Efficient Data Archiving Utilizing the "Yin-Yang" Codec System

MGI's DNBSEQ Sequencing Platform Enables Novel DNA based Data Storage

The "Yin-Yang" codec system is a unique codec system based on MGI DNBSEQ sequencing platform to solve the current technical problems in the field of DNA data storage[1]. The research is led by the BGI-Shenzhen, with the participation of several research teams from the Shenzhen National Gene Bank, Capital Normal University and Harvard University.

The research results were published in *Nature Computational Science* on April 25, 2022, entitled with "*Towards practical and robust DNA-based data archiving using the yin-yang codec system*".

Recommended application: Novel technology - DNA storage technology

Recommended model: DNBSEQ-T7RS

- **Efficient and high-quality sequencing data output**

  DNBSEQ sequencing technology exhibits many excellent features such as high accuracy, low repeat rate and low index hopping rate.

- **High stability of data recovery**

  The "Yin-Yang" codec can significantly improve the stability of data recovery while ensuring high data density, data conversion efficiency and high technical compatibility.

Background

DNA storage technology refers to the usage of the molecular structure of DNA for data storage, which is similar to the traditional information storage "information write-save-read" steps. DNA storage flow chart is shown in Figure 1. We are in the time of an unprecedented information explosion and traditional standard storage media can no longer satisfy the exponential growth of data storage needs. As an ancient and efficient information carrier in living organisms, DNA, with its exclusive natural advantages of ultra-high information density, ultra-long standby time, and super biocompatibility, has great potential to overturn existing technologies in terms of information density, copy and maintenance costs, and service life [2-4].

Codec of DNA storage is one of the most important aspects of DNA storage, which not only determines the efficiency of data conversion (data density), but also directly affects the stability and reliable recovery of stored data.

Since 2012, the development of DNA storage technology has mainly focused on enhancing information density, while the consideration of technical compatibility and stable recovery of original information is not yet comprehensive. The codec technology has failed to achieve full technical compatibility until 2017. In 2017, the DNA fountain code developed by a research team at Columbia University almost solved the previous technical bottleneck[5], but the issue of flexibility and applicability also arises in practical applications. Therefore, how to substantially improve the stability of information recovery while ensuring information conversion efficiency and technology compatibility has become a major challenge in achieving DNA storage.

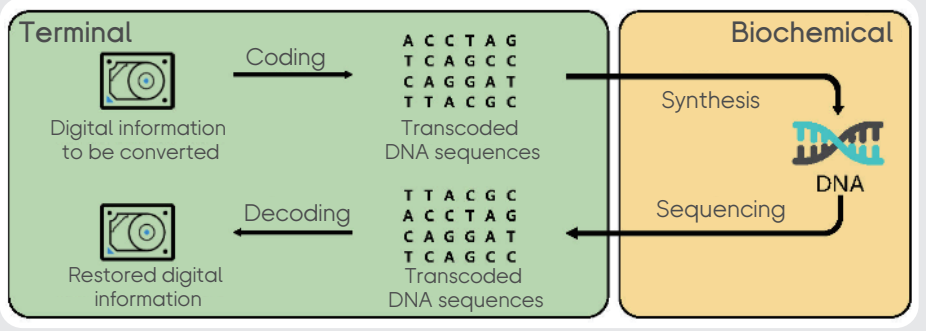

Figure 1. DNA storage process

Research Description

The "Yin-Yang" codec system based on the MGI DNBSEQ sequencing platform cleverly applies the Chinese "Yin-Yang" idea, which originated thousands of years ago, to the DNA codec system to convert two sets of binary information by "one-to-one" with two different rules. At the same time, this system take the part of the unified intersection of the two as the final solution to realize the unification of two independent data combinations into one DNA sequence, and its "Yin-Yang" codec rules are shown in Figure 2. The system has proven to have great advantages in various aspects such as data density, technical compatibility, and data recovery stability.

Materials and Methods

Encode files using YYC codecs and DNA fountain codes

The binary information of the file was converted into DNA sequences using the 888th coding rule of the YYC codec, and the feasible DNA sequences were selected using the "YYC-screener", resulting in an oligonucleotide pool containing 10,103 single-stranded 200 nt DNA sequences.

