Sustainable DNA data storage with low biased manipulation

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 repeatable and efficient information reading. Here, we demonstrated that biased amplification is serious matter for stable information retrieve from synthetic oligo pool comprising over ten thousand strands encoding over 2.85 MB digital information. Therefore, we adapted isothermal DNA amplification as a new biochemical framework, termed as isothermal DNA reading (iDR) for low biased manipulation of DNA pool with huge sequence complexity. iDR was built to work in a priming-free and error-spreading proof manner and achieved low-biased amplification. Finally, we demonstrated that immobilized DNA materials read by iDR is able to stably read data deeply and repeatedly, the first reported sustainable DNA storage system. Necessary amount of sequencing reads for perfect

oligos retrieve significantly decreased. These advanced features enable the iDR chemical framework being ideal alternative 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 applications, e.g., probe blend, DNA origami assembly, and genome synthesis[1]. Because of both the location on microchip and DNA sequence context, the copy number for each oligo strand from array synthesis is distinct[5]. 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 DNA materials, over a few hundred nanograms, at the concentration of micromolar is necessary for high quality DNA sequencing covering all oligos in the pool on commercial DNA sequencing platform, e.g., Illumina[8]. Therefore, amplification is crucial to boost the signal for the subsequent DNA sequencing for stably 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]. The unevenness of copy number generated a huge complexity that caused serious problem for DNA molecule retrieval and data decoding[4, 5, 12, 13]. Thus far, in 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 by great effort to develop high encoding, physical and sequencing redundancy[13, 15]. However high redundancy sacrificed storage density and increased cost in 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 amplification[16]. Moreover, successive repeated amplification significantly skewed the copy number distribution and caused large dropout[12]. Even the sequences were carefully designed to minimize the sequence context difference, biased amplification still caused chaos from the huge sequence complexity of oligo pool for DNA storage and then imperiled the information decoding. And for this pure engineering approach of DNA storage, it is valuable to develop alternative technology with distinct features.

For practical data storage, crucial issues have to been addressed. 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. The current thermal cycling process in DNA amplification 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 under more simple thermal condition.

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%. For amplification of oligo pool with high sequence complexity, we challenge to develop alternative system for low biased amplification and construction of repeatable DNA storage. Statistics analysis demonstrated that oligo pool was more prone to dropout than base error and oligo dropout significantly increased with the amplification depth and repeated times in biased amplification of PCR. Therefore, we recruited the isothermal DNA replication to challenge the stable manipulation of DNA pool with huge sequence complexity. Isothermal DNA reading (termed as iDR) was specifically built to address the crucial issues of inefficient-priming, product-as-template that caused the replication bias and error-spreading and oligo dropout remained very stable at low level. Finally, we demonstrated that immobilized DNA pool read by iDR as an efficient and repeated DNA system for sustainable information storage. For the development of practical DNA “hardware”, and iDR could work with PCR together in the different process, e.g., material preparation and stably reading respectively. To the best of our knowledge, this is the first sustainable DNA storage system successfully achieved the stable information “deep reading”.

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). Therefore, DNA storage can be considered as distributed storage system built on biomaterial. Due to this nature, we adapted a BASIC coding system that is one well optimized regenerating code with reducing computational complexity for distributed storage [18, 19](Ref.). Generally, the encoding process started by dividing the target file into non-overlapping groups and the split information was encoded into DNA sequence following an optimized encoding process (fig. S1 and Supplementary Note 1) with two adjustable parameter K (corresponding to oligo number in one non-overlapping group, K was set as constant value of 256) and L (corresponding to the length of oligo), which is flexible for various DNA oligo length from different commercial synthesis platform. Besides the normal file types, for the first time we tested the storage of genome sequence including human mitochondrion and one artificial bacteria cell (Supplementary Note 2) with a different encoding strategy. Genome nucleotide sequence was converted to binary file following a simple principle of T to 00, G to 01, C to 10 and A to 11, by which higher information storage density could be obtained as one nucleotide base to 2-bits binary in comparing with simply treating nucleotide letter as text contents. Because in general text computer system, e.g., ASCII, one letter is encoded by 8-bit binary. In nature 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. Furthermore, oligo pools with various payload length have been designed to store 2.85 MB files

