementary Figure 31

Errors containing mismatch and indel in #1 PCR and #1 iDR.

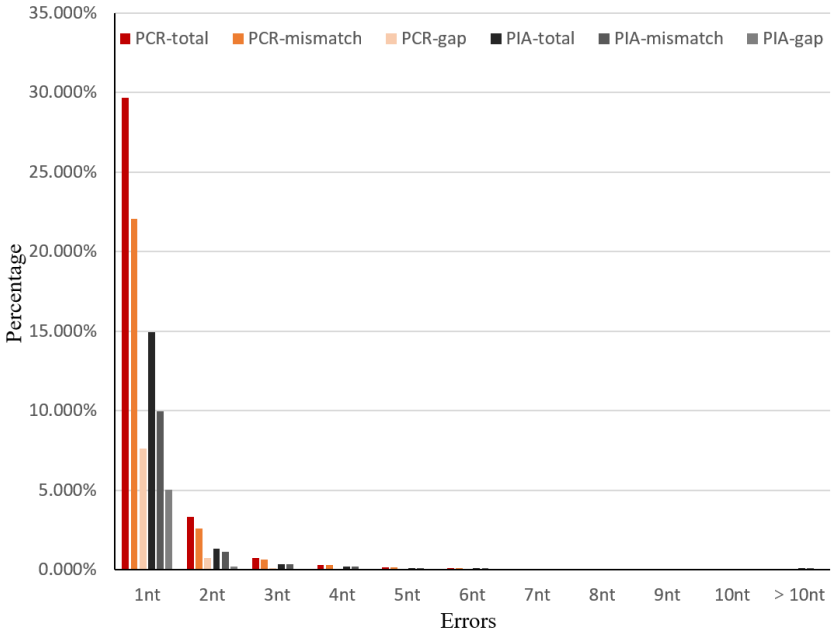
The percentage of the reads with errors containing base substitution and indel among total reads respectively in #1 PCR and #1 iDR (Pool 1, Twist Bioscience).



Supplementary Figure 32

Errors containing mismatch and indel in #5 PCR and #5 iDR.

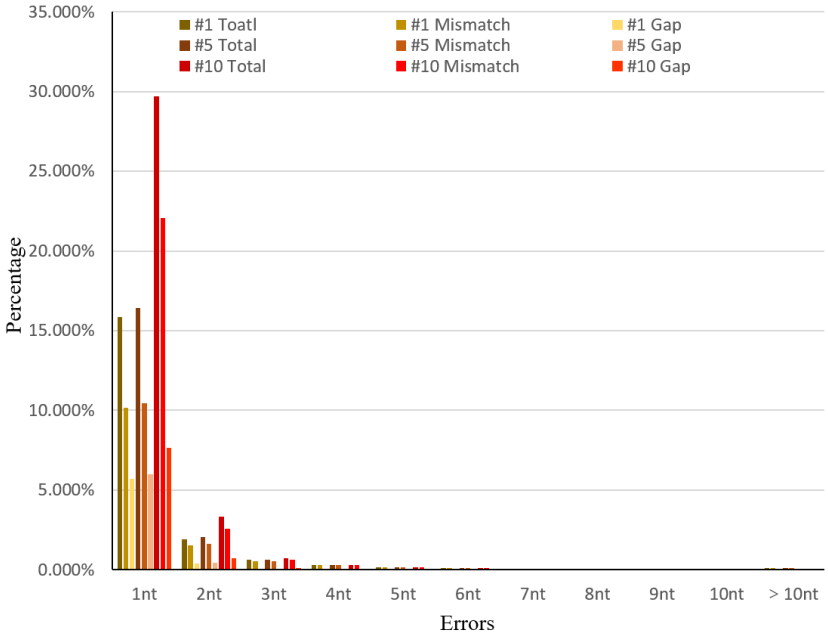
The percentage of the reads with errors containing substitution and indel among total reads respectively in #5 PCR and #5 iDR (Pool 1, Twist Bioscience).



Supplementary Figure 33

Errors containing mismatch and indel in #10 PCR and #10 iDR.

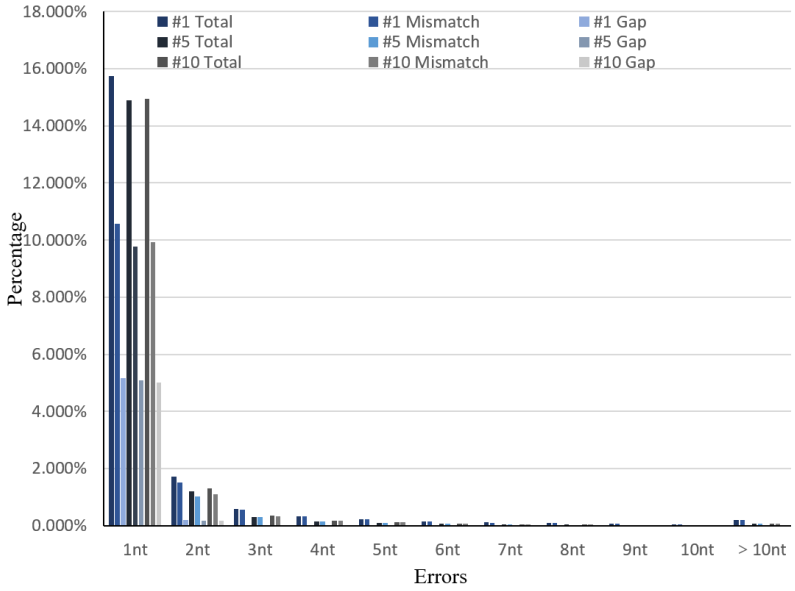
The percentage of the reads with errors containing substitution and indel among total reads respectively in #10 PCR and #10 iDR (Pool 1, Twist Bioscience).



Supplementary Figure 34

Errors containing mismatch and indel in #1, #5 and #10 PCR.

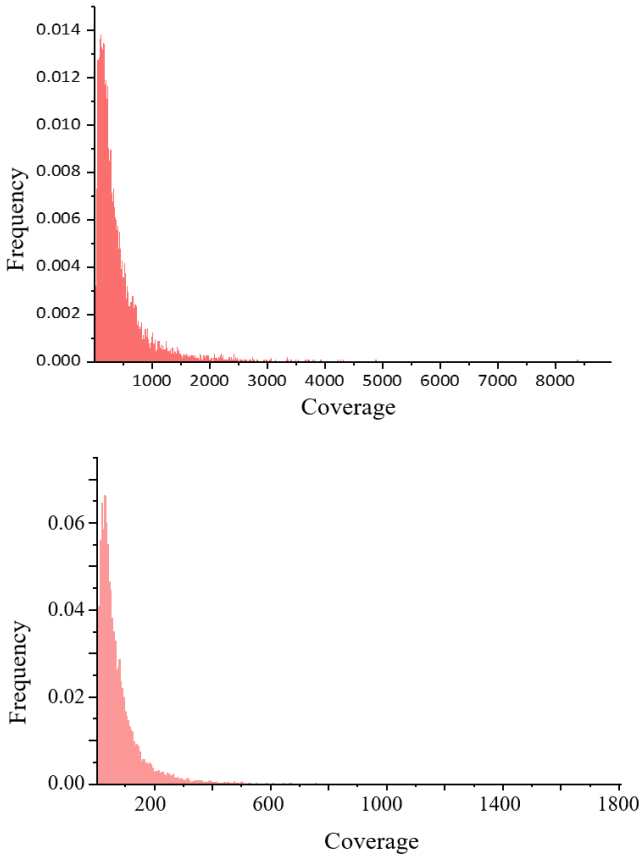
The percentage of the reads with errors containing substitution and indel among total reads respectively in repetitive #1, #5, #10 PCR (Pool 1, Twist Bioscience).



Supplementary Figure 35

Errors containing mismatch and indel in #1, #5 and #10 PCR.

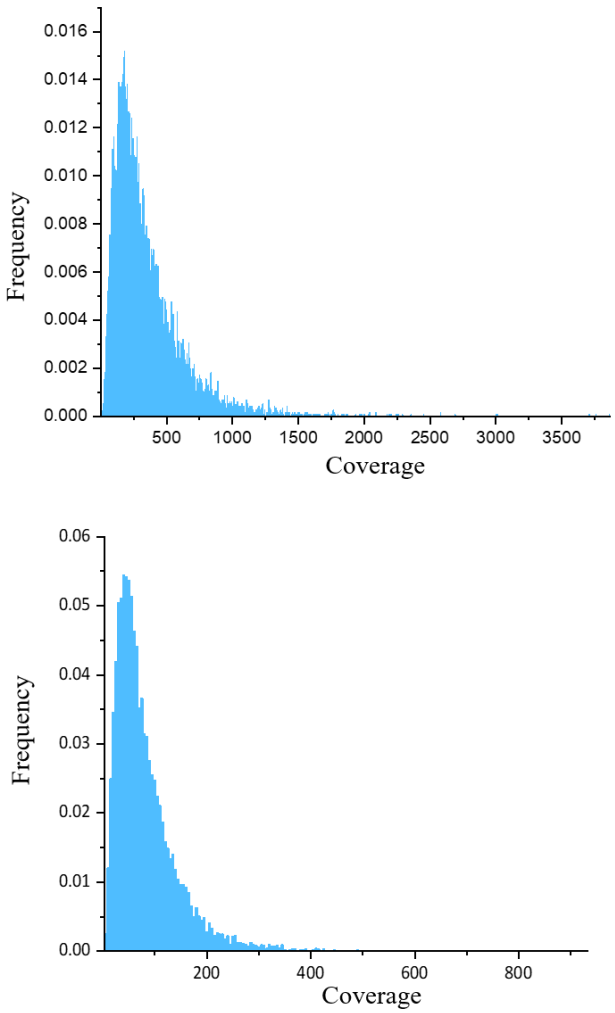
The percentage of the reads with errors containing substitution and indel among total reads respectively in #1, #5, #10 iDR (Pool 1, Twist Bioscience).



Supplementary Figure 36

The distribution of the number of reads per each given sequences of PCR.

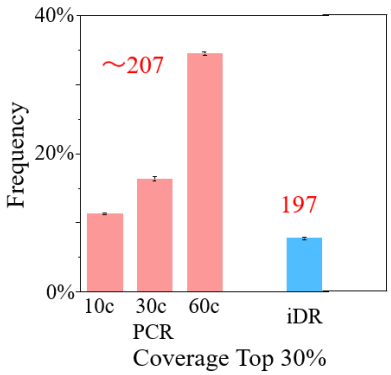
The distribution of the number of reads per each given sequence of total sequenced reads (the upper portion) and per million sequenced reads (the lower part) of PCR-amplified the oligo pool (Pool 1, Twist Bioscience).



Supplementary Figure 37

The distribution of the number of reads per each given sequence of iDR.

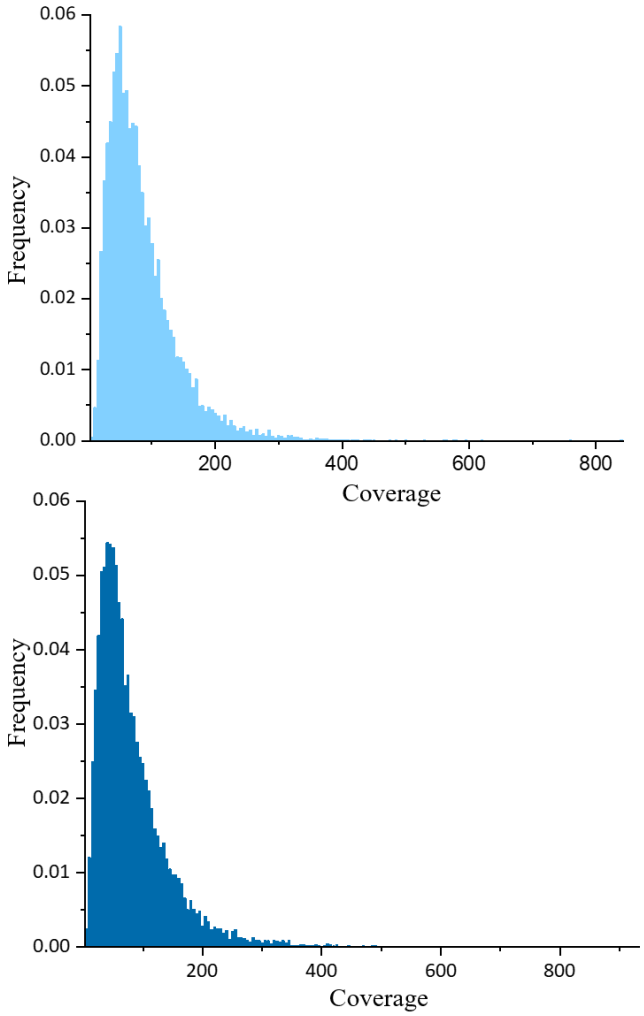
The distribution of the number of reads per each given sequence of total sequenced reads (the upper portion) and per million sequenced reads (the lower part) of iDR-amplified the oligo pool (Pool 1, Twist Bioscience).



Supplementary Figure 38

The percentage of reads with having a large coverage (the top 30% of coverage).

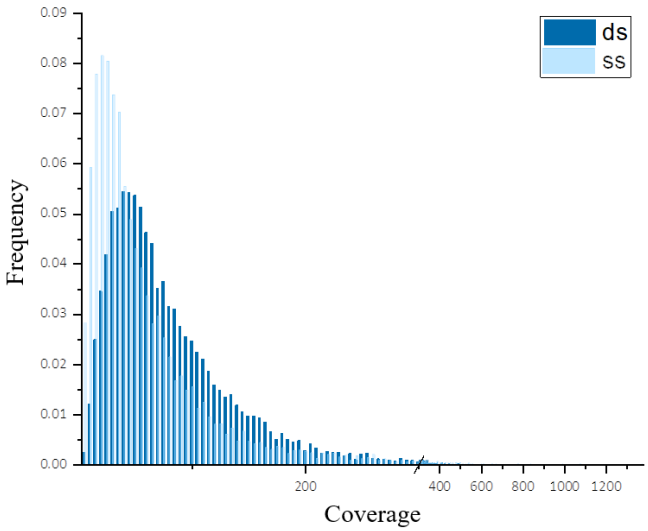
The percentage increased significantly from 11.31% to 34.52% as the number of PCR cycles increased from 10 to 60 while the number of these reads stabilized at 207. Meanwhile the number of these reads with a large coverage was 197 and these reads accounted for 7.77% of per million sequenced reads. Error bars represent the mean ±s.d., where n = 3.



