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)

Abstract:

In DNA data storage, the huge sequence complexity is challenging for repeatable and efficient information reading from massive DNA oligo pool. Here, we demonstrated that synthetic oligo pool comprising over ten thousand strands was largely skewed by PCR process due to its inherent mechanism of inefficient priming, product-as-template, error-spreading prone, which caused serious oligos dropout, over 50% oligos lost in repeated PCR amplification, much more fatal than base mutant for correct information retrieve. Therefore, we developed a new biochemical framework isothermal DNA reading (iDR) from large-scale oligo pool normalization (OPN) and isothermal amplification. Due to its priming-free and error-spreading proof, it achieved low-biased and stable amplification even for successive 10 times deep-reading without performance lost, the

first repeatable DNA storage. Furthermore, the skewed oligo pool with uneven molecule copy number of each oligos was rectified by a total 1280 chemically synthesized OPN-probes, the largest scale oligo normalization so far, and the necessary amount of sequencing reads for perfect oligos retrieve was further largely decreased. These advanced features enable the iDR chemical framework being ideal for manipulating the huge sequence complexity and building the sustainable and efficient DNA storage “hardware”.

Introduction:

In DNA data storage, technology originally developed for bioengineering approach including array oligo synthesis, PCR and DNA sequencing were integrated to construct the "hardware" for DNA storage[1-5]. Array synthesized DNA oligo pool comprising of from thousands up to millions of oligo strands with several hundred bases in length has been utilized in many advanced bioengineering applications, e.g., probe blend, DNA origami assembly, and genome synthesis[1]. Due to both the location on microchip and DNA sequence difference[5], the copy number for each oligo strand from array synthesis is distinct. Furthermore, the sheer number of synthesized oligos from array on microchip is very small, roughly from 10^5 to 10^{12} at the concentration of femtomolar depending on the synthesis platform[6, 7]. Generally, amount of a few hundred nanograms DNA at the concentration of micromolar is necessary for high quality DNA sequencing covering all oligos in the pool on commercial high throughput DNA sequencing platform, e.g., Illumina[8]. Therefore, amplification is crucial to boost the signal for the subsequent DNA sequencing for reading data.

The oligo pool size for DNA data storage is much larger at least several orders of magnitude than that in other bioengineering applications[9-11]. Furthermore, the unevenness of copy number generated a huge complexity that caused serious problem for DNA molecule retrieval and data decoding[4, 5, 12, 13]. In current reported systems, the information reading was achieved by PCR amplification and NGS, but the copy number unevenness originally stemmed from microchip synthesis was further skewed by highly biased amplification process[5, 12]. Therefore, more amplified DNA materials was necessary to fetch the minor oligo strands from the skewed oligo pool. The length and sequence context, GC content and secondary structure of DNA molecule are well known to introduce large amplification bias in PCR for amplification of multiple templates in parallel[14]. Minor oligos could be excluded easily in a few amplification cycles because of its

replication disadvantage. However, DNA storage requires amplification of from tens of thousands up to million distinct DNA strands at the same time. Thus far, this problem was dealt with by high encoding, physical and sequencing redundancy[13, 15], but paid price of losing storage density and increasing cost in all the respects of synthesis, sequencing and decoding calculation. In previous studies, it has been demonstrated that deep PCR amplification (over 60 thermal amplification cycles) increased the unevenness of copy number largely and required several orders of magnitude more sequencing reads for decoding[12]. For extreme case, only a few oligos (less than 10 % of the total oligos) wiped out almost all others after PCR amplification[16]. Moreover, successive repeated PCR amplification significantly skewed the copy number distribution, caused large positive shift[12]. Even the sequences were carefully designed to minimize the sequence context difference, PCR still caused significant chaos from the huge sequence complexity of oligo pool for DNA storage and then imperiled the information decoding.

For practical data storage, crucial issues have to been addressed. First, information reading method should be redesigned to handle amplification of oligo pool with high complexity of DNA sequence at low bias and also support the repeated reading for long-term storage. Second, it is required to flatten the unevenness of oligo copy number following microchip synthesis for decoding with less sequencing resource. Last, the thermal cycling process in PCR not only consumes energy but also will trouble the operation of storage device. It has been reported that long thermal treatment at around 65°C resulted to oligo decay and higher temperature caused more damage to longer DNA molecule[16, 17]. All these issues come down to developing biosystem which could stably and repeatedly handle oligo pool with high sequence complexity and uneven copy number.

Here, we adapted a BASIC code system[18, 19] recently developed for digital distributed file system for DNA storage, in which a high information density was achieved with a low encoding

redundancy of 1.56%, the lowest one ever reported so far. Above all, especially for amplification of oligo pool with high sequence complexity, advanced isothermal DNA reading, termed as iDR, was designed from a novel strand displacement isothermal reaction that drives amplification with distinct mechanism to PCR. Deep statistics analysis demonstrated that oligo pool was more prone to dropout than base error during biased amplification. Oligo pool comprising over ten thousand strands was skewed even in a very light 10 cycles PCR amplification resulting to almost 4 times of oligo dropout than iDR. Moreover, dropout enlarged with the amplification depth and repeated times increased, 13.90% and 53.19% of oligo were wiped out in a 60-cycle deep and 10-times repeated PCR amplification respectively. In contrast, the letter error, either base substitution or indel mutation, remained relatively stable, but the depth of same error increased, namely error-spreading prone. These observations demonstrated that the biased amplification skewed the evenness of oligo pool largely, but base error majorly depended on the total amount of molecular replication. However, due to its molecular nature, iDR addressed these crucial issues, inefficient-priming, product-as-template that caused the bias and error-spreading in PCR amplification. For iDR, oligo dropout remained very stable at low level of around 1.33%, even after successive 10 times repeated amplification. Over 70% of uniformity of oligo pool has been lost, decreased to 0.15 from 0.50 in multiple PCR, but iDR remained almost same at 0.58. To the best of our knowledge, this is the first DNA storage system successfully achieved the deep repeated information reading.

Furthermore, we developed large-scale oligos pool normalization (OPN) to stoichiometrically rectify the oligo copy number unevenness for further improvement of information decoding. Total 1280 OPN probes were designed from 5 distinct anchors and one set of 256 barcode. The large OPN probe mixture was demonstrated being able to improve the Gini index of oligo pool

97 comprising over thousands oligo stands, the most large-scale oligo normalization reported so far.
98 Oligos was 100% retrieved from sequencing reads much less than original oligo pool read by PCR.
99 The new biochemical framework was demonstrated being able to handle high sequence complexity,
100 by which stable multiple reading was achieved based on error-spreading proof and low bias
101 isothermal amplification, large scale oligo normalization and information could be decoded from
102 at least two orders of magnitude less necessary DNA material and sequencing resource and also
103 highly compatible with any other encoding software system for sustainable and efficient DNA data
104 storage.

105

Results:

BASIC code for DNA mediated distributed storage.

