nature nanotechnology

Article

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

A primordial DNA store and compute engine

Received: 28 October 2023

Accepted: 19 July 2024

Published online: 22 August 2024

Check for updates

Kevin N. Lin[1], Kevin Volkel[2], Cyrus Cao[1], Paul W. Hook[3], Rachel E. Polak[1,4], Andrew S. Clark[1], Adriana San Miguel[1,4], Winston Timp[3,5], James M. Tuck[2], Orlin D. Velev[1] ✉ & Albert J. Keung[1,4] ✉

Any modern information system is expected to feature a set of primordial features and functions: a substrate stably carrying data; the ability to repeatedly write, read, erase, reload and compute on specific data from that substrate; and the overall ability to execute such functions in a seamless and programmable manner. For nascent molecular information technologies, proof-of-principle realization of this set of primordial capabilities would advance the vision for their continued development. Here we present a DNA-based store and compute engine that captures these primordial capabilities. This system comprises multiple image files encoded into DNA and adsorbed onto ~50-µm-diameter, highly porous, hierarchically branched, colloidal substrate particles comprised of naturally abundant cellulose acetate. Their surface areas are over 200 cm^2 mg^{-1} with binding capacities of over 10^{12} DNA oligos mg^{-1}, 10 TB mg^{-1} or 10^4 TB cm^{-3}. This 'dendricolloid' stably holds DNA files better than bare DNA with an extrapolated ability to be repeatedly lyophilized and rehydrated over 170 times compared with 60 times, respectively. Accelerated ageing studies project half-lives of ~6,000 and 2 million years at 4 °C and −18 °C, respectively. The data can also be erased and replaced, and non-destructive file access is achieved through transcribing from distinct synthetic promoters. The resultant RNA molecules can be directly read via nanopore sequencing and can also be enzymatically computed to solve simplified 3 × 3 chess and sudoku problems. Our study establishes a feasible route for utilizing the high information density and parallel computational advantages of nucleic acids.

Current technologies cannot scale indefinitely to feed the ever-growing demands for data storage and computation. New classes of technologies are needed and are being explored with the potential for orders-of-magnitude leaps in capabilities, including quantum computers and molecular information systems[1]. In this work we motivate the potential of DNA-based molecular information by creating a system that captures the common features and functions of a classical store and compute engine[2]. These include a substrate analogous to a tape or hard drive that carries data, the abilities for all or specific portions of these data to be erased, replaced, read and computed upon, and the ability to execute functions in a relatively continuous and programmable manner[3]. We build upon and leverage an important and growing body of work that is advancing relevant unit processes. These include accelerating DNA synthesis and sequencing, developing encoding and decoding algorithms, and architectures for storing, organizing, accessing or computing upon DNA molecules[4–13].

There are two general classes of nucleic-acid-based molecular information systems to date. In one class, the molecules themselves

[1]Department of Chemical and Biomolecular Engineering, North Carolina State University, Raleigh, NC, USA. [2]Department of Electrical and Computer Engineering, North Carolina State University, Raleigh, NC, USA. [3]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD, USA. [4]Genetics Program, North Carolina State University, Raleigh, NC, USA. [5]Department of Molecular Biology and Genetics, Johns Hopkins University, Baltimore, MD, USA. ✉e-mail: odvelev@ncsu.edu; ajkeung@ncsu.edu

are simultaneously the data and information substrate. These systems forgo anchoring of DNA on a substrate and thereby exhibit extremely high information and computational density[14–19]; yet, reading and computing on the data typically destroys the DNA. The second class leverages changes in nucleic acid structure to store information that can be read out using electrical, optical or physical-probe-based signals[20–25]. These systems are potentially more compatible with repeated use but are harder to scale and execute other functions such as generating copies via polymerase chain reaction (PCR).

In this paper, the central innovation is the use of high-surface-area materials to create a hybrid system exploiting the advantages of both classes of systems (Fig. 1a). We discover that DNA can stably adsorb to specific types of dendricolloidal materials. By incorporating multiple distinct RNA promoters into the DNA design, transcription can be used to non-destructively copy data into RNA. These combined innovations unlock the ability to leverage the richness of molecular biology, including diverse and programmable enzyme activities, to execute other functions common to classical computers. We demonstrate that multiple distinct image files can be individually or completely erased, new data can be loaded onto the same dendricolloidal substrate, specific files can be read, and simplified 3 × 3 chess and sudoku problems can be computed and solved without destroying the original data. Furthermore, this system is implementable in a continuous microfluidic format and compatible with direct RNA nanopore sequencing.

Results

DNA adsorbs onto soft dendritic colloids

We envisioned a system in which DNA is immobilized on a high-surface-area substrate and RNA is transcribed from it non-destructively (Fig. 1a). We exploited a new class of polymer particles known as soft dendritic colloids (SDCs)[26–29] which have very large surface area to volume ratios derived from their hierarchical microscale–nanoscale fibrillar morphology[30]. We considered two modes of immobilizing DNA onto SDCs, covalent or adsorption, and decided that although adsorption may not immobilize the DNA as stably on the SDC, it would be easier to implement and provide a more flexible platform for imbuing functions such as erasing and rewriting[31–33]. Chosen for their compatibility with biological materials in general, ease of fabrication and low cost, we prepared SDCs made from cellulose acetate (caSDC), cellulose (ceSDC) and agarose (agSDC) and incubated them with 200 nucleotide (nt) double-stranded (ds) DNA in amounts ranging over six orders of magnitude, followed by two washes. Our results showed that DNA adsorbed to all three types of SDCs in a monotonic and concentration-dependent manner (Fig. 1b,c and Extended Data Table 1)[34]. We further confirmed the adsorption of DNA to the SDCs by using 200 nt dsDNAs that were 5′ labelled with fluorescein isothiocyanate (FITC) or ATTO550 dyes. We observed green fluorescence only on SDCs mixed with FITC-labelled dsDNA (caSDC-fitcDNA) and red fluorescence only on SDCs mixed with ATTO550-labelled dsDNA (caSDC-atto550DNA) (Fig. 1d). Additionally, we observed that the zeta-potential[35] of the SDCs became more negative when mixed with DNA (Fig. 1e).

Transcription of SDC-bound DNA in a microfluidic channel

The next goal was to develop a method to access the data in a non-destructive manner. We prepared 200 nt dsDNA containing synthetic T7 promoters and adsorbed them onto SDCs that were bound with magnetic beads (Extended Data Fig. 1a). We confirmed that DNA was not adsorbing to the magnetic beads and only to the SDC (Extended Data Fig. 1b). This complex was loaded into polytetrafluoroethylene microtubing using a syringe pump and anchored spatially by placing a paramagnetic cube on the exterior of the microtubing (Fig. 2a). We then copied the DNA-based information into RNA using in vitro transcription (IVT) by simply flowing IVT reagents through the system[36].

We tested the three different SDC materials and two other commercially available systems for non-covalently immobilizing DNA for comparison: streptavidin-functionalized magnetic beads binding biotinylated DNA (SpBioDNA) and solid-phase reversible immobilization beads binding unmodified DNA (SPRI-DNA). Equal masses of substrate and DNA were used across all experimental groups. We asked which material would be most suitable for repeated data access using IVT. Although the streptavidin-functionalized and SPRI beads initially yielded greater RNA per mass of substrate, their yields dropped significantly after each of five repeated rounds of IVT (Fig. 2b). Moreover, both materials are expensive per mass and per data stored (Extended Data Table 2)[36–38]. In contrast, all three SDC materials started with lower but still substantial RNA yields, and the caSDCs were able to maintain a substantial RNA yield over ten successive rounds of IVT (Fig. 2c,d).

The caSDC RNA yield of ~175 ng was sufficient to theoretically encode 46 TB of data; however, we found this yield could be increased further simply by extending the IVT incubation time (Fig. 2e). We also found that a minimum DNA amount of 200 ng was needed to generate detectable RNA yield (Fig. 2f). Interestingly, we observed a non-linear relationship between DNA length and RNA yield, with a substantial drop off in RNA yield for 140 nt and shorter DNAs at the same molar amounts (Fig. 2g). Furthermore, increasing the length of DNA adsorbed to SDCs to 1.5 kb did not significantly affect RNA yield, providing the potential for more efficient data storage and computation with lower encoding overhead devoted to indices and error correction (Fig. 2h). Altogether, these initial experiments narrowed our focus for subsequent studies to the use of caSDCs, a microfluidic system for the IVT reactions and adsorption of DNA at least 200 nt in length.

Repeated access of files from SDCs is robust and stable

We demonstrated the ability to adsorb many copies of a single DNA sequence onto SDCs and transcribe RNA from them. However, practical 'real' files comprise libraries of many distinct DNA strands. A key challenge when scaling to practical data is that the population distribution of the distinct strands comprising a file or database is not uniform across strands, even at the stage of DNA synthesis, and can be further skewed by downstream manipulations such as PCR and Illumina sequencing, leading to loss of strands and impacting the efficiencies and costs of decoding and accessing data[39]. Therefore, we not only tested if complex files could be stored and accessed in this new system, but also if and how IVT-based data access might affect strand distributions. We also checked if any alterations were cumulative with the number of access attempts, a major limitation of PCR-based systems. We designed and ordered 2,775 distinct DNA oligos, each 243 nt long, that encoded three digital JPEG files (Fig. 3a and Supplementary Notes 1 and 2). We first checked whether the simple act of adsorbing the DNA to SDCs would impact the distribution of reads via IVT-based file access; we performed IVT on DNA bound to caSDC and on unbound DNA and sequenced the resulting cDNA by Illumina chemistry. We found that the read distributions for the IVT-based samples were similar (Fig. 3b and Extended Data Fig. 1c,d).

We next tested whether repeated IVT of File1 bound to SDCs would alter the sequenced strand distribution. We washed the SDC–DNA complex, performed IVT and repeated the process ten times to simulate ten file accesses. RNA generated in each round was collected, converted to complementary (cDNA) and sequenced by Illumina chemistry. We observed a gradual decrease in cDNA quantity with each IVT round (Fig. 3c), suggesting some material loss from the SDC surface, probably due to the washing steps. However, strand distributions remained highly consistent (Fig. 3d and Extended Data Fig. 2a). Furthermore, error rates remained low and did not increase with repeated IVT reactions (Extended Data Fig. 2b). In all IVT rounds, there was a very low percentage of unique strand sequences missing in each IVT, and File1 was also accurately decoded (Fig. 3e and Extended Data Fig. 2c). In addition, the efficiency of the sequencing, represented by the amount of

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

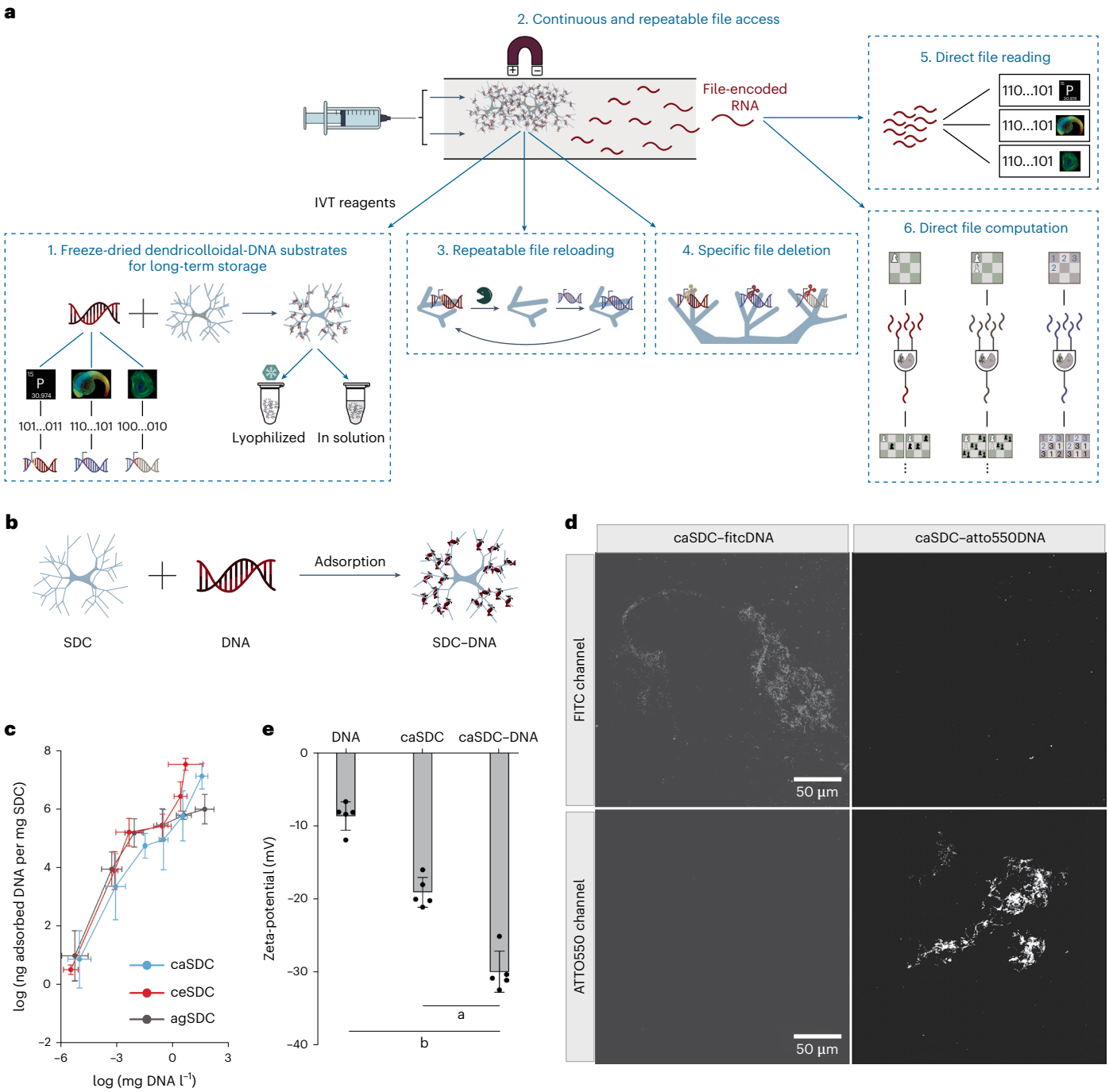


Fig. 1 | A primordial DNA store and compute engine is enabled by adsorbing DNA onto soft dendritic colloids. a, Immobilizing file-encoded DNA to high-surface-area dendricolloidal materials enables continuous and repeatable file access in a microfluidic-based device through in vitro transcription. The generated file-encoded RNA can be used for direct file reading and file computation. Additionally, DNA files stored on the dendricolloidal materials can be directly erased and reloaded with the same or new information, and stored lyophilized with minimal loss in repeated freeze–thaws. **b**, Schematic of DNA binding to an SDC particle. **c**, The amount of DNA adsorbed to the SDCs was quantified by the amount of DNA depleted from the solution phase as measured using real-time quantitative PCR. Plotted values represent the arithmetic mean, and error bars represent the s.d. of three independent experiments for binding of DNA to the SDCs. **d**, Images of caSDC bound with fluorophore-conjugated DNA. fitcDNA, FITC-labelled DNA; atto550DNA, DNA labelled with ATTO550 flurophore. **e**, Zeta-potential measurement for DNA, caSDC and caSDC bound with DNA. Plotted values represent the arithmetic mean, and error bars represent the s.d. of five independently syntheiszed SDC samples. '–' denotes non-covalent bonding. Samples in **c** were imaged using the same microscope settings, and adjusted identically for quantification purposes. Statistics were calculated using one-way analysis of variance (ANOVA) with Tukey–Kramer post hoc test for **e**. a, $P = 1.14 \times 10^{-4}$; b, $P = 6.87 \times 10^{-7}$.