Supplementary Figure 39

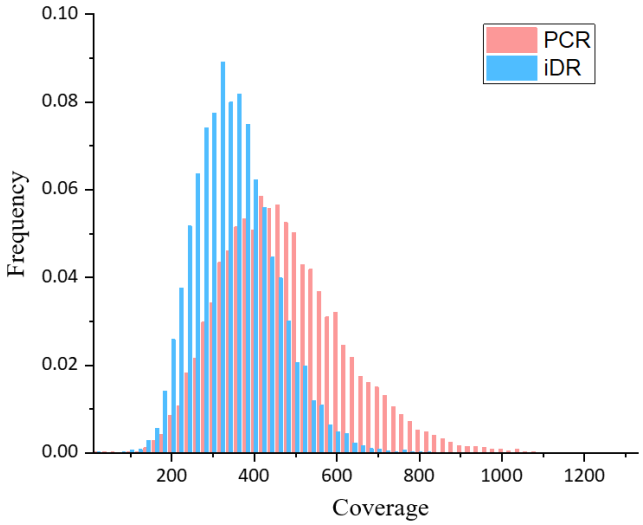
The distribution of the number of reads per each given sequence per million sequenced reads of free-iDR and iDR.

The distribution of the number of reads per each given sequence of free-iDR (without Streptavidin magnetic beads, the upper portion) and iDR-amplified (the lower part) the oligo pool (Pool 1, Twist Bioscience).



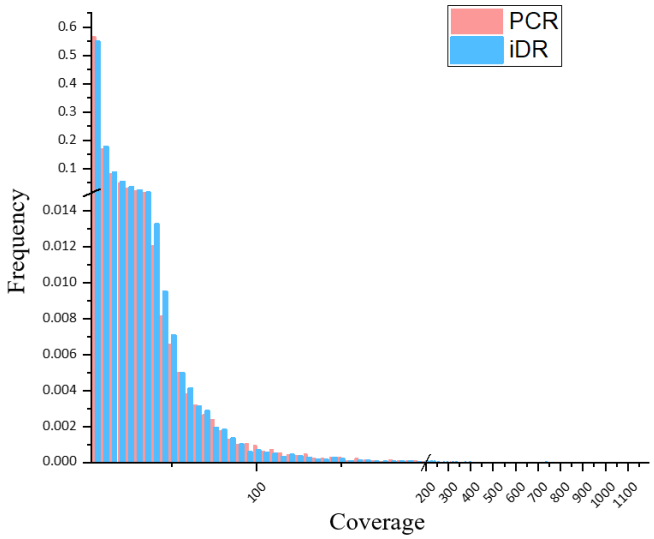
Supplementary Figure 40

The distribution of the number of reads per each given sequence per million sequneced reads of double-strand DNA product (blue) and single-strand DNA product (light blue) of iDR amplified the oligo pool (Pool 1, Twist Bioscience).



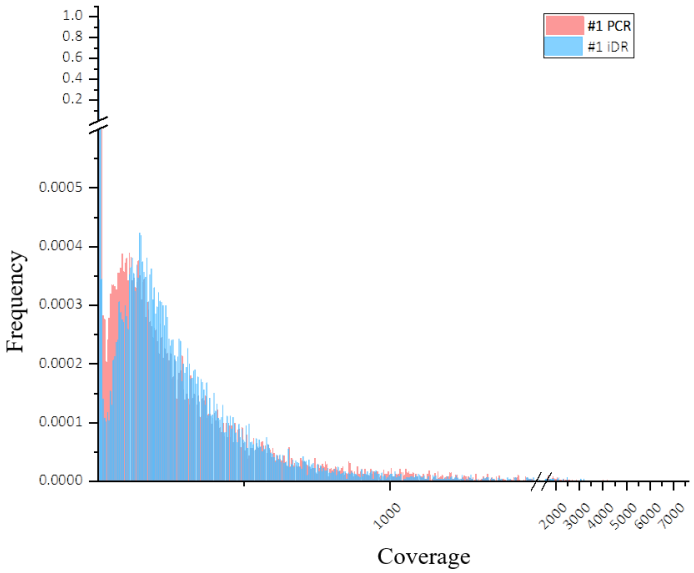
Supplementary Figure 41

The distribution of the number of reads per each given sequence of PCR and iDR-amplified the oligo pool (Pool 2, Twist Bioscience).



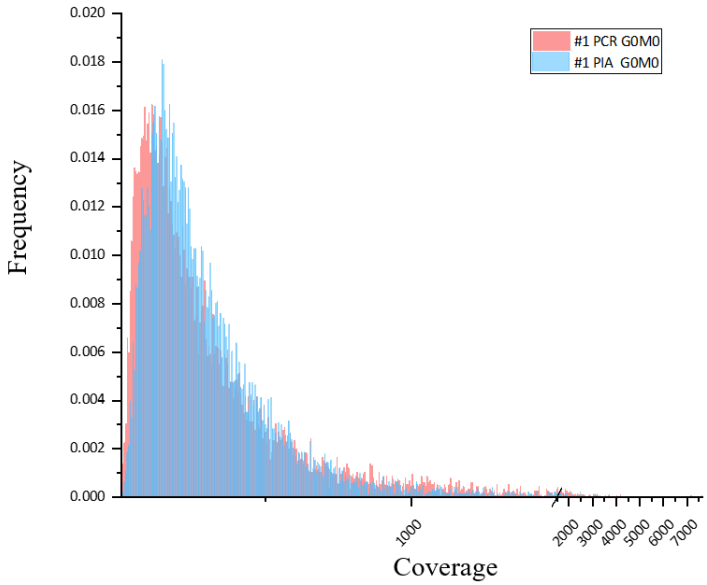
Supplementary Figure 42

The distribution of the number of reads per each given sequence of PCR and iDR-amplified the oligo pool (Pool 3, CustomArray).



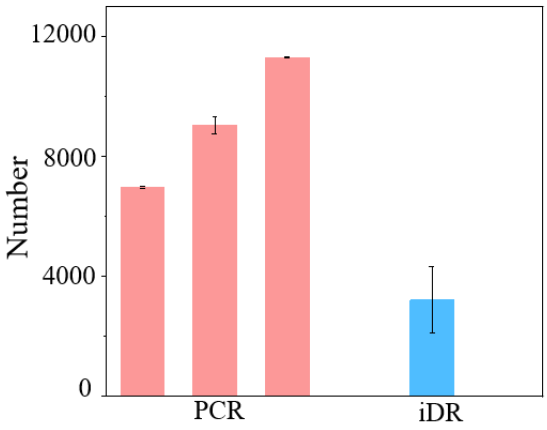
Supplementary Figure 43

The distribution of the number of reads of PCR and iDR –amplified the oligo pool (Pool 1, Twist Bioscience).



Supplementary Figure 44

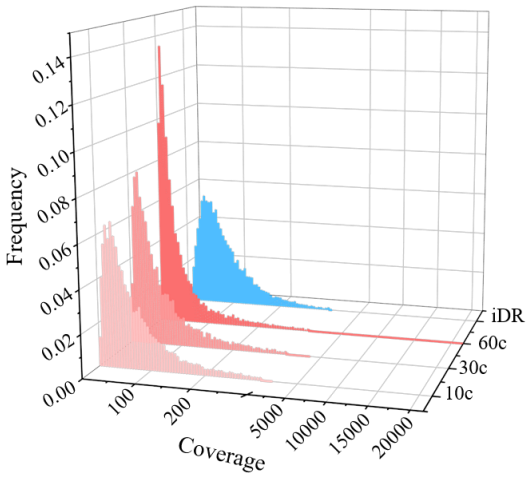
The distribution of the number of M0G0 reads of PCR and iDR –amplified the oligo pool (Pool 1, Twist Bioscience).



Supplementary Figure 45

The number of sequences affected by the M1G1 in the process of decoding.

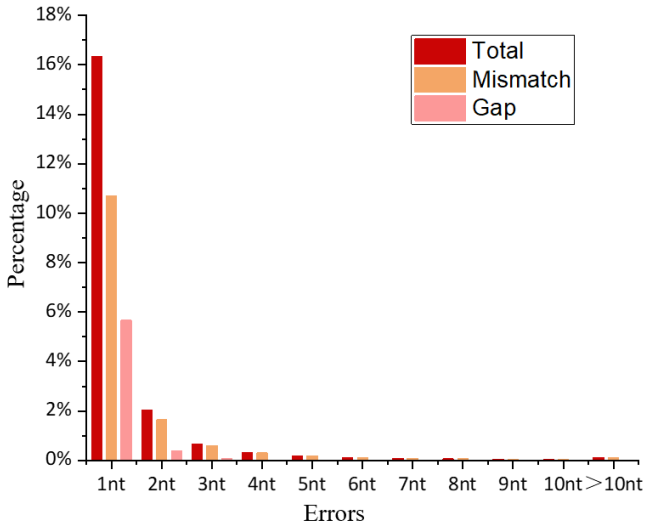
The number increased significantly from 6958±32 to 11289±8 as the number of PCR cycles increased from 10 to 60. The number of iDR was kept at 3216±1111. Error bars represent the mean ±s.d., where n = 3.



Supplementary Figure 46

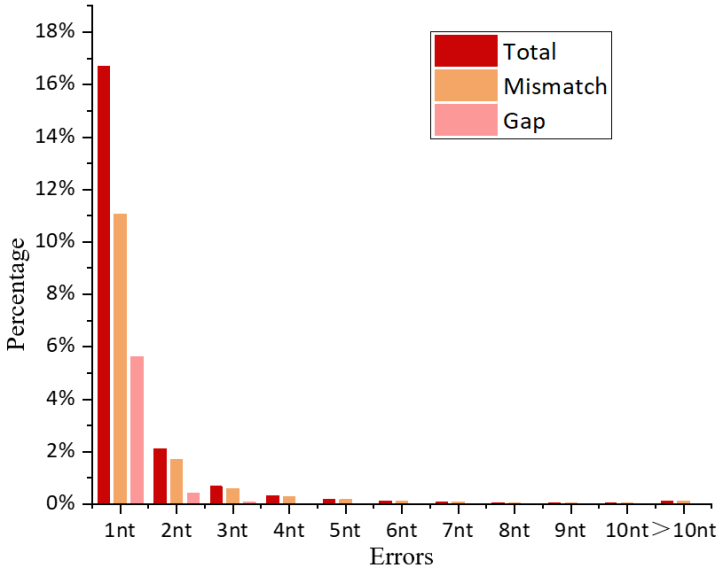
The distribution of the number of reads per each given sequence per million sequenced reads of PCR and iDR-amplified the oligo pool (12K-1 oligo pool, Twist Bioscience).

The distribution of the number of reads per each given sequence per million sequenced reads of PCR with 10 cycles (light red), PCR with 30 cycles (red), PCR with 60 cycles (dark red) and iDR (blue)-amplified the oligo pool (Pool 1, Twist Bioscience).



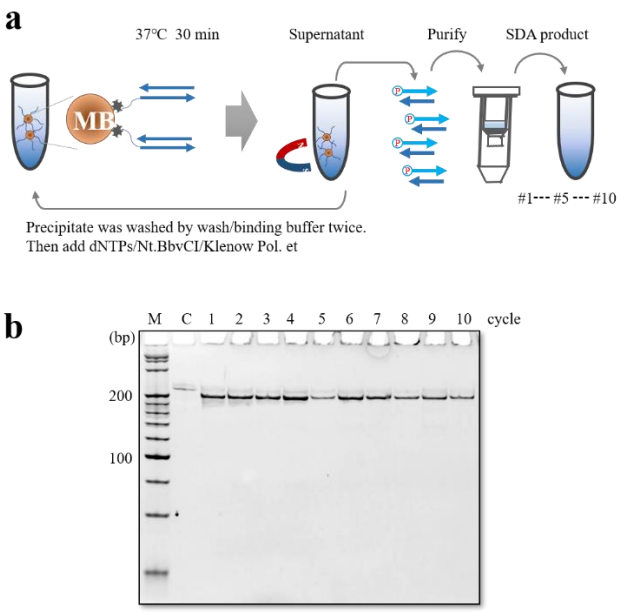
Supplementary Figure 47

The percentage of the reads with errors containing substitution and indel ranging from 1 to 10 nt and more than 10 nt among total reads respectively in PCR with 30 cycles.



Supplementary Figure 48

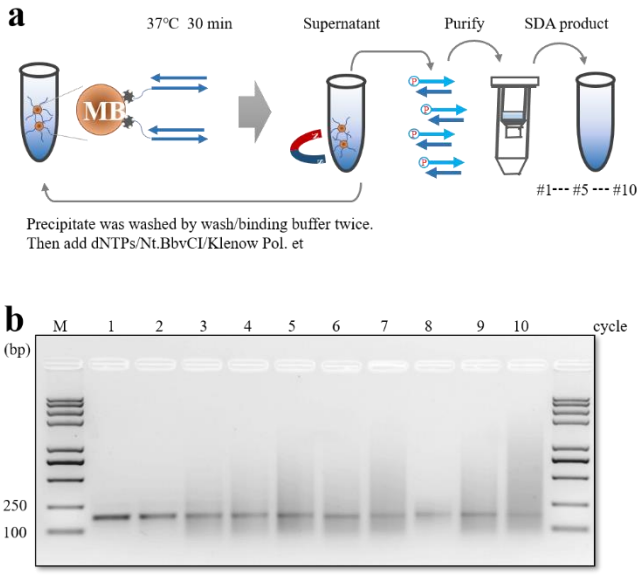
The percentage of the reads with errors containing substitution and indel ranging from 1 to 10 nt and more than 10 nt among total reads respectively in PCR with 60 cycles.



