nature nanotechnology

Supplementary information

https://doi.org/10.1038/s41565-024-01771-6

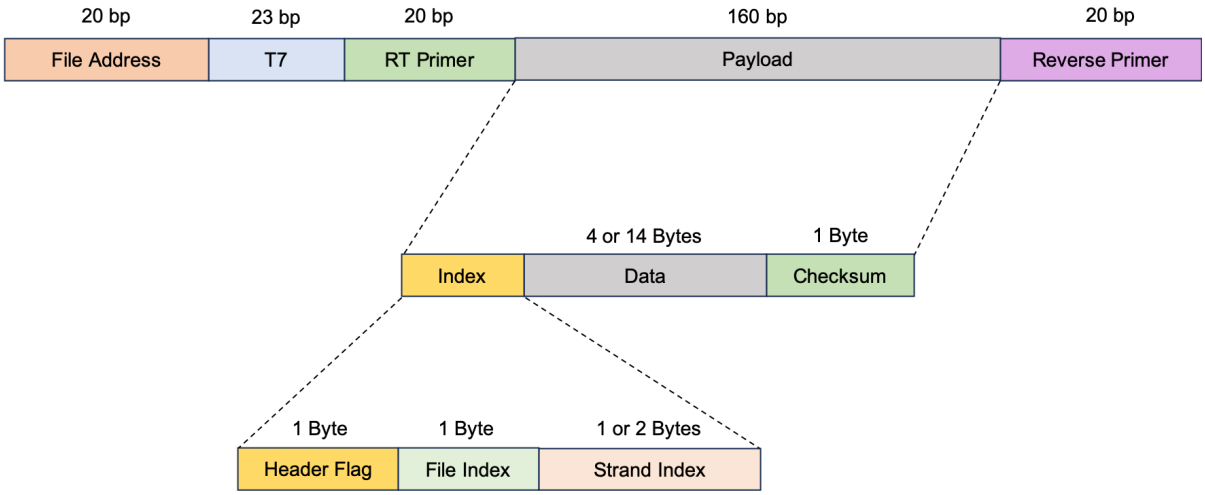
A primordial DNA store and compute engine

In the format provided by the authors and unedited

Supplementary Note 1. File encoding and decoding for phosphorous, zebrafish embryo and muscle cell image files

Image files were each encoded into DNAs of lengths approximately 243 nt. In our design we considered both errors that occur within strands and dropouts that may erase information from our data set. Thus, for each file we combined an inner code to tolerate intra-strand errors and an outer code to tolerate drop outs.

Our inner code implementation is a HEDGES convolutional code[1], and we chose a Reed-Solomon on the field of 8-bit symbols for our outer code. For each file, we encoded two sets of information. One being the payload information of the file, and the other being a meta-data header describing how the payload was encoded for version control purposes. When encoding the payloads, we utilize 2 different HEDGES code rates. Namely, for the phosphorous symbol we used a 0.25 rate code (0.5 bits/base), and for the other two we used a 0.5 rate code (1 bits/base). We chose two different code rates so we could study if there were any impacts of error correction strength on the data we collected. For each file, we included 3-4 bytes to be used as an index that allowed us to correctly order strands during decoding along with an additional 1-byte checksum appended to the end of each strand. By including a checksum along with error correction capabilities of HEDGES, we were able to have high confidence in the origin of a sequenced strand without the need to align every sequencing read with our encoded library set. By encoding 4 data bytes per strand for the 0.25 code rate and 14 bytes per strand for the 0.5 rate our entire strand reached 243 nt. To implement our outer code, we needed to organize data into appropriate sets that matched with the field GF ($2^8 = 256$). To achieve this, we divided all strands into subpackets. For the 0.5 rate codes we divided all strands in the files into 50 payload sub-packets and 50 error correction sub-packets, and for the 0.25 rate codes we used 58 payload sub-packets and 108 error correction sub-packets.



Our choices of code rate for the HEDGES inner code, and the choices of Reed-Solomon outer codes were based on in-silico simulations using an identically and independently distributed error model. In this error model, each base is injected with an error with a certain overall probability, with the type of error (deletion, insertion, substitution) being equally likely. According to our simulations, the 0.25 HEDGES rate design is capable of decoding the file for conditions where the error rate is between 10-15%, each strand has a probability of 0.3 of being dropped out, and each strand that does not drop out is read only once. By increasing the number of reads for each strand not dropped out to 5x we can decode this file even at a 20% error rate. Likewise, for the 0.5 rate HEDGES file designs, we can decode at a 30% drop out rate with 5 reads per strand kept at an error rate of 10%. Thus, our designs provide variable robustness, allowing us to be cost effective in the number of strands synthesized and ensuring data can be extracted in the event of

extremely high error rates. All of our fault injection simulations and sequencing data analysis is performed by utilizing FrameD[2], a framework that allows for the construction and analysis of DNA encoding designs.