'junk' sequencing reads, remained consistent, suggesting the quality of RNA transcribed was maintained over repeated IVT reactions (Fig. 3f).

Another important consideration for data storage in DNA is its stability over time. Lyophilization is commonly used to preserve DNA but can degrade DNA upon each freeze–thaw cycle[40]. In contrast, DNA in solution can be accessed multiple times but experiences degradation over time[40]. We therefore investigated how multiple rounds of lyophilization of SDC–DNA would affect the yield of cDNA compared

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

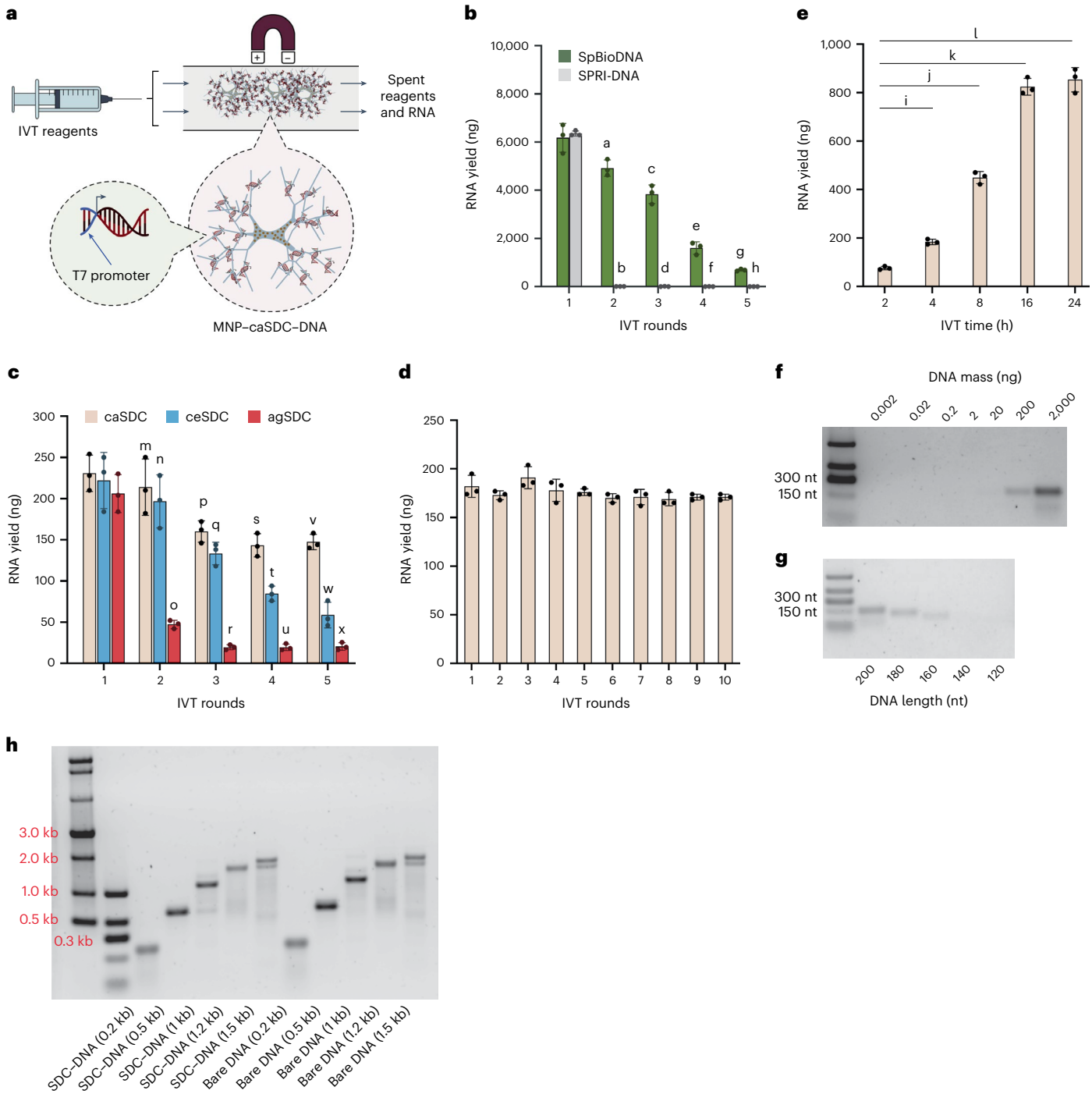


Fig. 2 | DNA bound to soft dendritic colloids can be repeatedly transcribed in a microfluidic channel. a, Schematic illustration of the process. MNP, magnetic nanoparticle. **b**, RNA yield from IVT of biotin-labelled DNA immobilized onto streptavidin-functionalized magnetic beads (SpBioDNA), and from IVT of DNA immobilized onto SPRI beads (SPRI-DNA). **c**, RNA yield from five sequential rounds of IVT of DNA adsorbed to caSDC, ceSDC and agSDC. **d**, RNA yield from ten sequential rounds of IVT of DNA adsorbed to caSDC. **e**, RNA yield from different IVT incubation lengths with DNA adsorbed to caSDC. **f**, RNA generated from IVT of different masses of DNA adsorbed to caSDC. **g**, RNA generated from IVT of DNA of different lengths adsorbed to caSDC. **h**, RNA generated from IVT of DNA with longer lengths adsorbed to caSDC, compared to RNA generated from IVT of the same DNA without adsorbing to caSDC. Plotted values represent the arithmetic mean, and error bars represent the s.d. of three independent IVT reactions. Statistics were calculated using one-way ANOVA with Tukey–Kramer post hoc test for **b**, **c** and **e**. Comparisons are relative to the first experimental condition and same type of substrate material in each plot for **b**, **c** and **e**. a, $P = 3.38 \times 10^{-4}$; b, $P = 5.89 \times 10^{-8}$; c, $P = 4.74 \times 10^{-3}$; d, $P = 5.91 \times 10^{-8}$; e, $P = 2.61 \times 10^{-4}$; f, $P = 5.87 \times 10^{-8}$; g, $P = 9.31 \times 10^{-5}$; h, $P = 5.86 \times 10^{-8}$; i, $P = 1.47 \times 10^{-4}$; j, $P = 1.48 \times 10^{-5}$; k, $P = 3.56 \times 10^{-6}$; l, $P = 1.20 \times 10^{-5}$; m, $P = 5.17 \times 10^{-1}$; n, $P = 4.01 \times 10^{-1}$; o, $P = 3.11 \times 10^{-4}$; p, $P = 8.90 \times 10^{-3}$; q, $P = 1.39 \times 10^{-2}$; r, $P = 1.56 \times 10^{-4}$; s, $P = 4.47 \times 10^{-3}$; t, $P = 2.52 \times 10^{-3}$; u, $P = 1.60 \times 10^{-4}$; v, $P = 3.88 \times 10^{-3}$; w, $P = 1.65 \times 10^{-3}$; x, $P = 1.66 \times 10^{-4}$.

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

to lyophilization of DNA alone or DNA in solution without lyophilization. We observed a negligible decrease in strand retention and in cDNA quantity over five rounds of lyophilization and IVT of File3 SDC–DNA (Fig. 3g). In contrast, DNA in solution and lyophilized DNA both exhibited greater loss in strand retention and cDNA yield. Interestingly, the complexation of DNA with the SDC appeared to protect the DNA from degradation or loss due to repeated freeze–thaws. In addition, the cDNA generated from the lyophilized SDC–DNA complex was also sequenced via Illumina chemistry and maintained similar strand distributions and strand retentions over five rounds of lyophilization and IVT (Fig. 3h,i). Error rates also remained low and did not increase with repeated rounds of lyophilization and IVT (Extended Data Fig. 2d). Further analysis showed that the lyophilized complex could theoretically provide up to 172 file accesses without degrading the decoding performance, whereas the SDC–DNA maintained in solution and lyophilized bare DNA could be accessed for 122 and 65 rounds, respectively (Supplementary Note 4). Comparison between the three conditions is of more direct utility because the absolute numbers can be arbitrarily augmented by increasing the copy number or total mass of each unique sequence.

Degradation can also be associated with the length of time in storage. Because it would be impossible to perform true long-term

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

Fig. 3 | Complex DNA files can be stored, lyophilized and protected from accelerated ageing and repeatedly accessed on soft dendritic colloids. a, File design. RT, reverse transcription. **b**, Strand distribution density (left), skewness (middle) and kurtosis (right) for File3. Illumina sequencing of synthesized File3 (Original File3), cDNA after IVT of unbound File3 (IVT of File3) and cDNA after IVT of File3 adsorbed to caSDC (IVT of SDC-File3). **c**–**f**, IVT was performed multiple times from caSDC-File1, with resultant cDNA yields, normalized to round 1 (**c**), violin plots of strand distributions (**d**), percentages of unique strands, with coloured boxes indicating all runs resulted in successful ('1') or unsuccessful ('0') decoding (**e**) and percentages of sequencing reads that are in File1 (**f**). **g**, cDNA amounts after each sequential round of lyophilization, reconstitution and IVT accessing of File3 DNA: adsorbed to caSDC, adsorbed to caSDC and maintained in solution, or lyophilized, normalized to the first round of IVT. **h**,**i**, Lyophilization, reconstitution and IVT was performed multiple times from caSDC-File3, with resultant violin plots of strand distributions (left) and percentage of unique DNA strands (right) (**h**), and strand distribution skewness (left) and kurtosis (right) (**i**). **j**–**l**, Lyophilized or solubilized SDC-File3 was incubated at 65 °C, followed by IVT, yielding amounts of cDNA (**j**), violin plots of the strand distributions of cDNA (left) and the percentage of unique strands (right) (**k**) and strand distribution skewness (top) and kurtosis (bottom) (**l**). Plotted values represent the arithmetic mean, and error bars represent the s.d. of three independent IVT reactions. Statistics were calculated using one-way ANOVA with Tukey–Kramer post hoc test for **b**, **c** and **g**–**l**. Comparisons are relative to the first experimental condition in each plot for **c**, **d** and **k**, and relative to the lyophilized SDC–DNA condition in each round in **g**. a, $P = 3.36 \times 10^{-4}$; b, $P = 1.24 \times 10^{-3}$; c, $P = 6.67 \times 10^{-1}$; d, $P = 7.47 \times 10^{-4}$; e, $P = 5.72 \times 10^{-3}$; f, $P = 5.19 \times 10^{-1}$; g, $P = 2.39 \times 10^{-2}$; h, $P = 4.78 \times 10^{-3}$; i, $P = 1.84 \times 10^{-3}$; j, $P = 3.39 \times 10^{-3}$; k, $P = 1.68 \times 10^{-3}$; l, $P = 1.57 \times 10^{-3}$; m, $P = 4.25 \times 10^{-4}$; n, $P = 1.26 \times 10{-}4$; o, $P = 1.18 \times 10^{-3}$; p, $P = 9.06 \times 10^{-1}$; q, $P = 1.10 \times 10^{-4}$; r, $P = 1.29 \times 10^{-1}$; s, $P = 2.16 \times 10^{-1}$; t, $P = 9.99 \times 10^{-1}$; u, $P = 9.99 \times 10^{-1}$; v, $P = 6.23 \times 10^{-1}$; w, $P = 5.29 \times 10^{-8}$; x, $P = 8.71 \times 10^{-1}$; aa, $P = 9.99 \times 10^{-1}$; ab, $P = 9.99 \times 10^{-1}$; ac, $P = 2.36 \times 10^{-1}$; ad, $P = 1.39 \times 10^{-1}$; ae, $P = 6.61 \times 10^{-2}$; af, $P = 5.75 \times 10^{-2}$; ag, $P = 1.20 \times 10^{-2}$; ah, $P = 1.35 \times 10^{-2}$; ai, $P = 8.19 \times 10^{-2}$; aj, $P = 2.49 \times 10^{-3}$; al, $P = 7.85 \times 10^{-1}$; am, $P = 9.97 \times 10^{-1}$; an, $P = 1.08 \times 10^{-1}$; ao, $P = 5.89 \times 10^{-1}$; ap, $P = 6.56 \times 10^{-1}$; aq, $P = 7.34 \times 10^{-1}$; ar, $P = 9.83 \times 10^{-1}$; as, $P = 7.43 \times 10^{-1}$; ba, $P = 9.99 \times 10^{-1}$; bb, $P = 3.56 \times 10^{-1}$; bc, $P = 3.06 \times 10^{-1}$; bd, $P = 1.37 \times 10^{-1}$; be, $P = 9.46 \times 10^{-2}$; bf, $P = 4.98 \times 10^{-4}$; bg, $P = 6.07 \times 10^{-4}$; bh, $P = 1.33 \times 10^{-1}$; bi, $P = 1.78 \times 10^{-3}$; bj, $P = 3.17 \times 10^{-3}$; bk, $P = 5.56 \times 10^{-3}$; bl, $P = 3.94 \times 10^{-3}$; bm, $P = 9.60 \times 10^{-2}$; bn, $P = 3.76 \times 10^{-2}$; bo, $P = 9.46 \times 10^{-3}$; bp, $P = 6.03 \times 10^{-1}$; bq, $P = 1.17 \times 10^{-1}$; br, $P = 1.94 \times 10^{-2}$; bs, $P = 6.53 \times 10^{-2}$; bt, $P = 3.26 \times 10^{-1}$; bu, $P = 1.81 \times 10^{-1}$; bv, $P = 7.28 \times 10^{-1}$.

experiments over decades or centuries, we modelled accelerated ageing with elevated temperatures[6]. We prepared SDC–DNA in both lyophilized and solubilized forms and incubated them in a thermocycler at 65 °C for 0, 8, 16, 24 and 48 h. Lyophilized SDC–DNA exhibited slower decay than solubilized SDC–DNA (decay rates of 4.00×10^{-7} s^{-1} and 9.33×10^{-7} s^{-1}, and half-lives of 0.0574 years and 0.0238 years, respectively; Fig. 3j and Supplementary Note 5). By fitting and extrapolating from previous models[6], this is equivalent to storing the DNA at 4 °C with half-lives of approximately 6,000 and 4,000 years, for the lyophilized and solubilized forms of SDC–DNA, respectively, or storing them in the Global Seed Vault (−18 °C) for 2 million and 0.8 million years, respectively. Strand distributions and strand retentions were maintained during the 48 h experiment (Fig. 3k,l) and error rates did not increase (Extended Data Fig. 2e). Finally, we also investigated if the efficiency of IVT was impacted by the microfluidic architecture versus being performed in bulk in a test tube. We observed minimal impact on RNA yield within the microfluidic system (Extended Data Fig. 2f). Overall, this system supports robust and stable file access repeatedly and over long periods of time.