In current DNA storage, the digital file was divided and written into a large group of small piece DNA oligos. Every single oligo function as individual information carrying unit. And then, the entire file can be read out from sequencing all the oligos (Fig. 1a and Supplementary Note 1). Therefore, DNA storage can be considered as a biomolecular distributed storage system, the "Software-Defined Storage" in semiconductor hard drive. Due to this nature, we adapted a BASIC coding system that is one well optimized regenerating code with reducing computational complexity for distributed storage. It is flexible for various DNA oligo length from different commercial synthesis platform. Generally, the encoding process started by dividing the target file into non-overlapping groups. And then, the split information was encoded into DNA sequence following an optimized encoding process (Supplementary Figs. 1-3) with two adjustable parameter K (corresponding to oligo number in one non-overlapping group) and L (corresponding to the length of oligo). Besides the normal file types, for the first time we tested the storage of genome sequence including human mitochondrion and one artificial bacteria cell with a different encoding strategy (Supplementary Fig. 4 and Supplementary Note 2). In genome sequence, there are many complicate structures hard for correct sequencing. Through the encoding process, the genome sequence was rewritten into nucleotide sequence again but with complicated sequence avoided and accurate sequencing guaranteed by error correcting code. Furthermore, oligo pools with different payload length have been designed to store 2.85 MB files in totally 109,568 oligo strands (Fig. 1a and Supplementary Figs. 6 and 8). Reed-Solomon code was used for error correction and coding redundancy for tolerance of missing entire oligo. In comparing with previous reported systems, relative high information density 1.65 bits/nt was achieved with a 1.56% of coding redundancy,

129 which allowing success decoding as long as oligo dropout less than 1.56%, randomly losing 4
130 strands from 256 (Supplementary Fig. 5 and Supplementary Note 6). Technically, higher encoding
131 redundancy tolerates losing more oligos and success decoding can be achieved from lower
132 sequencing coverage. However, higher redundancy requires more DNA synthesis. Considering the
133 synthesis cost is higher than sequencing, represented over 90% of the total cost of DNA storage,
134 it is more practical to trade-off encoding redundancy for synthesis cost.

135 Chip-based synthesis only produces very small amount of oligo, roughly from 10^5 to 10^{12} at the
136 concentration of femtomolar depending on the synthesis platform[7]. In the current workflow of
137 DNA data storage, the quality of oligo pool majorly impacted the storage performance[4].
138 Particularly, both the heterogenous oligo sequences and the unevenness of oligo copy number
139 generated a huge sequence complexity and then caused the PCR amplification bias. Minor oligo
140 molecules are prone to drop-out and more sequencing coverage was required for decoding oligo
141 pool with largely skewed copy number distribution (Fig. 1b). The material of skewed oligo pool
142 made stable and repeatable information decoding a huge challenge.

143 **Low-bias and error-spreading proof isothermal amplification**

144 It well known that PCR generated biased amplification from its inherent mechanism, i.e., product-
145 as-template, priming and thermal cycling dependent amplification[20]. In order to address these
146 problems that easily skewed the oligo copy number distribution, we designed a method from novel
147 isothermal DNA replication. In comparing with PCR depending on thermal cycling to drive DNA
148 replication, sequence specific nickase and DNA polymerase with processive strand displacement
149 activity were recruited for DNA amplification under consistent low temperature[21]. After
150 systematical optimization (Supplementary Figs. 12-24 and Supplementary Note 4), Nt.BbvCI and
151 exonuclease deficient DNAP I Klenow fragment (3'→5' exo$^-$) were used to amplify oligos

152 immobilized on magnetic microbeads. The immobilized isothermal DNA replication system was
153 designated as iDR, isothermal DNA reading (Fig. 2a).

154 The intrinsic features of iDR specifically facilitate DNA storage application. In theory, for one 50
155 ul reaction, 10 PCR thermal cycles required about 177.8J energy only for thermal regulation at
156 least two orders of magnitude more than 30 mins iDR of 2.52J (Fig. 2b and Supplementary Note
157 7) and it may be a huge issue at large scale of operation. Real time monitoring indicated that
158 amplification rate of iDR is very close to PCR, even generally it was considered as a linear
159 replication (Supplementary Figs. 25 and 26). Single-stranded or double-stranded DNA can be
160 produced in a controlled manner (Fig. 2c). Specially, ssDNA was amplified in a primer-free
161 manner and allows iDR to be a universal method for reading information with no sequence
162 information needed in advance. The nickase mediated site-specific phosphodiester bond cleavage
163 initiated the iDR amplification and generated a 5' terminal phosphate group, its function was
164 verified by direct ligation to a FAM labeled probe (Fig. 2d and Supplementary Figs. 27 and 28).
165 This inborn phosphate group is very convenient for subsequent functional adapter linking.

166 Amplified oligos were sequenced on commercial Illumina Hiseq 4000 platform with 150 paired-
167 end cycles and then deeply analyzed by a set of statistics methods developed from bioinformatic
168 BLAST program (Supplementary Fig. 30). Sequenced reads with various number of letter error
169 including substitution and indel were counted with significant different amount. Among them,
170 single letter error accounted for the vast majority in mutant sequenced reads for both PCR (80.1%)
171 and iDR (81.7%) (Fig. 2e, Supplementary Figs. 31-35 and Supplementary Note 11). The total indel
172 (0.03%) and substitution (0.2%) base rate were consistent with previous studies[13]. Although, RS
173 code is able to correct multiple errors in same DNA strand but will increase the computation
174 complexity largely[17, 22]. Therefore, both sequenced reads with no or single substitution/indel error

were collected as valid reads for further analysis. The distribution of the number of reads per each given sequence were different between iDR and light PCR of just 10 thermal cycles. The distribution normality was quantified by a modified function (Supplementary Note 9). In theory, oligo pool of higher value distribution normality require less sequencing reads to recover all synthesized oligos[12]. The coverage distribution of both 10 cycles PCR and iDR were positively skewed with a long tail, which comprising of high copy number oligos, but the normality of 10 cycles PCR decreased about 16% than iDR (Fig. 2f). The proportion of oligos with high copy number in the tail, the top 30% of high coverage, enlarged with the PCR cycles increased (Supplementary Fig. 38 and Supplementary Note 9). No obvious difference was observed between the coverage distribution of iDR amplification from free oligo pool and oligo pool immobilized on magnetic beads (Supplementary Figs. 39 and 40). These results demonstrated that even just 10 thermal cycles largely skewed the oligo pool due to the PCR amplification bias and the amplification of iDR was much low-biased.

Deep errors, reads with mutant in high copy number, impeded information decoding. Sequenced reads were further sorted out as group of M0G0 (with no letter error) and M1G1 (with single letter substitution or indel error) by developed BLAST programs. The distribution of the number of reads in M0G0 and M1G1 overlapped for both PCR and iDR (Fig. 2g). The extent of overlap between coverage distribution of M0G0 and M1G1 correlated with the potential of mutant reads affecting retrieval of correct reads for decoding. The max coverage for M1G1 of 10 cycles PCR was counted as 63 and 6956 oligos in PCR M0G0 was identified with copy number lower than it, accounting for 60.38% of total oligos. However, the corresponding number for iDR were counted respectively as 28 and 2011, accounting for 17.46% of total oligos, 71% less deep error than PCR. It indicted that in PCR 60.38% of oligos will be infected if retrieved by principle of law of large number-

makers, but only 17.46% for iDR. Because same master pool used as template, it demonstrated that error accumulated much more in PCR than iDR. And even only 10 thermal cycles of PCR caused deep error-spreading that would cause huge calculation in identifying the majority (over 60%) of correct oligos and the proportion enlarged with PCR cycle number increased (Supplementary Fig. 45). Moreover, both the total rate for substitution or indel error and proportion of valid reads (combined M0G0 and M1G1 group) in all noisy sequenced reads remained stable for both iDR and PCR of thermal cycles from 10 to 60 (Fig. 2h and Supplementary Figs. 47 and 48). In comparison, the dropout rate increased significantly from 4.18% to 13.80% with PCR amplification becoming deeper, much higher than iDR of 1.33%. Low replication fidelity of DNA polymerases and around 1% miss reading coming along with NGS sequencing process[23] largely contributed to the massive error reads with low copy number, e.g. 1-2 copy number, but it is relatively easy to identify them from correct reads in high copy number. Considering the mechanism, product-as-template and inefficient priming process caused the high amplification bias and made PCR amplification prone to error-spreading and oligo dropout. In contrast, iDR was designed to synthesize new oligo only from the original templates without priming process for replication initiation and therefore iDR achieved a low biased and error-spreading proof amplification and will require much less calculation resource in decoding process. Additionally, high temperature treatment resulted to DNA oligo decay especially for long strands[17, 24]. The decay lost rate in a 10 cycles PCR was calculated as 21.8‰ and 0.035‰ for iDR (Fig. 2b and Supplementary Note 8) from a plotted DNA half-life graph (Supplementary Fig. 29).

