Towards Practical and Robust DNA-based Data Archiving by Codec System Named 'Yin-Yang'

Zhi Ping[1,2,3,†], Shihong Chen[1,2,4,†], Xiaoluo Huang[1,†], Sha Joe Zhu[6], Chen Chai[1,4], Haoling Zhang[1,2,3], Henry Lee[3,7], Guangyu Zhou[3,7], Tsan-Yu Chiu[1,3], Tai Chen[1,3,4], Huanming Yang[1,5], Xun Xu[1,2,4,*], George M. Church[3,7,*], Yue Shen[1,2,3,4,*]

†These authors contributed equally to this work.

*To whom correspondence should be addressed (shenyue@genomics.cn, gchurch@genetics.med.harvard.edu, and xuxun@genomics.cn)

[1]Guangdong Provincial Key Laboratory of Genome Read and Write, Shenzhen Engineering Laboratory for Innovative Molecular Diagnostics, Guangdong Provincial Academician Workstation of BGI Synthetic Genomics, BGI-Shenzhen, Guangdong, China

[2]Shenzhen Institute of Synthetic Biology, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, People's Republic of China

[3]George Church Institute of Regenesis, BGI-Shenzhen, China

[4]China National GeneBank, BGI-Shenzhen, Jinsha Road, Shenzhen, 518120, China

[5]James D. Watson Institute of Genome Sciences, Hangzhou 310058, China

[6]Big data Institute, University of Oxford, Li Ka Shing Centre for Health Information and Discovery, Old Road Campus, Oxford, OX3 7LF, United Kingdom.

[7]Department of Genetics, Harvard Medical School, Boston, Massachusetts, USA.

Abstract

Motivation: DNA has been reported as a promising medium of data storage for its remarkable durability and space-efficient storage capacity. Here, we propose a robust DNA-based data storage method based on a new codec algorithm, namely 'Yin-Yang'.

Results: Using this strategy, we successfully stored different formats of files in one synthetic DNA oligonucleotide pool. Compared to most DNA-based data storage coding schemes presented to date, this codec system can efficiently achieve a variety of user goals (e.g. reduce homopolymer length to 3 or 4 at most, maintain balanced GC content between 40% and 60% and simple secondary structure with the Gibbs free

energy above -30 kcal/mol). We tested this codec by an end-to-end experiment including encoding, DNA synthesis, sequencing and decoding. We demonstrate successful retrieval of 2.02 Megabits /3 files using this method. The original information was fully retrieved after sequencing and decoding. Compared to the previously reported methods, our strategy exhibits great potential at achieving high storing capacity per nucleotide (230 PB/gram) and high fidelity of data recovery.

Introduction

DNA is the most ancient information carrier designed by nature. At present, it is considered to have great potential to be a novel storage medium, as the orthodox storage media can no longer meet the exponentially increasing data archiving demand. DNA molecule, compared to orthodox information carriers, exhibits multiple advantages; including extremely high storage density (455 EB or 455×10^{18} bytes per gram theoretically[1]), extraordinary durability (an estimated half-life of DNA in bone was 521 years[2]), and the capability of cost-efficient information amplification.

Multiple strategies were put forward for digital information storage using organic molecules, including oligopeptides[3] and metabolomes[4]. Recently, 'DNA memory' using DNA origami technology was proposed to be a novel strategy to store digital information. However, it is well-acknowledged that data archiving using synthetic DNA is still the most practical and efficient strategy currently in this field. The binary information extracted from files is transcoded directly into DNA bases and then synthesized and stored in the form of oligonucleotide or fragment. Since DNA sequences will be difficult for accurate synthesis and sequencing while they have specific patterns, e.g. abnormal GC content or long single-nucleotide repeats (homopolymers)[5]. Therefore, it is critical to maximize the coding efficiency while maintain the feasibility of synthesis and sequencing. In the past few years, several different coding schemes were developed to achieve high coding efficiency. Some well-accepted algorithms, including Church's 'simple code'[1], Goldman's code[6], Grass' code[7] and Erlich's DNA fountain[8], also improved practical feasibility of synthesis and sequencing and data fidelity by reducing homopolymers, maintaining balanced GC content, while allowing for the introduction of error-correction codes and other redundancies[9].

In order to achieve robust DNA-based data storage with high data density as well as reliability, we propose a novel binary-DNA codec algorithm and demonstrate its application in DNA-based data storage. It is named after a traditional Chinese concept, 'Yin-Yang'. 'Yin' and 'Yang' represent two different coding rules, which are incorporated subsequently to transcode binary information to sequences of 'A/T/C/G' (Supplementary information). This codec algorithm can effectively reduce long homopolymers and maintain balanced GC content with reduced secondary structures for synthesized DNA sequences. It could also achieve a coding efficiency of ~2 bits/nt, reaching the theoretical maximum.

The general principle of 'Yin-Yang' codec algorithm (referred as 'YYC' hereinafter) is to incorporate two pieces of information in binary and transcode them into one DNA sequence according to two independent rules, 'Yin' and 'Yang'. Both rules can derive to 256 and 6 different specific rules respectively. The incorporation of these two rules can derive a pool of 1536 different ways to encode the binary data and thus be possible to broaden the applications (Fig. 1A, Fig. S1).

Taking Rule No. 888, a random selected incorporated rule, as an example, in 'Yang' rule, {'A', 'T'} represents binary digit '0' and {'G', 'C'} represents binary digit '1'. In 'Yin' rule, the local nucleotide (current nucleotide to be encoded) is represented by the incorporation of the previous nucleotide (or 'supporting nucleotide') and the corresponding binary digit (Fig. 1B). In practice, the two rules are applied respectively for two independent binary segment and transcoded into one exclusive DNA sequence, and vice versa. Taking the first nucleotide in Fig. 1C as a simple example, the corresponding binary digits from segment 'a' and 'b' are '1' and '0' respectively. Therefore, according to 'Yang' rule, the local nucleotide will have two options, {'C', 'G'}. On the other hand, according to 'Yin' rule, the virtual supporting nucleotide is set as 'A' by default, so previous nucleotide 'A' and corresponding binary digit '0' will give the local nucleotide two options, {'A', 'G'} (Fig. 1B). Therefore, the result encoded local nucleotide will be the intersection from 'Yang' rule, {'C', 'G'} and 'Yin' rule, {'A', 'G'}, which provides the only option {'G'}, and so on. Moreover, it is also crucial to pay attention to the corresponding rules of each segment, because the switch of binary segments will also change the transcoded result (Fig. 1C), which means, {a: 'Yang', b: 'Yin'} and { b: 'Yang', a: 'Yin'} will result in completely different result in

this case. The simple example of transcoding demonstration process is also illustrated in supplementary information (Fig. S3).