Finally, our headers were all stored within 90-byte blocks of data with a 0.25 rate HEDGES code with 5 bytes per strand. We then provide 36 error correction strands for each 90-byte block so as to protect this small piece of important information.

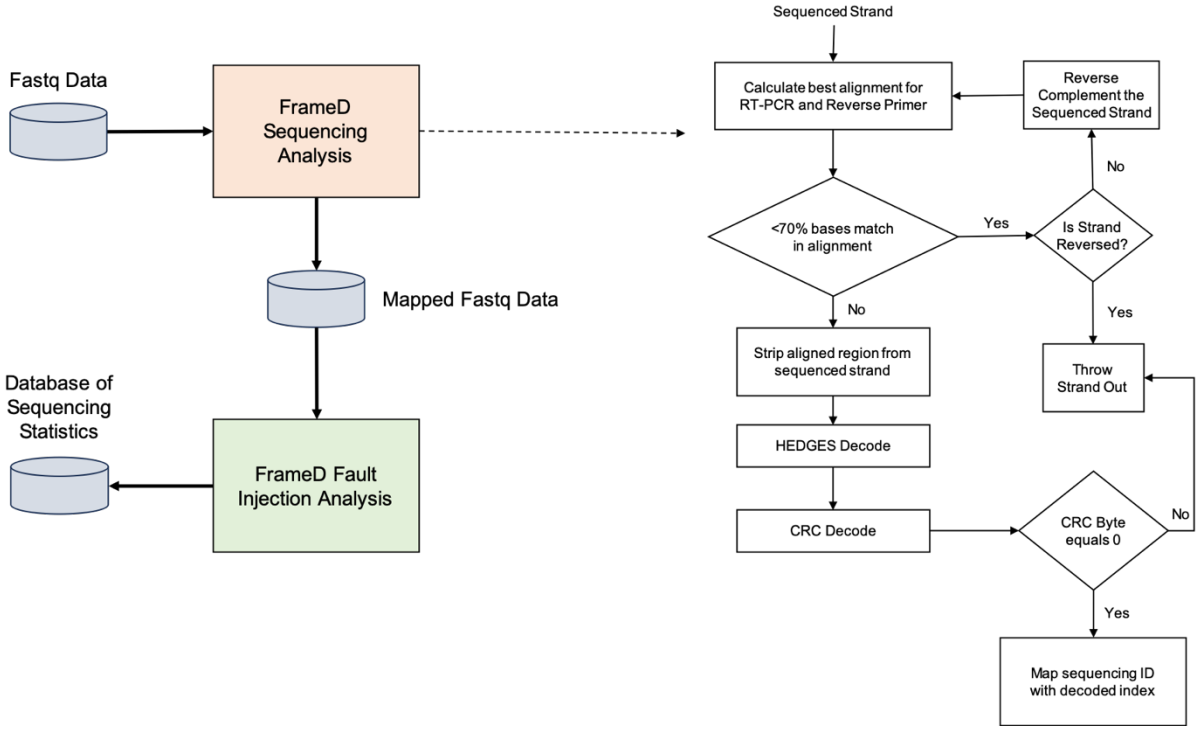
Supplementary Note 2. Primer design and cross-talk inspection

The primers used for the File Address and Reverse Primer (Supplementary Note 1) were taken directly from previous work[3]. RT Primer contains restriction sites 'GAATTC' (EcoRI) in the File1 oligo sequence, 'GAAGAC' (BbsI) in the File2 oligo sequence and 'GCTAGC' (NheI) in the File3 oligo sequence for the specific deletion. These primers meet several constraints: GC content between 40 and 60%, the last base is held to be a constant G while ensuring the last 5 bases do not exceed 60% GC content, each primer has a Hamming distance of at least 10 to reduce non-specific binding, and NUPACK simulations were performed to avoid homodimers, hairpins, and heterodimers that have a Gibbs free energy less than -10 kcal/mol at 50 °C.

To ensure that there was no cross talk between primers and payloads, we checked the Hamming distance between each primer associated with every file ('File Address', 'T7 Promoter', 'RT Primer' and 'Reverse Primer') and the payload of all synthesized strands. Our Hamming distance check slides each functional site of interest across the payload and determines the minimum Hamming distance that was observed during this procedure. Of all functional sites, the minimum Hamming distance we found was 6, which indicates there is likely a small chance of cross talk between payloads and primers.