For DNA fountain codes, in addition to redundancy, a DNA oligonucleotide library was generated using the default parameter settings from the original report to determine the minimum redundancy for file decoding. A final oligonucleotide library encoding a .tar archive zip file (9,185 sequences) and an oligonucleotide library encoding a mix of three individual files (10,976 sequences) were obtained.

Library preparation and sequencing

The three oligonucleotide pools were generated and delivered as DNA powder for sequencing by Twist Biosciences.

For in vivo storage, the DNA fragments are first split into subfragments with overlapping regions, and then further split into building blocks.

The 80 nt oligonucleotides were assembled onto blocks using Q5 high-fidelity DNA polymerase and cloned into the vector for Sanger sequencing. Sequencing-validated blocks were released by enzymatic vector cleavage, followed by PCR amplification, gel purification, and finally integration of the fully assembled fragment into the YBR150C gene in yeast chromosome II using the natural homologous recombination of yeast. Positive yeast colonies were screened for genomic DNA extraction and sequencing by a tagged LEU2 marker.

Library preparation for oligonucleotide pools: The DNA powder was dissolved in double-distilled water (ddH_2O) to obtain standard solutions, which were serially diluted 10-fold to obtain seven working solutions (WS6-WS0) with average concentrations ranging from 10^6 to 10^0 DNA molecules per microliter. The amplification products of P2 and each of the three different files of P1 and P3 were next obtained by PCR amplification of each working solution. The concentration of the products was measured using gel electrophoresis and a Qubit fluorometer. All amplified DNA libraries were sequenced using a DNBSEQ-T7 sequencer.

Yeast genomic DNA preparation and library preparation: Yeast cells were cultured in YPD medium for two days, and the particles were collected by centrifugation. Resuspension and

precipitation were performed by adding Bacteria Breaking Buffer. Add glass oobeads and PCI, vortex mix, and centrifuge. The aqueous layer was then transferred to a test tube, isopropanol was added to precipitate genomic DNA, left to stand, centrifuged, and finally genomic DNA was resuspended in TE buffer. Genomic DNA was fragmented using the Covaris instrument, followed by end-repair, fragment screening using the magnetic bead method, followed by the addition of an A-tail at the 3' end, ligation of double-ended sequencing junction with tag, and finally the ligated product was enriched using PCR. Samples were sequenced using the DNBSEQ-G400 and DNBelab sequencing platforms.

Bioinformatics analysis

A total of >3G PE150 reads were generated for the validation of in vitro storage experiment. Random secondary sampling of sequencing data with an average depth of 100 times was performed for information retrieval. The reads were first clustered and assembled to complete the sequence of each oligonucleotide. Flanking primer regions were removed, DNA sequences were decoded into binary fragments using coded inverse operations, and errors were corrected using RS codes. Binary segments were reordered according to the address area. In this process, the "pseudo-binary" segment is removed based on the address. The complete binary information was then converted to the digital file. Data recovery rate was calculated. For error analysis, six random secondary sampling of sequencing data with average depth of 100×, 300×, 500×, 700×, and 900× were performed using different random seeds.

In total, >50M PE-100 reads were generated for in vivo storage, in which SOAPnuke filtered 10% of the low-quality reads (Phred fraction <20). After BWA mapping, reads from the host genome were removed using samtools, and then short reads were assembled into overlapping clusters by SOAPdenovo. Blastn was used to discover the connections between contigs. A Python script was written to merge overlapping groups and to obtain the assembled sequences of each strain. Multiple sequence alignment was performed to identify structural variants, insertions and deletions by comparing the assembled sequences for the majority voting process. Pre-added RS codes were used for error correction of the replacement. The complete DNA sequence was decoded to recover the binary information by reversing the coding operation.