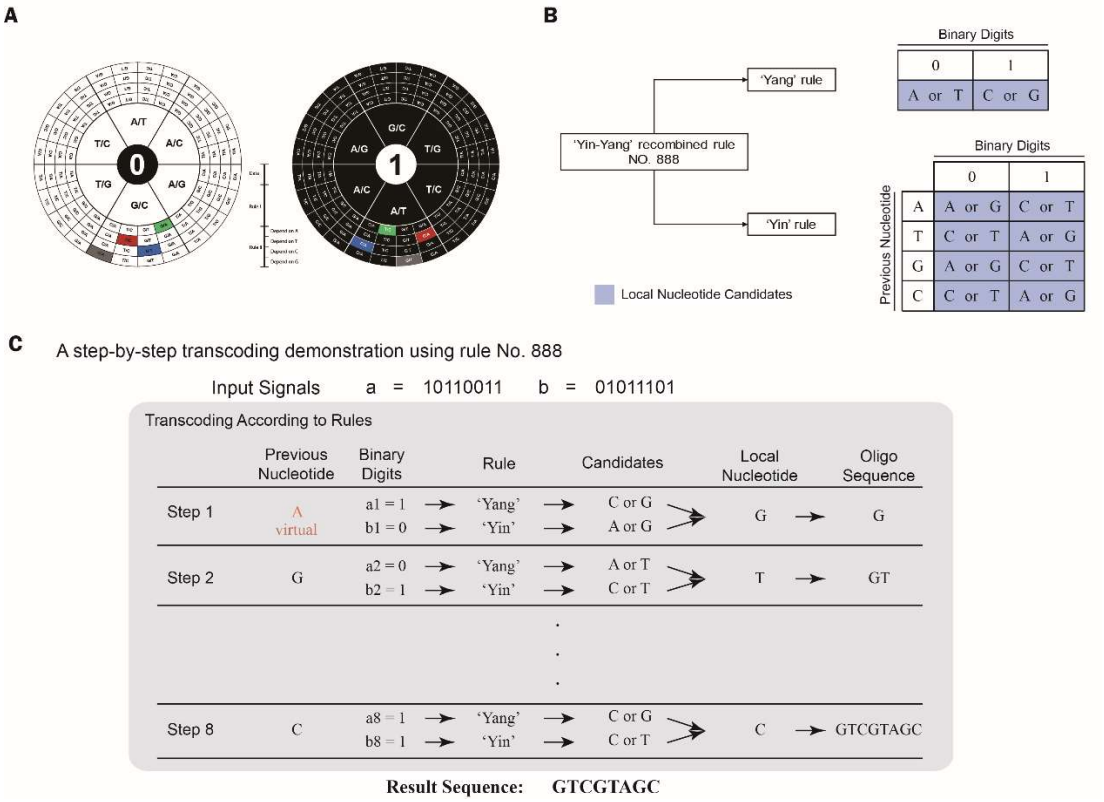


Figure 1 (A) 'Yin-Yang' incorporation rules can provide a pool of 1536 (6×256) different incorporated rules. See Supplementary Information for details. (B) An illustration of 'Yin-Yang' incorporated rule No. 888. (C) A step-by step demonstration of YYC transcoding process, rule No. 888 is used in this example. 'a1', 'b1' represents the first-position binary digit in segment 'a' and 'b', and so on. Virtual base A means this base is used only for determination of the first output base and will not be appear in the result sequence.

Results

High coding efficiency

Coding efficiency is always one of the key features of DNA-based data storage coding scheme evaluation. For current DNA codec algorithms based on natural nucleic acids (Adenine, Cytosine, Guanine and Thymine), the theoretical limit of coding efficiency is 2 bits/nt according to Shannon's information theory[10]. The novel strategy of 'Yin-Yang' codec algorithm (Methods) provides an extremely high coding efficiency. In Table.1, we compare 'Yin-Yang' codec with other current coding algorithms under

identical constrains, including length of oligo, index, flanking sequence, error-correction code and copy-number of oligos in DNA synthesis,. Apparently, 'Yin-Yang' coding scheme exhibits the highest coding efficiency. In practice, in this work, redundancy was introduced in case of oligo loss, resulting a real information coding density of 1.27 bits/nt. However, in this demonstration, the proportion of index, error correction and primer bases (non-data region) in one oligo was $\frac{14+8+\ \times 2}{182} = 34\%$ (Fig. 3B). Therefore, with length of synthesized DNA increasing and length of non-data region remaining identical, the net coding efficiency will increase and ideally reach its maximum potential (Supplementary Fig.S4). Therefore, the eventual coding potential of 'Yin-Yang' compiling system is 2 bits/nt, which achieves the theoretical limit. Beside theoretical information efficiency, we also calculated the physical density using different encoding algorithms. The physical density (Bytes/gram) can be calculated through the following formula:

$$\rho = n \times (M\ molecules \times \frac{200nt}{molecule} \times \frac{320Dalton}{nt} \times \frac{1.67 \times 10^{-24} g}{Dalton})^{-1},$$

where n is the number of bytes each oligo carries, and M is the average copy number of each oligo. Erlich et. al reported that in their working scheme, the minimum average copy number for data retrieval was around 1.3×10^3. In this work, we demonstrated that the minimum average copy number for data retrieval is 7×10^3 (Supplementary information), which is consistent with the previous report. Therefore, for the calculation of physical density, 7000 is used as average copy number.

Table 1 Comparison of information efficiencies of current coding algorithms used in DNA-based data storage. The information efficiency is calculated with identical parameters: length of oligo length (200nt), index (16nt), flanking sequence (24nt×2), error-correction code (8nt) and copy-number of oligos in DNA synthesis (7000 copies per oligo type).

Coding Algorithm	***Information efficiency***	***Physical density PB/g (Copy* No. *= 7000)***	***Physical density PB/g (Copy* No. *= 1300)***	***Theoretical Coding Potential***
Simple code	0.64 bits/nt	21.39	115.2	1 bit/nt

Goldman *et. al*	1.008 bits/nt	33.68	181.4	1.58 bits/nt
Grass *et. al*	1.138 bits/nt	38.03	204.8	1.78 bits/nt
Blawat *et. al*	1.024 bits/nt	34.22	184.3	1.6 bits/nt
DNA Fountain	1.196 bits/nt	40.00	215.2	1.98 bits/nt*
'Yin-Yang' code	1.28 bits/nt	42.71	230.3	2 bits/nt

*** based on length of seed.**

Better biochemical features for synthesis and sequencing

To prove the feasibility of 'Yin-Yang' codec algorithm and quantify its featured parameters compared to other well-accepted DNA-based data storage coding schemes[1, 6, 7, 8], a collection of different formats of files including text, image, audio and video files, with the total size of ~1 GB, were transcoded using different algorithms in silica. Statistics of GC content, free energy of secondary structure, and data retrieval efficiency based on the number of read DNA oligo, are obtained and analyzed. For previous coding schemes except DNA fountain, most of the DNA oligos possess relatively normal GC content (40% - 60%) (Fig. 2A). However, sequences with high or low GC ratio still exist, which might cause difficulties for synthesis and sequencing. In contrast, YYC and DNA fountain can maintain the GC content in certain range (40% - 60% in this work). Because in both coding schemes, the incorporation of binary segments is flexible, which provides more possibilities to obtain DNA sequences with desired GC content. Furthermore, a procedure called 'validity-screening' helps to filter the DNA sequences which do not meet the criteria (Fig. 2B, Supplementary information Fig. S2).