Erasing and loading data on SDCs

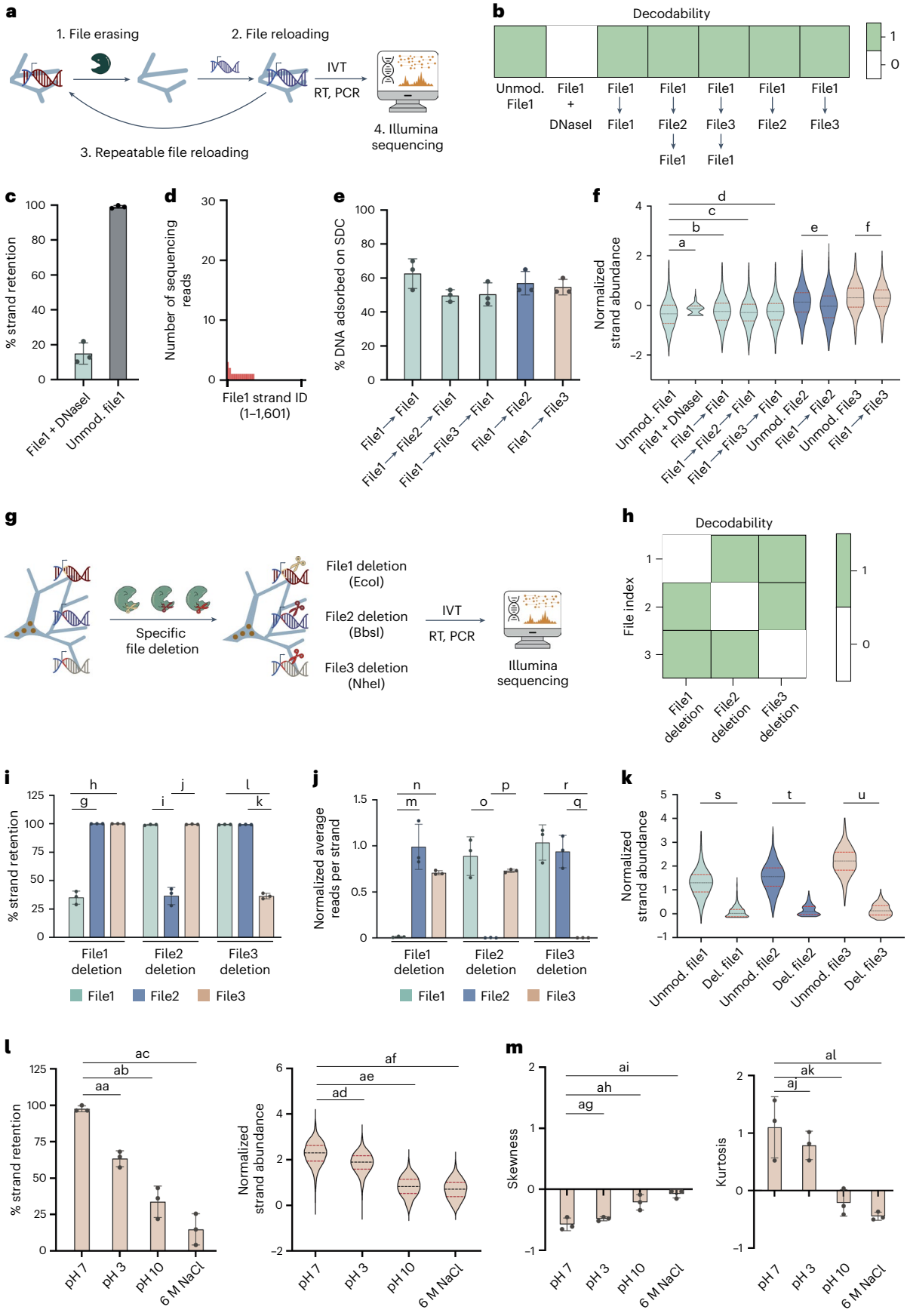
A core feature of traditional computers is the ability to store and work with different sets of data, including deleting specific files, adding new data and erasing an entire hard disk. The SDC–DNA system could enable a physical instantiation of such functions (Fig. 4a). To test this, we immobilized File1 on SDCs, added DNaseI[41], and then adsorbed File1, File2 or File3 individually. As expected, DNaseI treatment rendered us no longer able to decode File1 (Fig. 4b and Supplementary Note 3). It removed most of the total DNA strands by mass, and >80% of unique DNA strands were no longer detected (Fig. 4c,d and Extended Data Fig. 3b)[42]. Illumina sequencing also indicated successful and repeated loading of new files onto the SDCs with similar strand distributions as the original unbound files, albeit with reduced efficiency of roughly 60% compared with pristine SDCs (Fig. 4e,f).

To implement selective deletion of specific files when all three files are simultaneously present, all strands of each distinct file were designed with a common restriction endonuclease recognition sequence that did not appear in the strands of any other file[43]. We immobilized all three files together onto SDCs (Fig. 4g). The endonucleases were able to specifically cleave each file such that only the cleaved file was no longer decodable (Fig. 4h). Overall, in both deleting specific files and the entire database, the deletion process removed most of the DNA strands by mass, and <40% of the unique sequences were retained (Fig. 4i–k and Extended Data Fig. 3a,b). Finally, the process did not affect the strand distributions and error rates obtained when decoding the remaining or reloaded files (Fig. 4k, Extended Data Fig. 3c,d,f and Supplementary Methods).

Although DNA can be removed enzymatically, we investigated whether it could be further reduced by altering the buffer pH and ionic strength[29,44–47]. We immobilized File3 on SDCs and incubated the complex in buffers of pH 3, pH 7, pH 10 or 6 M NaCl (at pH 7). After IVT and reverse transcription we observed that the alkaline and high salt

Fig. 4 | Data files can be specifically erased from and reloaded onto soft dendritic colloids. a, Schematic of erasing, reloading and reading files. **b**, '→' denotes erasing the indicated File DNA from SDCs and reloading with the new specified File. Unmod., unmodified. '1' represents that all decoding runs were successful and '0' represents that none were successful. **c**,**d**, After DNaseI treatment, the percentage of unique strands of File1 remaining (**c**) and the number of sequencing reads per each unique strand in File1 (**d**), as measured by Illumina sequencing, were determined. **e**, Percentage of new File DNA adsorbed to caSDC after erasing the original file on SDCs. Values were quantified by the amount of DNA depleted from the solution phase using quantitative PCR and plotted as a percentage of the original file amount. **f**, Violin plots of File strand distributions after the indicated erasures and reloadings. **g**, Schematic of specific file deletion from a three-file database adsorbed to caSDCs. **h**–**k**, After restriction enzyme erasure of each file followed by subsequent reading of each file (**h**), we determined the percentage of unique strands of each file plotted as a percentage of the total unique strands (**i**), the number of sequencing reads for strands in each File normalized to the number of sequencing reads found in untreated File DNA prior to IVT (**j**), and violin plots of the strand distributions of File cDNA before and after deletion (**k**). Prior to reading, caSDC-File3 was incubated in various buffer conditions: pH 7 (control), pH 3, pH 10 or 6 M NaCl. **l**,**m**, The resultant percentage of unique strands remaining of File3 (left) and violin plots of the strand distributions of File 3 cDNA (right), normalized to the pH 7 condition (**l**), and skewness and kurtosis values (**m**). Plotted values represent the arithmetic mean across all strand sequences in a File, and error bars represent the s.d. of three independent replicate IVT reactions. Statistics were calculated using one-way ANOVA with Tukey–Kramer post hoc test for **f** and **i**–**m**. a, $P = 5.30 \times 10^{-8}$; b, $P = 5.78 \times 10^{-4}$; c, $P = 4.63 \times 10^{-1}$; d, $P = 8.43 \times 10^{-3}$; e, $P = 1.54 \times 10^{-5}$; f, $P = 6.71 \times 10^{-1}$; g, $P = 3.91 \times 10^{-5}$; h, $P = 3.92 \times 10^{-5}$; i, $P = 1.35 \times 10^{-4}$; j, $P = 1.31 \times 10^{-4}$; k, $P = 1.47 \times 10^{-6}$; l, $P = 1.47 \times 10^{-6}$; m, $P = 2.33 \times 10^{-3}$; n, $P = 7.05 \times 10^{-7}$; o, $P = 1.82 \times 10^{-3}$; p, $P = 1.92 \times 10^{-7}$; q, $P = 7.74 \times 10^{-4}$; r, $P = 7.45 \times 10^{-4}$; s, $P = 7.01 \times 10^{-8}$; t, $P = 6.98 \times 10^{-8}$; u, $P = 7.02 \times 10^{-8}$; aa, $P = 6.39 \times 10^{-3}$; ab, $P = 2.19 \times 10^{-5}$; ac, $P = 2.07 \times 10^{-5}$; ad, $P = 4.76 \times 10^{-10}$; ae, $P = 4.58 \times 10^{-10}$; af, $P = 4.65 \times 10^{-10}$; ag, $P = 2.22 \times 10^{-1}$; ah, $P = 1.86 \times 10^{-2}$; ai, $P = 2.23 \times 10^{-3}$; aj, $P = 4.06 \times 10^{-1}$; ak, $P = 1.71 \times 10^{-2}$; al, $P = 7.54 \times 10^{-3}$.

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



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

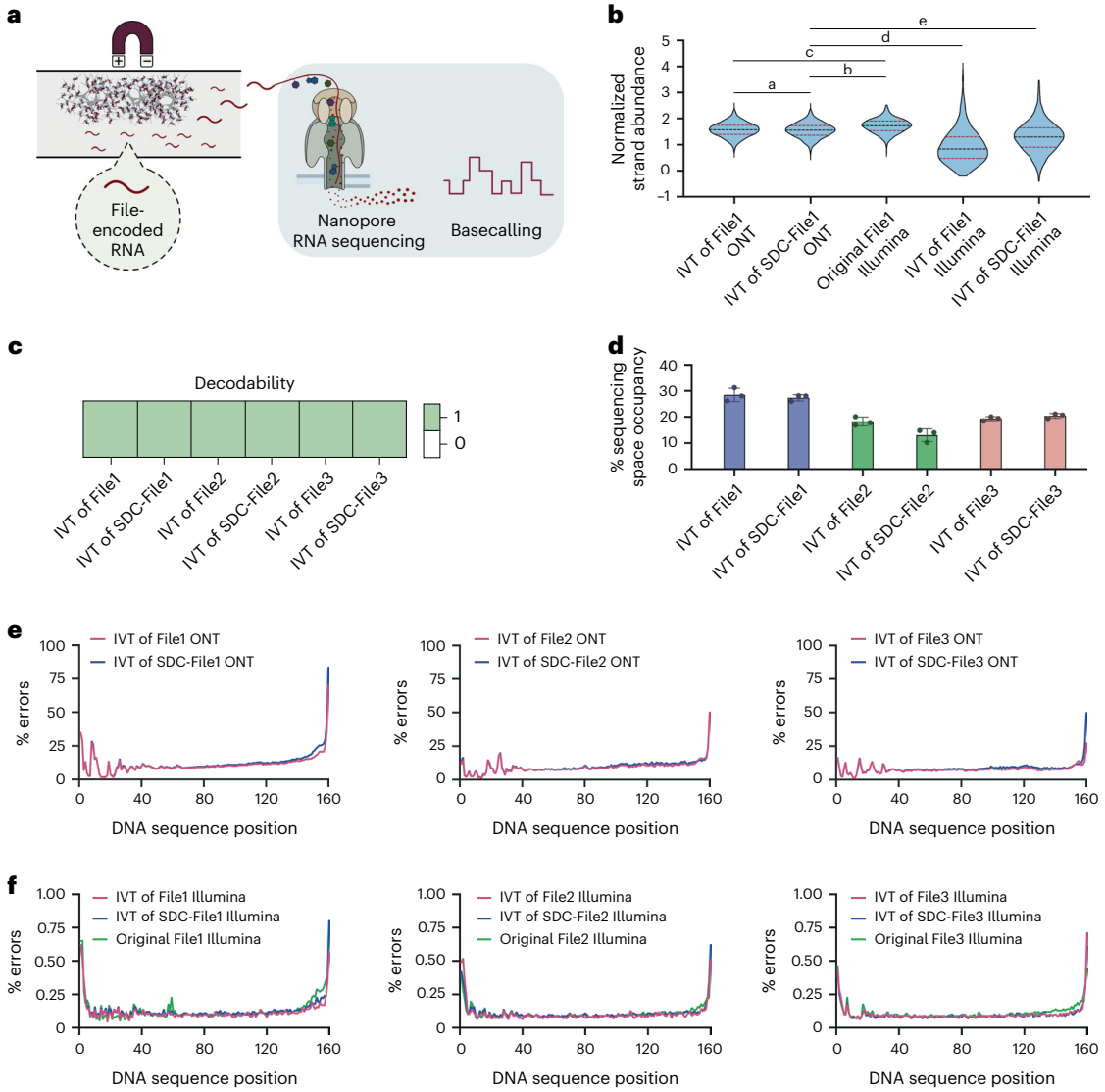


Fig. 5 | RNA nanopore sequencing promotes continuous data processing and reduces skewing of strand distributions. a, Schematic of direct nanopore sequencing of complex files adsorbed onto SDCs. **b**, Violin plots of the strand distributions for experimental samples of File1. These samples include direct sequencing of the File1 DNA obtained from the DNA synthesis provider (Original File1), RNA and cDNA obtained after IVT of File1 DNA adsorbed to caSDC (IVT of SDC-File1), and RNA and cDNA obtained from unbound File1 DNA (IVT of File1). RNA samples were processed with ONT, and cDNA samples were processed with Illumina sequencing. **c**, Alignment of nanopore sequencing reads obtained from RNA after IVT of SDC–DNA or unbound DNA. **d**, The percentage of all sequencing reads for a targeted file, obtained from RNA after IVT of unbound File DNA or SDC–DNA. Values were measured by ONT sequencing and plotted as a percentage of the total sequencing reads. **e**, The percentage error for each DNA sequence position in File RNA obtained after IVT of samples processed with ONT sequencing. **f**, The percentage error for each DNA sequence position in File cDNA obtained after IVT of samples processed with Illumina sequencing. The error rate was calculated by dividing the number of errors of a given type occurring at a nucleotide position by the total number of reads for that sequence (Methods). Plotted values represent the arithmetic mean, and error bars represent the s.d. of three independent IVT reactions. Statistics were calculated using one-way ANOVA with Tukey–Kramer post hoc test for **b**. a, $P = 6.87 \times 10^{-1}$; b, $P = 2.22 \times 10^{-11}$; c, $P = 2.23 \times 10^{-11}$; d, $P = 2.21 \times 10^{-11}$; e, $P = 2.22 \times 10^{-11}$.

conditions achieved greater reduction in strand retention than through enzymatic digestion (Fig. 4l,m and Extended Data Fig. 3e).

Towards continuous operation with RNA nanopore sequencing

Reducing latency may improve downstream automation and the coordination of unit processes. Nanopore sequencing technologies present the potential to address the most significant bottlenecks by providing live readouts of data as nucleic acids are being sequenced, and bypassing the need to convert the RNA to DNA by directly sequencing RNA[48–50]. Here we used Oxford Nanopore Technologies (ONT) to directly read the RNA generated from the SDC–DNA system (Fig. 5a). We first checked if ONT sequencing would negatively affect the recovery of the data by skewing the read distributions. Reassuringly, RNA reads obtained after IVT of File1, File2 and File3 DNA had similar distributions as the original synthesized DNA libraries sequenced by Illumina chemistry (Fig. 5b and Extended Data Fig. 4a–d), and all files were successfully decoded whether accessed from free DNA or from DNA adsorbed to SDCs (Fig. 5c–f). Moreover, we observed that by skipping the cDNA generation step, the distributions were more monodisperse using direct RNA ONT sequencing compared with Illumina chemistry.