Supplementary Note 3. File Decoding

Our decoding pipeline was carried out in multiple stages for both MiSeq and ONT data. Sequencing analysis using FrameD follows the flow of the following figure[2]. First, fastq files storing the sequencing data information is run against FrameD's sequencing analysis tool. We set this tool up to create a mapping between each read identifier in the fastq file with each unique index that we encoded for each strand. The logic flow of creating this mapping is also included in this figure. In order to quickly bin sequencing reads with encoded strands without doing pairwise alignments on entire sequencing files with the complete encoded strand set, we utilize the HEDEGES inner encoding error correction capability and the checksum to keep only strands that pass the checksum within our mapping. Our checksum provides us with additional confidence that the strand was correctly decoded and thus the identity of the strand is known. With this mapping, we now know what each sequencing read belongs to which encoded strand. We then process this mapping in FrameD's fault injection tool by setting the tool to inject sequenced strands instead of an in-silico error model. By doing this, we can generate statistics about our sequencing data such as error rates, strand distributions, and file decode success rates.



Supplementary Note 4. Extrapolation of the number of file accesses possible.

To estimate the maximum number of IVT rounds (i.e. file accesses) from the SDC-File complex and the minimal mass of DNA required for decoding, we first conducted qPCR quantification of cDNA obtained from multiple rounds of IVT of the SDC-File complex. We calculated the average rate of decrease in cDNA between IVT rounds. We then estimated the maximum number of IVT rounds possible using following equation:

$$N_f = N_i * k^n \quad (1)$$

Where N_f and N_i are the initial and final sequencing reads of each strand, respectively; k is the average unique sequence drop rate obtained from cDNA measurement using qPCR; n represents the maximum number of IVT rounds.

As our DNA libraries are designed to tolerate 30% dropout of unique oligo sequences, which means a file can still be decoded if only 70% of its unique sequences have sequencing reads that are equal or greater than 1. Following this principle, we sorted and ranked the number of reads for each sample from largest to smallest in the sequencing data, and recorded the sequencing reads of the last strand within the dropout range (e.g. for File3, it's the number of reads for the 355th strand). We then assumed each unique strand has the same probability of being dropped during physical operations of the system. To allow successful decoding, the minimal number of sequencing reads should be equal or greater than 1 (in our case, we assume $N_f = 1$). To calculate the maximum number of IVT rounds, the equation can be rearranged as:

$$n = \frac{\ln\left(\frac{1}{N_i}\right)}{\ln(k)} \quad (2)$$

Once the maximum number of IVT rounds has been obtained, we can estimate the sequencing reads for each strand at the maximum number IVT rounds by plugging the values into the equation (1). This was followed by calculating the total number of sequencings reads for strands within the dropout range:

$$\text{Total number of strand} = \sum_1^i N_i \quad (3)$$

Where N_i is the initial sequencing reads of each ranked strand, and i is the last strand ID within the dropout range.

We then assumed each strand has 100,000 copies for practical purposes, and calculated the minimal decoding mass using the following equation:

$$\text{Minimal decoding mass (pg)} = \frac{\text{Total number of strand} * 100000 * \text{M.W.} * 1E{+}09 * 1000}{6.022E{+}23} \quad (4)$$

Where $M.W.$ represents the average molecular weight of the file-encoded DNA library.

The summarized data for samples presented in Figure 2G was presented as:

Sample	Maximum Number of IVT Rounds	Minimal Decoding Mass	Source
SDC-File Lyophilized	172 +/- 5	28 +/- 1.37	Figure 2G
SDC-File Solution	122 +/- 8	41 +/- 3.97	Figure 2G
File-Lyophilized	65 +/- 1	71 +/- 2.52	Figure 2G

Supplementary Note 5. Extrapolation of DNA stability over time

To estimate the DNA stability of the file stored on the SDC-DNA complex shown in Figure 2H, we first used qPCR to measure cDNA obtained from IVT of the SDC-File complex incubated at 65 ˚C for the time points of 0, 8, 16, 24 and 48 hours. In the analysis, we assumed the decay followed a first-order reaction and plotted the logarithmic form of strand concentration measured by qPCR at each time point as:

$$\ln\left(\frac{[C]}{[C_0]}\right) = -kt \quad (5)$$

where $[C_0]$ is the initial strand concentration at the 0-hour time point and [C] is the strand concentration measured at each time point.