in totally 109,568 oligo strands (Fig. 1A, fig. S2A and table S1). Reed-Solomon code was used for error correction and coding redundancy for tolerance of missing entire oligo, causing dropout. Relative high information density was 1.65 bits/nt with a 1.56% of coding redundancy, which was calculated by previous reported method without modification (Supplementary Note 3)[13]. Success decoding could be achieved as long as oligo dropout less than 1.56%, randomly losing 4 strands from 256 (fig. S1B). Technically, higher encoding redundancy tolerates losing more oligos and success decoding can be achieved from lower sequencing coverage. However, higher redundancy requires more DNA synthesis. Considering the synthesis cost is higher than sequencing, represented over 90% of the total cost of DNA storage in our experiment, it is more practical to trade-off encoding redundancy for synthesis cost.

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

Low-bias and error-spreading proof isothermal amplification

It well known that PCR generated biased amplification from its inherent mechanism, i.e., product-as-template, priming and thermal cycling dependent amplification[3, 14, 20, 21](Ref.再多引用一些纯biology 的好的文章). In order to address these problems that easily skewed the oligo copy number

distribution, we recruited isothermal DNA replication, which are majorly used in systems for biomolecular analysis or complex DNA circuit network [22-24]. To the best of our knowledge, strand displacement replication has not been tested for stable amplification of large DNA library, not the size of library in DNA storage. It is still unknown if huge sequence complexity could be stably amplified by this novel DNA replication mechanism. Therefore, we challenge to build low biased isothermal amplification of DNA pool with high sequence complexity. After systematical optimization of reaction temperature, reagent component, replicon length and nickase recognition site sequence (figs. S2B-E), Nt.BbvCI nickase and exonuclease deficient DNAP I Klenow fragment (3'→5' exo$^-$) were used and the amplification efficiency of iDR is around 70% of PCR (fig. S2F). The immobilized isothermal DNA replication system was designated as iDR, isothermal DNA reading for large DNA pool amplification (Fig. 2A). The Nt.BbvCI nickase recognition site is only 7nts in length and its sequence has been avoided from the payload part of oligo pool in the file encoding step. By comparison, it will not cause much extra burden on synthesis.

The intrinsic features of iDR specifically facilitate DNA storage application. In theory, for one 50 ul reaction, 10 PCR thermal cycles required about 177.8J energy only for thermal regulation at least two orders of magnitude more than 30 mins iDR of 2.52J (Fig. 2B and Supplementary Note 4) and it may be a huge issue at large scale of operation. Real time monitoring indicated that 30 mins is the optimal amplification time (fig. S2G) and the amplification time was set as 30 mins for all experiments in this manuscript. Single-stranded or double-stranded DNA can be produced in a controlled manner (Fig. 2C). Specially, ssDNA was amplified in a primer-free manner and allows iDR to be a universal method for reading information with no sequence information needed in advance. The nickase mediated site-specific phosphodiester bond cleavage initiated the iDR amplification and generated a 5' terminal phosphate group, its function was verified by direct

ligation to a FAM labeled probe (Fig. 2D). This inborn phosphate group is very convenient for subsequent functional adapter linking.

Amplified oligos were sequenced on commercial Illumina Hiseq 4000 platform with 150 paired-end cycles (Supplementary Note 5) and then deeply analyzed by a set of statistics methods developed from bioinformatic BLAST program (fig. S3 and Supplementary Note 6). Sequenced reads with various number of letter error including substitution and indel were different. Among them, single letter error accounted for the majority in mutant sequenced reads for both PCR (80.1%) and iDR (81.7%) (Fig. 2E and fig. S4). The total indel (0.03%) and substitution (0.2%) rate were consistent with previous studies[13]. Although, RS code embedded in BASIC code is able to correct multiple errors in same DNA strand but will increase the computation complexity largely[17, 25]. Therefore, both sequenced reads with no or single substitution/indel error were collected as valid reads for further analysis and decoding (fig. S1B).