Multiple-repeated information reading.

The amplification capability for successive deep reading oligo pool containing from 11,520 to 89,088 DNA strands was examined. PCR amplification was successively performed 10 times from

aliquot of previous reaction and iDR amplification was repeated 10 times from immobilized oligos pool (Fig. 3a, Supplementary Figs. 49-51). In #1, #5 and #10 of successive PCR, the proportion of amplified oligos with up to total 10 substitution or indel letter error decreased from 89.09% (±10 of PCR #1) to 49.97% (±10 of PCR #10) and from 86.77% (±1 of PCR #1) to 48.58% (±1 of PCR #10). In comparison, the repeated iDR remained very consistent over 90% (Fig. 3b). These statistics results indicated that large noise was introduced during PCR procedure and amplified oligo with imperfect length increased from inefficient replication or miss priming. However, no obvious difference was observed in proportion of M1G1 in total sequenced reads between PCR and iDR, but the mean copy number in M1G1 reads increased from 1.10 of #1 PCR to 1.95 of #10 PCR, but only slightly changed for iDR, 1.08 of #1, 1.08 of #5 and 1.05 of #10 (Fig. 3c). This result was in agreement with previous experiment in figure 2g and indicated that error accumulated to high copy number during PCR and iDR achieved error-spreading proof amplification.

Interestingly, it was observed that the distribution of the number of reads per each given sequence changed differently. Normality of #1, #5 and #10 PCR decreased significantly from 0.50 to 0.15. In contrast, #1, #5 and #10 iDR gave a consistent normality of 0.58 (Fig. 3d, Supplementary Figs. 52-55). For successive PCR, the coverage distribution was largely positively skewed and the proportion of both low copy number and high copy number oligos significantly increased. It indicated that large enrichment driven by amplification bias for part of oligos efficiently occurred with the successive PCR, but not in repeated iDR. Furthermore, it was observed that only top 1% oligo of high coverage increased its proportion significantly and both 1% oligos of middle and low coverage decreased while the oligo pool was successively read by PCR. In contrast, all the 1% oligos remained steady in repeated iDR (Supplementary Fig. 56 and Supplementary Note 9). The dropout rate in random valid reads set with 10x coverage was quantified to further assess the

amplification bias. For PCR, the dropout rate increased sharply from 4.18% (#1 PCR) to 53.19% (#10 PCR) (Fig.3e), but iDR remained steady at about 2%. Due to the tolerance for 1.56% dropout in BASIC encoding algorithms, we also calculated the coverage depth for random valid reads set with 1.56% dropout (Supplementary Note 6), the crucial parameter for the theoretical minimum decoding coverage. For PCR, the minimum coverage depth was quantified as 17.2 (#1 PCR) and 167 (#5 PCR). Because #10 PCR lost too many oligos, the minimum coverage depth was calculated as 426 from calculation (Supplementary Fig. 61). For iDR, the minimum coverage depth was quantified as 11 for #1, 12 for #5 and 12.5 for #10. Thus far, for PCR based sustainable DNA media reading, one strategy is deep amplification from trace DNA material with large number of thermal cycles and another is successive amplification from aliquot of previous reaction. However, we demonstrated that both strategies are not practical. The nature of PCR including especially inefficient priming and product-as-template generated large amplification bias that significantly skewed the copy number distribution and resulted huge dropout in both 60 cycles deep amplification and 10 times successive amplification, especially for oligo pool with huge sequence complexity. Based on these deep statistical analyses, we pointed out that iDR amplification was more stable, robust and suitable for repetitive read than PCR in DNA data storage.

Large-scale oligo pool normalization improved information reading.

Besides the amplification procedure, the unevenness of oligo copy number originally stem from microchip-based synthesis. However, there is still short of handy technology to normalize oligo pool with high sequence complexity. To address this problem, we developed a simple oligo pool method based on a previous reported study[25] but with more simple and flexible procedure and without expensive DNA chemical modification. Two-parts probe was designed, in which constant

anchor part provide basic efficient binding and short variable 9nts barcode part generating 256 specific targets recognition with no expensive modification. In particular, oligo pool normalizing (OPN) sequence was synthesized at the 3' terminal end, which comprising of barcode (9 nts) and universal fragment (17 nts) (Fig. 4b and Supplementary Fig. 8). The barcode guided the OPN probe to identify its target while universal fragment binding as an anchor (Fig. 4b). The unique barcode sequence "GWSWSWS", alternating strong (G or C) and weak (A or T) bases (e.g. CACTGT or GTCTGA), has been proven being able to generate 256 high specific binding (Supplementary Note 5). After perfect binding, OPN probe was extended by DNA polymerase, turning oligo into double-stranded form and then the remaining ssDNA oligos were removed by exonuclease I degradation (Fig. 4a). Due to the precisely selected degradation of normalization and the low productivity of microchip oligo synthesis, it is necessary to amplify the oligo pool to a large quantity but keeping its single-stranded form. Therefore, we developed a simple single-stranded oligo pool amplification protocol (SOA), by which ssDNA oligo pool was amplified by PCR with normal forward primer and 5'-phosphated reverse primer and then only strand extended from forward primer was removed by lambda exonuclease and the intact complementary strand was left (Supplementary Figs. 62 and 63).

First, we tested the normalization of three oligos of different length with 3' terminal FAM label, which was mixed at input molar ratio of 1:5:25. After OPN normalization, the output oligos were quantified as molar ratio of 0.95:1:1.25 (Fig. 4d and Supplementary Fig. 64). Next, we proceeded to test normalization improved information decoding. As proof-of-concept, OPN 1.0 with 256 probes, which were synthesized separately and mixed at equal molar ratio, were used to normalize 256 oligos synthesized from microchip with other 10752 oligos together (Fig. 4c). Following the SOA protocol, 256 single-stranded oligos pool was prepared and normalized, each OPN probe was

equal to the average molar concentration of the oligo pool. The coverage depth of each oligo in OPN 1.0-iDR was significantly improved in comparing with iDR amplified oligo pool without OPN normalization (Fig. 4e). The PCR-amplified oligo pool showed skewed normality than OPN-iDR (Supplementary Figs. 65 and 67) and high copy number oligo in PCR was obviously normalized in OPN-iDR (Supplementary Fig. 66) with decreased standard deviation in quantification of molecule number of each oligo strands. And then, dropout rate for random valid reads set with various coverage depth was plotted (Supplementary Fig. 68). All of the dropout rate decreased as coverage increased. At the coverage of about 25, the dropout rate of PCR became lower than 1.56%, the limit for decoding. However, all of dropout rate of OPN-iDR was lower than 1.56% and became 0 at around coverage of 20. The dropout rate of iDR was always higher than OPN-iDR but much lower than PCR. At coverage of 10, the dropout rate of OPN-iDR was almost one order of magnitude lower than PCR. The recognition capability of OPN1.0 was extended simply by recruiting more anchor design. In OPN2.0 (Fig. 4f), 1024 specific target oligo recognition was generated from 4 distinct anchor sequences and the one set 256 barcode. All of these 1024 oligos was perfect retrieved from 87x sequencing reads less than that 223x of PCR amplified oligos (Fig. 4g and Supplementary Figs.69 and 71). Technically, this method can be expanded by using a combination of a group of specific universal probes[13] and the 256 barcodes to manipulate oligo pool comprising of up to 3 Million oligos (Supplementary note 5). Although minimum decoding coverage depth previously reported (Supplementary Fig. 72) is lower than the result presented here, we have to point out that the minimum decoding coverage is highly dependent on both the encoding system with different encoding redundancy and the quality of oligo pool from different synthesis platform (Supplementary Figs. 73-76) and it is hard to compare the results cross different systems.