This was followed by fitting the linear regression model to the data points. As indicated by the equation, the slope of the linear model represents the decay rate (k), and we used the followed equation to calculate the half-life for the first-order reaction:

$$t_{\frac{1}{2}} = \frac{\ln(2)}{k} \quad (6)$$

We then summarized decay rate and half-life for the lyophilize and in-solution forms of SDC-DNA as:

Sample	Decay Rate k (s^{-1}) at 65 ˚C	Half-Life $t_{1/2}$ (years)
SDC-File Lyophilized	4.00E-07	0.0574
SDC-File Solution	9.33E-07	0.0238

When applying the stability model developed in Grass et al. 2015 to this work, these measurements are equivalent to storing the DNA information at 4 ˚C with half-lives of approximately 6000 and 4000 years for the lyophilized and in-solution forms of SDC-DNA, respectively, or storing them at the Global Seed Vault (at-18 ˚C) for 2 million and 0.8 million years, respectively.

We scaled the half-life and decay rate for a sequence of 158 bp reported by Grass et al[4]. to 250 bp used in our experiments by $t_{\frac{1}{2}}^{250} = t_{\frac{1}{2}}^{158} * \frac{158}{250}$, and observed the values are within a similar scale.

Sample	Rescaled Equivalent Decay Rate k (s^{-1}) at 65 ˚C	Rescaled Equivalent Half-Life $t_{1/2}$ (years)
DNA on filter	5.14E-06	0.00428
DNA in polymer	2.66E-06	0.00826
DNA in silica	1.31E-06	0.0167

Supplementary Methods

Synthesis of MNP-SDCs and activation of caSDC for DNA adsorption

The synthesis of SDCs followed a previously published protocol[5]. Specifically, 1 wt. % SDC precursor (cellulose acetate, cellulose or agarose, Sigma-Aldrich) was dissoved in 3 wt. % acetic acid along with 1 wt. % magnetic nanoparticles (Fe2O3-gamma, 99% pure, average particle size = 20 mm, M K Impex Corp.) The precursor was then injected directly into the shear zone of the IKA Magic Lab device (IKA Works Inc.) operating at 20,000 rpm. These formulated dendricolloids were further activated using a previously described protocol for DNA adsorption. Specifically, 60 µg of caSDC particles infused with MNP were collected and washed twice with 190 proof ethanol (Fisher Scientific, AC615110010) on a magnetic stand. The SDC particles was then mixed with 40 µL of 190 proof ethanol, 160 µL of 1% Tween 20 and 4 µL of 5 M sodium hydroxide, followed by incubating the mixture in a thermocycler at room temperature for 30 mins at 1500 rpm. After thirty minutes, SDC particles in the mixture were washed three times with water to ensure the removal of excessive Tween 20. Next, the NaOH treated SDC particles were then mixed with 200 µL TEMPO-solution and incubated at room temperature for 10 min, followed by washing three times with water. After removing excessive TEMPO-solution, the treated SDC particles was then resuspended in 10mM sodium acetate with 0.1 M 1-ethyl-3-[3-dimethylaminopropyl]-carbodiimide (Sigma Aldrich, 341006) and 0.4 M N-Hydroxy-succinimide (Sigma Aldrich, 130672-5G). The mixture was incubated at room temperature for 30 min, followed by washing three times with water and two times with Dulbecco's Phosphate-Buffered Saline (DPBS, Thermo Fisher, 14200075). The surface modified SDC was resuspended in DPBS.

Creation of unlabeled dsDNA

Unlabeled ssDNA and primer oligos were purchased from Azenta Inc. 0.1 ng of ssDNA was mixed with unlabeled primer oligos at a final concentration of 1 µM using 0.5 µL of Q5 High-Fidelity DNA Polymerase (NEB, M0491S) in a 50 µL reaction containing 1x Q5 polymerase reaction buffer (NEB, B9072S) and 0.2 mM each of dATP (NEB, N0440S), dCTP (NEB, N0441S), dGTP (NEB, N0442S), dTTP (NEB, N0443S). The reaction conditions were 98 °C for 30 s and then 35 cycles of: 98 °C for 10 s, 53 °C for 20 s, 72 °C for 10 s, with a final 72 °C extension step for 2 min. The generated unlabeled dsDNA strands were purified using AMPure XP beads (Beckman Coulter, A63881) and eluted in 40 µL of water.

Creation of fluorophore-labeled dsDNA

5' FITC-labeled primer oligos and 5' ATTO550-labeled primer oligos were purchased from Integrated DNA Technologies Inc. 0.1 ng of ssDNA template was mixed with 1 µM final concentration of fluorophore-labeled oligos, and a PCR reaction was run with 0.5 µL of Q5 High-Fidelity DNA Polymerase (NEB, M0491S) in a 50 µL reaction containing 1x Q5 polymerase reaction buffer (NEB, B9072S) and 0.2 mM each of dATP (NEB, N0440S), dCTP (NEB, N0441S), dGTP (NEB, N0442S), dTTP (NEB, N0443S). The reaction conditions were 98 °C for 30 s and then 35 cycles of: 98 °C for 10 s, 53 °C for 20 s, 72 °C for 10 s, with a final 72 °C extension step for 2 min. The generated fluorophore-labeled dsDNA strands were purified using AMPure XP beads (Beckman Coulter, A63881) and eluted in 40 µL of water.