Implementation of addressable in-storage computation

The final property we addressed in this work is non-destructive computing. We took advantage of the fact that the data are accessed by making RNA copies of the DNA[51], providing the opportunity to compute

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

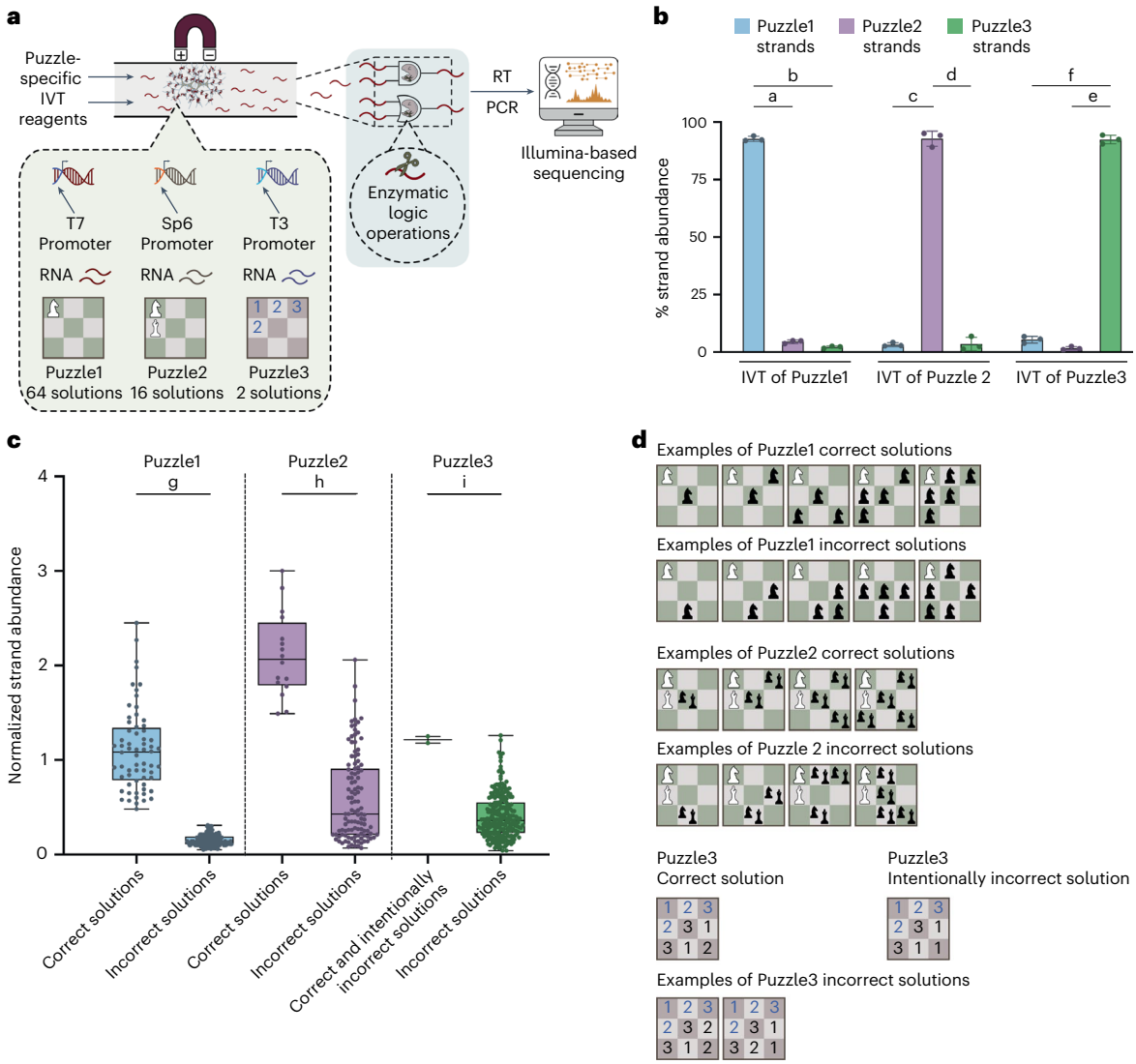


Fig. 6 | Soft dendritic colloids support addressable in-storage computation. **a**, Schematic of direct molecular computation for complex files adsorbed onto caSDC. The puzzle database contains the DNA for all three puzzles. **b**, The percentage of unique strands of each Puzzle cDNA after IVT of the Puzzle database adsorbed to caSDC using puzzle-specific polymerases. Values were measured by Illumina sequencing and plotted as a percentage of the total sequencing reads of the Puzzle database (Methods). Plotted values represent the arithmetic mean, and error bars represent the s.d. of three independent IVT reactions. **c**, Box plots of the strand distributions of Puzzle cDNA obtained after IVT of the Puzzle database adsorbed to caSDC using puzzle-specific polymerase, followed by enzymatic computation for each puzzle. The relative frequency at which each oligo sequence appeared was measured by Illumina sequencing and plotted as individual dots. Correct Solutions and Incorrect Solutions are oligos that do not and do violate the puzzle rules, respectively. Intentionally Incorrect Solutions are oligos that do violate the puzzle rules but were expected to be obtained as a solution by intentionally altering the physical computational steps. The box plots display the arithmetic mean, ±s.d., and maximum and minimum values of three independent puzzle computation reactions. **d**, Schematic of correct and incorrect solutions for Puzzle1, Puzzle2 and Puzzle3, and the intentionally incorrect solution for Puzzle3. Statistics were calculated using one-way ANOVA with Tukey–Kramer post hoc test for **b** and **c**. a, $P = 3.66 \times 10^{-8}$; b, $P = 2.04 \times 10^{-8}$; c, $P = 1.37 \times 10^{-6}$; d, $P = 3.90 \times 10^{-6}$; e, $P = 1.63 \times 10^{-7}$; f, $P = 3.70 \times 10^{-7}$; g, $P = 2.23 \times 10^{-10}$; h, $P = 2.22 \times 10^{-10}$; i, $P = 1.06 \times 10^{-3}$.

upon the RNA without destroying the DNA. We took inspiration from previous work that used RNaseH to selectively degrade RNAs that hybridized with 20 nt DNA oligos[52]. This property of RNaseH allows logical operations to be performed by adding DNA oligos in different combinations and temporal sequences to degrade all RNAs containing certain complementary sequences. For example, it could degrade all sequences representing states of a puzzle that are not valid solutions.

We ordered 1,000 distinct 250 nt DNA oligos comprising all possible correct and incorrect configurations of three different puzzles (Fig. 6a). Each DNA oligo is comprised of nine distinct 20 nt positions with each position representing the state of a square of the 3 × 3 puzzles. Each position can be one of two or three possible sequences, with each sequence representing no piece present, a knight present or a bishop present (or the number 1, 2 or 3 in the case of the sudoku board). Each oligo therefore represents a series of nine specific 20 nt sequences representing one potential board configuration. The computational approach is to take these oligo pools that represent all possible board configurations and eliminate those oligo sequences that violate the puzzle rules, leaving behind only oligos representing correct puzzle solutions (Extended Data Fig. 5 and Supplementary Methods). All DNA oligos related to each puzzle contained a common RNA polymerase promoter sequence distinct from the other puzzles. All Puzzle1 DNA oligos contained a T7 promoter; Puzzle2 oligos contained an Sp6 promoter; Puzzle3 oligos contained a T3 promoter. Puzzle1 is a chess problem for which solutions are all the board configurations with knights, where a white knight is present at the top left corner, and no

new black knights are placed on the board that can be attacked by the white knight. In Puzzle2, a white knight and bishop occupy the top left and middle left positions, respectively, and solutions are sought in which no new black knight or bishop can be placed on the board that is attacked by either white piece. Puzzle3 is a simplified sudoku problem in which the first four positions are defined with the digits 1, 2, 3 and 2, and each row and column must have each digit appear only once. To further test the accuracy of these physical computational steps, we intentionally altered one physical step so that the ninth position of Puzzle3 could intentionally but erroneously allow both the correct digit 2, but also the incorrect digit 1, to be present.

Strands related to individual puzzles were transcribed from SDC–DNA by adding their corresponding RNA polymerase (T7, Sp6 or T3) (Fig. 6b). The resultant RNA represented all potential game board configurations for that puzzle. We then designed an algorithm (Supplementary Methods), comprised of adding combinations of DNA oligos complementary to the RNA along with RNaseH, to destroy all RNA strands that violated the puzzle rules, leaving behind only RNAs representing the correct solutions. Following computation, the large majority of the surviving strands were of the expected solutions (Fig. 6c,d, Extended Data Fig. 5b and Extended Data Table 3).

Conclusion

Here we demonstrated the ability to implement core primordial features of information systems, including the ability to stably store, erase, reload, read and compute on specific data from a substrate in a non-destructive manner, and the overall ability to execute such functions in a relatively seamless, programmable and continuous manner. It is important to consider both the limitations and future possibilities of this system.

The current system impacts theoretical idealized estimations for information density and energy efficiency in both positive and negative ways. By adsorbing DNA onto SDCs, substantial information density is immediately sacrificed by the extra volume required by the SDCs. However, despite this sacrifice, the information density remains very high at 10^4 TB cm^{-3} (Extended Data Table 2), and could be improved further by optimizing the dendritic structure of the SDCs[26] or by using another high-surface-area material that could be produced by the expanded family of liquid–liquid techniques[30].

Often ignored are the physical handling steps required in molecular computation. For example, logical operations using nucleic acids often require splitting, pooling and executing multiple distinct enzymatic or chemical reactions. The SDC–DNA system presented here provides a format that is compatible with automated liquid handling and could incorporate more complex valving and mixing in a space-efficient format in the future[53–55].

This system is also compatible with an information management system using nucleic-acid-based logic gate operations[56,57], sitting between archival storage and computation[58,59]. Interestingly, complexation with SDCs seems to provide enhanced protection of DNA both over time and to repeated lyophilization and reconstitution, suggesting another way in which it may serve as a universal data substrate for archival storage through to computational applications.

Additional challenges specific to this current system involve primarily the efficiencies and completeness of each processing step, including deletion and reloading of data. Just like programming and design strategies by-pass physical defects that arise in electronic systems, analogous strategies could be developed to deal with incomplete molecular processes. For example, computational filtering or winner-takes-all strategies could be leveraged to handle incomplete molecular processing steps[60].

Finally, the restriction enzymes and RNA polymerases used in this work are limited in number and cannot be used to address data at scale. However, they provide a proof of principle that biomolecular machinery could be leveraged towards this challenge. Future work might implement creative instantiations of more scalable approaches such as toehold switches or invent new forms of CRISPR technologies that can recruit polymerases or act as nucleases in a customizable and sequence-specific manner.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41565-024-01771-6.

References

1. Ceze, L., Nivala, J. & Strauss, K. Molecular digital data storage using DNA. *Nat. Rev. Genet.* **20**, 456–466 (2019).
2. Copeland, B. J. in *The Stanford Encyclopedia of Philosophy* Winter 2020 edn (ed. Zalta, E. N.) (Stanford Univ., 2020).
3. Ceruzzi, P. E. A history of modern computing. *Choice Rev. Online* **36**, 36-4531–36-4531 (1999).
4. Goldman, N. et al. Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. *Nature* **494**, 77–80 (2013).
5. Church, G. M., Gao, Y. & Kosuri, S. Next-generation digital information storage in DNA. *Science* **337**, 1628 (2012).
6. 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. *Angew. Chem. Int. Ed.* **54**, 2552–2555 (2015).
7. Blawat, M. et al. Forward error correction for DNA data storage. *Procedia Comput. Sci.* **80**, 1011–1022.
8. Erlich, Y. & Zielinski, D. DNA Fountain enables a robust and efficient storage architecture. *Science* **355**, 950–954 (2017).
9. Lee, H. H., Kalhor, R., Goela, N., Bolot, J. & Church, G. M. Terminator-free template-independent enzymatic DNA synthesis for digital information storage. *Nat. Commun.* **10**, 2383 (2019).
10. Palluk, S. et al. De novo DNA synthesis using polymerase-nucleotide conjugates. *Nat. Biotechnol.* **36**, 645–650 (2018).
11. Lopez, R. et al. DNA assembly for nanopore data storage readout. *Nat. Commun.* **10**, 2933 (2019).
12. Mao, C., LaBean, T. H., Reif, J. H. & Seeman, N. C. Logical computation using algorithmic self-assembly of DNA triple-crossover molecules. *Nature* **407**, 493–496 (2000).
13. Adleman, L. M. Molecular computation of solutions to combinatorial problems. *Science* **266**, 1021–1024 (1994).
14. Organick, L. et al. Random access in large-scale DNA data storage. *Nat. Biotechnol.* **36**, 242–248 (2018).
15. Tabatabaei Yazdi, S. M. H., Yuan, Y., Ma, J., Zhao, H. & Milenkovic, O. A rewritable, random-access DNA-based storage system. *Sci. Rep.* **5**, 1–10 (2015).
16. Yazdi, S. M. H. T., Gabrys, R. & Milenkovic, O. Portable and error-free DNA-based data storage. *Sci. Rep.* https://doi.org/10.1038/s41598-017-05188-1 (2017).
17. Bornholt, J. et al. A DNA-based archival storage system. In *Proc. 21st International Conference on Architectural Support for Programming Languages and Operating Systems—ASPLOS '16* (eds Conte, T. & Zhou, Y.) 637–649 (ACM Press, 2016).
18. Bögels, B. W. A. et al. DNA storage in thermoresponsive microcapsules for repeated random multiplexed data access. *Nat. Nanotechnol.* **18**, 912–921 (2023).
19. Benenson, Y. et al. Programmable and autonomous computing machine made of biomolecules. *Nature* **414**, 430–434 (2001).
20. Bell, N. A. W. & Keyser, U. F. Digitally encoded DNA nanostructures for multiplexed, single-molecule protein sensing with nanopores. *Nat. Nanotechnol.* **11**, 645–651 (2016).