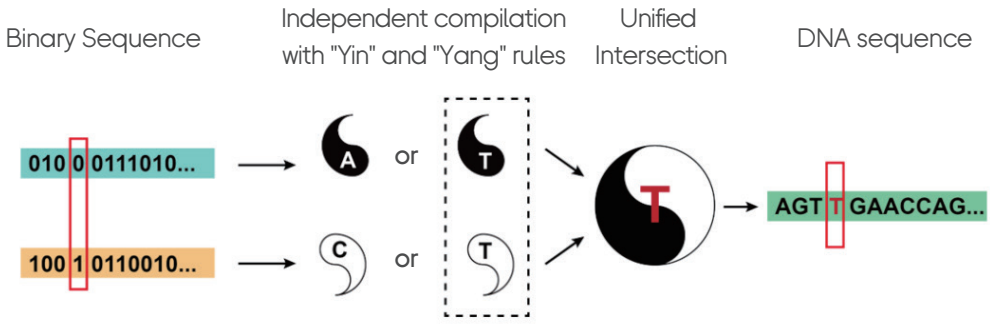

Figure 2. Principle of "Yin-Yang" codec rule

Results

Wide range of applications of YYC system

Inspired by the double-stranded model of DNA, this study combined with the principle of unity of opposites of "Yin-Yang" in Chinese culture, and cleverly applied it to DNA codec system to convert two binary information by "one-to-one" with two different rules, and then take the part of the unified intersection of the two as the final solution to realize the unification of two independent information combinations into one DNA sequence, as shown in Figure 3. On the other hand, by the screening mechanism introduced, the sequences that are not compatible with existing synthetic sequencing technologies were filtered by pre-set screening conditions. Depending on the combination method, the system can provide a total of 1536 different combinations of coding rules, which greatly extends its range of application scenarios.

Excellent data recovery capability of YYC system storage

The study tested the robustness of YYC to systematic errors by randomly introducing the three most common errors in DNA sequences at an average rate of 0.01% to 1%. In addition, the corresponding data recovery rates were analyzed in comparison with the DNA Fountain encoding scheme without error correction mechanism introduced. The results are shown in Figure 4. It can be seen that the data recovery performance of YYC is superior to that of DNA Fountain regardless of the presence of Indels or SNVs, and the data recovery rate remains at a fairly stable level of over 98%.