Supplementary Figure 49

Experimental procedures of repeated iDR-amplified the oligo (SDA28).

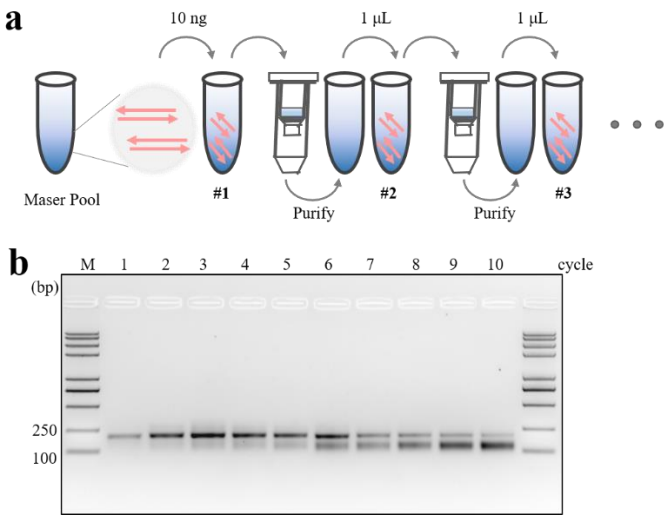
(a) Workflow of 10 consecutive amplification of iDR. **(b)** 10% native PAGE results of iDR products**.** Lane M: 20bp DNA ladder; lane C: DNA template of iDR; lane 1: iDR #1; lane 2: iDR #2; lane 3: iDR #3; lane 4: iDR #4; lane 5: iDR #5; lane 6: iDR #6; lane 7: iDR #7; lane 8: iDR #8; lane 9: iDR #9; lane 10: iDR #10.



Supplementary Figure 50

Experimental procedures of repeated iDR-amplified the oligo pool (Pool 1, Twist Bioscience).

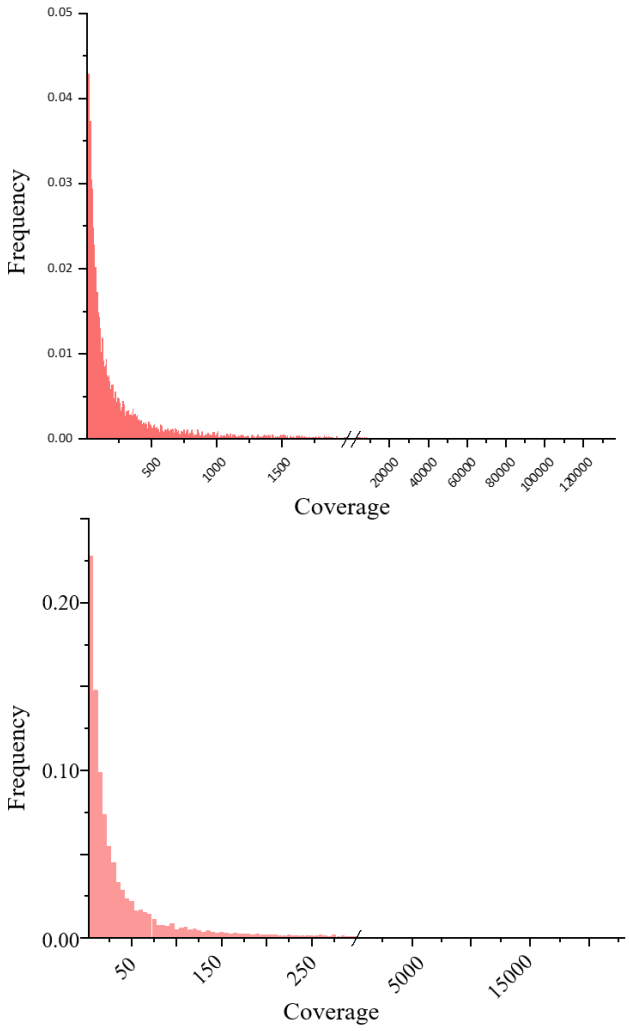
(a) Workflow of 10 consecutive amplification of iDR. **(b)** 2% agarose results of iDR products**.** Lane M: Trans2K Plus II DNA Maker; lane 1: iDR #1; lane 2: iDR #2; lane 3: iDR #3; lane 4: iDR #4; lane 5: iDR #5; lane 6: iDR #6; lane 7: iDR #7; lane 8: iDR #8lane 9: iDR #9; lane 10: iDR #10.



Supplementary Figure 51

Experimental procedures of repeated PCR-amplified the oligo pool (Pool 1, Twist Bioscience).

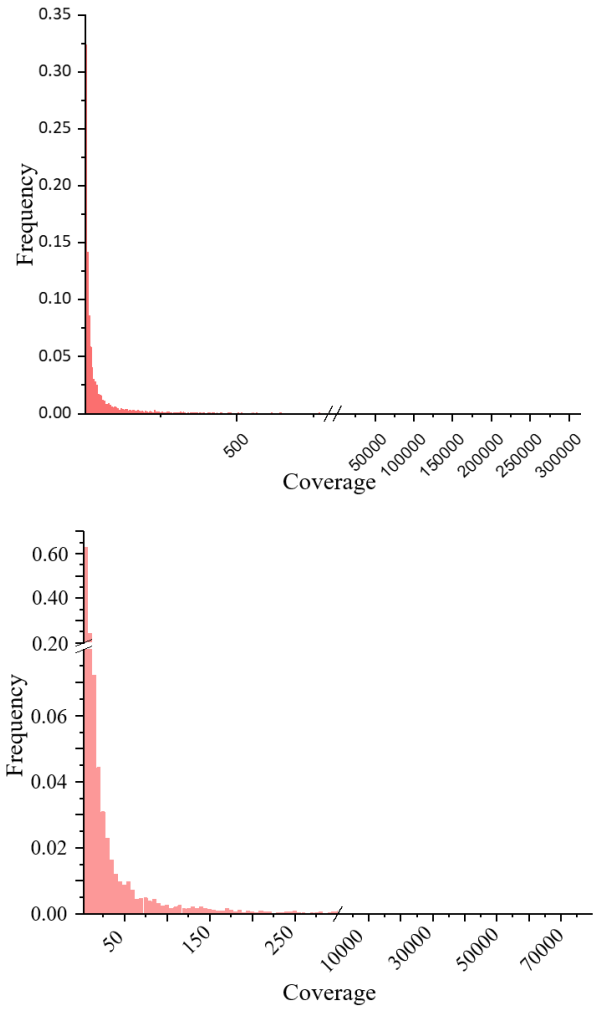
(a) Workflow of 10 consecutive amplification of PCR. **(b)** 2% agarose results of PCR products**.** Lane M: Trans2K Plus II DNA Maker; lane 1: PCR #1; lane 2: PCR #2; lane 3: PCR #3; lane 4: PCR #4; lane 5: PCR #5; lane 6: PCR #6; lane 7: PCR #7; lane 8: PCR #8lane 9: PCR #9; lane 10: PCR #10.



Supplementary Figure 52

The distribution of the number of reads per each given sequence of #5 PCR.

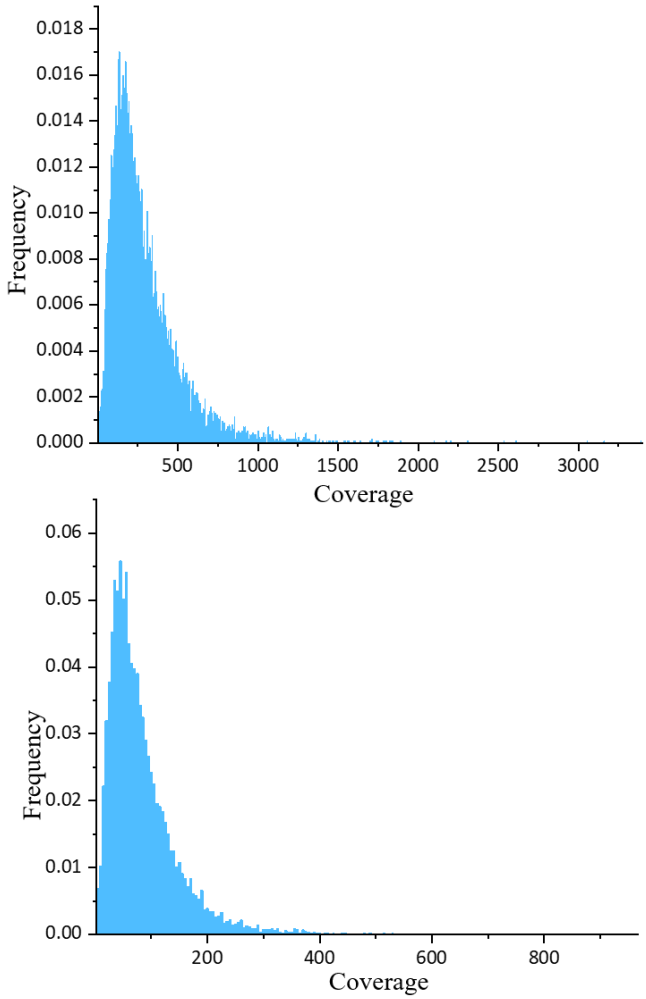
The distribution of the number of reads per each given sequence of total sequenced reads (the upper portion) and per million sequenced reads (the lower part) of PCR-amplified the oligo pool (Pool 1, Twist Bioscience).



Supplementary Figure 53

The distribution of the number of reads per each given sequence of #10 PCR.

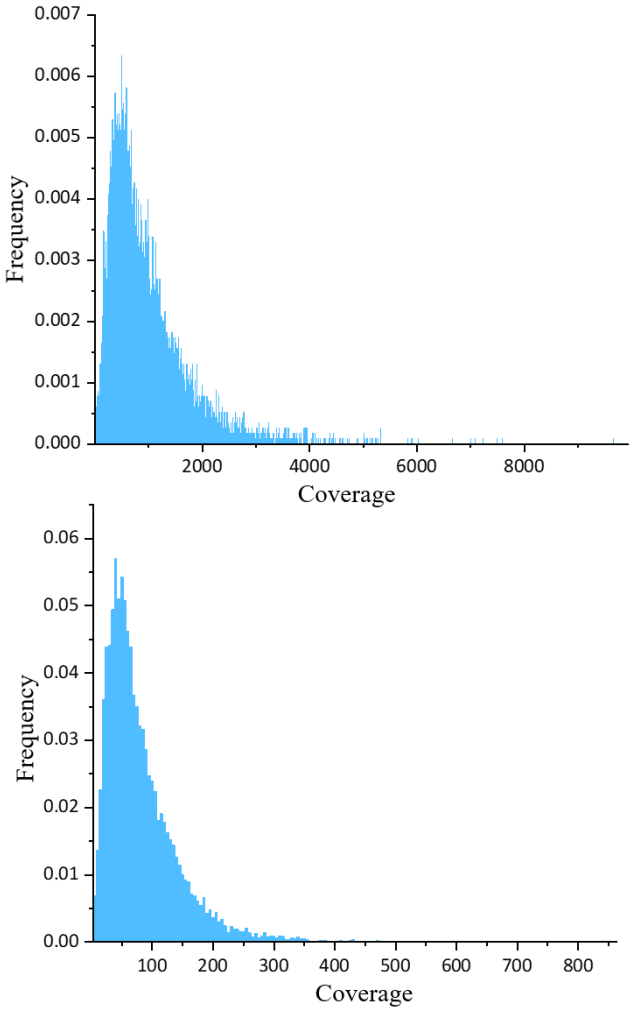
The distribution of the number of reads per each given sequence of total sequenced reads (the upper portion) and per million sequenced reads (the lower part) of PCR-amplified the oligo pool (Pool 1, Twist Bioscience).



Supplementary Figure 54

The distribution of the number of reads per each given sequence of #5 iDR.

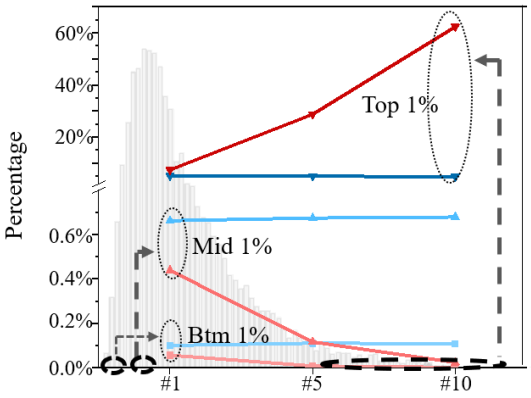
The distribution of the number of reads per each given sequence of total sequenced reads (the upper portion) and per million sequenced reads (the lower part) of PCR-amplified the oligo pool (Pool 1, Twist Bioscience).



Supplementary Figure 55

The distribution of the number of reads per each given sequence of #10 iDR.

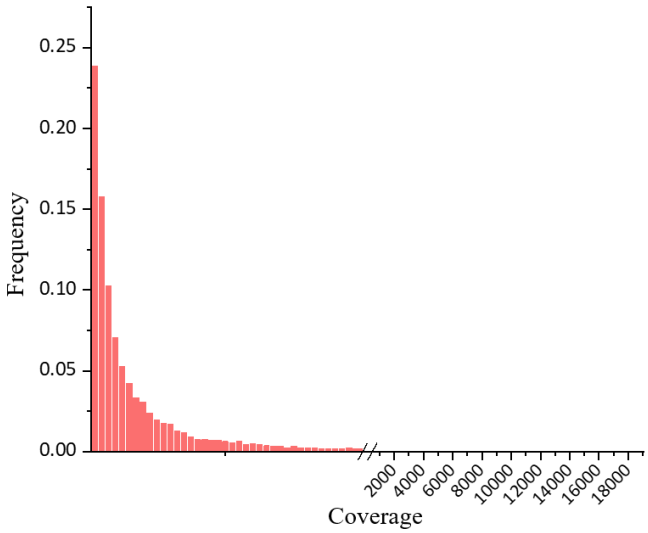
The distribution of the number of reads per each given sequence of total sequenced reads (the upper portion) and per million sequenced reads (the lower part) of PCR-amplified the oligo pool (Pool 1, Twist Bioscience).