21. Dickinson, G. D. et al. An alternative approach to nucleic acid memory. *Nat. Commun.* **12**, 2371 (2021).
22. Chen, K. et al. Digital data storage using DNA nanostructures and solid-state nanopores. *Nano Lett.* **19**, 1210–1215 (2019).
23. Chen, K., Zhu, J., Bošković, F. & Keyser, U. F. Nanopore-based DNA hard drives for rewritable and secure data storage. *Nano Lett.* **20**, 3754–3760 (2020).
24. Zhang, Y. et al. DNA origami cryptography for secure communication. *Nat. Commun.* **10**, 5469 (2019).
25. Numajiri, K., Kimura, M., Kuzuya, A. & Komiyama, M. Stepwise and reversible nanopatterning of proteins on a DNA origami scaffold. *Chem. Commun.* **46**, 5127 (2010).
26. 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).
27. Williams, A. H. et al. Printable homocomposite hydrogels with synergistically reinforced molecular-colloidal networks. *Nat. Commun.* **12**, 2834 (2021).
28. Bergenstråhle, M., Wohlert, J., Himmel, M. E. & Brady, J. W. Simulation studies of the insolubility of cellulose. *Carbohydr. Res.* **345**, 2060–2066 (2010).
29. Lindman, B., Medronho, B., Alves, L., Norgren, M. & Nordenskiöld, L. Hydrophobic interactions control the self-assembly of DNA and cellulose. *Q. Rev. Biophys.* **54**, e3 (2021).
30. Bang, R. S., Roh, S., Williams, A. H., Stoyanov, S. D. & Velev, O. D. Fluid flow templating of polymeric soft matter with diverse morphologies. *Adv. Mater.* **35**, 2211438 (2023).
31. Ali, M. E. et al. in *Reference Module in Materials Science and Materials Engineering* (Elsevier, 2016); https://doi.org/10.1016/B978-0-12-803581-8.04075-3
32. Paul, A. & Bhattacharya, S. Chemistry and biology of DNA-binding small molecules. *Curr. Sci.* **102**, 212–231 (2012).
33. Koch, J. et al. A DNA-of-things storage architecture to create materials with embedded memory. *Nat. Biotechnol.* **38**, 39–43 (2020).
34. 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**, 1700701 (2018).
35. Stetefeld, J., McKenna, S. A. & Patel, T. R. Dynamic light scattering: a practical guide and applications in biomedical sciences. *Biophys. Rev.* **8**, 409–427 (2016).
36. Lin, K. N., Volkel, K., Tuck, J. M. & Keung, A. J. Dynamic and scalable DNA-based information storage. *Nat. Commun.* **11**, 2981 (2020).
37. Fisher, S. et al. A scalable, fully automated process for construction of sequence-ready human exome targeted capture libraries. *Genome Biol.* **12**, R1 (2011).
38. DeAngelis, M. M., Wang, D. G. & Hawkins, T. L. Solid-phase reversible immobilization for the isolation of PCR products. *Nucleic Acids Res.* **23**, 4742–4743 (1995).
39. Chen, Y.-J. et al. Quantifying molecular bias in DNA data storage. *Nat. Commun.* **11**, 3264 (2020).
40. Matange, K., Tuck, J. M. & Keung, A. J. DNA stability: a central design consideration for DNA data storage systems. *Nat. Commun.* **12**, 1358 (2021).
41. Lauková, L., Konečná, B., Janovičová, Ľ., Vlková, B. & Celec, P. Deoxyribonucleases and their applications in biomedicine. *Biomolecules* **10**, 1036 (2020).
42. Robinson, P. K. Enzymes: principles and biotechnological applications. *Essays Biochem* **59**, 1–41 (2015).
43. Loenen, W. A. M., Dryden, D. T. F., Raleigh, E. A., Wilson, G. G. & Murray, N. E. Highlights of the DNA cutters: a short history of the restriction enzymes. *Nucleic Acids Res.* **42**, 3–19 (2014).
44. Allemand, J. F., Bensimon, D., Jullien, L., Bensimon, A. & Croquette, V. pH-dependent specific binding and combing of DNA. *Biophys. J.* **73**, 2064–2070 (1997).
45. Vandeventer, P. E. et al. Multiphasic DNA adsorption to silica surfaces under varying buffer, pH, and ionic strength conditions. *J. Phys. Chem. B* **116**, 5661–5670 (2012).
46. Cai, J. & Zhang, L. Rapid dissolution of cellulose in LiOH/urea and NaOH/urea aqueous solutions. *Macromol. Biosci.* **5**, 539–548 (2005).
47. Jiménez-Ángeles, F. & Firoozabadi, A. Hydrophobic hydration and the effect of NaCl salt in the adsorption of hydrocarbons and surfactants on clathrate hydrates. *ACS Cent. Sci.* **4**, 820–831 (2018).
48. Workman, R. E. et al. Nanopore native RNA sequencing of a human poly(A) transcriptome. *Nat. Methods* **16**, 1297–1305 (2019).
49. Soneson, C. et al. A comprehensive examination of nanopore native RNA sequencing for characterization of complex transcriptomes. *Nat. Commun.* **10**, 3359 (2019).
50. Smith, M. A. et al. Molecular barcoding of native RNAs using nanopore sequencing and deep learning. *Genome Res.* **30**, 1345–1353 (2020).
51. Qiu, M. et al. RNA nanotechnology for computer design and in vivo computation. *Philos. Trans. R Soc. A* **371**, 20120310 (2013).
52. Faulhammer, D., Cukras, A. R., Lipton, R. J. & Landweber, L. F. Molecular computation: RNA solutions to chess problems. *Proc. Natl Acad. Sci. USA* **97**, 1385–1389 (2000).
53. 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).
54. Newman, S. et al. High density DNA data storage library via dehydration with digital microfluidic retrieval. *Nat. Commun.* **10**, 1706 (2019).
55. Luo, Y. et al. Integrated microfluidic DNA storage platform with automated sample handling and physical data partitioning. *Anal. Chem.* **94**, 13153–13162 (2022).
56. Gerasimova, Y. V. & Kolpashchikov, D. M. Towards a DNA nanoprocessor: reusable tile-integrated DNA circuits. *Angew. Chem.* **128**, 10400–10403 (2016).
57. Guz, N. et al. Bioelectronic interface connecting reversible logic gates based on enzyme and DNA reactions. *ChemPhysChem* **17**, 2247–2255 (2016).
58. Polak, R. E. & Keung, A. J. A molecular assessment of the practical potential of DNA-based computation. *Curr. Opin. Biotechnol.* **81**, 102940 (2023).
59. Yang, S. et al. DNA as a universal chemical substrate for computing and data storage. *Nat. Rev. Chem.* **8**, 179–194 (2024).
60. Cherry, K. M. & Qian, L. Scaling up molecular pattern recognition with DNA-based winner-take-all neural networks. *Nature* **559**, 370–376 (2018).

Methods

Adsorption of DNA to surface-modified SDC

First, 300 ng of fluorophore-labelled dsDNA was mixed with 60 µg of surface-modified SDC in a 100 µl reaction containing Dulbecco's phosphate-buffered saline (DPBS). The mixture was then placed in a tube rotator (VWR) and gently rotated at 4 C overnight. The next day, the mixture was briefly centrifuged and placed on a magnetic stand for 5 min. The supernatant was collected to evaluate the binding capacity, and the magnetized SDC particles were gently washed twice with DPBS.

Confocal imaging

SDC bound with fluorophore-labelled dsDNA was imaged with a Nikon A1R laser scanning confocal microscope using a 20× objective (numerical aperture, 0.75; working distance, 1 mm; field of view, 25 mm; CFI Plan Apo, Nikon Instruments). FITC was imaged with a 488 nm laser and 525/50 560DCXR 2FW emission filter. ATTO550 was imaged with a 561 nm laser and 600/50 640DCXR emission filter. Identical acquisition settings and post-processing were used for all images.

Binding assay of unlabelled DNA to surface-modified SDC

Surface-modified SDC was mixed with unlabelled dsDNA in a 100 µl reaction containing DPBS, at different DNA amounts ranging from 2 pg to 2 µg or at different DNA lengths ranging from 200 nt to 120 nt. The mixture was placed in a tube rotator (VWR) and gently rotated at 4 °C overnight. The next day, the mixture was briefly centrifuged and placed on a magnetic stand for 5 min. The supernatant was collected and processed with real-time PCR to quantify the bound and unbound DNA amounts.

Real-time quantitative PCR

Quantitative PCR was performed in a 6 µl, 384-well plate format using SsoAdvanced Universal SYBR Green Supermix (Bio-Rad, 1725270). The amplification conditions were 95 °C for 2 min and then 50 cycles of 95 °C for 15 s, 53 °C for 20 s, and 60 °C for 20 s. Quantities were interpolated from the linear ranges of standard curves performed on the same quantitative PCR plate.

IVT of SDC–DNA

A DNA oligo with a T7 promoter sequence was purchased from Azenta and used as a template to create dsT7-DNA using PCR, followed by purification with AMPure XP beads (Beckman Coulter, A63881) and elution in 40 µl of water. First, 300 ng of dsT7-DNA was bound to 60 µg of SDCs. The next day, the mixture was briefly centrifuged and placed on a magnetic stand for 5 min. The complex was washed twice with DPBS and directly mixed with 30 µl of in vitro transcription buffer (NEB, E2050) containing 2 µl of T7 RNA Polymerase Mix and ATP, TTP, CTP and GTP, each at 6.6 mM. The mixture was incubated at 37 °C for 16 h and purified by a Monarch RNA Cleanup Kit (NEB, T2040L) following the manufacturer's instructions. The newly generated RNA transcripts were measured using a NanoDrop Spectrophotometer and Fragment Analyzer HS RNA Kit (Agilent Technologies, DNF-472-0500). After transcription, the beads with the bound DNA were washed twice with 100 µl of low-salt buffer containing 20 mM Tris–HCl, 0.15 M NaCl and 2 mM EDTA pH 8, and the transcription process was repeated four times.

IVT of biotinylated DNA complexed with streptavidin beads

First, 60 µg of streptavidin magnetic beads (NEB, S1420S) were pre-washed using a high-salt buffer containing 20 mM Tris–HCl, 2 M NaCl and 2 mM EDTA pH 8, and incubated with 300 ng of biotinylated dsDNA at room temperature for 30 min. The supernatant was then decanted, and the beads were washed with 100 µl of high-salt buffer and directly mixed with 30 µl of in vitro transcription buffer (NEB, E2050) containing 2 µl of T7 RNA Polymerase Mix and ATP, TTP, CTP and GTP, each at 6.6 mM. The mixture was incubated at 37 °C for 16 h and purified by a Monarch RNA Cleanup Kit (NEB, T2040L) following the manufacturer's instructions. The newly generated RNA transcripts were measured using a NanoDrop Spectrophotometer and Fragment Analyzer HS RNA Kit (Agilent Technologies, DNF-472-0500). After transcription, the beads with the bound DNA were washed twice with 100 µl of low-salt buffer containing 20 mM Tris–HCl, 0.15 M NaCl and 2 mM EDTA pH 8, and the transcription process was repeated four times.

IVT of DNA on AMPure beads

First, 300 ng of unlabelled dsDNA was resuspended in 30 µl of water and mixed with 60 µg of AMPure XP beads (Beckman Coulter, A63881). The mixture was then washed twice with 200 µl of 80% ethanol (Fisher Scientific, AC615110010). After drying off excessive moisture, the beads were directly mixed with 30 µl of in vitro transcription buffer (NEB, E2050) containing 2 µl of T7 RNA Polymerase Mix and ATP, TTP, CTP and GTP, each at 6.6 mM. The mixture was incubated at 37 °C for 16 h and purified by Monarch RNA Cleanup Kit (NEB, T2040L) following manufacturer's instructions. The newly generated RNA transcripts were measured using a NanoDrop Spectrophotometer and Fragment Analyzer HS RNA Kit (Agilent Technologies, DNF-472-0500). After transcription, the beads with the bound DNA were washed twice with 200 µl of 80% ethanol and the transcription process was repeated four times.

Zeta-potential

The zeta-potentials of materials were measured on a Zetasizer Nano ZSP (Malvern Instruments). Pure dsDNA, caSDC-DNA and surface-modified caSDC were individually prepared and resuspended in 1 ml of DPBS. Ultaviolet-transparent disposable cuvettes (Cole Parmer, #759150) were used for measuring zeta-potential. The experiments were performed at room temperature (25 °C) with an equilibration time of 50 s. Each sample was tested three times with 100 runs per single measurement.

Lyophilization-IVT process

First, 500 ng of File3 DNA was bound to 60 µg of caSDCs, followed by washing twice with DPBS and resuspending in 40 µl DPBS. Prior to lyophilization, five samples were prepared for both SDC–DNA complex and bare DNA (both resuspended in 40 µl DPBS). A freeze-dryer device (FreeZone Freeze Dryer, Labconco) was used to lyophilize the samples following the guidance from manufacturer's user manuals. To prevent a burst from the initial pressure stabilization after loading the sample, a parafilm sealing film (HS234526B, Sigma Aldrich) was used to seal the Eppendorf tube with an open lid, and a pinhole was created in the sealing film. The lyophilization was processed until all moisture has been removed from the complex. A matrix was design to ensure samples were lyophilized in designated rounds for both SDC–DNA and bare DNA formats. After each round of lyophilization, unfinished samples were resuspended in the 40 µl DPBS and the process was repeated, whereas the finished dried samples were directly resuspended and mixed with 30 µl of IVT buffer as previously described. After IVT, the RNA was quantified using a NanoDrop Spectrophotometer, followed by reverse transcription using each file's specific reverse primer.

Thermal incubation and IVT

First, 500 ng of File3 DNA was bound to 60 µg of caSDCs, followed by washing twice with DPBS and resuspending in 40 µl DPBS. This was used as the wet form of the SDC–DNA complex. To create the dried form, a freeze-dryer device (FreeZone Freeze Dryer, Labconco) was used to lyophilize the samples following the guidance from manufacturer's user manuals. To prevent a burst from the initial pressure stabilization after loading the sample, a parafilm sealing film (HS234526B, Sigma Aldrich) was used to seal the Eppendorf tube with an open lid, and a pinhole was created in the sealing film. In the thermal stability testing, five samples were prepared, placed in a thermocycler and incubated at 65 °C for 48 h for both dried and wet forms of SDC–DNA complex. At each time point of 0, 8, 16, 24 and 48 h, dried samples were directly

mixed with 30 µl of IVT buffer as previously described, whereas the wet samples were placed on the magnetic stand to remove the DPBS, followed by mixing with 30 µl of IVT buffer. After IVT, the RNA was quantified using a NanoDrop Spectrophotometer, followed by reverse transcription using each file's specific reverse primer.

Reverse transcription

First-strand cDNA synthesis was generated by mixing 5 µl of RNA with 500 nM of reverse primer in a 20 µl reverse-transcription reaction (Bio-Rad, 1708897) containing 4 µl of reaction supermix, 2 µl of GSP enhancer solution and 1 µl of reverse transcriptase. The mixture was incubated at 42 °C for 60 min, followed by heat deactivation at 85 °C for 5 min. The cDNA of each digital file was quantified using real-time PCR for evaluating file recovery performance. To maximize the DNA quantity for gel electrophoresis or next-generation sequence, the resultant cDNA was diluted 25-fold, and 1 µl was used as the template in a PCR amplification containing 0.5 µl of Q5 High-Fidelity DNA Polymerase (NEB, M0491S), 1× Q5 polymerase reaction buffer (NEB, B9072S), 1 µM of forward and reverse primer, and 0.2 mM each of dATP (NEB, N0440S), dCTP (NEB, N0441S), dGTP (NEB, N0442S) and dTTP (NEB, N0443S) in a 50 µl total reaction volume. The amplification conditions were 98 °C for 30 s and then 35 cycles of 98 °C for 10 s, 55 °C for 20 s, 72 °C for 10 s with a final 72 °C extension step for 2 min. The products were assayed by gel electrophoresis and their concentrations were measured with a Fragment Analyzer HS NGS Fragment Kit (Agilent Technologies, DNF-474-0500).