The distribution of the number of reads per each given sequence were different between iDR and PCR of various thermal cycles of from 10 to 60. The distribution normality was quantified by a modified mathematical statistics function (Supplementary Note 6). In theory, oligo pool of higher value distribution normality require less sequencing reads to recover all synthesized oligos[12]. The coverage distribution of both PCR and iDR were positively skewed with a long tail, which comprising of high copy number oligos, the normality of PCR decreased from 0.50 to 0.18 with thermal cycles increased from 10 to 60 (Fig. 2F). However, 0.60 was obtained for iDR, it indicated that most balanced DNA pool was amplified by iDR. The proportion of oligos with high copy number, the top 30% high coverage, enlarged with the PCR thermal cycles increased (Fig. 2G). In comparison, the dropout rate increased significantly from 4.18% to 13.80% with PCR amplification becoming deeper, much higher than 1.33% of iDR (Fig. 2G). No obvious difference

was observed between the coverage distribution of iDR amplification from free oligo pool and oligo pool immobilized on magnetic beads (fig. S5).

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. Therefore, for same oligo, if the coverage in M0G0 is smaller than the coverage in M1G1, it indicated that for this oligo sequencing reads with mutant is more than correct one in the pool. It is hard to identify the correct oligo from sequencing reads by principle of law of large number-makers and will tend to influence the correct decoding. We identified this as number of deep-error, it means how many oligos with more mutant sequencing reads than correct reads. It was observed that PCR with various thermal cycles gave average number of 177 for 10 cycles, 350 for 30 cycles and 1175 cycles for 60 cycles, all of them is larger than iDR of 46 (Fig. 2H). It indicated that biased amplification spread and accumulated error reads.

Low replication fidelity of DNA polymerases and around 1% miss reading coming along with NGS sequencing process[26] 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[16, 17]. The decay lost rate in a 10 cycles PCR was calculated as 21.8‰ and 0.035‰ for iDR

(Fig. 2B and Supplementary Note 7). Finally, all the files in DNA pool was successfully decoded following the BASIC decoding process and the minimal mean coverage depth of sequencing reads were calculated (Table S2)

DNA storage for repeated reading

Repeatable reading is crucial for practical DNA storage system. So far, different repeated reading systems have been proposed and achieved, in which generally DNA pool was amplified from aliquot of previous amplified product (Fig. 3A). However, error rate and oligo dropout were significantly accumulated through this repeated amplification[12, 13, 16]. Due to the low biased amplification performance, we believe that DNA storage for stable repeated reading could be achieved from iDR system. In this isothermal system (Fig. 3B), both the low biased iDR amplification performance and DNA pool immobilization on microbeads could contribute to the stability in repeated reading, but either of them has not been systematically tested so far. Therefore, we propose the isothermal reading of immobilized DNA pool as stable and robust DNA storage system.

The amplification capability for repeated 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. 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. 3C). 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. 3D). This result was in agreement with previous experiment and indicated that error accumulated to high copy number during PCR and iDR achieved error-spreading proof amplification, in which amplified DNA strand with error was not further amplified as template.

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. 3E). 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 (fig. S6). 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.3F), 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 3), the crucial parameter for the theoretical minimum decoding coverage. For PCR, the minimum coverage depth was quantified as 17.2x (#1 PCR) and 167x (#5 PCR) and 426x (#10 PCR). For iDR, the minimum coverage depth was quantified as 11 for #1, 12 for #5 and 12.5 for #10. Therefore, for 445KB file

encoded in 11520 DNA oligo pool, 10,101,935 and 149, 919 minimal sequencing reads were necessary for perfect decoding (Fig. 4 and Supplementary Note 8).

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 of immobilized DNA pool was stable, robust and suitable for repeated reading in DNA data storage.

Discussion:

DNA data storage is one very artificial application and actually the concept has been proposed for a long time[27]. 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[28], 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. 4 and Supplementary Note 8).

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. Each of these will raised the cost for handling the large DNA pool. Although, iDR is adapted from stand displacement replication, but this is the first system applying this amplification mechanism on DNA pool with high sequence complexity for artificial information storage and demonstrated that immobilized DNA pool read by iDR is able to stably read data deeply and repeatedly. Originally, the core thing of DNA storage is adapting current biotechnology for artificial purpose of storing information at a large scale and accuracy which is distinct for bioengineering approaches. Undeniably, PCR possesses power in DNA amplification, and iDR could work with PCR together in the different steps of DNA storage, e.g., material preparation and stably reading respectively.