Discussion:

DNA data storage is one very artificial application and actually the concept has been proposed for a long time[26]. But till lately, the significantly increasing capability of DNA synthesis and sequencing start making it possible. In many respects, practical DNA data storage requires more powerful capability for all related biotechnologies, including synthesis, sequencing and manipulation of oligo pool with high sequence complexity. The size of oligo pool used in DNA storage is already several orders of magnitude larger than that in other applications. However, there is still short of hardware technology which is designed for DNA storage, especially manipulation of DNA oligo pool with huge sequence complexity and unevenness of molecule copy number. In studies to date, PCR is still the only method for amplification of oligo pool for reading information. However, with deep bioinformatic and statistics analysis, we demonstrated that PCR amplification skewed the oligo pool even after a very few thermal cycles (10 cycles) and the skewness largely increased, deep error spreading and massive dropout, as the function of amplification cycle numbers. Deep error, mutant sequenced reads in high coverage, interfered decoding and caused significant increased calculation, but decoding fatally crashed due to massive dropout. We contend that the features of low temperature, priming-free, and enzyme-mediated double helix unzipping in iDR amplification presented here, overcame the major amplification bias related issues of PCR including the inefficient priming, product-as-template, sequence context dependent, and high temperature heating. iDR achieved very stable amplification performance, which efficiently prevented the mutant error from spreading, decreased over 70% deep error, and achieved successive repeated deep reading with consistent outcome quality. Actually, the physical storage density, the most significant advantage of DNA storage with a calculation of a few kilograms DNA material for storage of data from the whole human society[27], highly depend on the quality of DNA

reading. Therefore, higher physical storage density could be achieved by iDR system, which could store information with at least two orders of magnitude less DNA material and sequencing resource than the current PCR method (Fig. 5 and Supplementary Note 12).

Although, there is many variants of PCR, such as emulsion PCR, digital PCR, and multiplex PCR, but none of them can avoid the crucial issues of product-as-template, inefficient priming, and complex thermal regulation. Furthermore, Replication fidelity of DNAP I Klenow fragment used in iDR system is much lower than Q5 DNA polymerase in PCR. There is still much room for improvement. Additionally, practical DNA storage system must move out of biochemical test tube to build device by highly integrating all related biochemical processes. Prototypes of DNA storage hardware have already been proposed with high density DNA material in microfluidic device[28, 29]. We contend that besides some crucial features of iDR make it more fit for hardware construction. In iDR, only low temperature, slightly higher than room temperature, was required. Considering the operation of DNA storage up to large scale, there will be a huge difference in energy consumption. Furthermore, iDR is able to work in a primer-free fashion, only defined protein enzyme mix is required no matter what information was encoded, which makes it possible for universal information reading without any sequence information required in advance. To the best of our knowledge, iDR is the first system for stable repeated DNA information reading which is crucial and necessary feature for practical and sustainable storage hardware.

The large-scale oligo normalization is another advanced feature for DNA storage. Comparing with the previous reported system, OPN was developed as a simple and very economical process without any expensive modification directly on oligo pool or OPN probe, such as dexoyuracl (dU) and biotin, none of them has been reported to be synthesized directly on chip-array. Then OPN could be applied to any chip-array oligo synthesis platform and the OPN probe synthesis cost is

also acceptable (Supplementary Fig.77). Additionally, SOA protocol was developed enabling OPN to manipulate oligo pool with very small amount DNA molecule, generally chip-array synthesized oligo pool is very small amount. Moreover, the capability of OPN could be easily extended. OPN 2.0 achieved specifically targeting 1024 oligos, 4 times of OPN 1.0, simply by combining 4 distinct anchor sequences, which is the most large-scale oligo pool normalization reported so far. 1024 array synthesized oligos was revised with improved Gini index (Supplementary Fig.70), and was perfect identified from sequencing reads (87x) less than that (223x) for PCR amplified oligos. Therefore, the necessary DNA material and sequencing resource for information decoding could be decreased at least two orders of magnitude (Fig. 5), which is a huge advantage in considering application of DNA storage in large scale comparing to current big data center. Following the parallel extension strategy, target recognition could be leveraged to million in an economical way (Supplementary note 5 and Supplementary Fig.63). Therefore, we believed that this new biochemical framework with stable repeated reading and large-scale normalization lays a foundation for development of practical and sustainable DNA storage.

ACKNOWLEDGMENTS

We would like to thank Professor Hanxu Hou from Dongguan University of Technology for advice and assistance with designing algorithm for BASIC code. We also thank Yixi Wang from College of intelligence and computing at Tianjin University for her help in test of encoding system. This work was supported by the National Science Foundation of China (Grant No.21476167, No.21778039 and No.21621004).

AUTHOR CONTRIBUTIONS

Y.G. and H.Q. designed and performed all experiments. X.C., J.H. and C.Z. designed and developed the encoding and decoding program. X.C. developed program for the bioinformatics statistics analysis. Y.G., H.Qiao and H.Q. collected and analyzed all experiment data. G.Y. T.H. and H.Q. wrote the manuscript. H.Q. designed experiments, analyzed data and supervised this work.

COMPETING FINANCIAL INTERESTS

H.Q. is the inventor of two patents application for the biochemical method described in this article. The initial filings were assigned Chinese patent application (201911086860.0 and 201911087247.0) and international patent application (PCT/CN2019/123916). The remaining authors declare no competing financial interests.

References:

1. Kosuri, S. & Church, G.M. Large-scale de novo DNA synthesis: technologies and applications. *Nat Methods* **11**, 499-507 (2014).
2. Goodwin, S., McPherson, J.D. & McCombie, W.R. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet* **17**, 333-351 (2016).
3. Kozarewa, I. et al. Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. *Nat Methods* **6**, 291-295 (2009).
4. Organick, L. et al. Experimental Assessment of PCR Specificity and Copy Number for Reliable Data Retrieval in DNA Storage. (2019).
5. Chen, Y.-J. et al. Quantifying Molecular Bias in DNA Data Storage. (2019).
6. Klein, J.C. et al. Multiplex pairwise assembly of array-derived DNA oligonucleotides. *Nucleic Acids Res* **44**, e43 (2016).
7. Jingdong Tian, H.G., Nijing Sheng, Xiaochuan Zhou, Erdogan Gulari, Xiaolian Gao & George Church Accurate multiplex gene synthesis from programmable DNA microchips. *Nature* **432**, 1050-1054 (2004).
8. Linnarsson, S. Recent advances in DNA sequencing methods - general principles of sample preparation. *Exp Cell Res* **316**, 1339-1343 (2010).
9. Kosuri, S. et al. Scalable gene synthesis by selective amplification of DNA pools from high-fidelity microchips. *Nat Biotechnol* **28**, 1295-1299 (2010).
10. Lipshutz, R.J., Fodor, S.P., Gingeras, T.R. & Lockhart, D.J. High density synthetic oligonucleotide arrays. *Nat Genet* **21**, 20-24 (1999).
11. Bonde, M.T. et al. Direct mutagenesis of thousands of genomic targets using microarray-derived oligonucleotides. *ACS Synth Biol* **4**, 17-22 (2015).
12. Erlich, Y. & Zielinski, D. DNA Fountain enables a robust and efficient storage architecture. *Science* **355**, 950-954 (2017).
13. Organick, L. et al. Random access in large-scale DNA data storage. *Nat Biotechnol* **36**, 242-248 (2018).
14. Aird, D. et al. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol* **12**, R18 (2011).
15. Goldman, N. et al. Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. *Nature* **494**, 77-80 (2013).
16. Heckel, R., Mikutis, G. & Grass, R.N. A Characterization of the DNA Data Storage Channel. *Sci Rep* **9**, 9663 (2019).
17. 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).
18. Hou, H., Shum, K.W., Chen, M. & Li, H. BASIC Codes: Low-Complexity Regenerating Codes for Distributed Storage Systems. *IEEE Transactions on Information Theory* **62**, 3053-3069 (2016).
19. Hou, H., Han, Y.S., Shum, K.W. & Li, H. A Unified Form of EVENODD and RDP Codes and Their Efficient Decoding. *IEEE Transactions on Communications* **66**, 5053-5066 (2018).
20. Mullis, K.B. & Faloona, F.A. Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction. *Methods Enzymol* **155**, 335-350 (1987).