YYC gives oligos which possess a higher free energy larger than -30 kcal/mol, while other coding schemes did not take free energy into consideration and thus provide different amount of result sequences possess a lower free energy smaller than -30 kcal/mol (Fig. 2C). Collecting from the empirical data of DNA supplier, oligos with

free energy predicted less than -30 kcal/mol tend to form more stable secondary structure and may cause troubles for molecular manipulation such as PCR amplification.

Unlike DNA fountain, which employs the algorithm originally designed for communication channel, other DNA encoding schemes exhibits a linear retrieval pattern in which the amount of recovered information is proportional to the number of sequences read[9]. In contrast, DNA fountain employs a different data retrieval strategy based on its grid-like topology of data segments, and will recover all the data if enough segments are read[8]. However, if the number of data segments read does not exceed necessary amount, the data will all lost (Fig. 2D).

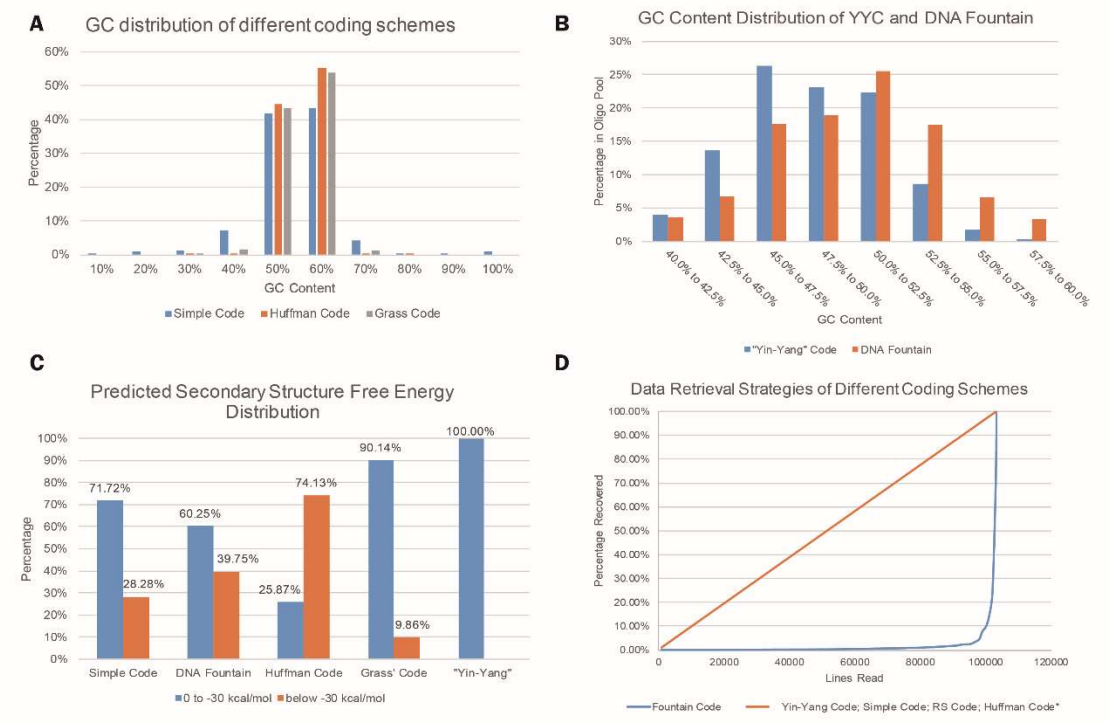


Figure 2. A) Distribution of GC content of generated oligos that transcoded using Simple code (blue), Huffman code (orange) and Grass' code (grey); B) Distribution of GC content of generated oligos that transcoded using YYC (blue) and DNA fountain (orange); C) Percentage of free energy of predicted secondary structure of generated oligos that transcoded using different codec algorithms, red: free energy between 0 to -30kcal/mol, blue: free energy below 30kcal/mol; D) Data retrieval strategies of different coding schemes, DNA fountain (blue) exhibits an exponential growth curve with a long lag-phase, while other coding methods (orange) suggest a linear growth curve.

Experiment suggests feasibility of YYC codec system

As a proof of concept, the 'Yin-Yang' codec system was applied to encode three digital files into a synthesized oligo pool, including a 113 kb JPG image showing the mechanism of DNA replication[11], a 97 kb English poem collection from the Shakespearian Sonnets, and a 36 kb ancient Chinese classic - Tao Te Ching (Fig. 3A). The oligos in the synthesized pool possess the same structure, including flanking primer regions (20 nt in both ends), data payload region (120 nt), hamming code region (8 nt) and index region (14 nt), like Fig. 3B shows. The transcoding gave 11098 oligo sequences with 182 nt each, including 33.3% of redundancy in case of sequence loss.

After the synthesized DNA oligo pool was obtained from the supplier, PCR amplification generated a double-stranded DNA library for sequencing afterwards. It is shown that by using different pairs of primers, we can amplify specific files from the pool. The quantities of each file are also of consistent quantity according to gel electrophoresis (Fig. 3C), which suggested the good synthesis quality as well as the feasibility of random-access. A sequential dilution (100/1000/10000x) was performed on each file to determine the average number of oligos for successful decoding (Supplementary Fig. 3).

The amplified DNA library was then deep-sequenced (average sequencing depth = 4000x) using BGI-Seq-T1 NGS sequencer. The sequencing data was obtained and analyzed afterwards. For accuracy, reads with length not equal to 182 was first removed and gave approximately half of the original sequencing data. For sequencing depth, the three different files exhibit a uniform distribution, which is consistent with common sequencing depth distribution (Fig. 3D). The sequencing data shows that, with sequencing in this instrument, even with high sequencing depth, loss of few oligos is inevitable. This might be caused from many processes, including synthesis, necessary manipulation like PCR amplification, probe capture assay or other molecular approaches, storage, transportation of samples, and even sequencing itself. It is concluded that to prevent loss of data, certain amount of redundancy is one of the requisites. Errors introduced in DNA synthesis and sequencing were considered to be a major issue in data retrieval. Hence, error correction codes are used to ensure the fidelity. However, it is interesting that, with the sequencing depth increasing, the errors can be corrected by mutual calibration and thus decrease. Although errors are still inevitable, since redundancy is necessary, it is possible to retrieve correct data by mutual calibration using redundancy data instead of error-correction codes. Therefore, it can

be suggested that with sufficient sequencing depth, extra redundancy is necessary for DNA-based data storage, but not error-correction codes (Supplementary Fig. 4).