Supplementary Figure 56

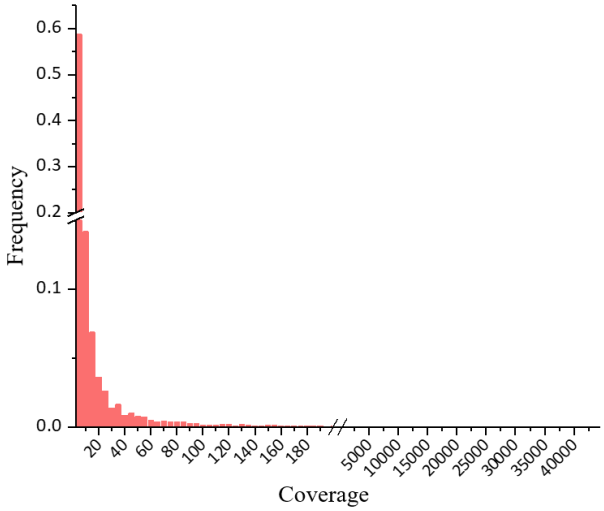
The percentage of some certain sequences (1% of total number) with the bottom, middle and top coverage in PCR #1 and iDR #1 was investigated in ten serial PCR and iDR separately (Supplementary Note 9).

The frequency of 116 oligos (1% of the total number, dark red) with the top coverage rose to 62.3% (#10 PCR), compared with 7.4% (#1 PCR). But, the frequency of 116 strands (1% of the total number) with the middle (Pink) and bottom (light pink) coverage was down to 0.011% and 0.0002% separately in #10 PCR. However, the frequency of 116 oligos whatever the coverage was had remained stable in iDR (blue-colored items).



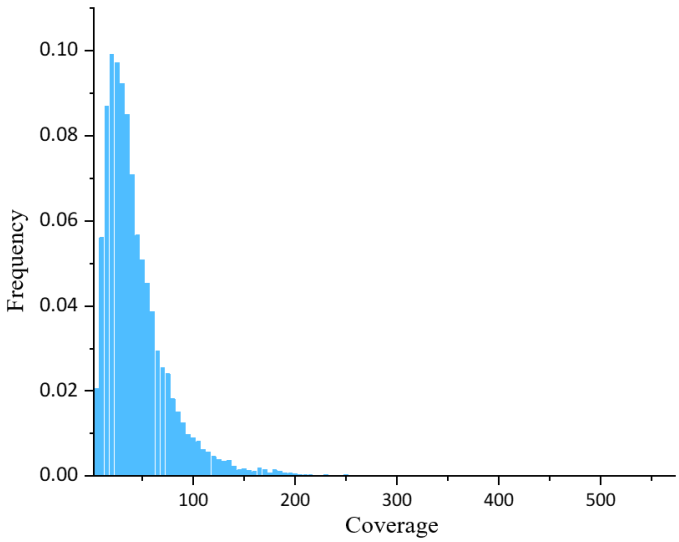
Supplementary Figure 57

The distribution of the number of M1G1 reads per each given sequence of #5 PCR amplified the oligo pool (Pool 1, Twist Bioscience).



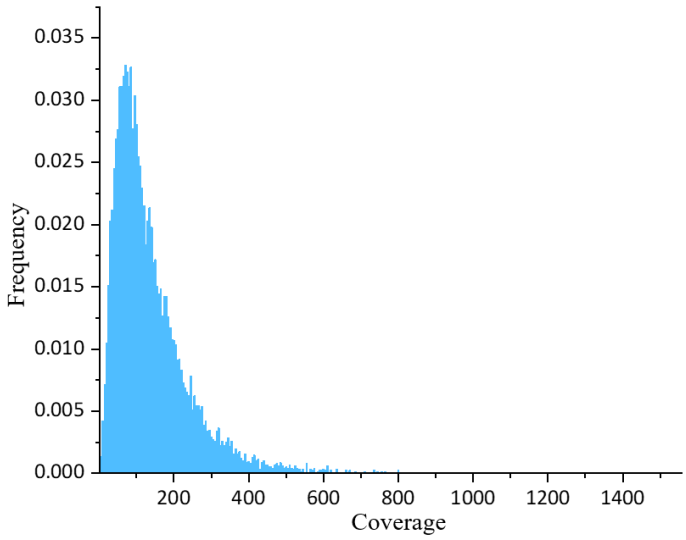
Supplementary Figure 58

The distribution of the number of M1G1 reads per each given sequence of #10 PCR amplified the oligo pool (Pool 1, Twist Bioscience).



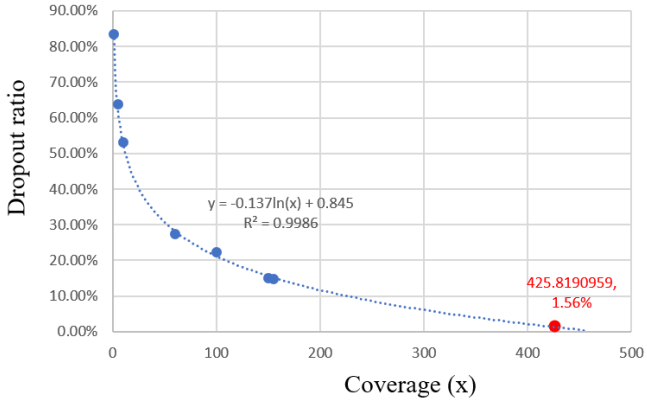
Supplementary Figure 59

The distribution of the number of M1G1 reads per each given sequence of #5 iDR amplified the oligo pool (Pool 1, Twist Bioscience).



Supplementary Figure 60

The distribution of the number of M1G1 reads per each given sequence of #5 PCR amplified the oligo pool (Pool 1, Twist Bioscience).



Supplementary Figure 61

The coverage – dropout ratio curve of #10 PCR**.**

The dropout ratio is 1.56%, the limit of our decoder, at an average coverage of 426 x (red dot).

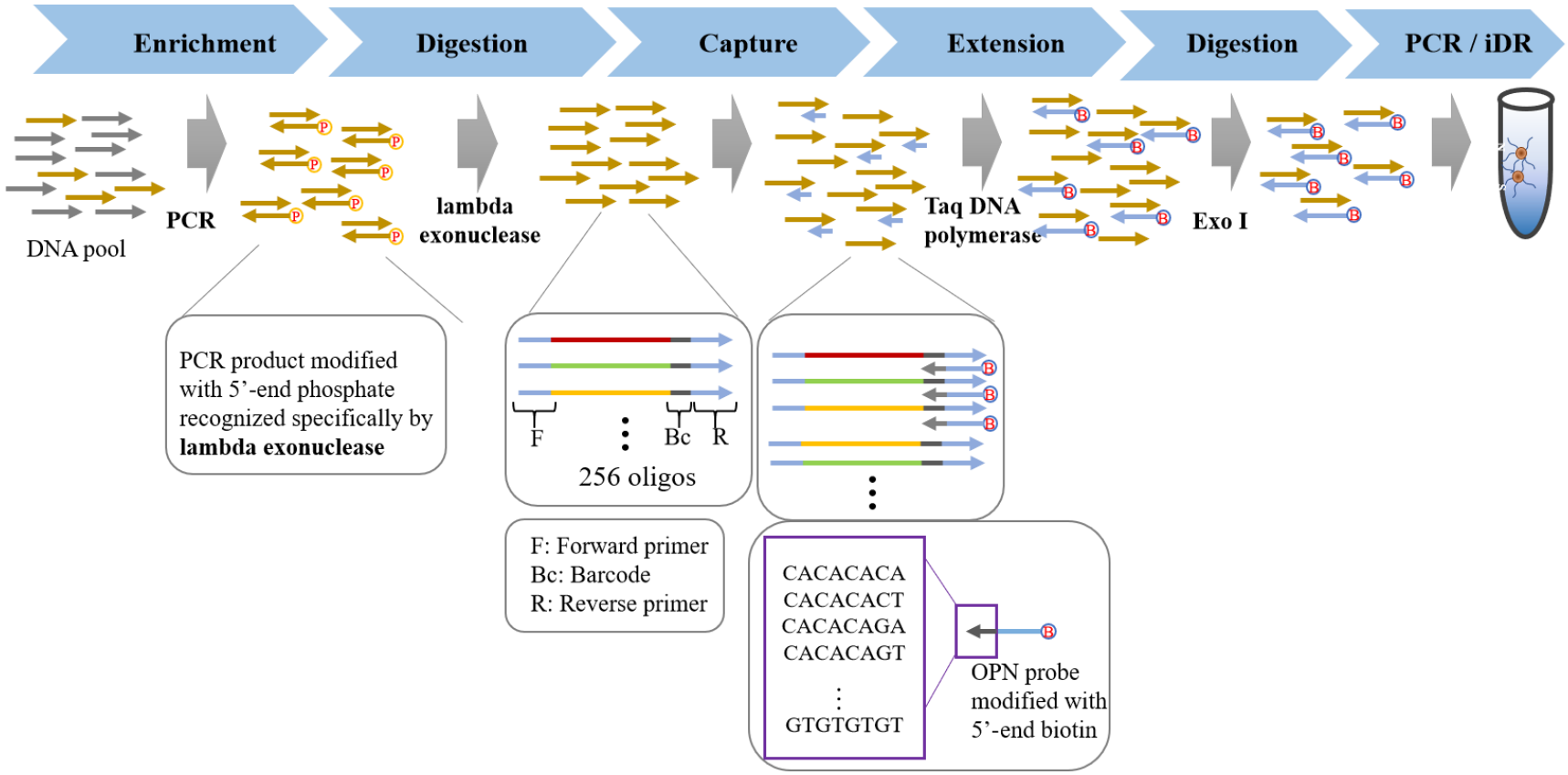
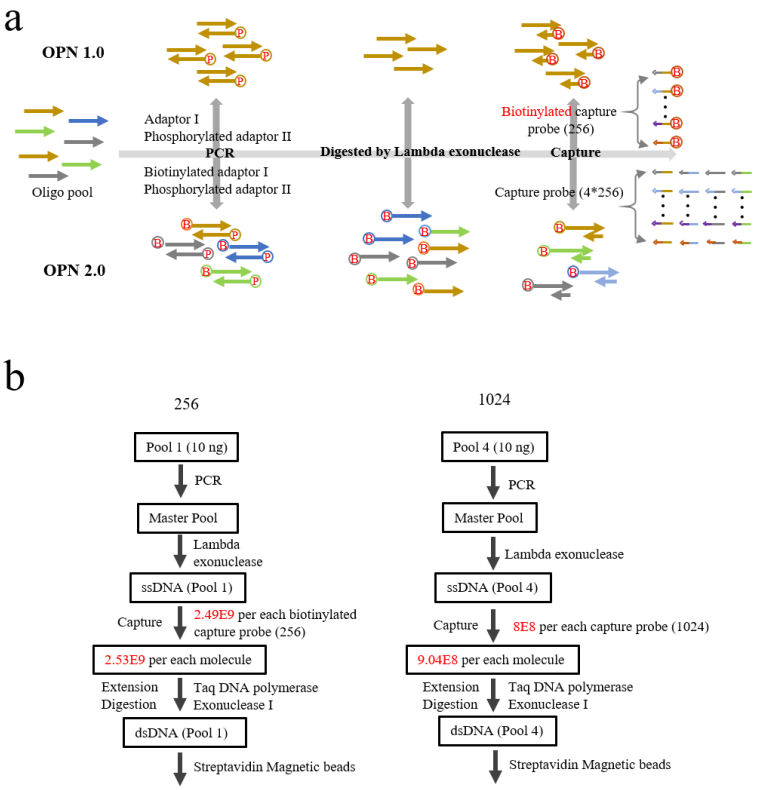


Figure S62

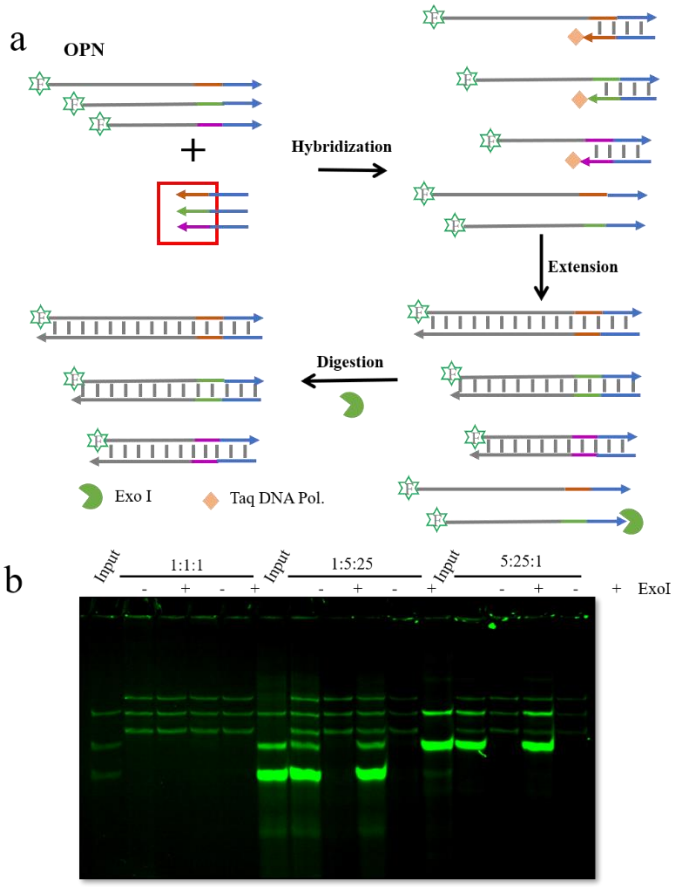
Workflow of single-stranded oligo pool amplification (SOA) and oligo pool normalizing (OPN).