21. Joneja, A. & Huang, X. Linear nicking endonuclease-mediated strand-displacement DNA amplification. *Anal Biochem* **414**, 58-69 (2011).
22. 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).
23. Yazdi, S.M., Yuan, Y., Ma, J., Zhao, H. & Milenkovic, O. A Rewritable, Random-Access DNA-Based Storage System. *Sci Rep* **5**, 14138 (2015).
24. A Characterization of the DNA Data Storage Channel. (2018).
25. Pinto, A., Chen, S.X. & Zhang, D.Y. Simultaneous and stoichiometric purification of hundreds of oligonucleotides. *Nat Commun* **9**, 2467 (2018).
26. Wallace, M.R. Molecular Cybernetics: The Next Step? *Kybernetes* **7**, 265-268 (1978).
27. Extance, A. How DNA could store all the world's data. *Nature* **537**, 22-24 (2016).
28. Newman, S. et al. High density DNA data storage library via dehydration with digital microfluidic retrieval. *Nat Commun* **10**, 1706 (2019).
29. Takahashi, C.N., Nguyen, B.H., Strauss, K. & Ceze, L. Demonstration of End-to-End Automation of DNA Data Storage. *Sci Rep* **9**, 4998 (2019).

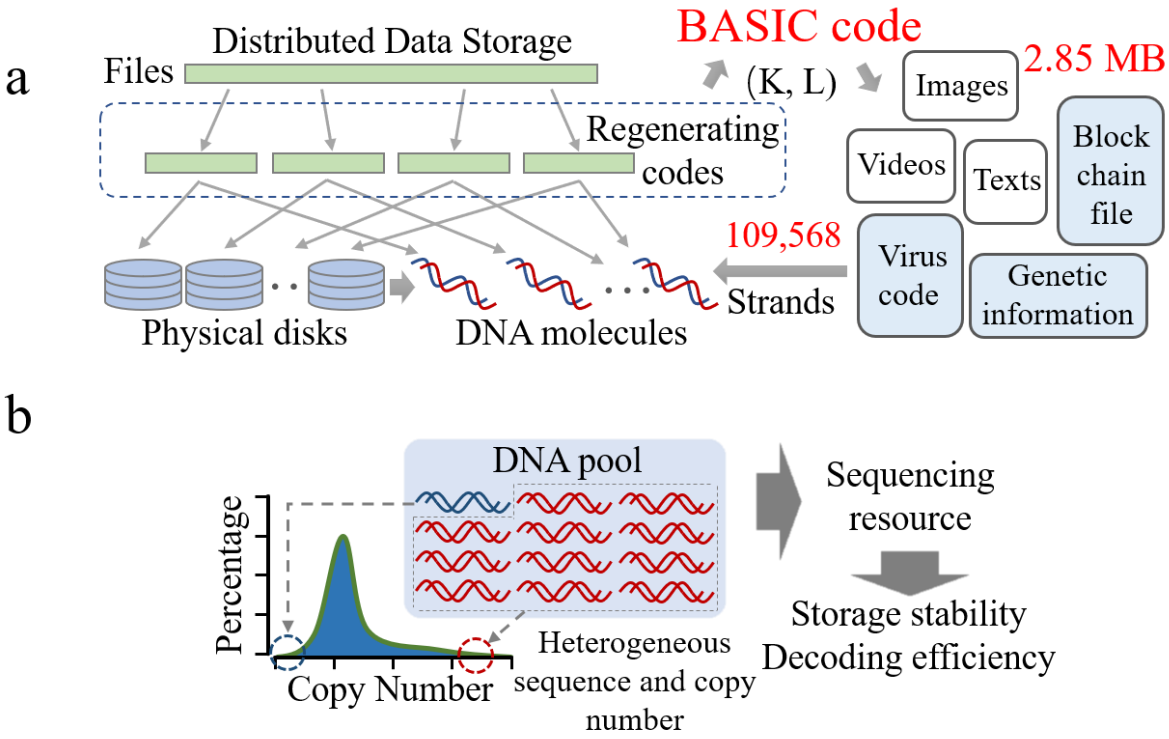

Figure 1: Overview of the DNA data storage.

(a) Schematic of the DNA mediated distributed data storage and totally 2.85 MB data, including text,s, images, videos, and block chain file, computer virus code, genetic information, were encoded into total 109,568 synthesis oligos using a adapted BASIC code system.

(b) Illustration of data storage in synthesis oligo pool. The huge sequence complexity and unevenness of oligo copy number is challenging the stable storage and precise decoding.