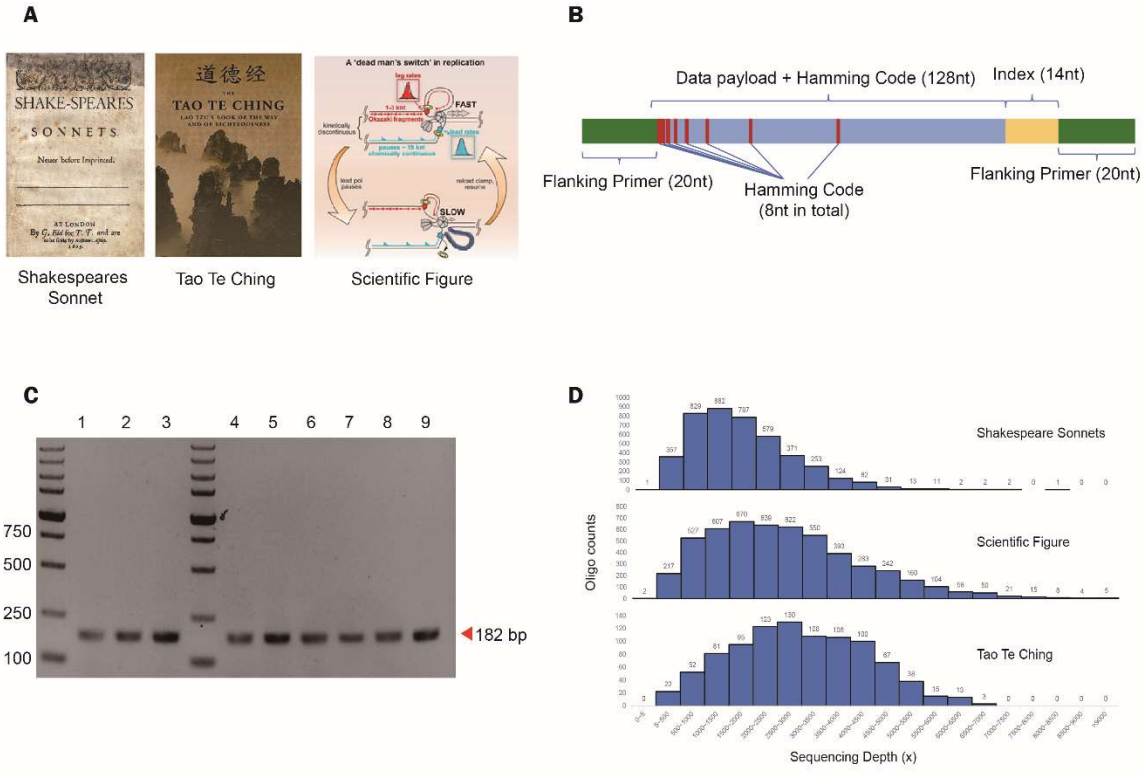


Figure 3 (A) Three files used for DNA-based data storage in this work, including Chinese and English text and image. (B) Structure of oligos. (C) Electrophoresis results of PCR amplification products using synthesized oligo pool as template but different primers: Line 1/4/5: file 1 with 100x/1000x/10000x dilution; Line 2/5/8: file 2 with 10x/100x/1000x dilution; Line 3/6/9: file 3 with 100x/1000x/10000x dilution. (D) Reads distribution of three files.

Discussion

Apart from the outstanding performance in coding efficiency and valid DNA sequence generation, YYC also shows broad application not only because of its variant rules but also the flexibility of incorporation. In the current pipeline, the flexible incorporation is used in screening process in order to obtain valid DNA sequences according to preset criteria. Furthermore, the unlimited incorporation strategies give multiple choices for data manipulation and broaden the application scenario. For example, classified information encryption, the incorporation pattern can be various and thus provide a more secured data archiving. Another application scenario could be combining several

related files into one DNA file, such as the image of museum exhibits and their audio narratives or literatures.

To summarize, in this work, we reported a robust and efficient codec system which displays outstanding feature in coding efficiency and generating sequences highly feasible for synthesis and sequencing. We tested out codec algorithm in silica with collection of relatively large number of different types of files. The results suggested a better overall performance for our strategies compared to previous ones in data density and fidelity. A wet experiment also proved the feasibility and also showed the capability of our pipeline for random-access of files. We expect that this codec system could provide more choices for people to use DNA-based data storage more efficiently in a broader field.

Moving forward, it should not be a major issue to improve the coding efficiency in current DNA-based data storage since DNA fountain and YYC are already capable of reaching the theoretical limit. Meanwhile, instead of file-level random-access, it is also important to develop other novel techniques in smaller scale. For example, Tomek et. al reported an emulsion-PCR based random-access technique[12], which can also be used for sub-file level random access if reasonable partition rules and primers are designed. Although the cost and speed of DNA synthesis and sequencing are still far behind silicon-based storage system, DNA-based data storage is currently considered as a potential solution for huge data long-term archiving. And with DNA synthesis and sequencing technology improving, we expect that DNA-based data storage could gradually be cheaper, faster and more secured way of immediate access storage for people's daily lives.

Currently, some other molecule storage medium was proposed including oligopeptides[3, 13] and metabolomes[4]. It is innovative for these techniques to use amino acids and metabolomes for binary transcoding. Their variety can achieve a greater data compressibility compared to DNA, which have four different kinds of bases. However, the stored information cannot be amplified for information backup, which is a significant function for data archiving. Moreover, both techniques use array chips for storage and mass spectroscopy for information retrieval. As a result, the data density could be a problem as well as the widespread application. As both works concluded, these techniques have great space for optimization. Hence, we may suggest that before

breakthrough techniques are developed, DNA-based information storage would still be a reliable storage strategy for mass data archiving in the not-so-far future.

Methods

'Yin-Yang' DNA-based data storage pipeline

The 'Yin-Yang' codec pipeline includes three major steps: segmentation, incorporation and validity-screening (Supplementary Fig. S2). 1) Binary information is extracted from source file and then partitioned into smaller segments according to requirements. Indices will be added to each segment and a pool of binary segments with identical length will be obtained. 2) Two segments will be selected randomly or following certain patterns，and then incorporated into one DNA sequence according to 'Yin-Yang' codec algorithm. Considering the features of incorporation algorithm, binary segments contains too much '0' or '1' tends to produce DNA sequences with abnormal GC content or undesired repeats. Therefore, for the convenience to reduce the calculation in this demonstration, binary segments contain high ratio of '0' or '1' (>80%) will be collected into a separate pool and will be firstly selected to incorporate with binary segments with normal ratio of '0/1'. 3) In order to prevent the difficulties of synthesis and sequencing, the incorporated sequence will be tested through a validity-screening process, including checking their GC content, maximal homopolymer length, secondary structure free energy, etc., and only the sequences meet the criteria will be considered as valid sequences. The details of validity-screening is described in the following section.