Supplementary Figure 63

A comparison of Oligo Pool Normalizing (OPN) 1.0 and 2.0 and the workflow of OPN 1.0 and 2.0 in detail.

(a) Improved the process of oligo pool normalizing. **(b)** The workflow of OPN 1.0 and OPN 2.0 containing some specific parameter in our experiments.

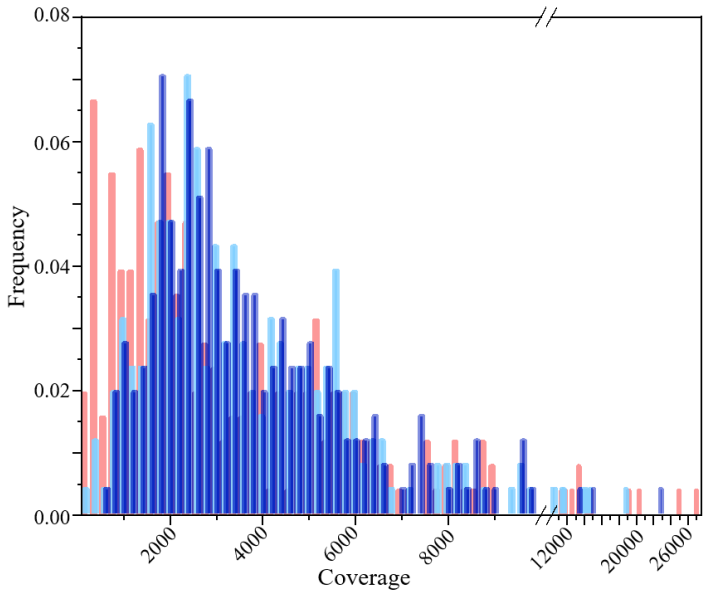


Supplementary Figure 64

Fluorescence PAGE to visualize the uniformity after the process of OPN through three oligos of different length bearing 3' FAM-label.

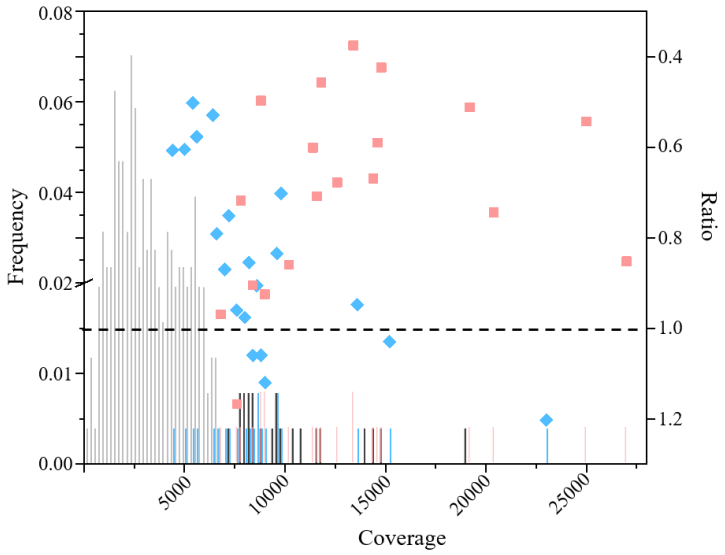
(a) Workflow of characterizing OPN probe using three FAM-labeled oligos of different lengths.

(b) 12% native PAGE results of OPN products. Input: the initial quantity of oligos.



Supplementary Figure 65

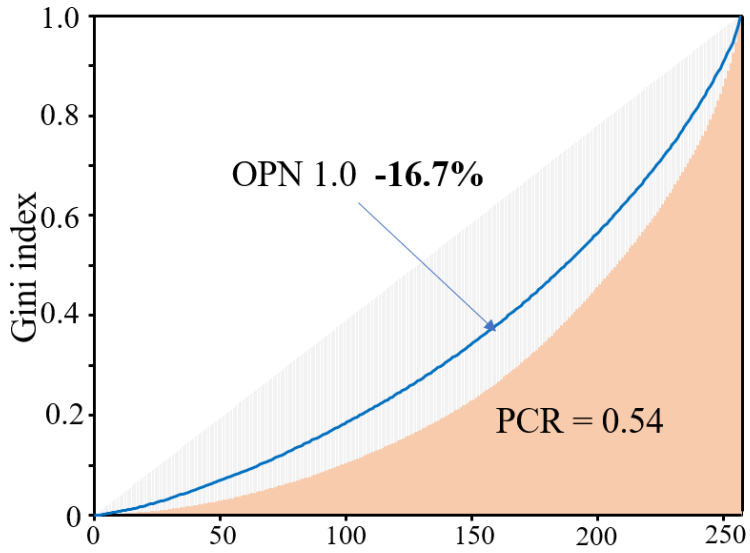
The distribution of the number of reads per each given sequence per million sequenced reads of PCR-amplified (red), iDR-amplified (cambridge blue) and OPN-iDR-amplified (dark blue) the oligo pool (Pool 1-OPN), normalized to 1 million sequenced reads.



Supplementary Figure 66

The coverage ratio of the sequences with a large coverage before and after OPN.

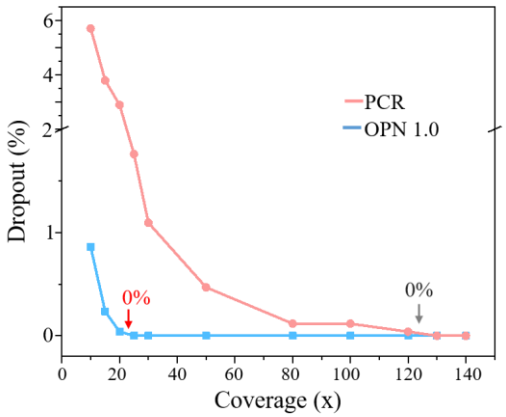
The distribution of the number of reads per each given sequence per million sequenced reads of iDR-amplified (gray) the oligo pool and 26 sequences with a large coverage of the given sequences are labeled black. The distribution of the number of reads per each given 26 sequences per million sequenced reads of PCR-amplified (red) and OPN-iDR (blue) the oligo pool (Pool 1- OPN). Blue diamond represents the coverage ratio of OPN-iDR to iDR; red square represents the coverage ratio of OPN-iDR to PCR.



Supplementary Figure 67

Lorenz curve showing the cumulative concentration distribution of the OPN-iDR products from grossly biased oligo pool.

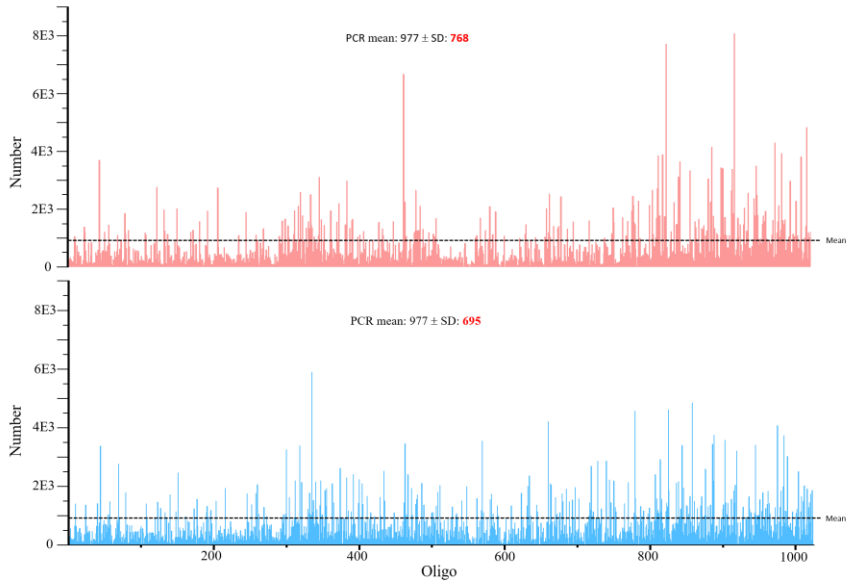
From this figure, we determined that the OPN process reduces the Gini inequality coefficient from 0.54 to 0.45.



Supplementary Figure 68

Coverage depth of random sequenced reads was plotted to its dropout rate.

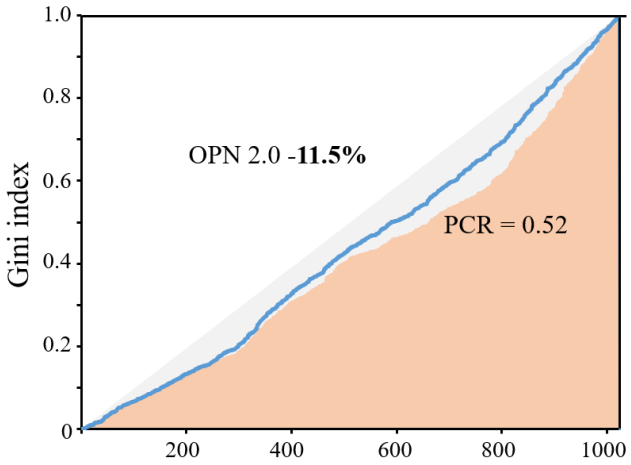
For 10x coverage depth, the dropout rate for PCR is 5.70% and 0.86% for OPN 1.0-iDR. The dropout rate became 0 at coverage of about 20 (red arrow) for OPN 1.0 iDR and 120 (gray arrow) for PCR.



Supplementary Figure 69

The copy number per oligo before (red) and after (blue) OPN 2.0 per million the noisy reads.

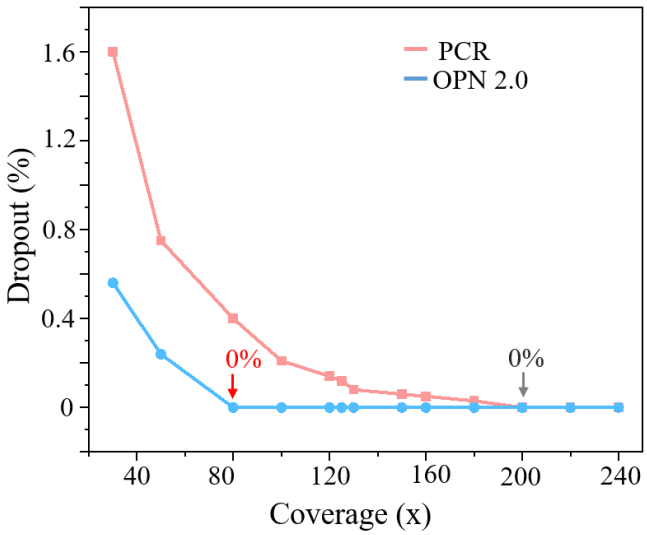
The standard deviation of coverage which the sequenced oligo pool did not proceed the process of OPN 2.0 was larger than that the sequenced oligo pool was carried out the process of OPN.



Supplementary Figure 70

Lorenz curve showing the cumulative concentration distribution of the OPN 2.0-iDR products from grossly biased oligo pool (1024).

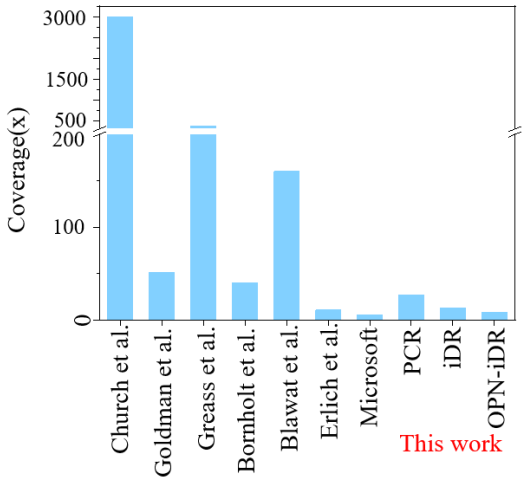
From this figure, we determined that the OPN process reduces the Gini inequality coefficient from 0.52 to 0.46.



Supplementary Figure 71

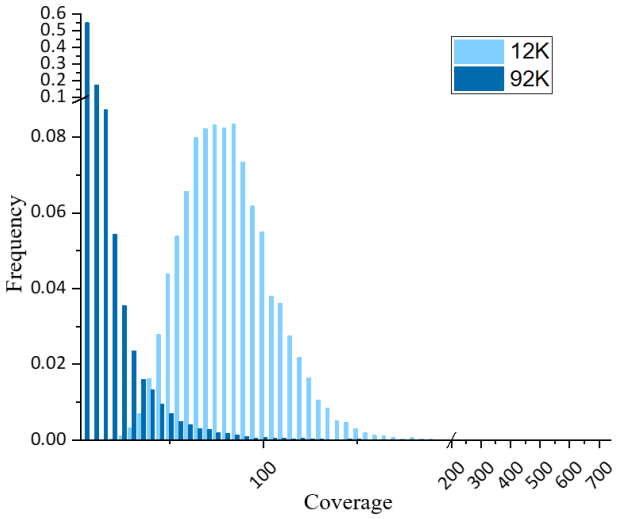
Coverage depth of random sequenced reads was plotted to its dropout rate.

The dropout rate became 0 at coverage of about 80 (red arrow) for OPN 1.0 iDR and 200 (gray arrow) for PCR.



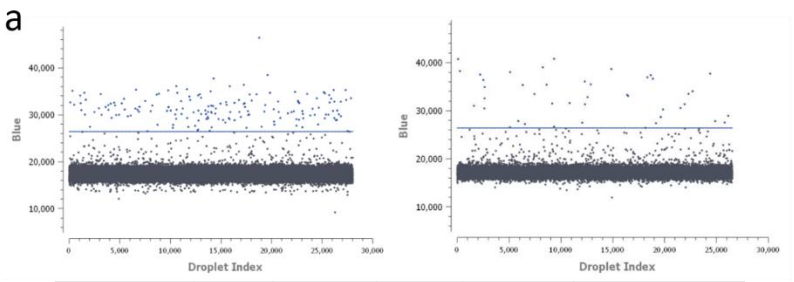
Supplementary Figure 72

Minimum coverages for decoding compared with prior work.