Creation of 5' biotinylated dsDNA

5′ biotin labeled oligos were purchased from Integrated DNA Technologies Inc. 0.1 ng of ssDNA was mixed with a final concentration of 1 µM 5' biotin-labeled oligos, and a PCR was run with

0.5 µL of Q5 High-Fidelity DNA Polymerase (NEB, M0491S) in a 50 µL reaction containing 1x Q5 polymerase reaction buffer (NEB, B9072S) and 0.2 mM each of dATP (NEB, N0440S), dCTP (NEB, N0441S), dGTP (NEB, N0442S), dTTP (NEB, N0443S). The reaction conditions were 98 °C for 30 s and then 35 cycles of: 98 °C for 10 s, 53 °C for 20 s, 72 °C for 10 s, with a final 72 °C extension step for 2 min. The generated biotin-labeled dsDNA strands were purified using AMPure XP beads (Beckman Coulter, A63881) and eluted in 40 µL of water.

Algorithm of board puzzles.
In Figure 6, the steps used to execute the in-storage RNA computation of each puzzle were:

1. Each board puzzle has a position assigned like:

a	b	c
d	e	f
g	h	i

2. Depending on the type of puzzle, each position contained either binary (0 or 1 for chess) or ternary digital information (0 or 1 or 2 for Sudoku).

3. The digital information at each position was represented by specific 20 nt sequences (bit sequence). For example, the sequence representing bit 1 at position a is orthogonal to the sequence representing bit 0 at the same position.

4. The mathematical operations for each puzzle were:

Puzzle1: $\neg f \wedge \neg h$
Puzzle2: $\neg b \wedge \neg f \wedge \neg h$
Puzzle3: the solution was split into two reactions (ternary digits of 0, 1 and 2 were used to represent the actual digital value of 1, 2 and 3, respectively).
reaction1: $\neg e_1 \wedge \neg e_0 \wedge \neg f_1 \wedge \neg f_2 \wedge \neg g_0 \wedge \neg g_1 \wedge \neg h_1 \wedge \neg h_2 \wedge \neg i_1 \wedge \neg i_2$;
reaction2: $\neg e_1 \wedge e_2 \wedge \neg f_0 \wedge \neg f_1 \wedge \neg g_0 \wedge \neg g_1 \wedge \neg h_0 \wedge \neg h_1 \wedge \neg i_1$

4. The chemical process of each operation included hybridizing the specific bit sequence DNA oligo associated with each algorithm step, followed by Rnase H digestion and RNA cleanup for collecting undigested RNA fragments. These fragments contained correct puzzle solutions and were further processed by reverse transcription and next-generation sequencing.

Error rates analysis
In Figures 4, 5 and Extended Data Figure 2, we calculated the total error rates by the following equation:

$$Total\ error\ rates = error\ rates\ of\ deletion + error\ rates\ of\ substitution + error\ rate\ of\ insertion \quad (7)$$

The error analysis was performed based on the payload sequence of each strand. To determine and error rate in the presence of alignment errors such as insertions and deletions, we calculate the edit operations required to transform each encoded strand to the strand observed from sequencing based on the Levenshtein distance calculation. This was achieved using the Levenshtein Python open-source module (https://pypi.org/project/python-Levenshtein/). With a list of edit operations, and where they occur relative

to the originally encoded strand, the error rate can be calculated by dividing the number of errors of a given type seen at a base position by the total number of strands considered for a given file in a given sequencing sample. Error rate values for specific samples are provided as a Source Data file.

Calculation of maximum binding capacity

In Figure 1C, we calculate the maximum binding capacity of 200bp dsDNA onto SDCs using a previously published protocol[6]:

a. An adsorption isotherm was developed by plotting the amount of adsorbed dsDNA onto SDCs against the corresponding concentration of free dsDNA in the supernatant at equilibrium.

b. Fit the adsorption isotherm to the Langmuir isotherm model using the equation below:

$$Q_e = \frac{K_L * C_e}{1 + K_L * C_e} * Q_{max} \quad (8)$$

Where, Q_e= amount of dsDNA adsorbed to the SDCs at equilibrium (ug/mg), C_e = concentration of dsDNA in the solution at equilibrium (mg/L), K_L = Langmuir adsorption constant (L/mg) and Q_{max} =maximum amount of dsDNA adsorbed to the SDCs (ug/mg).

c. A linearized form of the Langmuir model can be formed by plotting the $1/Q_e$ as a function of $1/C_e$ in the equation shown below:

$$\frac{1}{Q_e} = \frac{1}{Q_{max} * K_L} * \frac{1}{C_e} + \frac{1}{Q_{max}} \quad (9)$$

d. the slope of the linear plot represents the reciprocal of the product of the maximum binding capacity and Langmuir constant, whereas the value of the maximum binding capacity can be determined from the intercept.