YYC-screener to tackle biochemical constraints

Biochemical features of DNA molecules including high GC or AT content, homopolymer and complex secondary structure are usually quite harsh for normal DNA synthesis and sequencing procedure, therefore results in difficulties of writing and reading process in DNA-based data storage applications. It has been reported that DNA sequence with over 6-mer homopolymers or a stable secondary structure (of low free energy) will cause many problems for sequencing[14, 15]. Additionally, DNA sequences with GC content below 40% or above 60% often bring troubles to DNA synthesis manufacturers. For practical DNA-based data storage usage, it is therefore important to

build a working scheme to reduce these situations. By harnessing the advantage of assembling two binary information to one DNA sequence in 'Yin-Yang' codec system, a working scheme named 'YYC-screener' is established here to obtain the valid sequences. As mentioned previously, two sequences with different complexity are selected and incorporated according to YYC. Generated sequences with GC content over 60% or less than 40%, or carrying >6-mer homopolymer, or possessing a predicted secondary structure of <-30 kcal/mol are abandoned and the two original segments are put back into the pool. A new pair of segments will then be selected to repeat the screening process until it passes the validity-screening test (Supplementary Fig. S2). We performed this working scheme on ~1GB data including articles, images, audios and videos. It is found that YYC-screener can generate DNA sequences which meet the criteria as expected.

File encoding

The 3 files' binary form (9.26×10^5 bits, 7.95×10^5 bits and 2.95×10^5 bits) were extracted respectively and segmented into three 120 bit-segment pools. First, we added 8-bit Hamming code for error correction, which is able to correct at most two mutation errors possibly caused by synthesis and sequencing. After this, three 128-bit binary segments (data payload + Hamming code) were used to generate a fourth redundant binary segment to prevent possible sequence loss during the experiment. Furthermore, 14-bit index information was added into each binary segment to infer their address in the digital file and in the oligo mixture for purpose of decoding. Rule 888 from 'Ying-Yang' coding scheme was then used to convert the binary information into DNA bases. To reduce the sequence complexity, the aforementioned 'YYC-screener' was used to screen the transcoded DNA sequence. Sequences passed the screening all satisfied the criteria of balanced GC content (40% to 60%), higher secondary structure free energy (> -30kcal/mol) and minimum (< 6) homopolymer repeats. These enabled us to generate 8323 DNA sequence segments with 142 nt each. 33.3% redundancy was added in case of possible sequence loss during manipulation through exclusive-or calculation[16]. For the future reading and random-accessing of each file, well-designed flanking sequence of 20 nt was added at both ends of each DNA segment. Finally, an oligo pool containing 11098 single-stranded DNA sequences of 182 nt was sent to the DNA supplier for synthesis (Fig. 3B).

Polymerase Chain Reaction (PCR) amplification of synthesized DNA oligo pool

Before NGS sequencing, the single-stranded DNA oligo pools were amplified by PCR reactions. Three independent reactions were conducted to obtain the amplified product of three different files respectively. Q5 polymerase was used in these reactions in order to ensure the fidelity and 18/20/25 cycles were set to test the uniformity of amplification. The experiment conditions are shown below.

Table 2. Solution component for PCR amplification.

Component	Volume（μL）
5 × Q5 Reaction Buffer	5
Q5 polymerase	0.5
dNTPs（10mM）	1
Forward Primer（10μM）	2
Reverse Primer（10μM）	2
Template DNA （×100/1000/10000）	1
ddH_2O	38.5

Table 3. Experiment condition for PCR amplification.

Reactions	Temperatures	Time	Cycles
Pre-denaturation	98℃	5min	×1
Denaturation	98℃	15s	
Annealing	62℃	30s	×18/20/25
Synthesis	72℃	10s	
Extension	72℃	5min	×1

Conflict of Interest

Sha Zhu is a currently an employee at the Sensyn Health Plc. This work was completed when Sha Zhu was working at the University of Oxford, and consulting for the BGI. George Church significant interests in Twist, Roswell, BGI, v.ht/PHNc, v.ht/moVD.

Acknowledgements

This work was supported by the Guangdong Provincial Academician Workstation of BGI Synthetic Genomics (No. 2017B090904014); Shenzhen Institute of Synthetic Biology, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, People's Republic of China; Guangdong Provincial Key Laboratory of

Genome Read and Write (No. 2017B030301011); and Shenzhen Engineering Laboratory for Innovative Molecular Diagnostics (DRC-SZ[2016]884). This work was also supported by the George Church Institute of Regenesis, BGI-Shenzhen, China.

Reference

1. Church GM, Gao Y, Kosuri S. Next-generation digital information storage in DNA. *Science* **337**, 1628 (2012).

2. Allentoft ME*, et al.* The half-life of DNA in bone: measuring decay kinetics in 158 dated fossils. *Proc Biol Sci* **279**, 4724-4733 (2012).

3. Cafferty BJ*, et al.* Storage of Information Using Small Organic Molecules. *ACS Cent Sci* **5**, 911-916 (2019).

4. Kennedy E*, et al.* Encoding information in synthetic metabolomes. *PLoS One* **14**, e0217364 (2019).

5. Kosuri S, Church GM. Large-scale de novo DNA synthesis: technologies and applications. *Nat Methods* **11**, 499-507 (2014).

6. Goldman N*, et al.* Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. *Nature* **494**, 77-80 (2013).

7. Grass RN, Heckel R, Puddu M, Paunescu D, Stark WJ. Robust chemical preservation of digital information on DNA in silica with error-correcting codes. *Angew Chem Int Ed Engl* **54**, 2552-2555 (2015).

8. Erlich Y, Zielinski D. DNA Fountain enables a robust and efficient storage architecture. *Science* **355**, 950-954 (2017).

9. Ping Z*, et al.* Carbon-based archiving: current progress and future prospects of DNA-based data storage. *Gigascience* **8**, (2019).

10. Zhirnov V, Zadegan RM, Sandhu GS, Church GM, Hughes WL. Nucleic acid memory. *Nat Mater* **15**, 366-370 (2016).

11. Graham JE, Marians KJ, Kowalczykowski SC. Independent and Stochastic Action of DNA Polymerases in the Replisome. *Cell* **169**, 1201-1213 e1217 (2017).

12. Tomek KJ*, et al.* Driving the Scalability of DNA-Based Information Storage Systems. *ACS Synth Biol* **8**, 1241-1248 (2019).