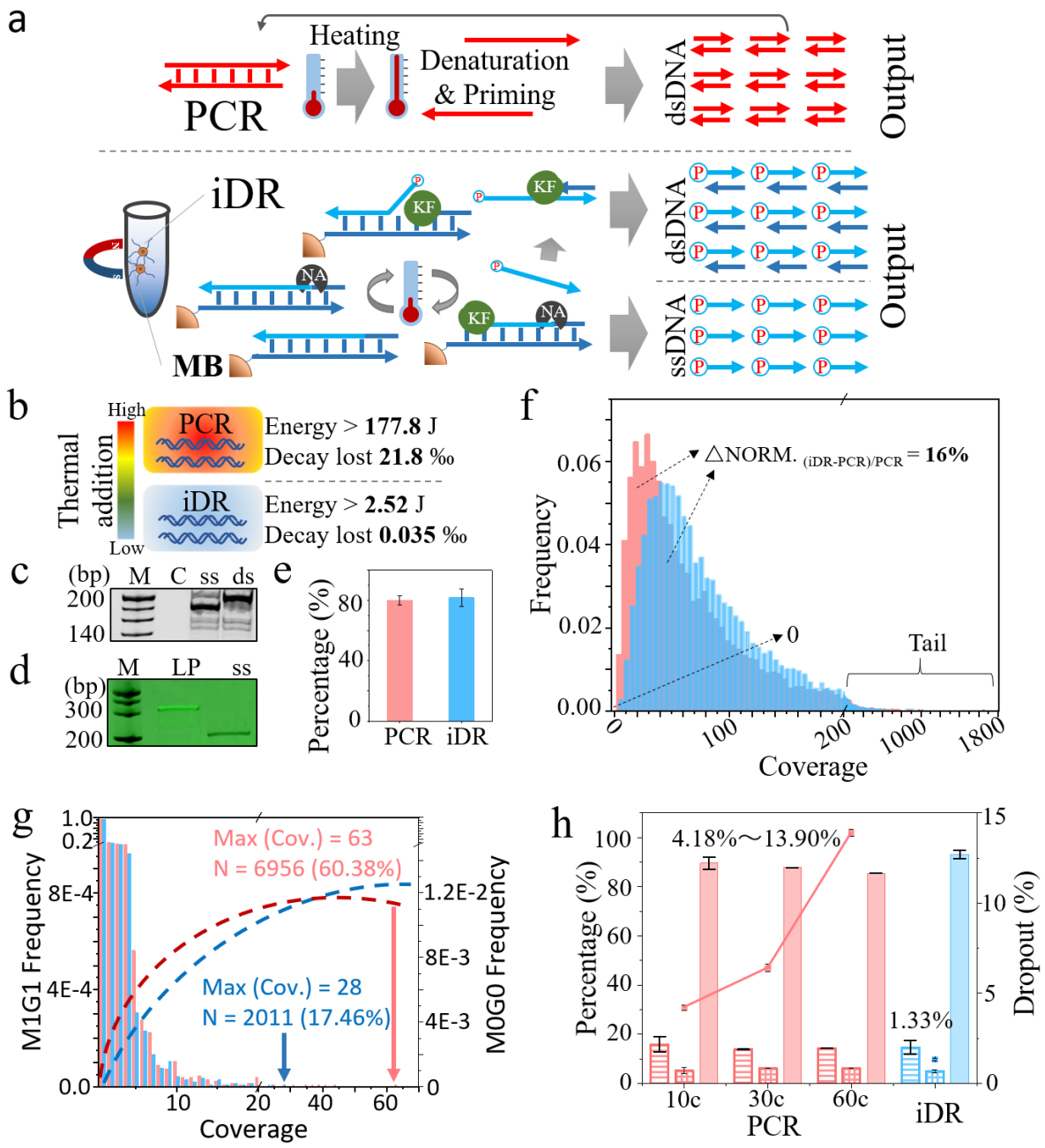

Figure 2: Designed isothermal DNA reading (iDR) for DNA data storage.

(a) Illustration of the DNA amplification of PCR (upper) and iDR (lower). PCR depend on thermal cycling to drive replication. By contrast, isothermal DNA reading was developed from

systematically optimized strand displacement mediated replication, in which nickase and specific DNA polymerase collaborate to drive amplification under consistent low temperature.

(b) Calculation of energy consumption and the oligo degradation decay for one 50ul liquid reaction amplified by 10 thermal cycles PCR or iDR. PCR and iDR required 177.8 J and 2.52 J respectively only for thermal regulation theoretically. Thermal treatment caused oligo decay rate was calculated as 21.8‰ for 10 cycles PCR and 0.035‰ for iDR respectively.

(c) Single-stranded and double-stranded oligo amplified by iDR from a 218nt dsDNA template were analyzed on 10% native PAGE gel. M: 20 bp DNA Ladder; C: negative control of amplification without input template; ss: single-stranded DNA product; ds: double-stranded DNA product.

(d) A 5' terminal FAM labeled 30nt single-stranded probe ligated to 198nt single-stranded DNA with an inborn 5' terminal phosphate group directly from iDR amplification was analyzed and imaged on a 12% UREA denature PAGE. M: 20 bp DNA Ladder; LP: ligation product; ss: ssDNA product of iDR.

(e) Sequenced reads with single letter error, substitution or indel, accounted for the vast majority in total mutant sequenced reads for both PCR (80.1%) and iDR (81.7%). Error bars represent the mean ±s.d., where n = 3.

(f) Coverage depth distribution of sequenced reads for 10 cycles PCR and iDR amplification were positively skewed but with 0 oligo dropout, but the distribution normality of PCR decreased about 16% than iDR.

(g) Sequenced reads were grouped as M0G0 (with no letter error) and M1G1 (with single letter substitution or indel error) by developed BLAST programs. In M1G1 of 10 cycles PCR the max

492 coverage was counted as 63 and 6956 oligos overlapped with M0G0 group, accounting for 60.38%
493 of total oligos. For iDR it was 28 and 2011 respectively, accounting for 17.46% of total oligos
494 with 71% of deep error decreased.

495 **(h)** The proportion of error reads and valid reads (total reads with no or single substitution/indel
496 error) per million noise sequenced reads were plotted, striped column for base substitution; grid
497 column for indel, and solid bar for valid reads. For 10, 30 and 60 cycles PCR, the substation were
498 15.92%, 13.90% and 14.27%, and indel were 5.28%, 6.16% and 6.15%, and valid reads were
499 89.58%, 87.73% and 85.53% respectively. For iDR, 14.69% for substitution error, 4.99 for indel
500 and 93.15% for valid reads. Error bars represent the mean ±s.d., where n = 3. Oligo dropout rate
501 calculated from random reads set with a mean 10x coverage depth was plotted. For 10, 30 and 60
502 thermal cycles PCR (red line), it was counted as 4.18%, 6.41% and 13.90% respectively, and 1.33%
503 for iDR (blue dot). Error bars represent the mean ±s.d., where n = 10.

504

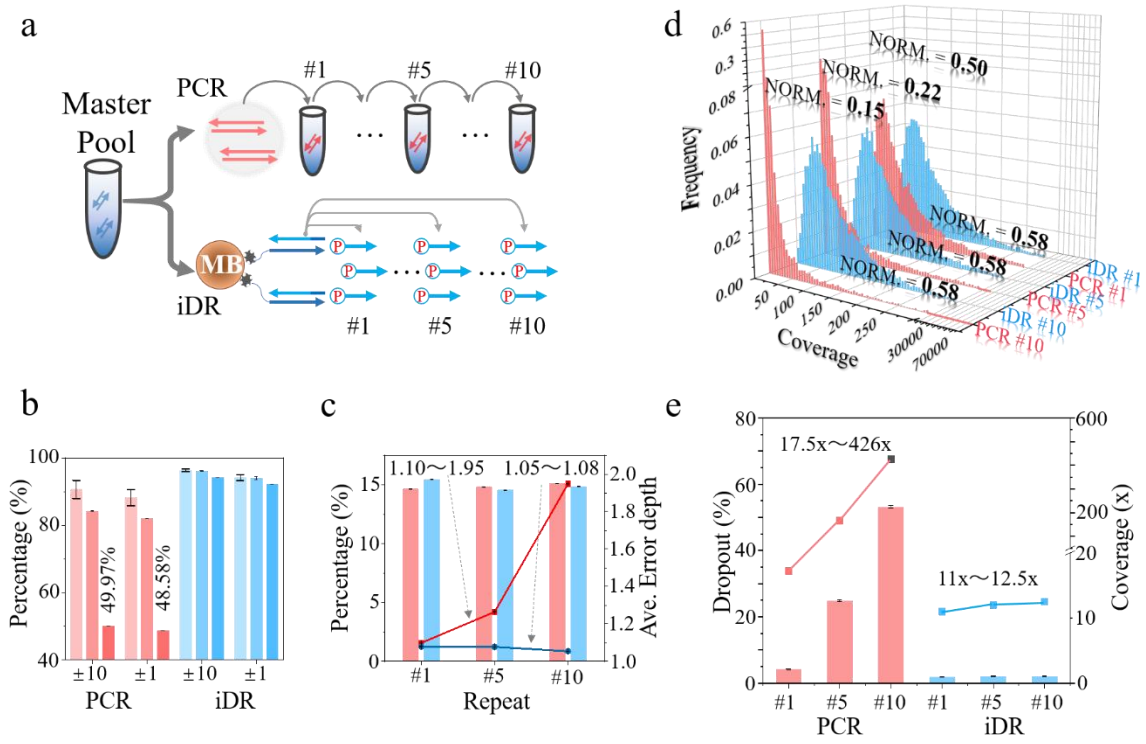

Figure 3: Multiple information reading from DNA storage.