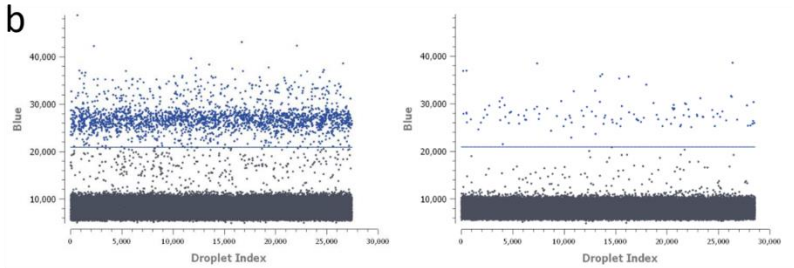


Supplementary Figure 73

The distribution of the number of reads per each given sequence per million sequenced reads of iDR-amplified the oligo pool (Pool 2, Twist Bioscience, light blue) and oligo pool (Pool 3, CustomArray, dark blue).



Sample	Total number	FAM (cp/μl)	"positive"	uncertainty	The concentration of S1 (cp/uL)
S1	28028	10.3	168	15%	$1.819*10^3$
S1(10^{-1} dilution)	26522	2.77	43	30%	$4.893*10^3$

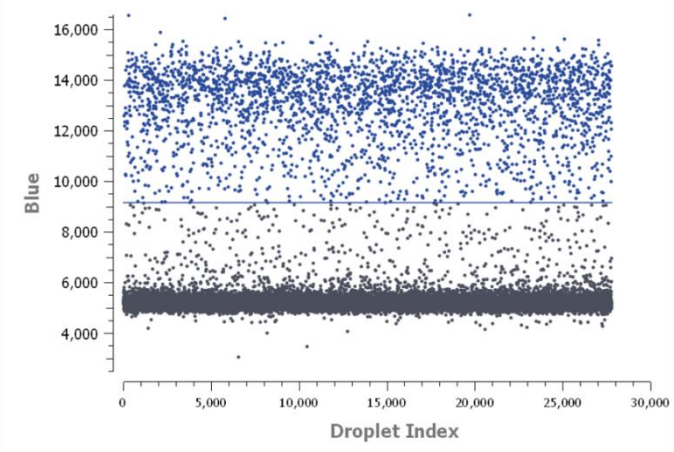


Sample	Total number	FAM (cp/μl)	"positive"	uncertainty	The concentration of S2 (cp/uL)
S2	27446	141.7	2186	4%	$4*10^4$
S2(10^{-1} dilution)	28558	7.01	117	18%	$1.5334*10^4$

Supplementary Figure 74

The number of molecules determined by digital PCR.

(a). The result of oligo pool (Pool 3, CustomArray, 170 nt). Sample without dilution on the left; Sample with being diluted 10-folds on the right. **(b).** The result of oligo pool (Pool 1, Twist Bioscience, 180 nt) in the bottom part. Sample without dilution on the left; Sample with being diluted 10-folds on the right. The average copy number of oligo pool 3 is an order of magnitude less than that in oligo pool 1 when the mass of DNA of both oligo pool was 7 ng. The results indicated that the quality of oligo pool from different synthesis platform is different.

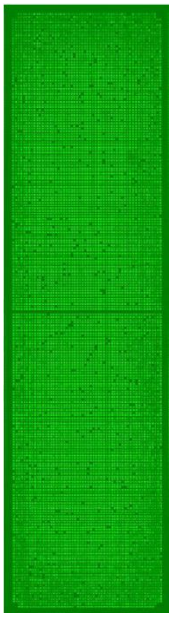


Sample	Total number	FAM (cp/μl)	"positive"	uncertainty	The concentration of S1 (cp/uL)
S1(dilution 10^3)	27772	180.6	2788	4%%	$2.26*10^6$

Supplementary Figure 75

The number of molecules of oligo pool (1024) determined by digital PCR.

The result of oligo pool (Pool 4, CustomArray). Sample with being diluted 10^3-folds. The average copy number of oligo pool 4 is $2.26*10^6$ when the mass of DNA of both oligo pool was 7 ng.



Supplementary Figure 76

The image of quality control of oligo pool 4 (1024).

D.Y. Zhang et al.*		Ours	
Component	Cost	Component	Cost
Capture probe (256 sequences)	$5,026.60	Capture probe (OPN 1.0)	$5,026.60
		Capture probe (OPN 2.0)	$1,290.97
Oligo pool (256 precursors)	$3,041.93	Oligo pool (12K)	$1,956.27
USER Enzyme (2U)	$3.23	Lambda exonuclease (7.5U)	$0.58
		Exonuclease I (5U)	$0.13
Total cost	$8,071.76	Total cost (OPN 1.0)	$6,983.58
		Total cost (OPN 2.0)	$3,247.95
Total base	17,920 nt	Total base (OPN 1.0)	46,080 nt
		Total base (OPN 2.0)	153,600 nt
Cost per each base	$0.450	Cost per each base (OPN 1.0)	$0.152
		Cost per each base (OPN 2.0)	$0.021

Supplementary Figure 77

The cost of oligo pool normalizing using OPN 1.0 and OPN 2.0 is inexpensive, compared to SNOP[1].

Supplementary Table 1. Sequences were avoided in the process of encoding.

Serial number	Sequences
1	GCTCTTC
2	GAAGAGC
3	GCTGAGG
4	CCTCAGC
5	CCTAAGC
6	GCTTAGG
7	CCTTAGC
8	GCTAAGG
9	CCTGAGC
10	GCTCAGG
11	CCTCAGT
12	ACTGAGG
13	CCTCAGG
14	GCGGCCGC
15	CTTAAAGCGCT
16	AGATAG
17	TGTTGG
18	GAGCTG
19	AGTCTG

Supplementary Table 2. Primer sequences used.

Name	Sequence (5'→3')	For oligo pool
Forward primer	*GTCCCGCTCATGCATCACCTA**CC***TCAGCTCAACTCACT	Pool 1
Adaptor 2	TCCACGACGATCAGACT	
Adaptor 2-1	Biotin-AAAAATCCACGACGATCAGACT	
Adaptor 2	AGCGCTTTAAGCCAACA	Pool 1-OPN
Adaptor 2-1	Biotin-AAAAAAGCGCTTTAAGCCAACA	
Adaptor 2-2	Phosphate-AGCGCTTTAAGCCAACA	
Adaptor 1	C*T*A*CTCCCACTCGTCTATCT	
probe	FAM-AGATCAATTAATACGATACCTGCGTTT	Ligation
splint	CTCGGAAGAGCTGAAAACGCAGGTATCG	
T1	GTCGCTAACAGAGTAACCTCCTCAGCTCTTCCGAGTCGGC AGCACTGCATAATTCTCTTACTGTCATGCCATCCGTAAGAT GCTTTCTGTGACTGGTGAGTACTCAACCAAGTCATTCTGAG AATAGTGTATGCGGCGACCGAGTTGCTCTTGCCCGGCGTC AATACGGGATAATACCGCGCCACATAGCAGAACTTATGAG TGGAGGTGTAAAGTG	
Forward primer	*TGCATCACCTA**CC***TCAGC	Pool 2
Adaptor 2	TCCACGACGATCAGACT	
Adaptor 2-1	Biotin-AAAAATCCACGACGATCAGACT	
Forward primer	*GTCGCCTTCT**CC***TCAGC	Pool 3
Adaptor 2	AGCGCTTTAAGCCAACA	
Adaptor 2-1	Biotin-AAAAAAGCGCTTTAAGCCAACA	
Forward primer	Biotin-AAAAACACTTTACACCTCCACTCAT	Pool 4
Adaptor 2-1	Phosphate-TCACCATCCACTCTAAACAC	
Adaptor 2-2	Phosphate-CTTCCGACCACTATACCTCT	
Adaptor 2-3	Phosphate-ACTCCCACTCACCTATATCC	
Adaptor 2-4	Phosphate-CTACTCCCACTACTACCACA	

Note: Forward primer were comprised of Adaptor I marked with italic and recognition sequences of nickase marked with bold fonts. Adaptor 1-1 was phosphorothioate modified, marked with '*', to prevent 5'-exonucleolytic degradation by Lambda exonuclease.

SUPPLEMENTARY INFORMATION

Supplementary Note 1. BASIC code strategies.

Distributed Storage Systems in application to DNA storage. DNA coding is a new type of distributed storage system. In a distributed storage system, data redundancy is the most basic strategy for ensuring system reliability and improving data availability. By storing multiple instances of the same data file with different nodes to ensure data availability, even some of the data is unavailable, the remaining nodes can still reconstruct the original data. Redundancy strategies must consider two points: firstly, how to create redundant data, and secondly, how to reconstruct data when some nodes fail. Currently, widely used redundancy strategies are replication and erasure coding. Replication distributes multiple copies of a file to different nodes in the system. As long as one of these copies is valid, the whole file can be obtained. This method has high reading and writing efficiency, but it has low storage utilization and is not suitable for DNA coding. Erasure Coding is another important redundancy strategy. (N, K) erasure code matrix divides an original file of size M into K blocks, each block with size of M/K; then the K block files are encoded into N code blocks and distributed to N nodes. The original file can be reconstructed from any K code blocks in the N code blocks. Erasure coding requires less storage than replication, whereas the calculation is relatively complex. BASIC code is a kind of distributed erasure code designed for DNA coding, aiming at maximizing storage utilization and effectively guaranteeing the reliability of the storage system.

This section briefly introduces the binary cyclic code. Let p be a prime number greater than N and the ring R_p is defined as $R_p \coloneqq \mathbb{F}_2[z]/(1+2^p)$, where the element $\sum_{i=0}^{p-1} a_i z^i$ in R_p is called polynomial, and the vector $(a_0, a_1, \dots, a_{p-1}) \in \mathbb{F}_2^p$ is a codeword of

polynomial $\sum_{i=0}^{p-1} a_i z^i$. A binary cyclic code of length p is a subset of additions and z-multiplier closures defined in R_p, where addition is an XOR operation and z multiplication is a cyclic right shift operation.

In this paper, we consider the parity code $\mathcal{C}_p$, which represents the set of polynomials for which all non-zero coefficient entries in R_p are even. $\mathcal{C}_p$ is formalized as:

$$\mathcal{C}_p \coloneqq \{\sum_{i=0}^{p-1} a_i z^i \in R_p | a_0 + a_1 + \cdots + a_{p-1} = 0\}$$
$$= \{a(z) \in R_p | a(z) \equiv 0 \; mod \; (1+z)\}$$
$$= \{a(z)(1+z) | a(z) \in R_p\}$$

The coefficient of the highest term of the element a(z) in $\mathcal{C}_p$ is the sum of the former m-1 coefficients, i.e., $a_{p-1} = \sum_{i=0}^{p-2} a_i$. It can be verified that $\mathcal{C}_p$ satisfies the addition and z multiplication closure. Because the operations in the domain $\mathcal{C}_p$ are only XOR and the right-shifted loop, it can be well applied to design the encoding system.

The addition defined in $\mathcal{C}_p$ is an XOR operation, which will not be described in detail here. The z-multiplication operation in $\mathcal{C}_p$ is defined as $\mathrm{z}_p: R_p \times C_p \to C_p$. $\mathrm{z}_p(2^i, \mathrm{c}) = 2^i * c \; mod(2^p + 1)$, that is, the loop shifts right by i bits. For example, the codeword is $\mathrm{z}_3(2^1, 101) = 011;\ \mathrm{z}_3(2^1 + 2^0, 101) = \mathrm{z}_3(2^1, 101) + \mathrm{z}_3(2^0, 101) = 011 + 101 = 110.$

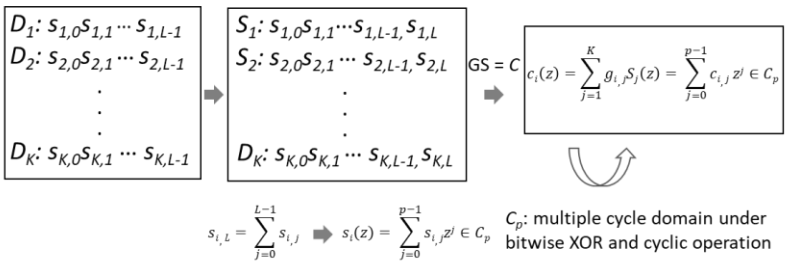
Vandermonde matrix is $\mathrm{V} = [\mathrm{v}_{i,j}]_{n \times k}$, where $\mathrm{v}_{i,j} = \alpha_j^{i-1}$. The determinant of an N-order Vandermonde square matrix can be represented as $\det(\mathrm{V}) = \prod_{1 \le i < j \le n}(\alpha_j - \alpha_i)$. When α_i are not the same, $\det(\mathrm{V})$ is not zero. The Vandermonde matrix has many interesting features. The most important thing here is that the sub-polynomials formed by any row and column are invertible. The Vandermonde matrix and its transformation matrix ensure that the encoding

799 data can be decoded. For convenience, note that $\mathrm{v}_i = (\mathrm{v}_{i,1}, \mathrm{v}_{i,2}, \ldots, \mathrm{v}_{i,n}) =$
800 $(\alpha_1^{i-1}, \alpha_2^{i-1}, \ldots, \alpha_n^{i-1})$ is the ith row vector of V.

801 **Constraints in encoding DNA digital information.** GC content was restricted in the range
802 between 45% and 55%. Meanwhile, long homopolymers (i.e., AAAA, TTTT, GGGG, CCCC)
803 were dropped. More than 6-bp self-sequence complementarity and 10bp inter-sequence
804 complementarity were avoided. Certain sequences were circumvented (Supplementary Table
805 1).