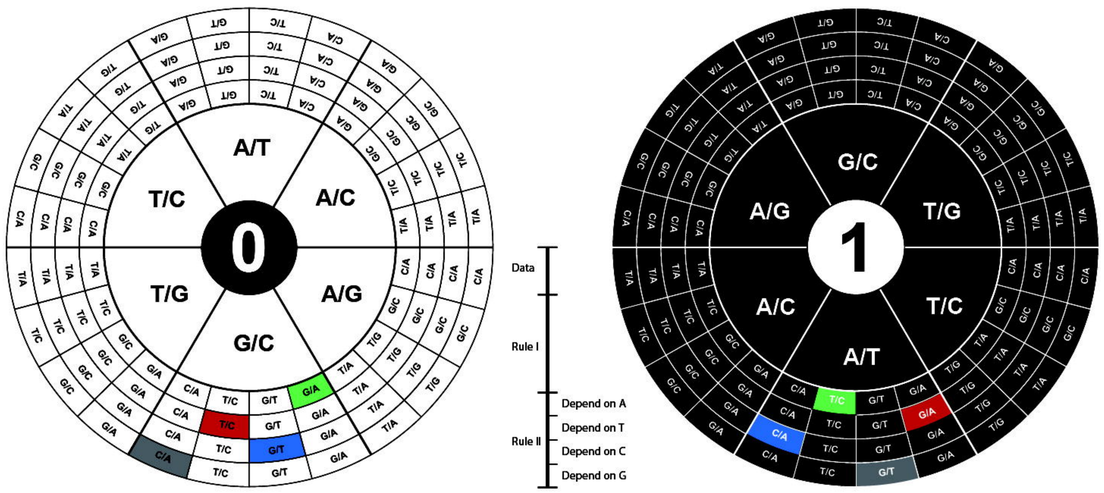
13. Cafferty BJ*, et al.* Storage of Information Using Small Organic Molecules. *ACS Central Science* **5**, 911-916 (2019).

14. Kulski JK. Next-generation sequencing—an overview of the history, tools, and "Omic" applications. *Next Generation Sequencing–Advances, Applications and Challenges*, 3-60 (2016).

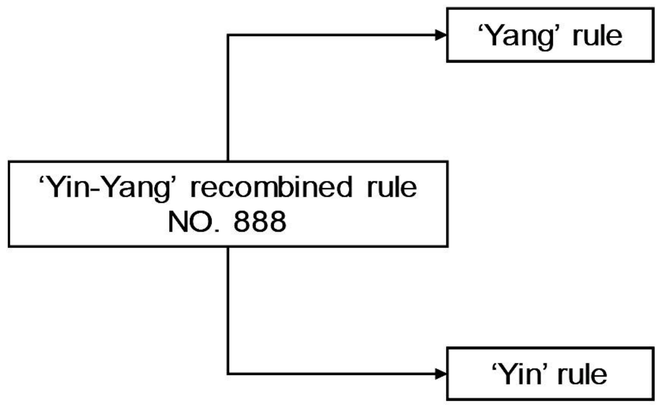
15. Kieleczawa J. Fundamentals of sequencing of difficult templates--an overview. *J Biomol Tech* **17**, 207-217 (2006).

16. Organick L*, et al.* Random access in large-scale DNA data storage. *Nat Biotechnol* **36**, 242-248 (2018).

A



B



Binary Digits

0	1
A or T	C or G

Binary Digits

Previous Nucleotide	0	1
A	A or G	C or T
T	C or T	A or G
G	A or G	C or T
C	C or T	A or G