(a) Illustration of successive DNA reading. Repeated PCR amplification was successively performed 10 times with aliquot of previous reaction as template and the oligos pool immobilized on magnetic beads was successive amplified 10 times by iDR. The same master oligo pool was used as the initial template and #1, #5 and #10 amplified oligos were analyzed.

(b) The proportion of reads with up to 10 substitution or indel letter error per million sequenced reads were counted as 90.65% (#1 PCR, light red), 84.26% (#5 PCR, red), 49.97% (#10 PCR, dark red), 96.35% (#1 iDR, light blue), 96.14% (#5 iDR, blue), and 94.12% (#10 iDR, dark blue), and

error reads with up to 1 error were counted as 88.18% (#1 PCR), 82.05% (#5 PCR), 48.58% (#1 PCR), 94.18% (#1 iDR), 94.04% (#5 iDR), and 92.21% (#10 iDR). Error bars represent the mean ±s.d., where n = 3.

(c) The proportion of M1G1 reads per million valid reads were, PCR (red column), 14.65% (#1 PCR), 14.82% (#5 PCR), 15.10% (#10 PCR), and iDR (blue column) 15.46% (#1 iDR), 14.55% (#5 iDR), and 14.86% (#10 iDR). The average error reads depth of PCR (red line, 0.5 for #1, 0.22 for #5, 0.15 for #10) increased and iDR (blue), but iDR (blue line, 1.08 for #1), 1.08 for #5), 1.05 for #10) remained stable. Error bars represent the mean ±s.d., where n = 3.

(d) The distribution of reads number per each given sequence per million sequenced reads of #1, #5, and #10 of PCR and iDR. The distribution normality was 0.5 (#1 PCR), 0.22 (#5 PCR), 0.15 (#10 PCR), 0.58 (#1, #5, #10 iDR) respectively.

(e) The dropout rate for random sequenced reads set with 10x coverage depth was plotted, PCR (red column, 4.18% for #1, 22.89% for #5 and 53.19% for #10) and iDR (blue column, 1.86% for #1, 2.08% for #5 and 2.09% for #10). Error bars represent the mean ±s.d., where n = 3. The coverage depth for random sequenced reads with 1.56% dropout was calculated as 17.2 (#1 PCR), 167 (#5 PCR), 426 (#10 PCR) and 11 (#1 iDR), 12 (#5 iDR), and 12.5 (#10 iDR) respectively. Error bars represent the mean ±s.d., where n = 10.

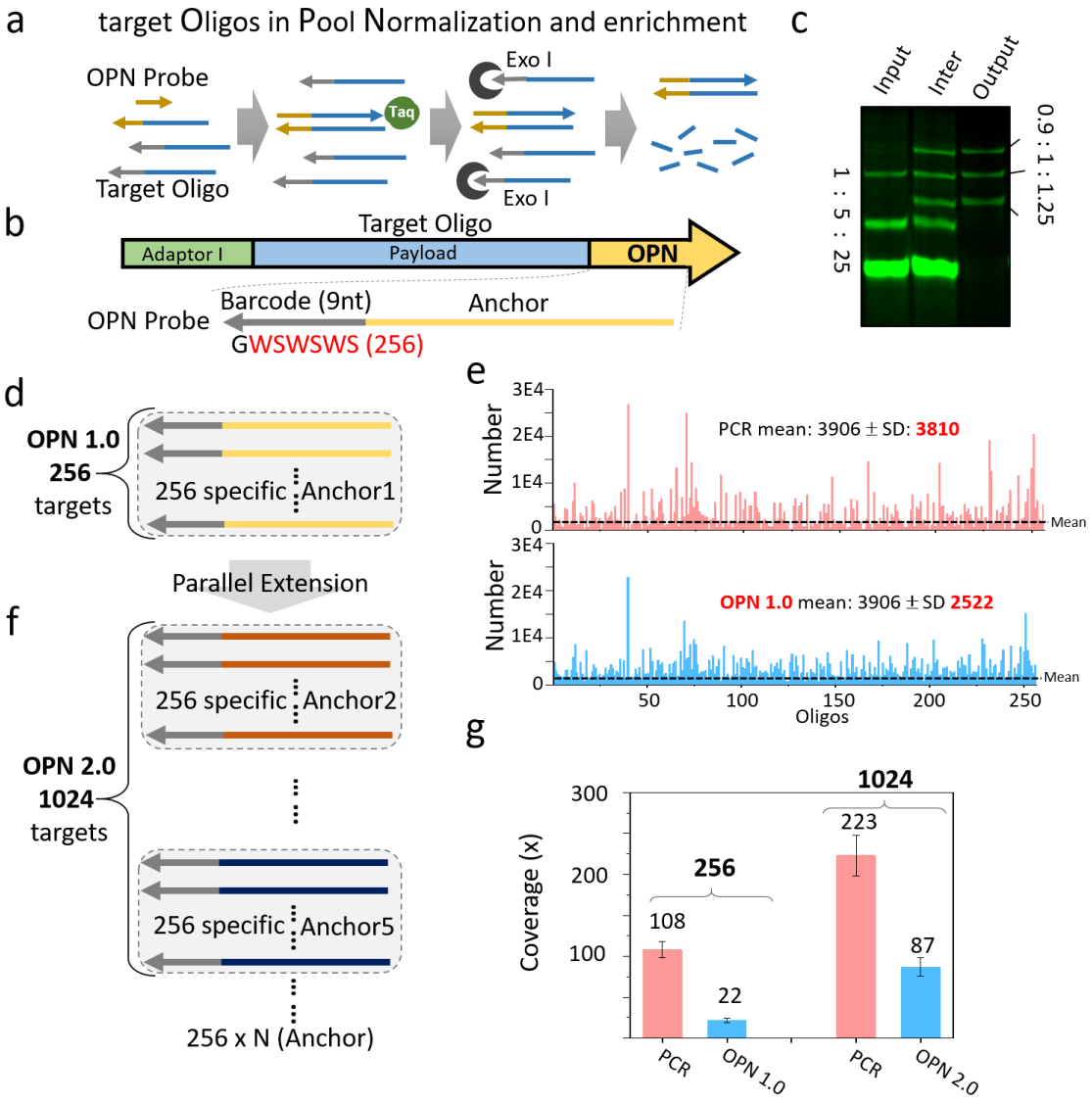

Figure 4 Large scale normalized oligo pool improved information read-out.

(a) Brief workflow for OPN probe mediated oligo pool normalization. Oligos are captured by the corresponding OPN probe and extended to dsDNA by Taq DNA polymerase. Then the remaining ssDNA oligos and probes are digested by exonuclease I.

(b) Structure of oligo for oligo pool normalizing (OPN). The adaptor I region, R (nickase recognition sequence) region, payload, and OPN region. In OPN region, barcode sequence with 9 nts in length and a unique "GWSWSWS" pattern with alternating strong (G or C) and weak (A or T) bases (e.g. CACTGT or GTCTGA), by which 256 high specific binding could be generated and universal fragment (17 nts).

(c) Normalization of 3' terminal FAM labeled three oligos, 60 nts, 73 nts and 90 nts in length, with input molar ratio of 1:5:25 and the output oligos were quantified as molar ratio of 0.95:1:1.25 on 15% native PAGE after OPN.

(d) OPN 1.0 constructed on one anchor sequence with 256 specific target oligo recognition.