DNA gel electrophoresis

Agarose-based DNA gels were made by mixing and microwaving 100 ml of 1× TAE buffer (Fisher Scientific, BP13324) with 1.5 mg of molecular biology grade agarose (Genesee Scientific, 20102); 0.1× SYBR Safe DNA Gel Stain was added to visualize DNA (Invitrogen, S33102). DNA samples and ladder (NEB, N3231S) were loaded with 1× DNA loading dye containing 10 mM EDTA, 3.3 mM Tris–HCl (pH 8.0), 0.08% SDS and 0.02% Dye 1 and 0.0008% Dye 2 (NEB, B7024S). Electrophoresis was performed with 1× TAE buffer in a Thermo Scientific Mini Gel Electrophoresis System (Fisher Scientific, 09-528-110B) at a voltage gradient of 16 V cm^{-1} for 45 min. Purification of DNA in the gel was achieved by using a Monarch Gel Extraction Kit (NEB, T1020S) following the manufacturer's instructions.

RNA gel electrophoresis

All equipment was cleaned with 10% bleach (VWR, 951384) and RNaseZap (Fisher Scientific, AM9780) to minimize nuclease contamination, particularly ribonuclease (RNase) contamination. The following procedures were performed in a PCR workstation with sterile pipetting equipment to further reduce ribonuclease contamination. Agarose-based RNA gels were cast by mixing and microwaving 100 ml of 1× TAE buffer (Fisher Scientific, BP13324) with 1.5 mg of molecular biology grade agarose (Genesee Scientific, 20102); 0.1× of SYBR Safe Gel Stain (Invitrogen, S33102) was added to visualize the RNA. RNA samples were treated with 2 U DNase I (NEB, M0303S) and incubated at 37 °C for 15 min, followed by a purification process using a Monarch RNA Cleanup Kit (NEB, T2030S) following the manufacturer's instructions. The purified samples and RNA ladder (NEB, N0364S) were mixed with 1× RNA loading dye containing 47.5% formamide, 0.01% SDS, 0.01% bromophenol blue, 0.005% xylene cyanol and 0.5 mM EDTA (NEB, B0363S). The mixtures were heated up at 65 °C for 2 min, followed by immediate cooling on ice for 5 min. RNA electrophoresis was performed at a voltage gradient of 15 V cm^{-1} for 50 min.

Restriction digestions

First, 500 ng of a DNA mixture containing three digital files was bound to caSDCs. The complex was then directly treated with 10 U (or 50 U) of restriction enzymes (EcoRI-HF for File1, BbsI-HF for File2 and NheI-HF for File3) in a 20 µl reaction containing 1× rCutSmart buffer (NEB, R3101S, R3505S). The reaction was incubated at 37 °C for 1 h, followed by gently washing the immobilization twice with DPBS and directly mixing with 30 µl of IVT buffer as previously described. After IVT, the RNA was quantified using a NanoDrop Spectrophotometer, followed by reverse transcription using each file's specific reverse primer.

DNase I digestions

First, 500 ng of File1 DNA was bound to caSDCs. The complex was then directly treated with 10 U (or 50 U) of DNase I in a 50 µl reaction volume. The reaction was incubated at 37 °C for 1 h, followed by gently washing the complex twice with DPBS and directly mixing with 30 µl of IVT buffer as previously described. After IVT, the RNA was quantified using a NanoDrop Spectrophotometer, followed by reverse transcription.

Direct RNA sequencing by ONT and subsequent data processing

Barcodes for direct RNA sequencing were contained in custom reverse transcription adapters (RTAs) designed for demultiplexing with DeePlexiCon[50]. Single-stranded DNA oligos needed for four, custom DeePlexiCon RTAs (https://github.com/Psy-Fer/deeplexicon) were ordered from Integrated DNA Technologies (IDT) and reconstituted to either 50 or 100 µM in water. Corresponding pairs of oligos were annealed at a final concentration of 1.4 µM in IDT nuclease-free duplex buffer (IDT, #11-01-03-01) by incubating mixtures at 95 °C for 2 min followed by cooling at room temperature for at least 2 h. Annealed RTAs were stored at −20 °C until use.

Samples for sequencing were prepared using the ONT direct RNA sequencing kit (ONT, SQK-RNA002) and sequence-specific protocols provided by ONT (protocol version DSS_9081_V2_REVQ_25MAY2022) with modifications recommended by the DeePlexiCon developers[50]. Briefly, a custom annealed RTA (1.4 µM) was ligated to 500 ng of each RNA sample using T4 DNA ligase (NEB, M0202) for 15 min at room temperature and reverse transcription was performed with SuperScript III Reverse Transcriptase (Thermo Fisher Scientific, 18080044) as recommended by ONT. Reverse transcription reactions were cleaned up using 1.8× Ampure XP beads (Beckman Coulter, A63881) and 70% ethanol. Reverse-transcribed samples were eluted in water and quantified using a Qubit 1× dsDNA high-sensitivity kit (ThermoFisher, Q33230). Next, 75 ng each of three barcoded samples were pooled and ligated to ONT RNA sequencing adapters using either T4 DNA ligase (NEB M0202) or Quick T4 DNA ligase (NEB E6057A) for 15 min at room temperature. Reactions were cleaned up using 1× Ampure XP beads and ONT wash buffer. Pooled libraries were eluted in ONT elution buffer and quantified using a Qubit 1× dsDNA high-sensitivity kit. Each pooled library contained one of each file type and different combinations of barcodes were used in each library to avoid any unexpected demultiplexing bias (Extended Data Table 4).

Each barcoded and multiplexed library was sequenced on a R9.4.1 MinION flowcell (ONT, FLO-MIN106D) run on a GridION Mk1 sequencing device (ONT, GRD-MK1) for a total of six runs. Sequencing was basecalled in real-time using the Guppy basecaller (v.6.3.9) in MinKNOW (v.22.10.7) with a direct RNA high accuracy basecalling model (v.2020-09-07_rna_r9.4.1_minion_256_8f8fc47b) and default settings. All reads regardless of quality were demultiplexed. FASTQs were demultiplexed using DeePlexiCon (v.1.2)[50] with a strict threshold of 0.5 using the standard DeePlexiCon model (resnet20-final.h5).

Puzzle designs

Puzzle1 and Puzzle2 were chess problems, asking for all board configurations possible that would not violate any piece attacking another piece. Puzzle3 was a sudoku problem, asking what board configurations would result in no number being repeated in any horizontal row or vertical column. Each DNA strand sequence represented one potential board configuration, with the sequence of each 20 nt position representing

the presence or absence of a board piece or a number. All DNA strands (all board configurations) of a particular puzzle had a distinct RNA polymerase promoter sequence.

RNase H digestion

First, 100 pmol of RNA was mixed with 400 pmol of computing oligos in a 50 µl reaction volume containing 10 mM Tris–HCl, pH 7.5, 50 mM NaCl and 1 mM EDTA. The reaction was heated to 95 °C for 30 s, followed by gradually cooling to 10 °C (1 °C s^{-1} temperature drop) for 30 s. After hybridization, 5 U RNase H endonuclease and 1× reaction buffer (NEB, M0297L) were added and incubated at 37 °C for 30 min. The digested sample was processed with a Monarch RNA Cleanup Kit (NEB, T2030S) following the manufacturer's instructions.

Gel imaging

Fluorescence imaging of both DNA and RNA gel samples was performed with a Li-Cor Odyssey Fc Imaging System, and the fluorescence intensity was quantified using FIJI software[61].

Illumina next-generation sequencing

Amplicons were purified with AMPure XP beads (Beckman Coulter, A63881) according to the TruSeq Nano protocol (Illumina, 20015965). The quality and band sizes of libraries were assessed using the HS NGS Fragment Analysis Kit (Advanced Analytical, DNF-474) on a Fragment Analyzer (Agilent Technologies). Samples were submitted to Azenta for Illumina-based next-generation sequencing (MisSeq v.2 10M reads). Samples related to the puzzle computations were prepared and submitted to Amplicon-EZ (Azenta) for Illumina-based next-generation sequencing. Ligation of Illumina sequencing adapters to the prepared samples was performed by Azenta. Data analysis was performed by using FLASH v.1.2.11 from Conda v.23.5 for QC and using Pandas v.2.0.2 from Python v.3.8 to sort the number of reads for each strand.

Statistics and reproducibility

No statistical method was used to predetermine sample size. No data were excluded from the analyses. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Raw sequencing data, sequences of DNA oligos and source data for all plots are available at https://doi.org/10.5281/zenodo.12169723 (ref. 62) and https://doi.org/10.5281/zenodo.12192541 (ref. 63) and https://github.com/keung-lab/Lin-et-al-2024.git, https://github.com/dna-storage/framed/tree/sdc_nature_submission and https://github.com/dna-storage/framed/releases. All other data are available upon reasonable request. Source data are provided with this paper.

Code availability

The software algorithms we developed to perform the reported analyses are available at https://doi.org/10.5281/zenodo.12169723 (ref. 62), https://github.com/dna-storage/framed/tree/sdc_nature_submission and https://github.com/dna-storage/framed/releases under a permissive open source license with instructions for installation. We implemented code in Python using many standard open-source packages, including biopython, primer3, numpy, scipy, pandas and others. These dependences are documented in the form of a Python requirements.txt file that guides installation of additional dependent software packages.

References

61. Schindelin, J. et al. Fiji: an open-source platform for biological-image analysis. *Nat. Methods* **9**, 676–682 (2012).
62. Keung Lab. keung-lab/Lin-et-al-2024: v1.0.1. *Zenodo* https://doi.org/10.5281/zenodo.12169723 (2024).
63. Lin, K. & Keung, A. FASTQ data for: a primordial DNA store and compute engine. *Zenodo* https://doi.org/10.5281/zenodo.12192541 (2024).

Acknowledgements

We thank K. J. Tomek, M. Lee and K. Matange for helpful discussions and R. M. Kelly for use of their vacuum concentrator. We thank the Biomanufacturing Training and Education Center (BTEC) at NCSU for the use of their lyophilizer device. We also thank S. Mukherjee for providing training on the use of the Zetasizer Nano ZSP for measuring the zeta-potential of the SDC–DNA samples. This work was supported by the National Science Foundation (ECCS-2027655 and CSR-1901324). K.N.L. was supported by a Department of Education Graduate Assistance in Areas of Need fellowship, P200A160061. R.E.P. was supported by T32GM133366. Some artwork was created with BioRender.

Author contributions

K.N.L., O.D.V. and A.J.K. conceived the study. K.N.L. planned and performed the wet-lab experiments with guidance from O.D.V. and A.J.K. K.N.L. and K.V. designed the three-file oligo library for SDC. K.V. performed file encoding and decoding simulation with guidance from J.M.T. P.W.H. performed nanopore RNA sequencing with guidance from W.T. K.N.L and R.E.P. planned and performed the simulation for computation library design. C.C. fabricated SDC materials made from cellulose acetate, cellulose and agarose with guidance from O.D.V. K.N.L. and A.S.C. designed the microfluidic system with guidance from A.S.M. K.N.L. performed zeta-potential measurements. K.V. and K.N.L. processed the next-generation sequencing and nanopore sequencing data. K.N.L. and A.J.K. wrote the paper with input from all authors.

Competing interests

The authors declare the following competing interests: A.J.K. and J.M.T. are cofounders of DNAli Data Technologies that has potential interest in translating and commercializing DNA-based information systems. A.J.K., K.V., J.M.T. and K.N.L. are inventors of patent WO 2020/096679 which has been licensed to DNAli Data Technologies and from which some of this work is derived. W.T. has two patents (8,748,091 and 8,394,584), licensed to Oxford Nanopore Technologies (ONT), which were used for direct RNA sequencing in this work. W.T. has received travel funds to speak at symposia organized by ONT. The other authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41565-024-01771-6.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41565-024-01771-6.

Correspondence and requests for materials should be addressed to Orlin D. Velev or Albert J. Keung.

Peer review information *Nature Nanotechnology* thanks Tom de Greef, Casey Platnich and Victor Zhirnov for their contribution to the peer review of this work.

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

Extended Data Table 1 | Maximum binding capacity for dsDNA bound to caSDC, ceSDC, and agSDC

caSDC-DNA	ceSDC-DNA	agSDC-DNA
Maximum Binding Capacity (in 10^{12} molecules DNA / mg SDC)		
1.93 +/- 0.81	3.51 +/- 0.54	1.29 +/- 0.05
Maximum Binding Capacity (in Terabytes / mg SDC)		
10.7 +/- 4.46	19.42 +/- 2.98	7.14 +/- 0.25
Maximum Binding Capacity (in 10^{4} Terabytes / cm^{3} SDC)		
1.01 +/- 0.42	1.83 +/- 0.28	0.67 +/- 0.03

The maximum binding capacity for double-stranded DNA binding to caSDC, ceSDC, and agSDC, in terms of molecules of DNA per mg SDC, terabytes per mg SDC, and terabytes per cubic centimetre SDC.

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

Extended Data Table 2 | Cost of substrate material that data is bound to

Binding Materials	DNA Binding Capacity (ug/mg)	Material Cost ($/mg)	Cost per Data (10^{-4} $/TB)
MNP-caSDC	6	0.00122	0.9
Streptavidin-Biotin	62	19	115
AMPure Beads	12000	76	4.69

The DNA binding capacity and associated materials cost per mg material or terabyte of data stored. Materials including caSDC, streptavidin-functionalized magnetic beads, and solid phase reversible immobilization beads (SPRI Beads).

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

Extended Data Table 3 | Accuracy of computation for each puzzle

ID	Promoter	Number of Correct Solutions	Accuracy
Puzzle1	T7	64	100%
Puzzle2	Sp6	16	88%
Puzzle3	T3	2	17%

The number of correct solutions for each puzzle, and the experimental accuracy accessing those correct solutions, with the definition of accuracy described in the Methods.

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

Extended Data Table 4 | Annotation of direct RNA sequencing by ONT and data processing

Sequencing Batch	File	ID	Barcode
batch_1	File1	id_1	bc_1
batch_1	File2	id_7	bc_2
batch_1	File3	id_13	bc_3
batch_2	File1	id_2	bc_2
batch_2	File2	id_8	bc_3
batch_2	File3	id_14	bc_4
batch_3	File1	id_3	bc_3
batch_3	File2	id_9	bc_4
batch_3	File3	id_15	bc_1
batch_4	File1	id_4	bc_4
batch_4	File2	id_10	bc_1
batch_4	File3	id_16	bc_2
batch_5	File1	id_5	bc_2
batch_5	File2	id_11	bc_1
batch_5	File3	id_17	bc_4
batch_6	File1	id_6	bc_4
batch_6	File2	id_12	bc_3
batch_6	File3	id_18	bc_2

The sequencing IDs and barcodes associated with each direct RNA sequencing batch and file.