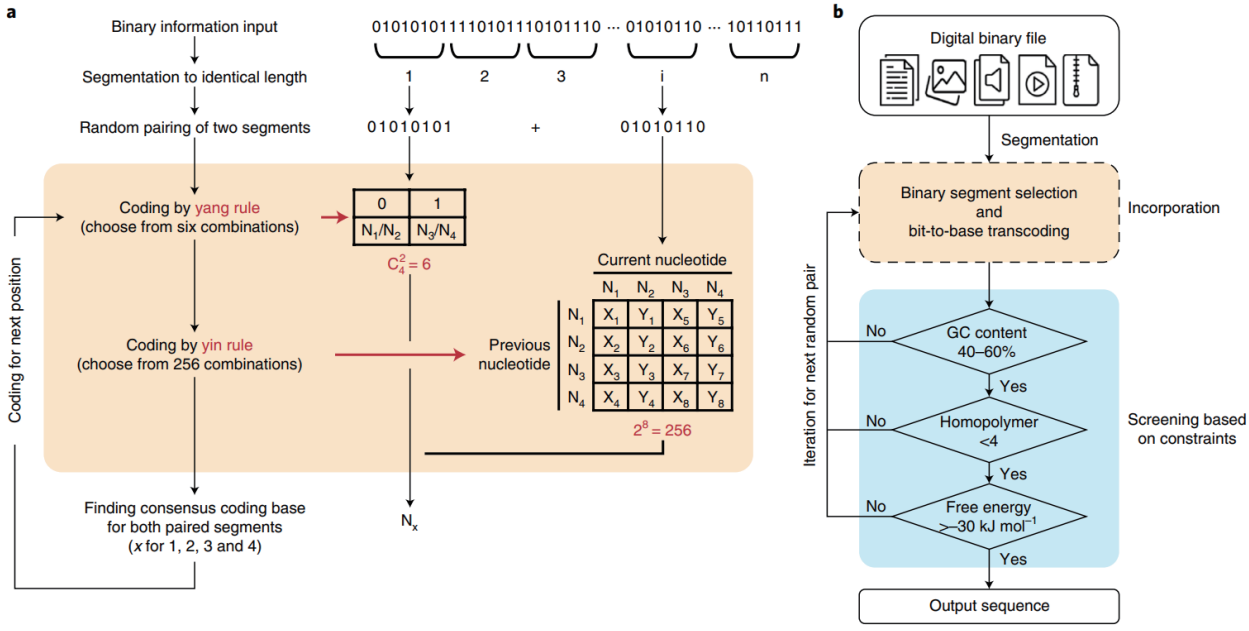

Figure 3. Schematic diagram of YYC

High recovery rate of YYC in vitro storage data

The minimum number of copies of oligonucleotides required for this study to evaluate successful file recovery and the robustness analysis for the loss of DNA molecules are shown in Figure 5b. Its sequencing results showed that the data corresponding to P1 could be recovered to 99.9% at AMC numbers above 10^3; when the number of AMCs was 10^2, the average data recovery rate dropped to 71.2%, ranging from 65.69% to 87.53% for each stored file. When the number of the AMC was less than 10^1, it dropped further to below 10%. In general, YYC showed a linear recovery trend and was positively correlated with the retention of DNA molecules encoded by the data, as shown in Figure 5c. For the DNA Fountain algorithm, the data recovery rate was comparable to YYC when the number of AMC was above 10^4, but dropped significantly to single-copy level when the number of AMC was below 10^3.

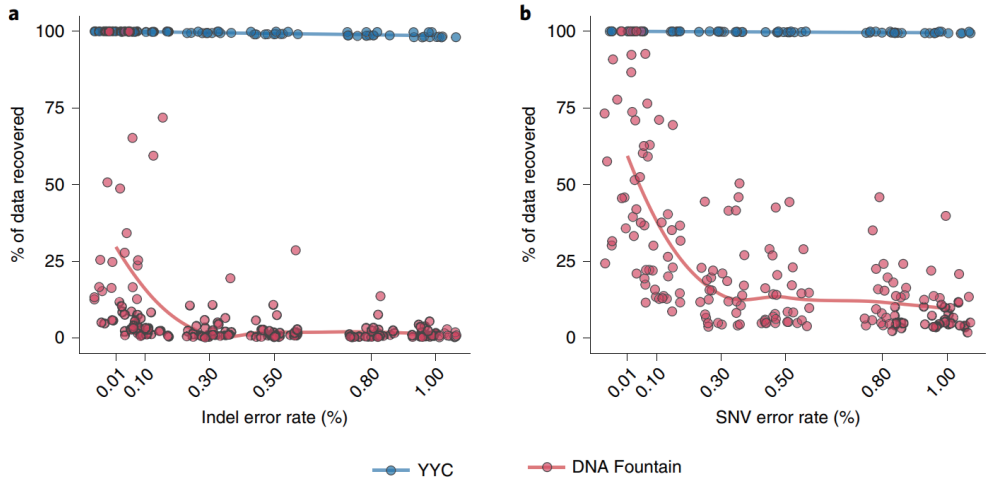

Figure 4. Data recovery capability analysis of YYC and DNA fountain codes

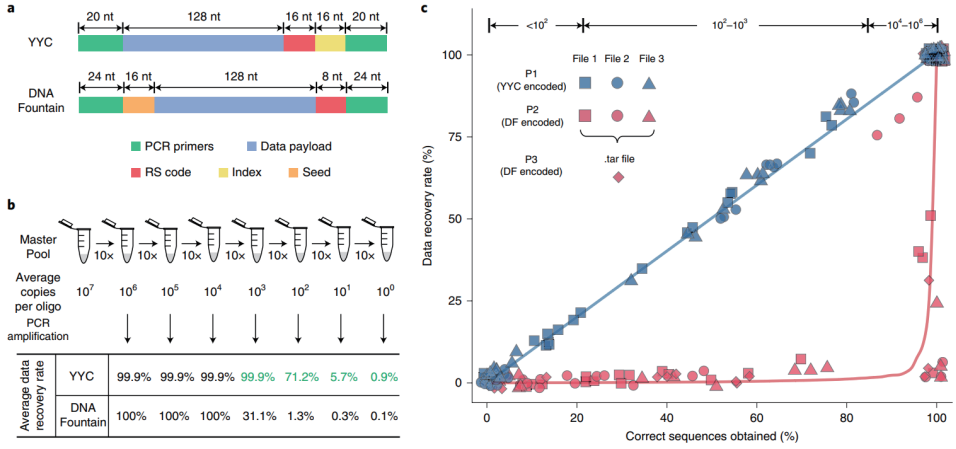

Figure 5. Validation results of YYC in vitro storage experiment.