C A step-by-step transcoding demonstration using rule No. 888

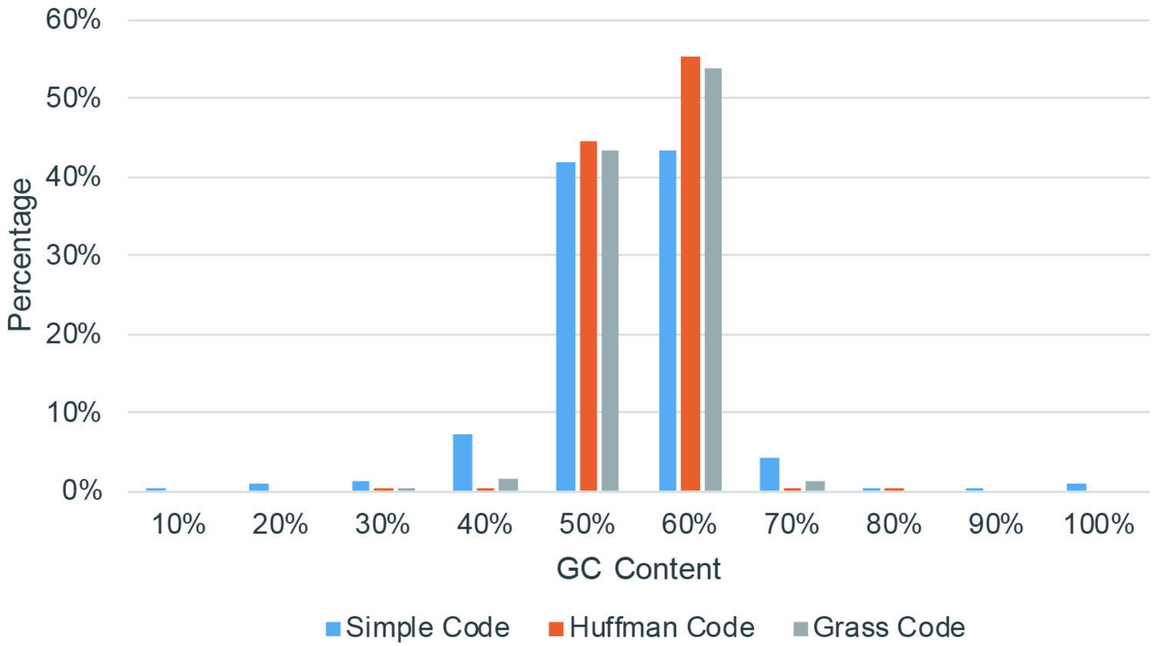
Input Signals a = 10110011 b = 01011101

Transcoding According to Rules

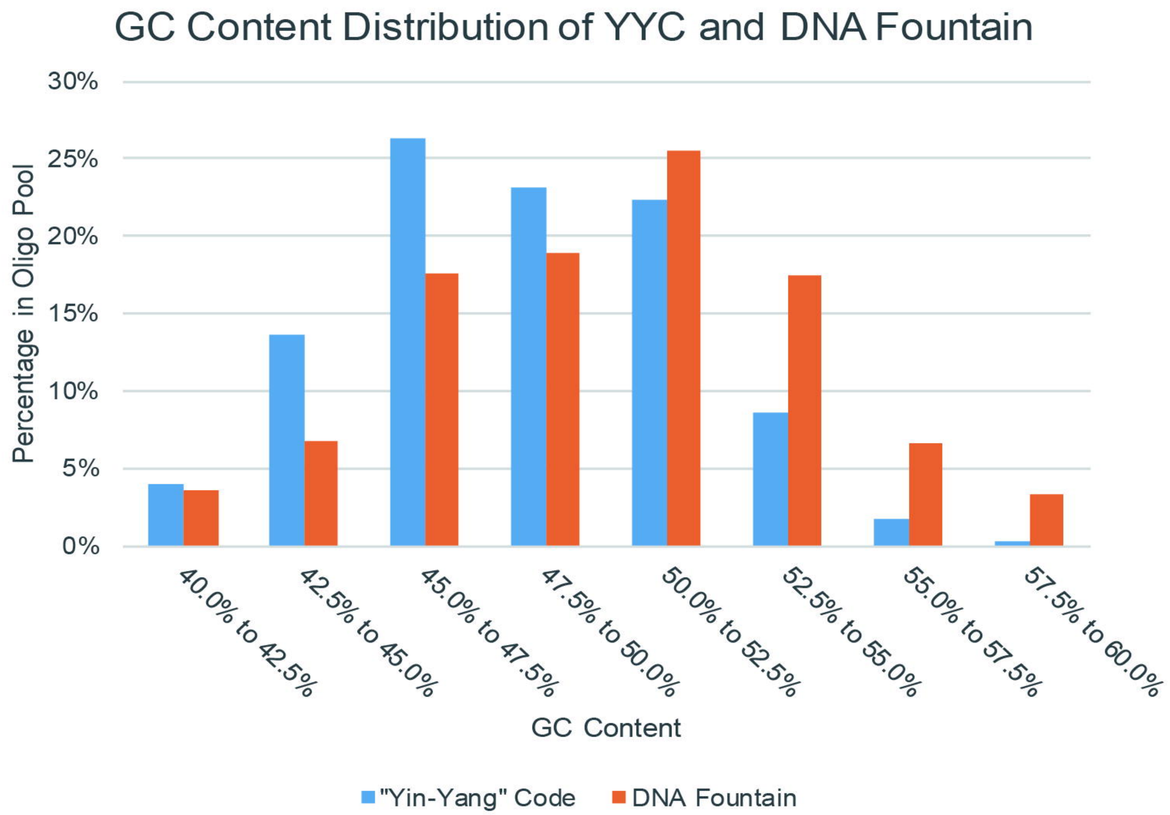
	Previous Nucleotide	Binary Digits	Rule	Candidates	Local Nucleotide	Oligo Sequence
Step 1	A virtual	a1 = 1 b1 = 0	'Yang' 'Yin'	C or G A or G	G	G
Step 2	G	a2 = 0 b2 = 1	'Yang' 'Yin'	A or T C or T	T	GT
			. . .			
Step 8	C	a8 = 1 b8 = 1	'Yang' 'Yin'	C or G C or T	C	GTCGTAGC