Conversion between DNA quantity to amount of digital information

In Extended Data Table 1, we converted the maximum number of DNA molecules bound on SDCs to the theoretical capacity of digital information by the following steps:

a. The encoding density of File3 (1 bit/base) was applied to the conversion. In File3, it required 507 unique strands to encode 2780 bytes of information.

b. Convert the moles of DNA on the SDC to the amount of digital information by the following equation:

$$Terabyte\ of\ data\ per\ mg\ SDC\ \left(\frac{TB}{mg}\right) = \frac{maximum\ binding\ capaity\ of\ 200\ bp\ dsDNA\ in\ molecules\ per\ mg\ of\ SDC * 2780}{507\ * n *\ (1E + 12)} \quad (10)$$

Where n is the copy number per each unique strand, which was assumed to be 1.

c. The SDC surface area was estimated using Brunauer–Emmett–Teller (BET) theory and the average value was 212 cm^2/mg.

d. The estimated volume of SDC per mg was calculated using the following equation:

$$Volume\ per\ mass\ of\ SDC\ \left(\frac{cm^3}{mg}\right) = pi * \left(\frac{d*10^{-7}}{2}\right)^2 * \frac{212}{2*pi*\left(\frac{d*10^{-7}}{2}\right)} \quad (11)$$

Where d is the average diameter of a single SDC particle, which was 200 nm.

e. Calculate the maximum binding capacity of data per volume of SDC by the following equation:

$$Terabyte\ of\ data\ per\ cubic\ centimeter\ of\ SDC\ \left(\frac{TB}{cm^3}\right) = \frac{terabyte\ of\ data\ per\ mg\ SDC}{Volume\ per\ mass\ of\ SDC} \quad (12)$$

Calculation of cost of data bound on magnetic beads made from different materials.
In Extended Data Table 2, we calculated the cost of data bound on materials made from MNP-caSDC, streptavidin beads, and AMPure beads using the following steps:

a. Material cost per unit mass of MNP, caSDC, streptavidin beads, and AMPure beads was estimated to be \$0.001/mg, \$0.000216/mg (Sigma Aldrich, 419028-500G), \$19.2/mg (NEB, S1420S) and \$152/mg (Beckman Coulter, A63881), respectively.

b. Maximum binding capacity of 200 bp dsDNA for each type of beads was:
MNP-caSDC = 6 μg/mg
Streptavidin Beads = 62 μg/mg
AMPure Beads = 12000 μg/mg

c. Calculate the binding capacity of data per unit cost by the following equation:

$$Cost\ of\ substrate\ material\ per\ TB\ of\ data\ \left(\frac{\$}{TB}\right) = \frac{cost\ of\ material\ per\ unit\ mass\ of\ the\ substrate}{\frac{maximum\ binding\ capaity\ of\ 200\ bp\ dsDNA\ in\ molecules\ per\ mg\ of\ substrate * 2780}{507 * n * 1E + 12}} \quad (13)$$

Where n is the copy number per each unique strand, which was assumed to be 1.

Percent of sequencing space used
In Figures 4, 6 and Extended Data Figure 3, we calculated the percent of sequencing space used by the following equation:

$$Percent\ of\ sequencing\ space\ used = \frac{number\ of\ sequencing\ reads\ for\ selected\ file}{number\ of\ total\ sequencing\ reads\ in\ the\ same\ sample} \quad (14)$$

Percent of strand retained
In Figures 4 and 5, we calculated percent of strand retained by the following equation:

$$Percent\ strand\ retention = 100 * (1 - \frac{number\ of\ unique\ strands\ missing\ in\ the\ targeted\ file}{number\ of\ total\ unique\ strands\ for\ the\ targeted\ file}) \quad (15)$$

Normalized averaged number of reads per strand

In Figures 5 and Extended Data Figure 2, we calculated normalized averaged number of reads per strand by the following equation:

$$Normalized\ reads\ per\ strand = \frac{Average\ number\ of\ reads\ for\ file\ after\ deletion}{Average\ number\ of\ reads\ for\ the\ same\ file\ without\ modification} \quad (16)$$