Good robustness of data storage in YYC

The study used YYC to encode a portion of the text file into a 54240 bp DNA fragment, as shown in Figure 6a. And taking advantage of the higher homologous recombination efficiency of yeast, these fragments were directly transformed into yeast strain BY4741 with the linearized low-copy of the mitophagy vector pRS416 to achieve in vivo one-step assembly of full-length DNA. After approximately 1000 generations of cell cultures were transferred in bulk, whole genome sequencing was performed on 15 single colonies to assess the stability period of the YYC regimen, as shown in Figure 6b. In addition, partial fragment loss was observed in all 15 single colonies, with loss ranging from ~21.1 kbps to ~51.4 kbps and data recovery levels ranging from 38.9% to 95.0%. Reconstruction and comparison of multiple colonies produced consistent sequences by using a single simple majority voting strategy. This study reconstructed a complete sequence containing 66 SNVs that could not be corrected by RS codes introduced into the data block and fully recovered the stored data.

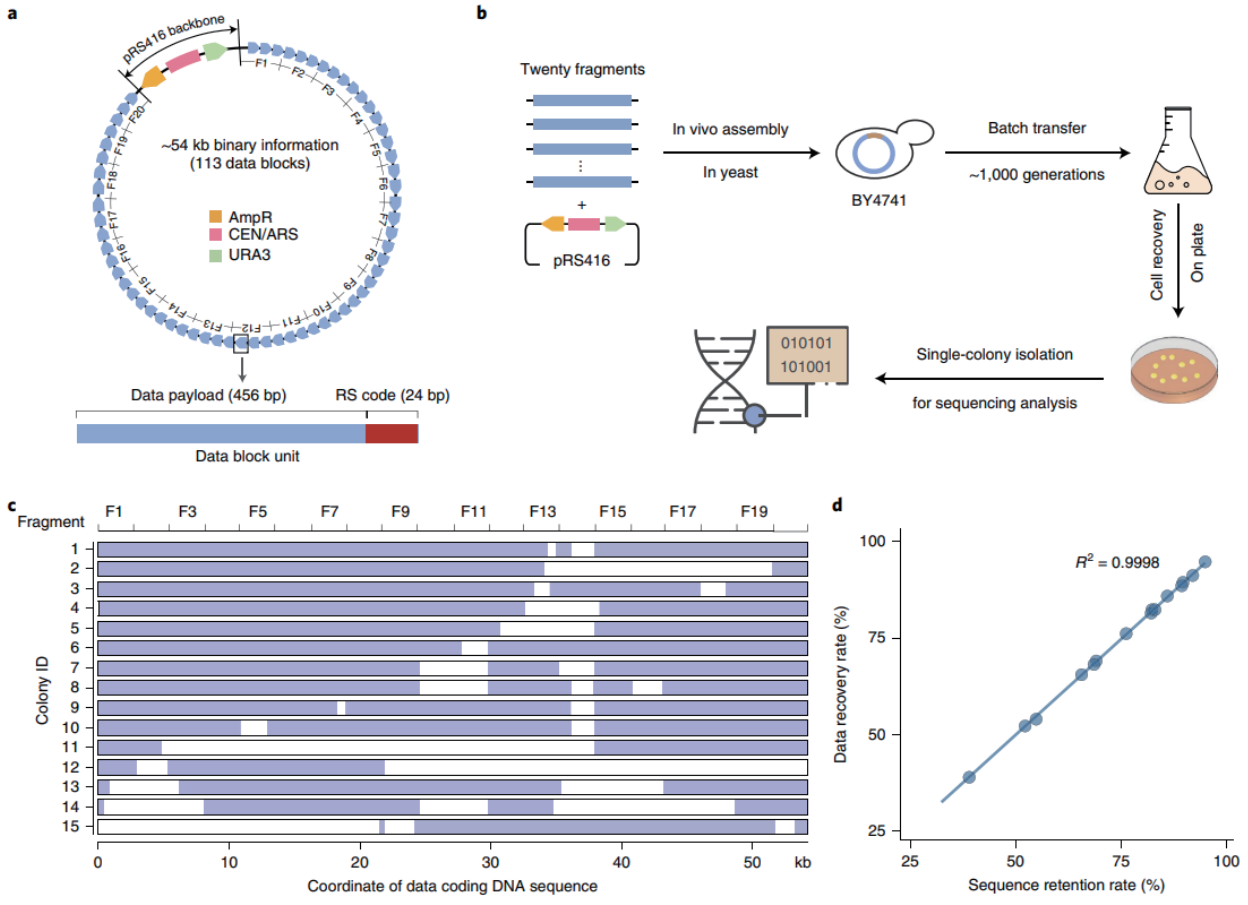

Figure 6. Validation results of YYC in vivo storage experiment

Conclusion

A unique "Yin-Yang" codec system based on the MGI DNBSEQ sequencing platform provides a high-density and high-stability bit-base codec method for DNA information storage applications, and has completed experimental validation of two modes of information storage *in vivo* and *in vitro*. This study developed a new DNA storage coding method, which provides an important tool for multiple types of DNA storage applications.

DNBSEQ sequencing platform with MGI full-process sequencing products is cost-effective and stable, which provides fully localized, efficient and reliable high-throughput technical support for the research and development of a new DNA storage and coding method. The research team expressed their confidence in the future popularity and development of the system, which is expected to play a positive role in the research of new media for long-term storage of massive data.

DNBSEQ-T7RS Genetic Sequencer

References

1. Ping Z., Chen S., Zhou G., et al. Towards practical and robust DNA-based data archiving using the yin–yang codec system[J]. *Nature Computational Science*, 2022, 2(4): 234-242.