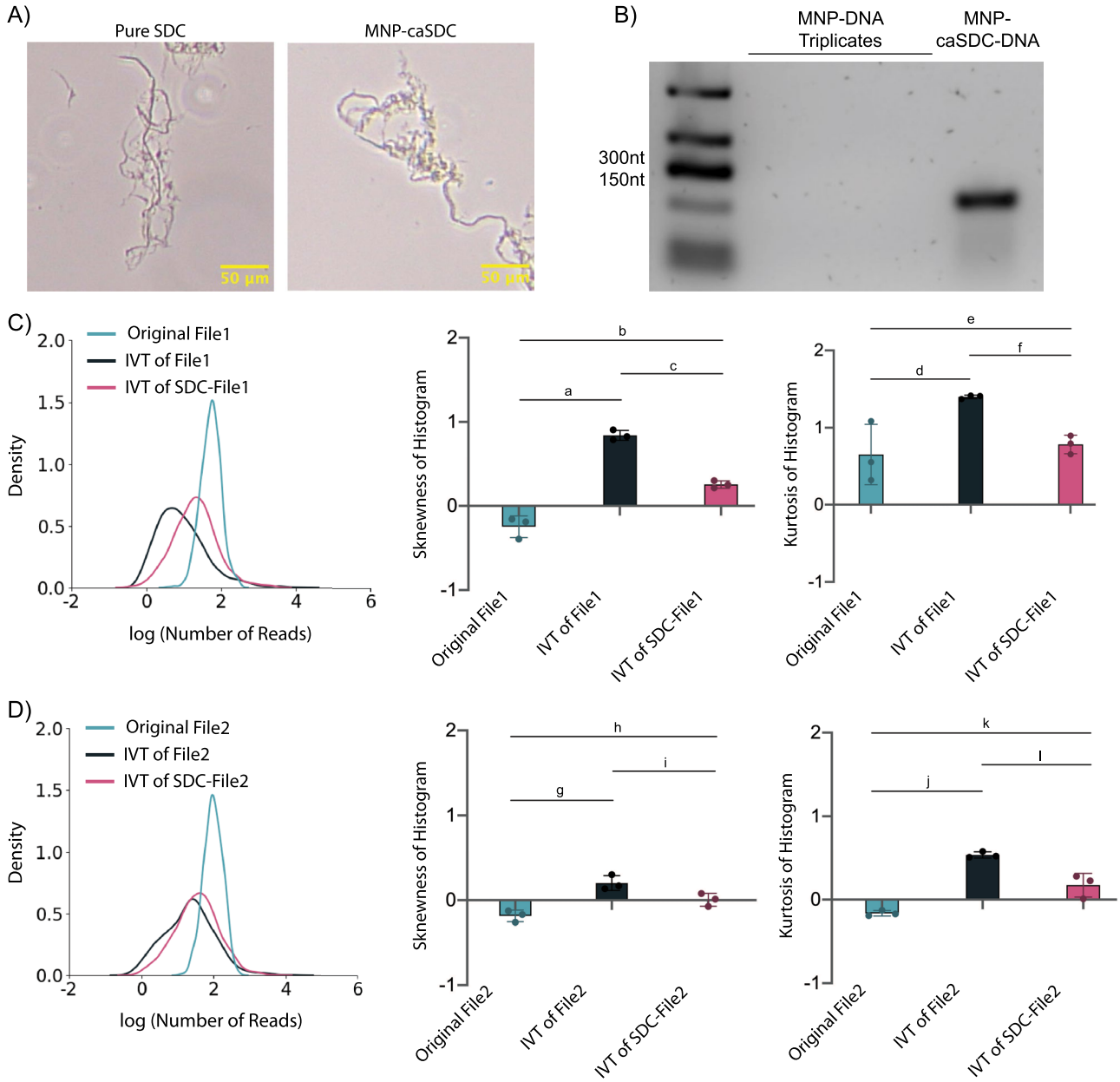
https://doi.org/10.1038/s41565-024-01771-6

Extended Data Table 5 | Annotation of each sample shown in Extended Data Fig. 4c,d

Extended Data Fig. 4c		
Experimental Condition	**File Operation**	**File Decoding ID**
1	Erasing File1 with Dnase I	File1
2	Erasing File1, Reloading File1	File1
3	Erasing File1, Reloading File2, Erasing File2, Reloading File1	File1
4	Erasing File1, Reloading File3, Erasing File3, Reloading File1	File1
5	Erasing File1, Reloading File2	File2
6	Erasing File1, Reloading File3	File3
Extended Data Fig. 4d		
1	Unmodified File1	File1
2	Erasing File1 with Dnase I	File1
3	Erasing File1, Reloading File1	File1
4	Erasing File1, Reloading File2, Erasing File2, Reloading File1	File1
5	Erasing File1, Reloading File3, Erasing File3, Reloading File1	File1
6	Unmodified File2	File2
7	Erasing File1, Reloading File2	File2
8	Unmodified File3	File3
9	Erasing File1, Reloading File3	File3

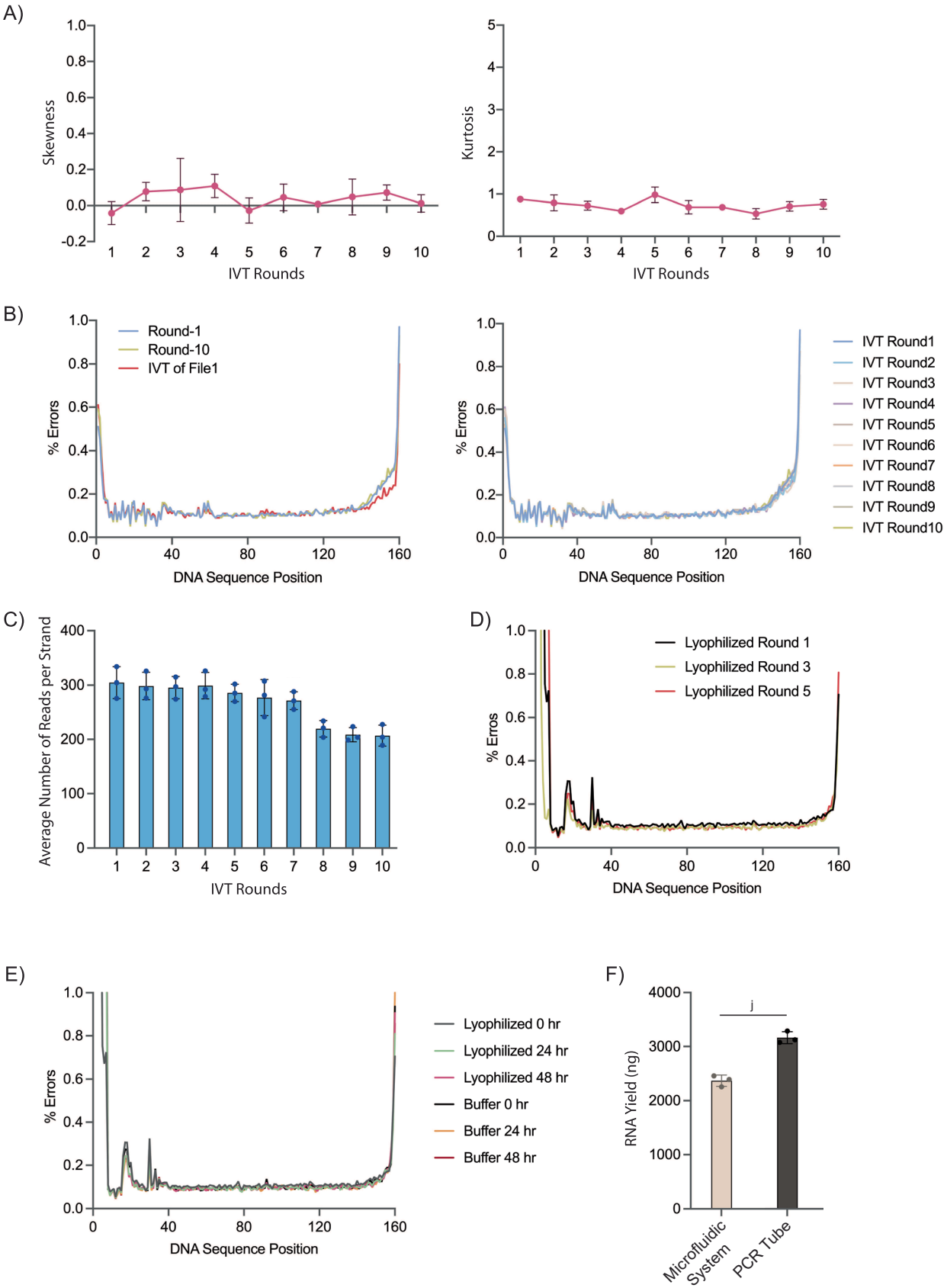
Detailed description of each experimental condition shown in Extended Data Fig. 4c (top) and Extended Data Fig. 4d (bottom), with the sequence of molecular operations, and the file decoded, as indicated.

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



Extended Data Fig. 1 | Adsorbing DNA onto SDC can lead to stable RNA generation. A) Morphology of SDC with (left) and without (right) magnetic nanoparticles. Images were taken using a Nikon Ts2 Inverted microscope. All samples were imaged using the same 10× objective. Samples were imaged using the same microscope settings, and adjusted identically for quantification purposes. **B**) Equivalent masses of magnetic nanoparticles alone (MNP) or caSDC infused with MNP (MNP-caSDC) were incubated with DNA (dsDNA), washed, and subjected to IVT. An RNA gel indicates no RNA generated from the magnetic nanoparticles. **C**) Representative strand distribution density (left), skewness (middle), and kurtosis (right) plots for experimental samples of File1. These samples included direct sequencing of the File1 DNA obtained from the DNA synthesis provider (original File1), cDNA obtained after IVT of File1 DNA adsorbed to caSDC (IVT of SDC-File1), and cDNA obtained after IVT of unbound File1 DNA (IVT of File1). **D**) Representative strand distribution density (left), skewness (middle), and kurtosis (right) plots for experimental samples of File2. These samples included direct sequencing of the File2 DNA obtained from the DNA synthesis provider (original File2), cDNA obtained after IVT of File2 DNA adsorbed to caSDC ('IVT of SDC-File2'), and cDNA obtained after IVT of unbound File2 DNA (IVT of File2). Plotted values represent the arithmetic mean, and error bars represent the standard deviation of three independent IVT reactions. Statistics was calculated using One-Way ANOVA with Tukey–Kramer post-hoc for panel C and D. a $p = 1.77\times10^{-4}$, b $p = 2.92\times10^{-3}$, c $p = 1.67\times10^{-4}$, d $p = 3.01\times10^{-2}$, e $p = 6.12\times10^{-1}$, f $p = 9.51\times10^{-4}$, g $p = 3.76\times10^{-3}$, h $p = 3.18\times10^{-2}$, i $p = 4.28\times10^{-2}$, j $p = 1.54\times10^{-5}$, k $p = 1.63\times10^{-2}$, l $p = 1.29\times10^{-2}$.

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

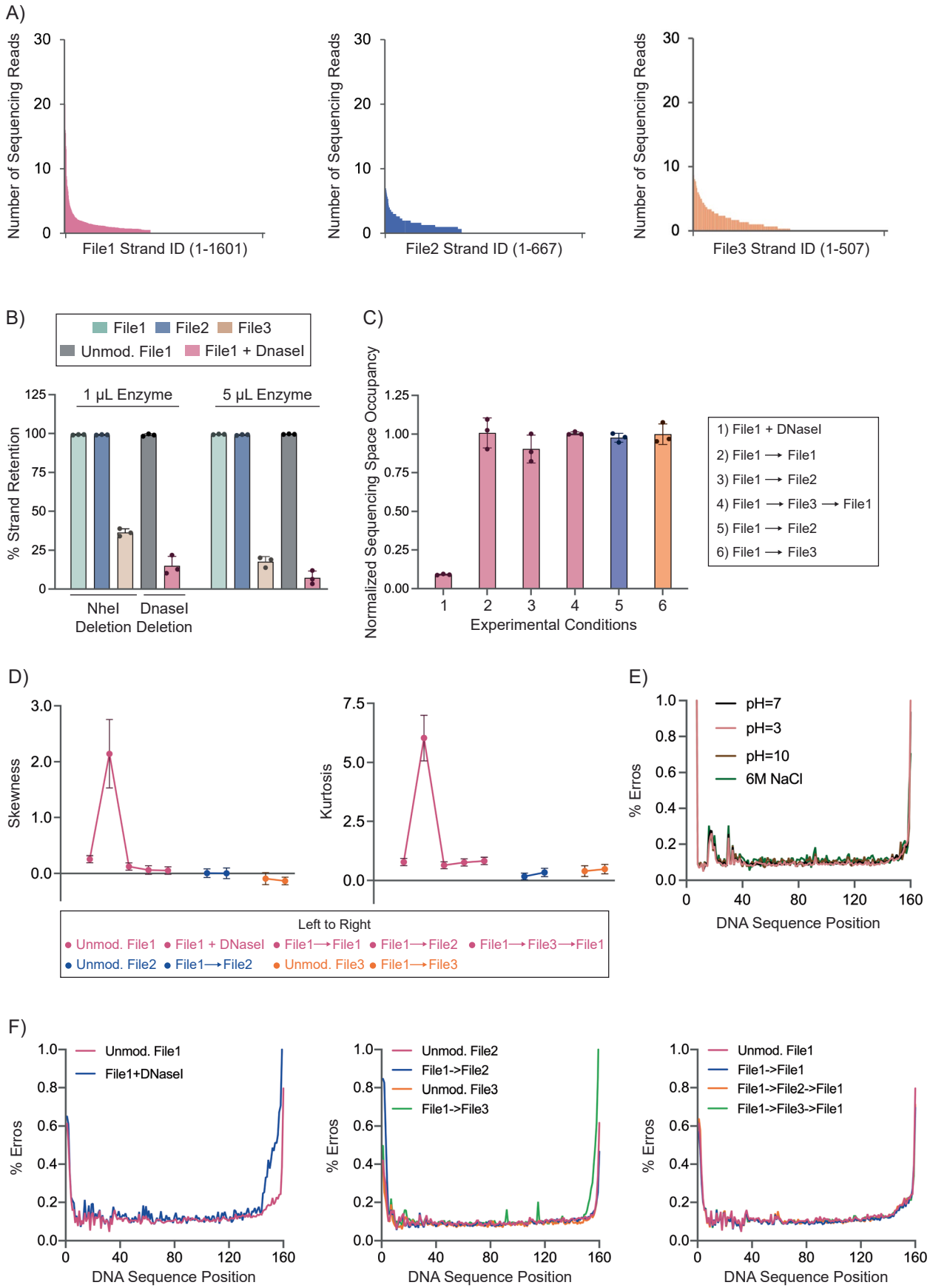


Extended Data Fig. 2 | See next page for caption.

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

Extended Data Fig. 2 | Complex files adsorbed on SDCs can be repeatedly accessed with robust and stable performance. A) Skewness (left) and kurtosis (right) plots for strand distribution density of cDNA obtained after each sequential round of IVT to access File1 DNA adsorbed to caSDC. The results demonstrated a consistent strand distribution for sequential rounds of accessing File1 DNA adsorbed to the SDCs. **B**) Percent error for each DNA sequence position in the cDNA obtained after IVT of File1 DNA adsorbed to caSDC after each sequential round of IVT, and cDNA obtained after IVT of unbound File1 DNA (IVT of File1). The error rate was calculated by dividing the number of errors of a given type occurring at a nucleotide position by the total number of reads for that sequence (Method). **C**) Number of sequencing reads for strands in cDNA obtained after each sequential round of IVT to access File1 DNA adsorbed to caSDC. **D**) Percent error for each DNA sequence position in the cDNA obtained from lyophilized File3-SDC after IVT of the first lyophilization (1st round), and after the 3rd and 5th rounds of lyophilization, as obtained by Illumina sequencing. **E**) Percent error for each DNA sequence position in the cDNA obtained from IVT after 0, 24 and 48 hours of incubation of lyophilized or solubilized File3-SDC at 65 ˚C. The error rate was calculated by dividing the number of errors of a given type occurring at a nucleotide position by the total number of reads for that sequence. **F**) 300 ng DNA was input into identical IVT reactions either in a microfluidic system placed in an incubator at 37 ˚C or in a PCR tube placed in a PCR machine held at 37 ˚C for overnight. Plotted values represent the arithmetic mean, and error bars represent the standard deviation of three independent IVT reactions. Statistics was calculated using One-Way ANOVA with Tukey–Kramer post-hoc. Comparisons are relative to the first experimental condition in panel C. a $p = 3.93\times10^{-1}$, b $p = 6.76\times10^{-1}$, c $p = 8.12\times10^{-1}$, d $p = 3.81\times10^{-1}$, e $p = 3.38\times10^{-1}$, f $p = 1.62\times10^{-1}$, g $p = 1.10\times10^{-2}$, h $p = 6.50\times10^{-3}$, i $p = 8.42\times10^{-3}$.