Result Sequence: GTCGTAGC

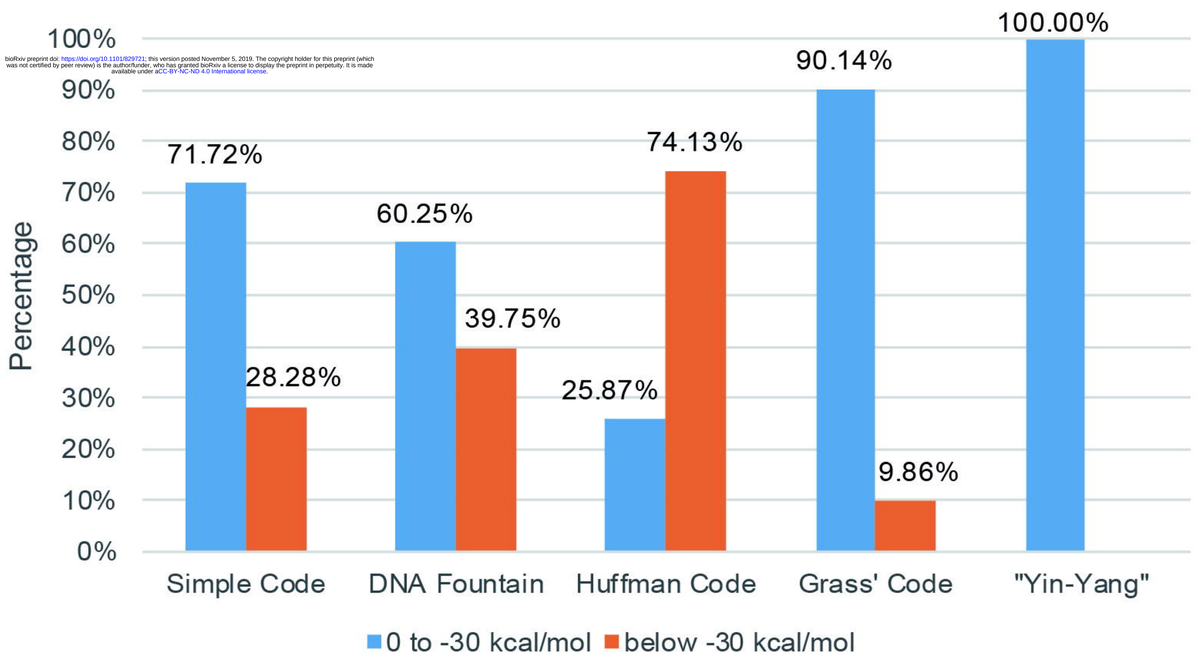
A



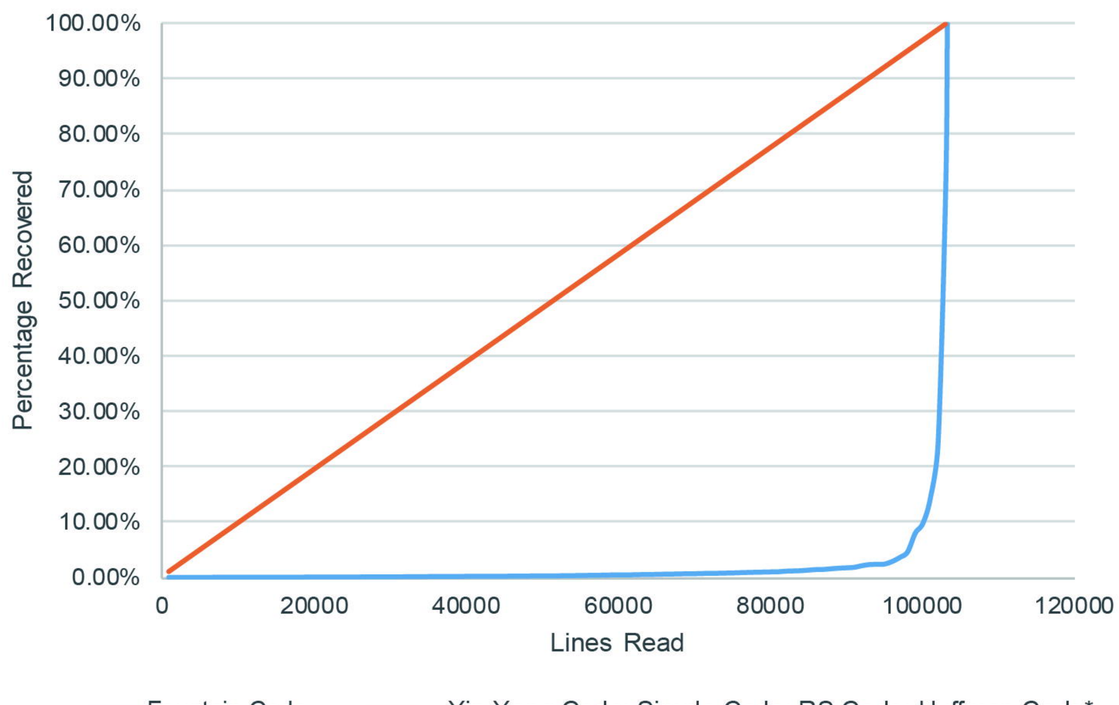
B



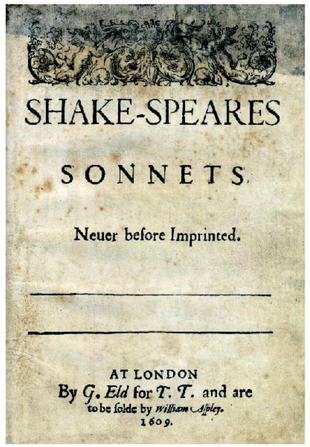
C



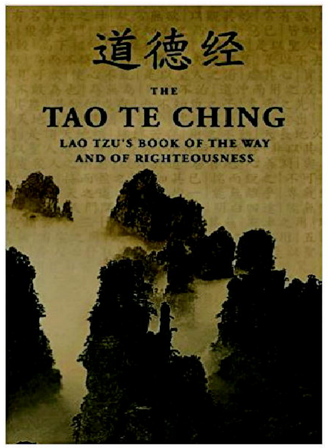
D



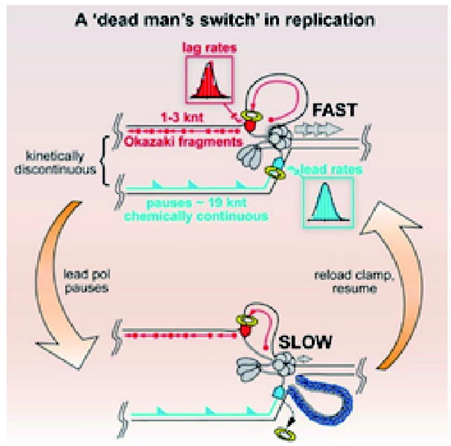
A



Shakespeares Sonnet

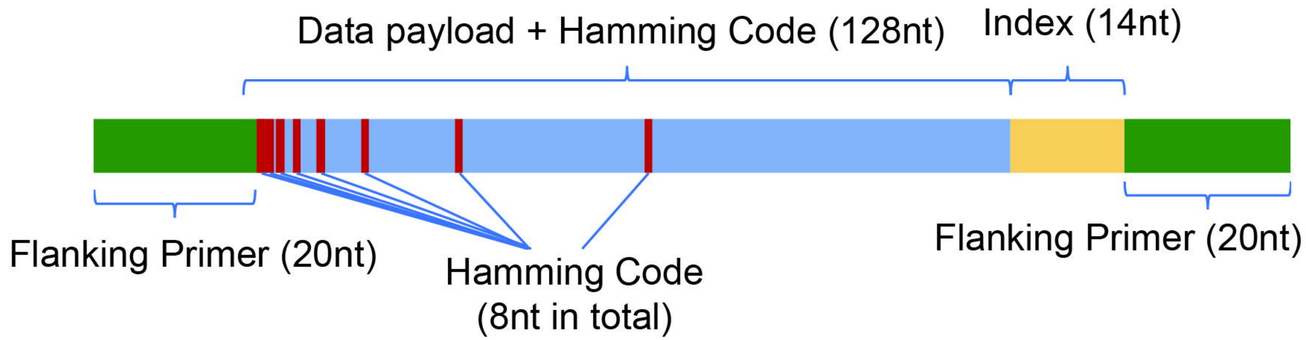


Tao Te Ching

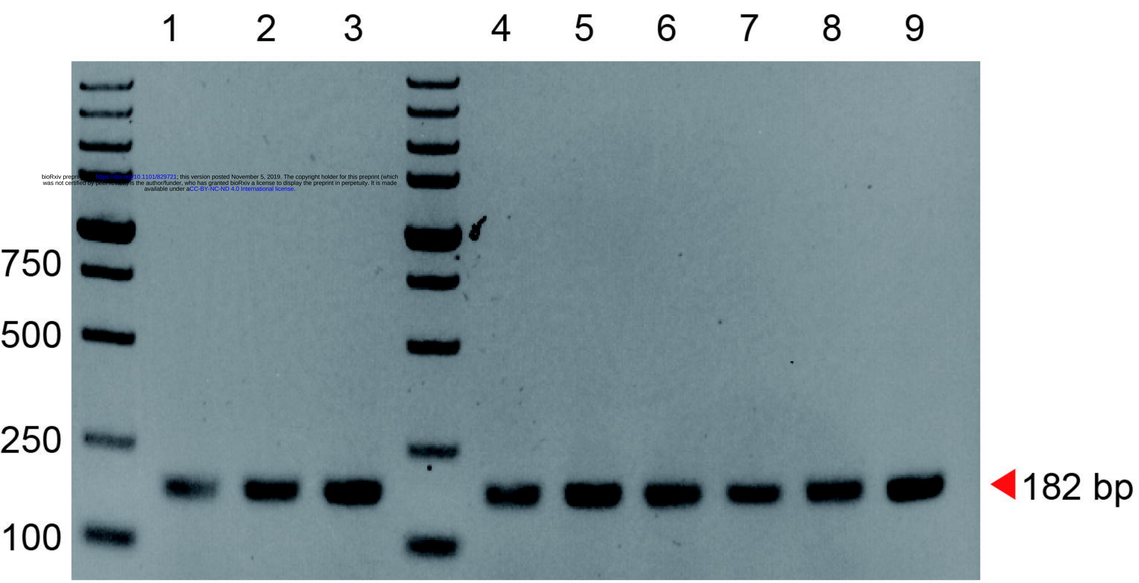


Scientific Figure

B



C



D