2. Church G. M., Gao Y., Kosuri S. Next-generation digital information storage in DNA[J]. *Science*, 2012, 337, 1628.
3. Allentoft M. E., Collins M., Harker D., et al. The half-life of DNA in bone: measuring decay kinetics in 158 dated fossils[J]. *Proceedings of the Royal Society B: Biological Sciences*, 2012, 279,4724-4733.
4. Bhat W. A. Bridging data-capacity gap in big data storage[J]. *Future Generation Computer Systems*, 2018, 87, 538-548.
5. Erlich Y., Zielinski D. DNA Fountain enables a robust and efficient storage architecture[J]. *Science*, 2017, 355, 950-954.

Recommended Ordering Information

Category	Product	Cat. NO.
Instruments	Genetic Sequencer DNBSEQ-T7RS	900-000242-00
	MGIDL-T7RS DNB Loader	900-000134-00
	Genetic Sequencer DNBSEQ-G400RS	900-000170-00
	Genetic Sequencer DNBSEQ-E25RS	900-000537-00
Software	Data Center Appliance	900-000444-00
	MegaBOLT Bioinformatics analysis accelerator	900-000555-00
Library Prep	MGIEasy Universal DNA Library Prep Set (16 RXN)	1000006985
Sequencing Reagents	DNBSEQ-T7RS High-throughput Sequencing Set (FCL PE150) V3.0	940-000268-00
	DNBSEQ-G400RS High-throughput Sequencing Set (FCL PE150)	1000016952
	DNBSEQ-E25RS High-throughput Sequencing Set(FCL PE150)	940-000567-00

MGI Tech Co.,Ltd

Building 11, Beishan Industrial Zone, Yantian District, Shenzhen, CHINA, 518083

Version: November 2023

+86-4000-688-114
en.mgi-tech.com
MGI-service@mgi-tech.com

Authors: Chen Liqin, Li Hanping
Editor-in-Charge: Wang Qiwei
Reviewer: Jiang Yao

1. For StandardMPS and CoolMPS: Unless otherwise informed, StandardMPS and CoolMPS sequencing reagents, and sequencers for use with such reagents are not available in Germany, Spain, UK, Sweden, Belgium, Italy, Finland, Czech Republic, Switzerland, Portugal, Austria and Romania. Unless otherwise informed, StandardMPS sequencing reagents, and sequencers for use with such reagents are not available in Hong Kong. No purchase orders for StandardMPS products will be accepted in the USA until after January 1, 2023.
2. For HotMPS sequencers: This sequencer is only available in selected countries, and its software has been specially configured to be used in conjunction with MGI's HotMPS sequencing reagents exclusively.
3. For HotMPS reagents: This sequencing reagent is only available in selected countries.