(e) The copy number of each oligo of 256 oligo pool per million valid sequenced reads. 256 oligo pool amplified by PCR (red) was counted with mean copy number 3906 ± 3810 and (blue) 3906 ± 2522 for oligo pool revised by OPN 1.0 and then read by iDR.

(f) OPN 2.0 constructed from 4 distinct anchor sequences and one set of 256 barcode with 1024 specific target oligo recognition.

(g) The minimal necessary sequencing reads for complete oligos retrieve without dropout for 254 oligos pool read by PCR and OPN 1.0-iDR respectively, and 1024 oligo pool read by PCR and OPN 2.0-iDR respectively.

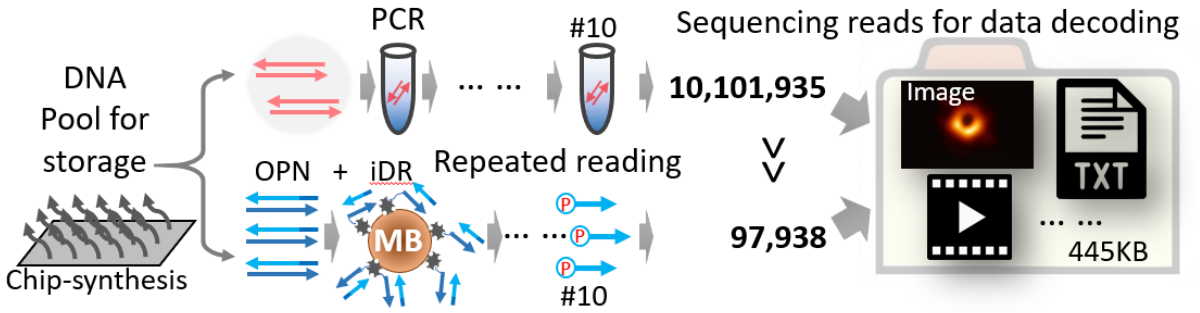

Figure 5. Sustainable and repeated DNA storage.

Chip-synthesized DNA oligo pool, in which digital information including image, video and text, were stored, was repeated read by PCR and OPN-iDR respectively. For current used PCR, 10,101,935 NGS noise reads was necessary for perfect decoding 445KB digital files, but two orders of magnitude less only 97,938 NGS noise reads were necessary for OPN-iDR.

ONLINE METHODS:

DNA Master Oligo Pool. The pool (Pool 1/2--Twist Bioscience; Pool 3/4--CustomArray) was resuspended in 1x TE buffer for a final concentration of 2 ng/uL. PCR was performed using Q5 High-Fidelity DNA Polymerase (NEB). Mix 10 ng of ssDNA pool (5 uL) with 2 μL of 100 μM of the forward primer and 2 μL of 100 μM of the Adaptor 2 (Adaptor 2-1), 10 μL 5x Q5 Reaction Buffer, 4 μL of 2.5 mM dNTPs, 0.5 μL Q5 High-Fidelity DNA Polymerase. Thermocycling conditions were as follows: 5 min at 98℃; 10 cycles of: 30 s at 98℃, 30 s at 56℃, 15 s at 72℃; followed by a 5 min extension at 72℃. The reaction was then purified according to the instructions in the Eastep Gel and PCR Cleanup Kit and eluted in 50 μL DNase/RNase-free water. This library was considered the master pool. All primers we used are in Supplementary Table 2.

iDR reaction. 10 ng of DNA oligo out of the master pool was attached to 1 μL of Streptavidin Magnetic Beads (NEB). The iDR reaction mixtures contained 1 μL of the DNA template attached to the beads (10 ng/μL), 0.25 mM dNTPs, 2.5 μL 10x NEBuffer 2, 0.08 U/μL Nt.BbvCI (NEB), 0.16 U/μL KF polymerase (exo$^-$) (Vazyme), 4 μM T4 Gene 32 Protein, 0.2 mg/mL BSA, 0.5 μM Adaptor 2 (For production of ssDNA, adaptor 2 was not added.). The mixtures were incubated at 37℃ for 30 min. After amplification, the specific amplicon and the template was isolated through magnetic pull-down. In the process of repeated iDR, the template attached to magnetic beads was washed by Wash/Binding buffer (0.5 M NaCl, 20 mM Tris-HCl (pH 7.5), 1 mM EDTA) twice and mixed with the above mentioned iDR reaction mixtures except for the DNA template. The process was proceeded for 10 times. To retrieve the information, the amplified products were purified and then sequenced on one Illumina Hiseq 4000 platform with 150 paired-end cycles in Novogene (Supplementary Note 3). The process of optimization in detail are given in Supplementary Note 4.

PCR reaction. 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)), 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℃, followed by extension at 72℃ for 5 min. In the repeated PCR, each subsequent PCR reaction consumed 1 μL of the prior PCR reaction and employed 10 cycles in each 50-μL reaction. The PCR product was purified by Eastep Gel and PCR Cleanup Kit and eluted in 50 μL DNase/RNase-free water. Then we sequenced the PCR product on Illumina Hiseq 4000 platform.

OPN probe. Each probe contains two parts, from 5' to 3': a universal sequence (adaptor 2) and a oligo-specific barcode sequence. Each barcode sequence is comprised of a number of commutative strong (C or G) and weak (T or A) nucleotides according to earlier report[25]. Here, the length of barcode is 8 nucleotides, corresponding to a total of $2^8 = 256$ barcode instances (Supplementary Note 5).

Single-stranded oligo pool amplification (SOA). The schematic of SOA is illustrated (Supplementary Fig. 61). We used PCR with reverse primer (adaptor 2) modified with 5' phosphate to amplify an oligonucleotide library with specific barcodes sequences. Then PCR product was degraded from 5' phosphate groups to 3' direction by lambda exonuclease, thus conversion of linear double-stranded DNA to single-stranded DNA (ssDNA). The mixture was purified by Eastep Gel and PCR Cleanup Kit. Then 10% denaturing (7 mol/L urea) PAGE was used to analyze the degraded products. 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. More detailed steps are in Supplementary Note 10.

Oligo pool normalizing (OPN). The schematic of OPN is illustrated (Supplementary Fig. 61 and 62). The OPN probes were synthesized respectively (Supplementary Table 4 and Table 5). An equimolar mixture of OPN probes (256 or 1024), which the number of each OPN probe was equivalent to the average molar concentration of the oligo pool, was applied to capture the corresponding oligo separately. The 256/1024 oligos were mixed with 256/1024 OPN probes and hybridization buffer (10 mM Tris-EDTA, 0.5 M NaCl, and 0.05% Tween-20 (volume / volume)) in 20 μL reaction. 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. Then, extension reaction was carried out when the target was captured by corresponding OPN probe. To ensure temperature uniformity, the pre-reaction mixture containing Taq DNA polymerse and dNTPs was also pre-heated to 60℃ before adding to the ssDNA/OPN probe mixture. The resulting mixture was incubated for another 15 min at 60℃ to obtain dsDNA product. Further, Exo I was added to the resulting mixture to digest the remaining ssDNA and OPN probes. After Exo I was inactivated, Streptavidin Magnetic Beads were added and incubated for another 30 min at 37℃ in shaker to isolate dsDNA product. The dsDNA product attached to magnetic beads was the template for OPN-iDR. More detailed steps are in Supplementary Note 10.

Data availability. The BASIC code for encoding and decoding for both Linux and Windows and bioinformatic analysis programs may be obtained via (https://github.com/xiaomingao/DNA-information-storage). Furthermore, the original sequencing FASTQ file and the designed sequence file may be obtained via (http://pan.tju.edu.cn:80/#/link/627DB5C9EB819984F1183D8D4A0B72E3, Code: oZ3D).