Rescaled normalized strand abundance
In Figures 4, 5, 6, 7 and Extended Data Figure 4, we calculated rescaled normalized strand abundance by the following equation:

$$Rescaled\ normalized\ strand\ abundance = \log(number\ of\ sequencing\ reads\ per\ strand * (1 - (number\ of\ sequencing\ reads - (\min(\text{number of sequencing reads}))/(num\ of\ reads))) \quad (17)$$

Percentage of DNA bound to SDCs
In Figure 4, we calculated the percentage of DNA bound to SDCs by the following equation:

$$Percentage\ DNA\ adsorbed\ on\ SDC = 100 * \frac{Original\ DNA\ quantity - Quantity\ of\ DNA\ measured\ in\ the\ solution}{Original\ DNA\ quantity} \quad (18)$$

Percentage of puzzle strand abundance
In Figure 6, we calculated the percentage of puzzle strand abundance by the following equation:

$$Percent\ strand\ abundance = 100 * \frac{number\ of\ reads\ for\ the\ specific\ puzzle}{number\ of\ reads\ for\ (Puzzle1 + Puzzle2 + Puzzle3)} \quad (19)$$

Accuracy of computation
In Figure 6 and Extended Data Table 3, we calculated the accuracy of computation for each puzzle using the following steps:

a) The sequences in the fastq file after each puzzle's computation were aligned with the puzzle's original oligo sequences.

b) The percent strand abundance for each oligo sequence was calculated using

$$Percent\ strand\ abundance = 100 * \frac{number\ of\ reads\ for\ each\ oligo}{total\ number\ of\ reads\ for\ all\ oligos} \quad (20)$$

c) The percentage was ranked from largest to lowest for each puzzle's computation.

d) The theoretical number of expected solutions is 64, 16 and 2 for Puzzle1, Puzzle2 and Puzzle3, respectively.

e) In each puzzle's ranking, the first 64 oligos sequences with the highest percent strand abundance was collected for Puzzle1. The first 16 oligo sequences with highest percent strand abundance was collected for Puzzle2. The first 2 oligo sequences with highest percent strand abundance was collected for Puzzle3.

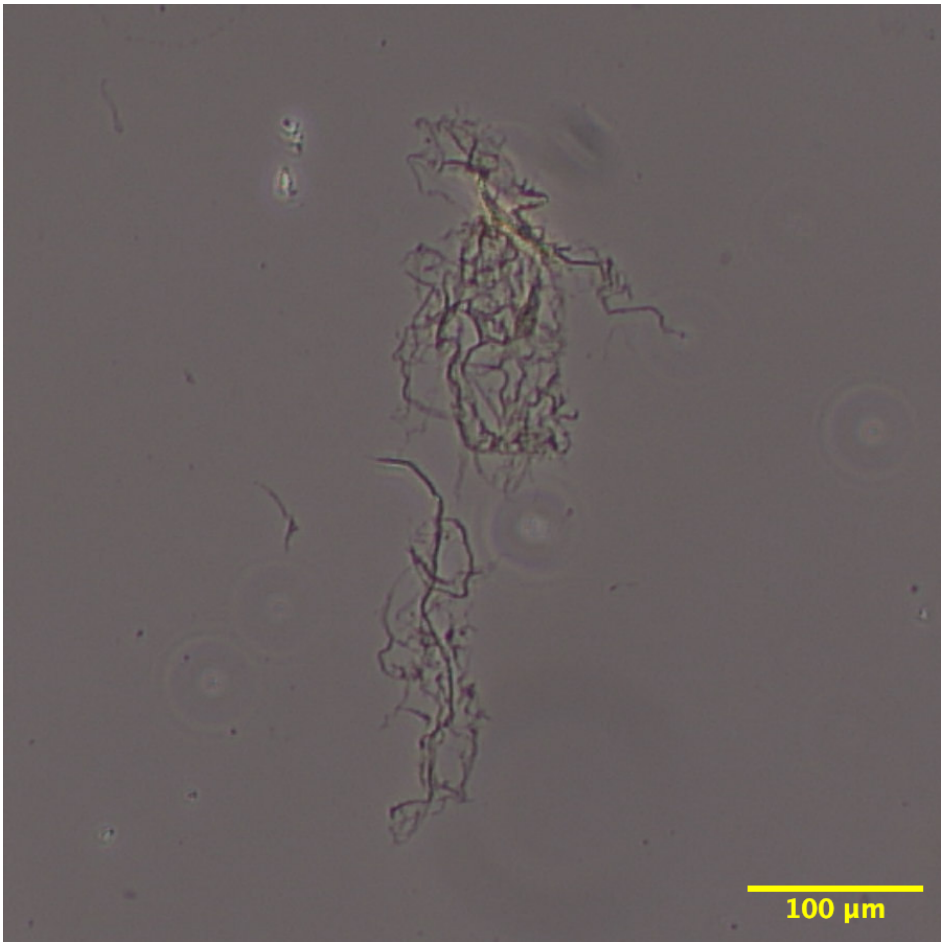
f) For each puzzle, the accuracy of computation was calculated using

$$Accuracy = \frac{number\ of\ oligo\ sequences\ of\ expected/correct\ solutions}{total\ number\ of\ expected/correct\ solutions\ in\ each\ puzzle} \quad (21)$$