806 **The distributed storage system coding strategy.** The goal was to transform the input file to
807 DNA sequence reads with biochemical constraints. DNA basic code should enable error-
808 detection, error-correction and full recovery. There were two key steps: (a) erasure coding, (b)
809 RS coding. Since the sequence reads needed to satisfy the biochemical constraints, both
810 processes included the step of filtering the sequences.

811 **Erasure coding.** The algorithm divided the file into non-overlapping groups of length K*L
812 bits, each group containing K binary data of length L. In the subsequent encoding step, the
813 algorithm processed the data in groups. K and L were arbitrarily defined. Here we used K=252,
814 L=256 bits (32 bytes) because this parameter setting was compatible with the standard
815 computing environment and was within the capabilities of actual manufacturing. The BASIC
816 encoding was shown as follows.



For each group, we proposed BASIC code to encode the K pieces of data one by one. The BASIC code used codeword as the encoding and decoding unit, a codeword was a binary data of L bits. K codewords of length L were encoded into N codewords by polynomial matrix operations on the polynomial cycle domain C_p, where N>K, N-K was the maximum number of error sequences per group. Actually, N was 256, which meant each group of data allowed a maximum of 4 reads to be lost or corrupted. For simplicity of discussion, a codewordis represented as D_i or C_i. The input data was regarded as a vector $\mathrm{D} = (\mathrm{D}_1, \mathrm{D}_2, \ldots, \mathrm{D}_K)$, and the encoded data was a vector $\mathrm{C} = (\mathrm{C}_1, C_2, \ldots, C_N)$. The generation matrix G is a matrix defined on R_p, which was an improvement of the Vandermonde matrix. The reason for using the Vandermonde matrix was that the BASIC data recovery algorithm required the encoding matrix to be reversible by any n*n submatrix. The specific process was as follows:

1) In order to ensure that the obtained $s_i(z)$ was in C_p, for a set of K*L data $\mathrm{D} = (\mathrm{D}_1, \mathrm{D}_2, \ldots, \mathrm{D}_K)$, first added a parity bit at the end of each sequence read. The parity bit was calculated as $s_{i,p-1} = \sum_{j=0}^{p-2} s_{i,j}$ (where the addition is an exclusive-OR operation). After that, the codeword was $\mathrm{S} = (S_1, S_2, \ldots, S_K)$. And p was 257 here.

2) Initialize the generation matrix G as follows:

$$G = \begin{bmatrix} 2^0 & 2^1 & \cdots & 2^{K-1} \\ 2^{0\times 2} & 2^{1\times 2} & & 2^{(K-1)\times 2} \\ \vdots & & \ddots & \vdots \\ 2^{0\times N} & 2^{1\times N} & \cdots & 2^{(K-1)\times N} \end{bmatrix} mod\ (2^p + 1) = [g_{i,j} = 2^{(j-1)\times i\ mod\ p}]_{N\times K}$$

First, a constant S was added to all codewords, because zero-by-zero multiplication was still zero, which could cause the algorithm to not terminate. Next, initialized C_i, got $C_i \leftarrow g_i S^T$,$g_i = (g_{i,1}, g_{i,2}, \ldots, g_{i,K})$ was the i-th row of G. Finally, for any C_i, the last parity bit was removed and its first L bits were saved.

RS Encoding. In order to ensure the accuracy of data storage in DNA, RS codes were used to increase error correction and repair capabilities. Here, for each sequence reads, 2 bytes were allocated for the RS codes, which could detect errors within 2 bytes and correct errors within 1 byte.

Filtering. Each codeword C*i* consisted of three parts: the sequence number of the group, the sequence number within the group, and the number of adjustments to G_i. The index part was allocated a total of 4 bytes, in which the first 8 bits represented the group address, the middle 8 bits represented the address within each group, and the last 16 bits represented the number of adjustments of G_i.

All sequence reads must meet the biochemical constraints, which meant that the sequence could not contain any avoidance sequences. For the sequences obtained after the erasure coding, we generated a number of random equal length sequences, so that the encoding sequences didn't contain any avoidance sequence after exclusive-OR operation with one of the random sequences. It could be proved by experiments that the upper 8 bits of the 16 bits, which represent the number of adjustments of G_i, were always 0. Thus, we used these bits to store

854 the sequence number of the random sequence that was XORed for each encoding sequence.
855 When decoding, it was only necessary to find a random sequence according to the serial number
856 for XOR. The same process was performed for the individual RS codes. A number of random
857 sequences with the same length of the RS code were generated for XOR operation. Since there
858 were no bits to store the sequence number of the RS XOR sequence, we built a mapping table
859 to store the RS code and its corresponding sequence number. It should be noted that all RS
860 codes in the mapping table were unique.

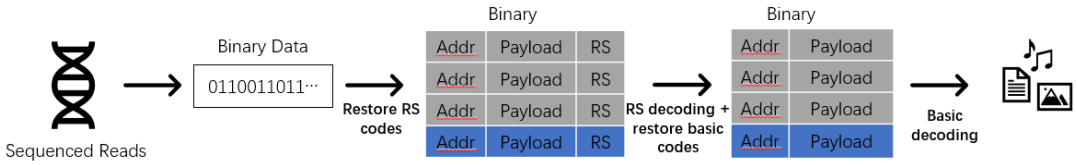
861 **Decoding.** The decoding process was reversed step by step according to the encoding process.
862 XOR operation was performed according to the mapping table to restore the RS code, and then
863 the RS code was used for error correction to ensure that each sequence was accurate. Restore
864 the BASIC code sequence. For each group of data, it was decoded according to the BASIC
865 decoding algorithm.

866 1) Find K lossless sequences for each group. A parity bit is added for each codeword,
867 which is denoted as $C' = (C'_1, C'_2, \dots, C'_K)$.

868 2) Construct a generation matrix G' according to the intragroup address and the
869 number of adjustments of G_i. Calculate G's inverse matrix $G'^{-1} = [f_{i,j}]_{K\times K}$.

870 3) Decode $D'_i = \sum_{j=1}^{k} f_{i,j} * C'_j$ according to the matrix G'^{-1}. Remove the last
871 parity bit and finally recover the original data.

872 The decoding processes are shown as follows.


Binary Data
0110011011···
Binary
Addr
Payload
RS
Restore RS codes
RS decoding + restore basic codes
Binary
Addr
Payload
Basic decoding
Sequenced Reads

873

874

875

Supplementary Note 2. Genome sequence were converted to binary information.

Human mitochondrial genome containing Chinese (16,570 bp), Italy (16,569 bp), Native American (16,570 bp), and South African (16,567 bp) were converted to binary data by a simple rule (Supplementary Fig. 4). Genome information comes from mtDB – Human Mitochondrial Genome Database (http://www.mtdb.igp.uu.se/sequences.php). And then the binary file was encoded to 12K oligo pool (Twist Bioscience).

The 1.08-Mbp *M. mycoides* JCVI-syn1.0 genome sequence (accession CP002027) were also converted and encoded to 92K oligo pool (CustomArray) (Supplementary Figs. 4, 9).

Supplementary Note 3. Sequencing on an Illumina Hiseq 4000 platform.

Sample collection and preparation. DNA degradation and contamination were monitored on 2% agarose gels. DNA purity was checked using the NanoPhotometer spectrophotometer (IMPLEN, CA, USA). DNA concentration was measured using Qubit DNA Assay Kit in Qubit 2.0 Flurometer (Life Technologies, CA, USA).

Library preparation for sequencing. A total amount of 700ng DNA per sample was used as input material for the DNA sample preparations. Sequencing libraries were generated using NEB Next® Ultra DNA Library Prep Kit for Illumina® (NEB, USA) following manufacturer's recommendations and index codes were added to attribute sequences to each sample. Briefly, the Chip DNA was purified using AMPure XP system (Beckman Coulter, Beverly, USA). After adenylation of 3' ends of DNA fragments, the NEB Next Adaptor with hairpin loop structure were ligated to prepare for hybridization. Then electrophoresis was used to select DNA fragments specified in length. 3 μL USER Enzyme (NEB, USA) was used with size-selected, adaptor-ligated DNA at 37°C for 15 min. At last, the products were purified (AMPure XP system) and library quality was assessed on the Agilent Bioanalyzer 2100 system.

Clustering and sequencing. The clustering of the index-coded samples was performed on a cBot Cluster Generation System using HiSeq 4000 PE Cluster Kit (Illumina) according to the manufacturer's instructions. After cluster generation, the library preparations were sequenced on an Illumina Hiseq 4000 platform and 150bp paired-end reads were generated.

Supplementary Note 4. Optimizing the reaction conditions for iDR.

To obtain the highest yield of iDR product, we investigated the effects of several factors including the combination of nickase and polymerase with strand displacement activity, nickase to polymerase ratio, reaction buffer, and additives et. al on iDR.

Effects of combination of nickase and polymerase with strand displacement activity on iDR. We chose these nickases (e.g. Nt.BstNBI, Nt.BspQI, and Nt.BbvCI) of which the length of recognition sequences is more than 5 nt for encoding in convenience (Supplementary Fig. 10). Meanwhile, we used the polymerases with strand displacement activity such as Bst 2.0 WarmStart DNA polymerase (Bst), Vent DNA polymerase (Vent), Bsu DNA polymerase (Bsu), Klenow fragment (3'→5' exo-, KF) and Phi29 DNA polymerase (Phi29) (Supplementary Fig. 11). Due to the different optimal temperature of nickases and polymerases, we chose the combination of Nt.BspQI and Bst/Vent which can be carried out at 55℃ (Supplemnetary Fig. 12). And when performed at 37℃, the Nt.BspQI and Bsu/KF, Nt.BbvCI and Bsu/KF were combined (Supplementary Fig. 13). In addition, we applied Nt.BbvCI and Phi29 to the iDR. iDR product was not observed on Native PAGE gel (Supplementary Fig. 14). From the Native PAGE gels results and taking into account the reaction temperature, we adopted the combination of Nt.BspQI and Bsu DNA polymerase/Klenow fragment (3'→5' exo-), Nt.BbvCI and Klenow fragment (3'→5' exo-) in subsequent iDR.

Effects of the concentration ratio of nickase and polymerase on iDR. We assessed the iDR product in the condition of different concentration ratio of nickase and polymerase (Supplementary Figs. 15-17). Each reaction was carried out in 25 μL total volume. All reaction

components but the nickase and polymerase were assembled. Then nickase and polymerase were added to the tube respectively based on the concentration ratio designed. The reaction was then incubated at 37℃ for 30 min. The enzyme in iDR were inactive by heating 80℃ for 20 min when the DNA template for iDR were not attached to streptavidin magnetic beads. While the DNA template for iDR were attached to streptavidin magnetic beads, the supernatant mixture was removed and the proteinase was added into the mixture, then incubated at 50℃ for 30 min to inactive the nickase and polymerase.

The results showed that the yield of iDR product was increased with increasing amount of Nt.BspQI while the by-product was also increased. But the iDR product were slightly decreased as Bsu DNA polymerase was increased. We considered that the optimal concentration of Nt.BspQI and Bsu DNA Polymerase, Large Fragment were 0.5U/μL and 0.25U/μL separately. Although the by-product occurred, the product of iDR was largely increased (Supplementary Fig. 15). Bsesides, the yield of iDR product was slightly increased with increasing amount of Nt.BspQI. But the iDR product were unchanged as Bsu DNA polymerase was increased. We considered that the optimal concentration of Nt.BspQI and Klenow Fragment (3'→5' exo-) are 0.25 U/μL and 0.25 U/μL separately (Supplementary Fig. 16). The highest yield product of iDR was observed at 0.08 U/μL Nt.BbvCI and 0.16 U/μL Klenow fragment (3'→5' exo-). The concentration of Nt.BbvCI and KF was applied in subsequent assays (Supplementary Fig. 17).

Reaction buffer. Given the effect of the reaction buffer on Nt.BspQI and Klenow fragment (3'→5' exo-) activity based on product specification from NEB and previous results[2-4], we proceeded to customize our reaction buffer (Supplementary Fig. 18). The result displayed

that the higher iDR product was obtained when the reaction buffer includes 0.5x NEBuffer2 (25 mM NaCl, 5 mM Tris-HCl, 5 mM $MgCl_2$, 50 μg/ml BSA) and 0.5x 3.1 Buffer (50 mM NaCl, 25 mM Tris-HCl, 5 mM $MgCl_2$, 50 μg/ml BSA). Because the activity of Nt.BbvCI in NEBuffer 2.1 which is the optimal reaction buffer for Klenow fragment (3'→5' exo-) is 100%. Thus, NEBuffer 2.1 was used as the iDR reaction buffer when Nt.BbvCI and Klenow Fragment (3'→5' exo-) were applied to iDR amplification.

Additives. We also explored the effects of additives including DMSO and SSB protein (T4 Gene 32 Protein) on iDR as previous reported[5]. Each reaction was carried out in 10 μL total volume. All reaction components but the additive were assembled in 9.4 μL. Then 0.6 μL of 100% DMSO (6%) and 67 μM T4 gene 32 protein (4 μM) was added into a corresponding tube respectively. The result show that the iDR product was largely increased and the by-product was decreased and even disappeared with addition of T4 gene 32 protein (Supplementary Fig. 19). The ssDNA product of iDR appeared "ladder" band. This was probably because T4 gene 32 protein bound to single-strand DNA.

Effects of DNA fragment (as DNA template) length on iDR. We studied the effect of DNA fragment (as template for iDR) length on iDR (Supplementary Fig. 20). The iDR reaction using Nt.BspQI and Bsu DNA polymerase was not related to the length of DNA fragment. However, the iDR reaction using Nt.BstNBI and Bst DNA polymerase, large fragment had a preference for the DNA with less than 200bp.

Adaptor I length. We investigated the effect of adaptor I length ("primer", sequence region at the front of the recognition site of nickase) on iDR (Supplementary Fig. 21). The

adaptor I length of booster was 13nt which was enough stable for avoiding the adaptor I dissociating.