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[29, 30]. 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. Therefore, we believed that this new biochemical framework with stable repeated reading 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. van Dijk, E.L., Jaszczyszyn, Y. & Thermes, C. Library preparation methods for next-generation sequencing: tone down the bias. *Exp Cell Res* **322**, 12-20 (2014).

21. Dabney, J. & Meyer, M. Length and GC-biases during sequencing library amplification: a comparison of various polymerase-buffer systems with ancient and modern DNA sequencing libraries. *Biotechniques* **52**, 87-94 (2012).

22. Zhao, Y., Chen, F., Li, Q., Wang, L. & Fan, C. Isothermal Amplification of Nucleic Acids. *Chem Rev* **115**, 12491-12545 (2015).

23. Reid, M.S., Le, X.C. & Zhang, H. Exponential Isothermal Amplification of Nucleic Acids and Assays for Proteins, Cells, Small Molecules, and Enzyme Activities: An EXPAR Example. *Angewandte Chemie* **57**, 11856-11866 (2018).

24. Zhang, D.Y. & Seelig, G. Dynamic DNA nanotechnology using strand-displacement reactions. *Nat Chem* **3**, 103-113 (2011).

25. 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).

26. Yazdi, S.M., Yuan, Y., Ma, J., Zhao, H. & Milenkovic, O. A Rewritable, Random-Access DNA-Based Storage System. *Sci Rep* **5**, 14138 (2015).

27. Wallace, M.R. Molecular Cybernetics: The Next Step? *Kybernetes* **7**, 265-268 (1978).

28. Extance, A. How DNA could store all the world's data. *Nature* **537**, 22-24 (2016).

29. Newman, S. et al. High density DNA data storage library via dehydration with digital microfluidic retrieval. *Nat Commun* **10**, 1706 (2019).

30. 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).

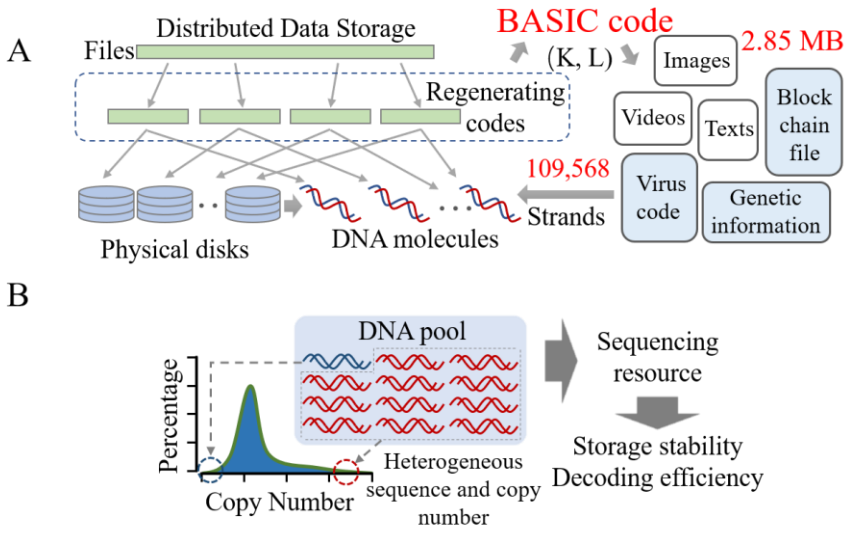

Figure 1: Overview of the DNA data storage.