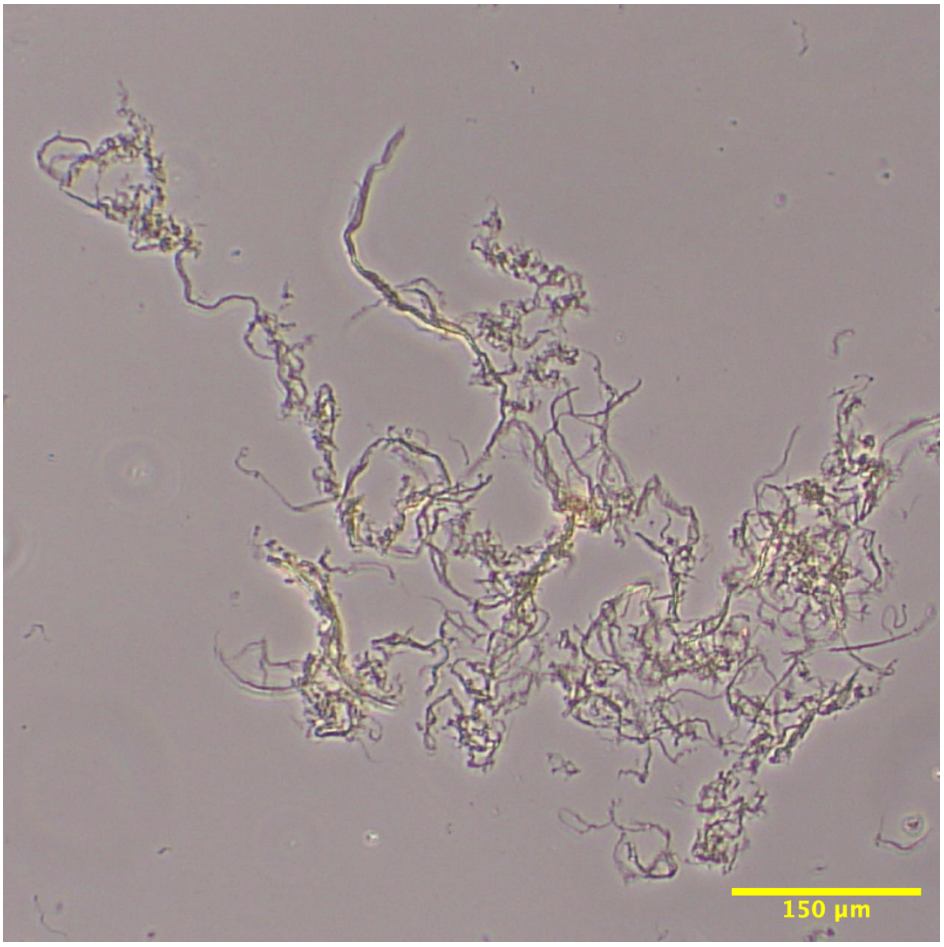
Supplemental References

1. Press, W. H., Hawkins, J. A., Jones, S. K., Schaub, J. M. & Finkelstein, I. J. HEDGES error-correcting code for DNA storage corrects indels and allows sequence constraints. *Proceedings of the National Academy of Sciences* **117**, 18489–18496 (2020).
2. Volkel, K. D. *et al.* FrameD: framework for DNA-based data storage design, verification, and validation. *Bioinformatics* **39**, (2023).
3. Tomek, K. J. *et al.* Driving the Scalability of DNA-Based Information Storage Systems. *ACS Synth Biol* **8**, 1241–1248 (2019).
4. 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 International Edition* **54**, 2552–2555 (2015).
5. Roh, S., Williams, A. H., Bang, R. S., Stoyanov, S. D. & Velev, O. D. Soft dendritic microparticles with unusual adhesion and structuring properties. *Nat Mater* **18**, 1315–1320 (2019).
6. Lin, K. N., Grandhi, T. S. P., Goklany, S. & Rege, K. Chemotherapeutic Drug-Conjugated Microbeads Demonstrate Preferential Binding to Methylated Plasmid DNA. *Biotechnol J* **13**, (2018).

Uncropped image for Figure 1A (Left)



Uncropped image for Figure 1A (Right)



Uncropped image for Figure 1B