Effects of variation in recognition region on the nicking efficiency of nickase. We explored fidelity of Nt.BspQI and Nt.BbvCI by nick the DNA templates containing a base error at the recognition region (Supplementary Figs. 22-23). The iDR product was observed when GCTCTTC, recognition sequence of Nt.BspQI, was converted into GCTCTTA. It demonstrated that Nt.BspQI was not specifically to recognize GCTCTTC (Supplementary Fig. 22). Meanwhile, the iDR product was observed when CCTCAGC, recognition sequence of Nt.BspQI, was converted into CCTAAGC. It demonstrated that Nt.BspQI was not specifically to recognize CCTCAGC (Supplementary Fig. 23). Thus, GCTCTTN and CCTNAGC was evaded in the sequence we encoded.

Effects of DNA templates attached to the streptavidin magnetic beads on iDR. DNA templates modified with biotin and attached to the streptavidin magnetic beads had no influence on iDR reaction (Supplementary Fig. 24).

Note: DNA sequences used in the process of systematic optimization are in Supplementary table 3.

Supplementary Note 5. OPN probe design.

OPN probe comprises two parts, from 5' to 3': a universal sequence (Adaptor 2/2-1/2-2/2-3/2-4) and a barcode sequence.

Each barcode sequence was comprised of a number of commutative strong (C or G) and weak (T or A) nucleotides according to earlier report[1]. Here, the length of barcode was 8 nucleotides, corresponding to a total of $2^8 = 256$ barcode instances.

Universal sequences (namely primer) can be designed based on previous report[6]. Here, a universal sequence was used and there are 256 probes in OPN 1.0 (Supplementary Table 4). And four universal sequences were used and there are 1024 (4*256) probes in OPN 2.0 (Supplementary Table 5). According to a number of heuristic design criteria, we can obtain an estimated maximum library size of 14,000 20-mer primer pairs. Thus, an oligo pool containing up to 3 million oligos can be manipulated by using a combination of universal sequences and the barcodes (14,000*256 = 3,584,000).

Supplementary Note 6. Calculating logical density and redundancy for DNA storage systems.

Logical density calculated for previous studies is taken from Organick et al.[6] Logical redundancy for previous studies is taken from Anavy, L. at al.[7] Redundancy for Organick et al.[6] is taken from their report.

In this work, due to adjustable system parameter K and L, we take standard system parameter (K=252, L=256) as example. Here, we have used 10,752 DNA sequences (a part of DNA sequences in Pool 2) of length 200 with payload of length 152 to store 329KB yielding a net density of approximately 1.65 bits per base (1.25 bits per base including primers). According to our encoding strategy which allows a maximum of 4 DNA sequences to be lost or corrupted in each group (each group contains 256 DNA sequences), the redundancy can be calculated ideally as: 4/256=1.56%.

Supplementary Note 7. Energy consumption.

We assumed that the system was thermally insulated and there was no energy consumption during the state of constant temperature. Energy consumption $Q = c * v * (T_2-T_1)$ (c is the specific heat capacity of the solution; the volume of the system is v; T_1, T_2 is the initial and the final temperature respectively) when the reaction is exothermic. Energy consumption $Q = c * v * (T_1-T_2)$ while the reaction is endothermic.

We took 25 μL of water as an example. We supposed that the specific heat capacity is equivalent to water's (c, $c = 4.2 * 10^3$ J/(kg * ℃)) and the initial temperature (T_0) is 25℃. The thermocycling conditions of PCR were as follows: 5 min at 98℃; 10 cycles of: 30s at 98℃, 30s at 58℃, 10s at 72℃, followed by a 5 min.

25-98 $Q = 4.2 * 10^3 * 25 * 10^{-6} * (98 – 25) = 7.665$ J

98-58 $Q = 4.2 * 10^3 * 25 * 10^{-6} * (98 – 58) = 4.2$ J

58-72 $Q = 4.2 * 10^3 * 25 * 10^{-6} * (72 – 58) = 1.47$ J

72-98 $Q = 4.2 * 10^3 * 25 * 10^{-6} * (98 – 72) = 2.73$ J

10 cycles $Q_{total} = 7.665 + (4.2+1.47) * 10+2.73 * 9 = 88.9$ J

The thermocycling conditions of iDR was as follows: 30 min at 37℃.

25-37 $Q = 4.2 * 10^3 * 25 * 10^{-6} * (37 – 25) = 1.26$ J

To highlight the difference, we've doubled the energy consumption.

Supplementary Note 8. DNA decay caused by thermal condition.

Half-life of DNA extrapolate according to the Arrhenius Equation with activation energies of 155 ± 10 KJ mol^{-1} and compared to literature data on DNA stability in solution previously reported[8] (Supplementary Fig. 29).

Supplementary Note 9. The bioinformatic statistical analysis.

We stitched the reads pairs using PEAR[9] used for oligo copy distribution, error and dropout ratio analysis.

The sequenced reads were aligned with the given sequences (synthesized by Twist Bioscience and CustomArray) by BLAST. Here, the reads without error containing substitution and indel were defined as M0G0 reads, and the reads with an error including substitution or indel and with two errors including substitution and indel were defined as M1G1 reads whose coverage, number and DNA sequences can be obtained via M0G0_M1G1.pl (Supplementary Fig. 30). The M0G0 and M1G1 reads were considered as valid reads.

The coverage and number could be achieved by Valid_Coverage_Number.pl (Supplementary Fig. 29). The frequency was achieved via the number dividing by total number of given sequences. Then the distribution of number of reads per each given sequence was displayed (Supplementary Figs.36-37, 41-42, 46,57-60).

The distribution of the number of reads was displayed through analyzing coverage and number obtained by Obtain_Payload.pl (for obtaining valid DNA sequence namely payload), Cluster.pl and Sort.pl (Supplementary Figs. 30, 43-44). The frequency was obtained through the number dividing by the sum of these numbers.

The coverage of aligned sequences was sorted from small to large and numbered them in sequence. Top 30% of the serial number was selected and the frequency of these reads was calculated (Supplementary Fig. 38).

115 sequences (1% * number of oligos in file) with top, bottom, and middle (mode value)

coverage of #1 PCR and #1 iDR acted as a reference database. Per million valid sequences of both #5, #10 PCR and #5, #10 iDR were aligned with the chosen sequences. Then the percentage of reads was calculated via dividing by 1 million valid reads (Supplementary Fig. 56).

To depict the normality of oligo distribution, we used the piecewise function.

$$M = F(a + \triangle) - F(a - \triangle)$$

or

$$F(a) = \int_{a-\triangle}^{a+\triangle} f(x)\,dx$$

Here, F(x) is distribution function, f(x) is probability density function, a is the mean of oligo distribution, M is offset by defining $\triangle$ as the distance ($\triangle$ can be different values under the condition of $\triangle \leq a$). The area under the histogram is 1. In our study, $\triangle$ is 50.

Supplementary Note 10. The process of single-stranded oligo pool amplification (SOA) and OPN.

SOA process. The schematic of SOA process is illustrated (Supplementary Fig. 62). We used PCR to amplify an oligonucleotide library with specific barcodes sequences. Mix 20 ng of ssDNA pool (10 μL, only 256 of 11776 strands) with 2 μL of 100 μM of the adaptor 1 and 2 μL of 100 μM of the adaptor 2-2, 10 μL 5x Q5 reaction buffer, 0.2mM dNTPs, and 21.5 μL of DNase/RNase-free water. Thermocycling protocol were as follows: (1) 98℃ for 3 min, (2) 98℃ for 30 s, (3) 58℃ for 30 s, (4) 72℃ for 8 s, (5) go to step 2 30 times, (6) 72℃ for 5 min. The reaction was then purified according to the instructions in Eastep Gel and PCR Cleanup Kit. Then, PCR products were degraded from 5' phosphate groups to 3' direction by lambda exonuclease, thus conversion of linear double-stranded DNA to single-stranded DNA (ssDNA). The reaction was performed in 30 μL reaction mixtures including 3 μL 10x Lambda Exonuclease Reaction Buffer, 600 ng dsDNA, and 7.5U lambda exonuclease. The reaction was incubated at 37℃ for 90 min, then stopping it by adding EDTA to 10 mM. The mixture was purified by Eastep Gel and PCR Cleanup Kit. Eventually, 10% denaturing (7 mol/L urea) PAGE was used to analyze the degraded products. The gel was stained with SYBR Gold for 20 min. Further, gel band quantitation was used to assess the yield of ssDNA. Azurespot software was subsequently used to perform band detection, background subtraction and band quantitation. The concentration of ssDNA was defined as C = (the gray value of ssDNA) * (the quality of the standard sample) / ((the gray value of the standard sample) * (the volume of ssDNA)). Further, we could calculate the mean of ssDNA molecules (n_1) according to the quality (m) and the number of the ssDNA pool (N) (namely $n_1 = m * N_A / (M * N)$, N_A - Avogadro's number;

M – the relative molecular mass of ssDNA).

OPN process. The schematic of OPN process is illustrated (Supplementary Fig. 61). The oligo pool normalizing probes (OPN probes) were synthesized respectively (Supplementary Table 4 and Supplementary Table 5). Take 256 as an example, an equimolar mixture of OPN probes (256), which the number of each OPN probe was less than the mean of ssDNA molecules (n_1), was applied to capture the corresponding oligo separately. The sample without the process of OPN was mixed with adaptor 2-1. The number of adaptor 2-1 was equal to the mean of ssDNA molecules (n1) * 256. In the reaction, 20 μL of the ssDNA (1.0769 pmol of ssDNA pool) was mixed with 5.3 μL of OPN probes (4.134 attor mol of each probe) and 24.7 μL of hybridization buffer (10 mM Tris-EDTA, 0.5 M NaCl, and 0.05% Tween-20 (volume / volume). The mixture was denatured at 95℃ for 3 min and slowly cooled to 60℃ at a ramp of 0.1℃/s, following kept for 2 h at 60℃ using an Eppendorf Mastercycler instrument. Extension reaction was carried out in 65 μL reaction mixture. Before use, the pre-reaction mixture was mixed containing 6.5 μL of 10x EasyTaq Reaction Buffer, 5.2 μL of 2.5 mM of dNTPs, 2.5 U EasyTaq, and 2.3 μL DNase/RNase-water. To ensure temperature uniformity, the pre-reaction mixture was also pre-heated to 60℃ before addition to the ssDNA/OPN probe mixture. The resulting mixture was incubated for another 15 min at 60℃ to extend completely. Then, 5 U Exo I was added to the resulting mixture. And the resulting mixture was incubated at 37℃ for 3 h, then 80℃ for 20 min to inactive Exo I. The resulting mixture after being digested by Exo I was incubated for another 30 min at 37℃ at shaker to allow the biotin-streptavidin capture reaction to proceed. Magnet was applied to side of tube for approximately 30 seconds, then the supernatant was removed and discarded. Finally, the precipitate was washed by Wash/Binding

Buffer twice and resuspended as the template for iDR using 20 μL 0.5x TE Buffer.

Supplementary Note 11. Calculating error rate.

All sequenced reads were aligned with the actual reference sequences by basic sequence alignment program BLAST to screen out these reads with errors containing substitution, insertion, and deletion (henceforth referred to simply as “errors”) at the payload of individual sequences. And the number of reads with an error, two errors, three errors, ……, ten errors, more than ten errors in individual sequences were counted in detail by Mismatch_Analysis.pl and Gap_Analysis.pl and the frequency were calculated through the number of these reads dividing by the total number of noisy reads (Supplementary Figs. 31-35).

Supplementary Note 12. Minimum sequencing resource for perfect decoding.

Taking pool 1 as an example. 12K oligo pool is composed of 11520 oligos. We theoretically recovered the file at a coverage of 426 x in #10 PCR. We successfully retrieved the file at a coverage of 12 x in #10 iDR and 8 x in OPN-iDR. The percentage of valid reads are 48.58%, 92.21%, and 94.1% among corresponding noisy reads of #10 PCR, #10 iDR, and OPN-iDR respectively. Therefore, the total noisy reads which recovered the information with 100% accuracy required 11520*426/48.58% = 10,101,935, 11520*12/92.21% = 149,919; and 11520*8/94.1% = 97,938 of PCR, iDR, and OPN-iDR separately. The ratio of noisy reads needed for successful decoding of #10 PCR to #10 iDR was 67-folds, and the ratio of #10 PCR to OPN-iDR was 103-folds.

References

1. Pinto, A., Chen, S.X. & Zhang, D.Y. Simultaneous and stoichiometric purification of hundreds of oligonucleotides. *Nat Commun* **9**, 2467 (2018).
2. Walker, G.T. Empirical aspects of strand displacement amplification. *PCR Methods Appl* **3**, 1-6 (1993).
3. Zhou, X. & Xing, D. Amplified electrochemiluminescence detection of nucleic acids by hairpin probe-based isothermal amplification. *Analyst* **137**, 4188-4192 (2012).
4. Ehses, S., Ackermann, J. & McCaskill, J.S. Optimization and design of oligonucleotide setup for strand displacement amplification. *J Biochem Biophys Methods* **63**, 170-186 (2005).
5. Mok, E., Wee, E., Wang, Y. & Trau, M. Comprehensive evaluation of molecular enhancers of the isothermal exponential amplification reaction. *Sci Rep* **6**, 37837 (2016).
6. Organick, L. et al. Random access in large-scale DNA data storage. *Nat Biotechnol* **36**, 242-248 (2018).
7. Anavy, L., Vaknin, I., Atar, O., Amit, R. & Yakhini, Z. Data storage in DNA with fewer synthesis cycles using composite DNA letters. *Nat Biotechnol* (2019).
8. Grass, R.N., Heckel, R., Puddu, M., Paunescu, D. & Stark, W.J. Robust chemical preservation of digital information on DNA in silica with error-correcting codes. *Angewandte Chemie* **54**, 2552-2555 (2015).
9. Zhang, J., Kobert, K., Flouri, T. & Stamatakis, A. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics* **30**, 614-620 (2014).