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

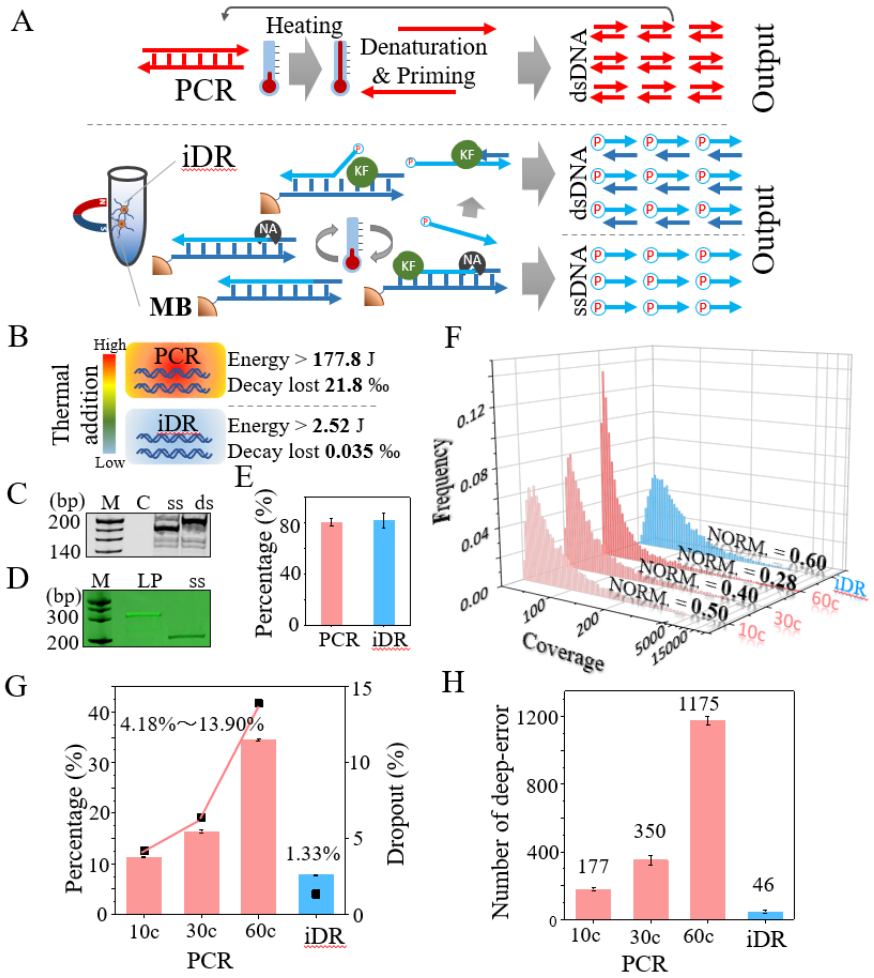

Figure 2: Designed isothermal DNA reading (iDR) for DNA data storage.

(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, 30 and 60 cycles PCR and iDR amplification were positively skewed but with 0 oligo dropout. The distribution normality was 0.50 (10c), 0.40 (30c), 0.28 (60c), 0.60 (iDR) respectively.

(G) 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. Oligo dropout rate calculated from random reads set with a mean 10x coverage depth was plotted. For 10, 30 and 60 thermal cycles PCR (red line), it was counted as 4.18%, 6.41% and

13.90% respectively, and 1.33% for iDR (blue dot). Error bars represent the mean ±s.d., where n = 10.

(H) The number of reads with mutant in a high copy number. The number increased significantly from 177±9 to 1175±26 as the number of PCR cycles increased from 10 to 60. The number of iDR was 46±10.