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

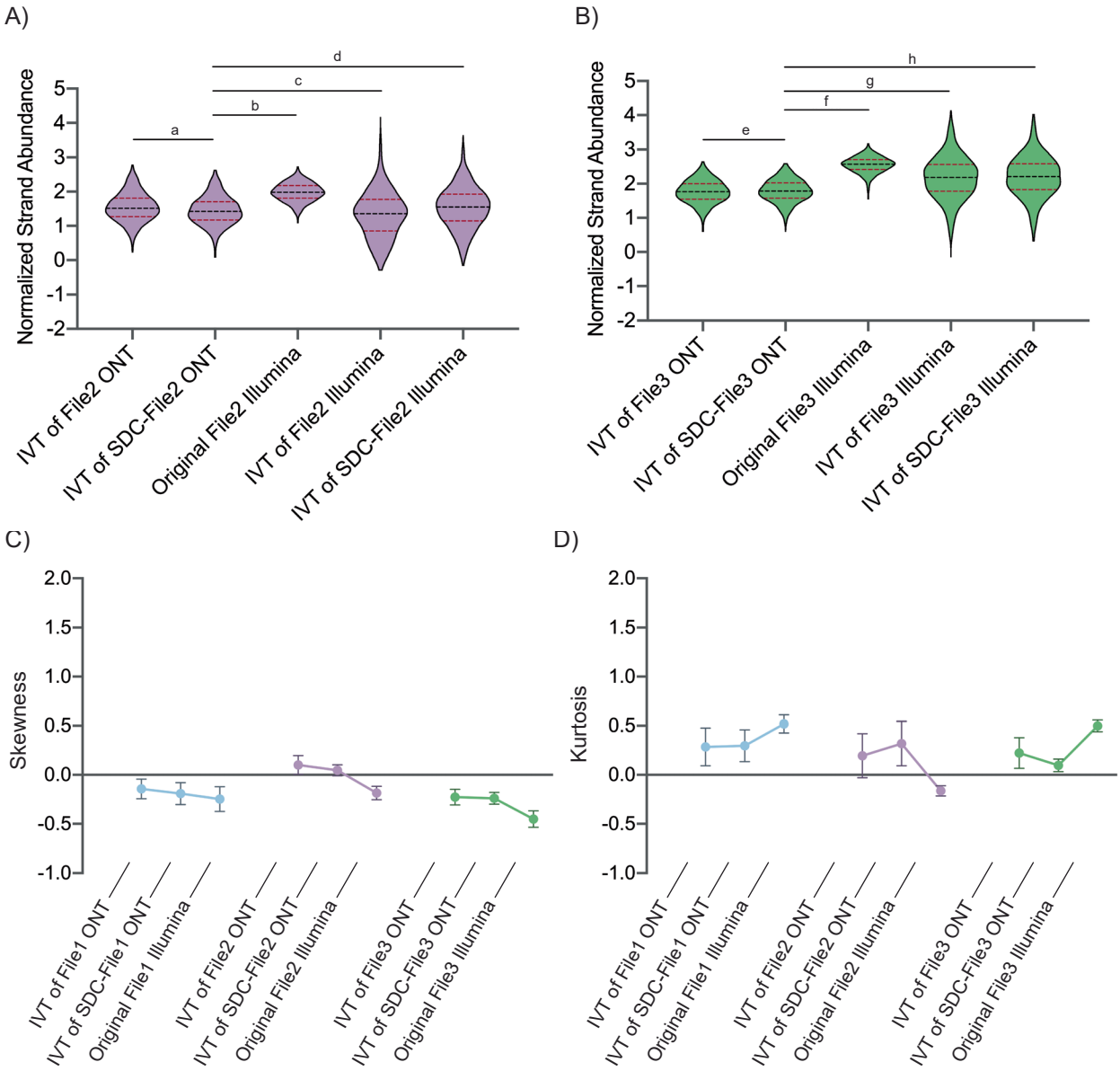


Extended Data Fig. 3 | See next page for caption.

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

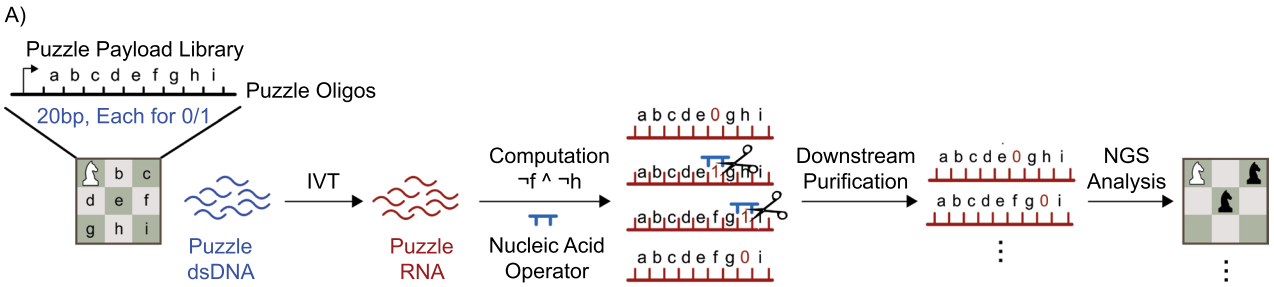
Extended Data Fig. 3 | Complex files adsorbed onto SDCs can be specifically erased, and new information can be reloaded onto SDCs. A) Number of sequencing reads for strands in cDNA obtained from IVT of File DNA adsorbed to caSDC after processed with restriction digestion as file deletion. The values were measured by Illumina sequencing. **B**) Percent of unique strands of each file found in cDNA after File3 was specifically deleted from the three-file database as measured by Illumina sequencing. The deletion was executed with 1 µL or 5 µL of restriction enzyme. Values were plotted as a percentage of the total unique strands. **C**) The fraction of all sequencing reads for a targeted file, obtained from cDNA after IVT of File DNA adsorbed to caSDC after reloading. Annotation of each operation is listed in Extended Data Table 5. FileX->FileY indicates FileX was deleted and FileY was then loaded, with IVT of FileY performed, measured, and plotted. Values were measured by Illumina sequencing and plotted as a percentage of the total sequencing reads. **D**) Skewness (left) and kurtosis (right) plots for strand distribution density of cDNA obtained after IVT of unmodified File DNA ('Unmod.'), erasing treated File DNA adsorbed to the SDCs (File1+DNaseI) and reloaded new File DNA after each operation on the SDCs. Values were measured by Illumina sequencing. '→' denotes removing current File DNA on SDCs and reloading with new file information. Annotation of each operation is listed in Extended Data Table 5. **E**) Percent error for each DNA sequence position in the cDNA obtained after incubating File3-SDC under various buffer conditions, followed by IVT. **F**) Percent error for each DNA sequence position in the cDNA obtained after IVT of unmodified File DNA (Unmod.) and reloaded File DNA. The error rate was calculated by dividing the number of errors of a given type occurring at a nucleotide position by the total number of reads for that sequence. Values were measured by Illumina sequencing and plotted after normalizing to its number of sequencing reads found in untreated File DNA prior to IVT. Plotted values represent the arithmetic mean, and error bars represent the standard deviation of three independent IVT reactions.

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

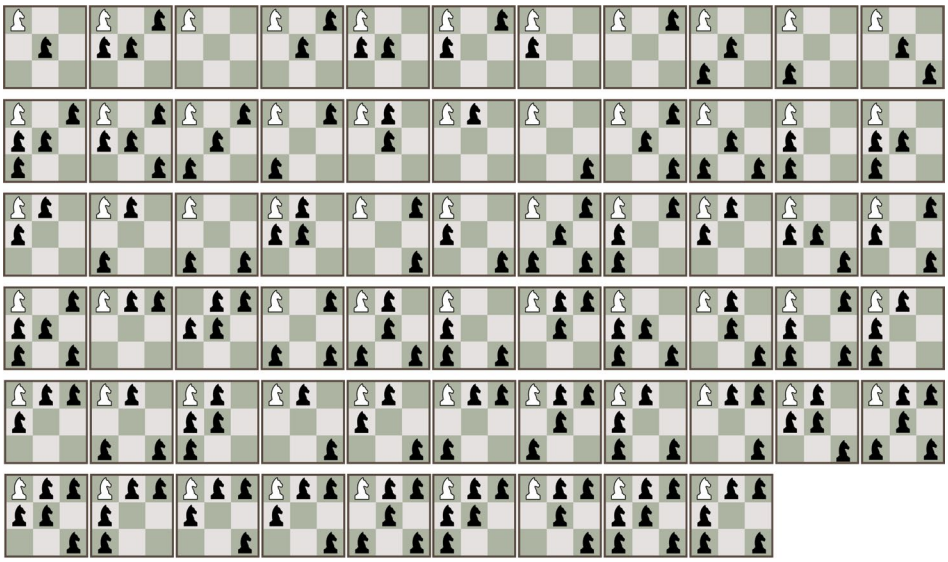


Extended Data Fig. 4 | Complex file DNA adsorbed on SDCs can be directly sequenced after IVT using Oxford nanopore sequencing. A) Violin plots of the strand distributions for experimental samples of File2. These samples include direct sequencing of the File2 DNA obtained from the DNA synthesis provider (Original File2), RNA and cDNA obtained after IVT of File2 DNA adsorbed to caSDC (IVT of SDC-File2), RNA and cDNA obtained after unbound File2 DNA (IVT of File2). **B**) Violin plots of the strand distributions for experimental samples of File3. These samples include direct sequencing of the File3 DNA obtained from the DNA synthesis provider (Original File3), RNA and cDNA obtained after IVT of File3 DNA adsorbed to caSDC (IVT of SDC-File3), RNA and cDNA obtained after unbound File3 DNA (IVT of File3). **C,D** (**C**) Skewness and **(D)** kurtosis plots for strand distribution density of RNA obtained after IVT of unbound File DNA, File DNA adsorbed to caSDC, and of DNA obtained after direct sequencing of File DNA from synthesis provider. Plotted RNA samples were processed with Oxford nanopore sequencing (ONT) and DNA samples were processed with Illumina Sequencing (Illumina). Each plotted value represents the arithmetic mean, and error bars represent the standard deviation of three independent IVT reactions. Statistics was calculated using One-Way ANOVA with Tukey–Kramer post-hoc for panel A and B. a $p = 2.71\times10^{-3}$, b $p = 1.56\times10^{-10}$, c $p = 3.58\times10^{-4}$, d $p = 1.59\times10^{-2}$, e $p = 8.14\times10^{-1}$, f $p = 1.52\times10^{-10}$, g $p = 1.50\times10^{-10}$, h $p = 1.55\times10^{-10}$.

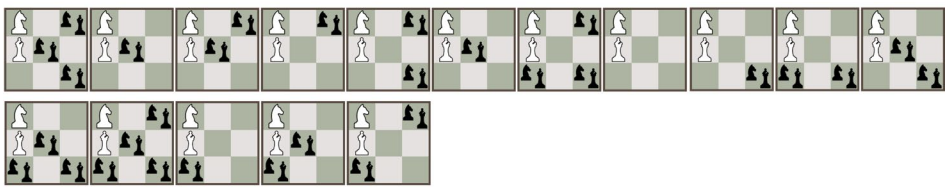
https://doi.org/10.1038/s41565-024-01771-6



B) Puzzle 1 Correct Solutions



Puzzle2 Correct Solutions



Puzzle3
Correct Solution

1	2	3
2	3	1
3	1	2

Puzzle3
Intentionally Incorrect Solution

1	2	3
2	3	1
3	1	1

Extended Data Fig. 5 | Implementation of addressable in-storage computation. **A**) Schematic of computation rules for Puzzle1. The payload of each oligo in the DNA library is divided into nine sections, with each section representing the a specific position on the puzzle board. Each position is composed of a specific 20 nt DNA sequence (bit sequence). Combination of these oligos completes the full starting configuration for Puzzle1. In computation, short DNA oligos are used as nucleic acid operators to hybridize to specific puzzle RNA strands which contain information violates puzzle rules. This process triggers endonuclease activities and leaves behind RNA strands representing correct puzzle solutions. These remaining strands are purified and retained for downstream processes and NGS analysis. **B**) Schematic of correct solutions for Puzzle1, Puzzle2, and Puzzle3, as well as the intentionally incorrect solution for Puzzle3.

nature portfolio

Corresponding author(s): Albert J. Keung, Orlin D. Velev

Last updated by author(s): Jun 19, 2024

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	The version of FIJI Image used for Figures 1D, Supplementary Figures 1A-B, was 2.0.0-rc-69/1.52p The version of qPCR analysis used for Figures 1B, 2B-E, 3G, 3J, 4E, Supplementary Figures 2F, was Bio-Rad CFX Maestro 4.1.2434.0124. The software used for Figures 2F-H, Supplementary Figures 1A-B was Licor Image Studio Lite 4.0. The NanoDrop software used for Supplementary Figure 2F was Thermo Scientific™ ND2000LAPTOP.
Data analysis	The software algorithms we developed to perform the reported analyses are available at 10.5281/zenodo.12169723, https://github.com/keung-lab/Lin-et-al-2024.git, https://github.com/dna-storage/framed/tree/sdc_nature_submission, and https://github.com/dna-storage/framed/releases under a permissive open source license with instructions for installation. We implemented code in python using many standard open source packages, including biopython, primer3, numpy, scipy, pandas, and others. These dependences are documented in the form of a python requirements.txt file that guides installation of additional dependent software packages.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Raw sequencing data, sequences of DNA oligos, and source data for all plots are available at 10.5281/zenodo.12169723, https://github.com/keung-lab/Lin-et-al-2024.git, https://github.com/dna-storage/framed/tree/sdc_nature_submission, and https://github.com/dna-storage/framed/releases. All other data are available upon reasonable request.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	n/a
Reporting on race, ethnicity, or other socially relevant groupings	n/a
Population characteristics	n/a
Recruitment	n/a
Ethics oversight	n/a

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No statistical method was used to predetermine sample size.
Data exclusions	No data were excluded from the analyses.
Replication	Experimental effect sizes were large so that replication were not pursued.
Randomization	The experiments were not randomized.
Blinding	The Investigators were not blinded to allocation during experiments and outcome assessment.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks	n/a
Novel plant genotypes	n/a
Authentication	n/a