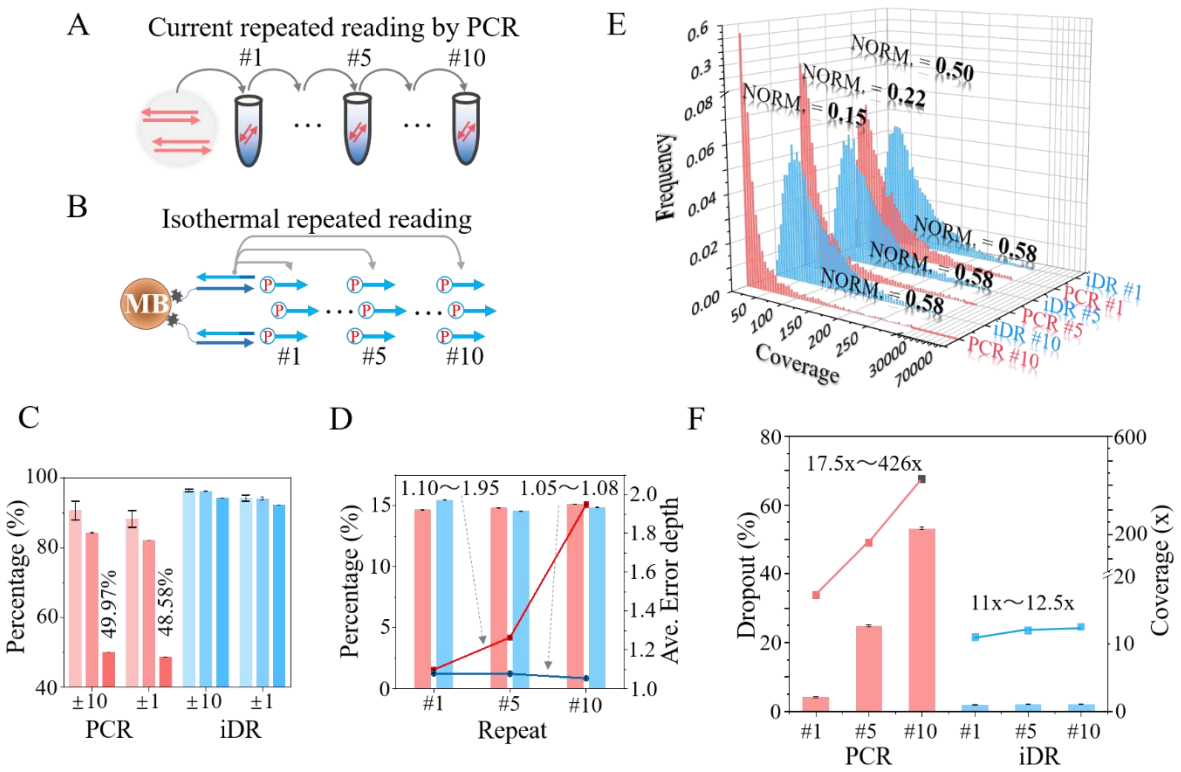

Figure 3: Multiple information reading from DNA storage.

(A) Illustration of successive DNA reading. Repeated PCR amplification was successively performed 10 times with aliquot of previous reaction as template.

(B) 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.

(C) 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.

(D) 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.

(E) 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.

(F) 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.

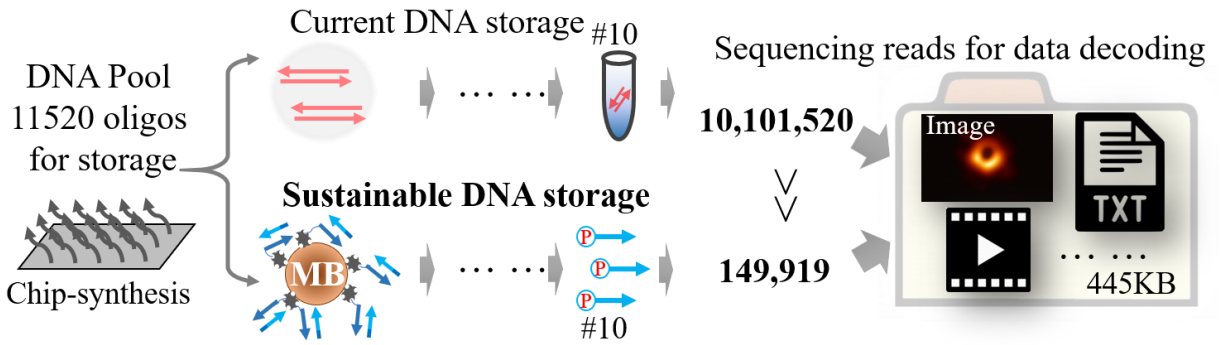

Figure 4. Sustainable and repeated DNA storage.

Chip-synthesized DNA oligo pool, in which digital information including image, video and text, were stored, was repeated read by PCR and iDR respectively. For current used PCR, 10,101,935 NGS noise reads was necessary for perfect decoding 445KB digital files, but only 149,919 NGS noise reads were necessary for iDR.

ONLINE METHODS:

DNA Master Oligo Pool. The pool (Pool 1/2--Twist Bioscience; Pool 3--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 table S3.

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 5).

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.

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).