nature biotechnology

ARTICLES

https://doi.org/10.1038/s41587-019-0240-x

Corrected: Author Correction

Data storage in DNA with fewer synthesis cycles using composite DNA letters

Leon Anavy[1*], Inbal Vaknin[2], Orna Atar[2], Roee Amit[2] and Zohar Yakhini[1,3*]

The density and long-term stability of DNA make it an appealing storage medium, particularly for long-term data archiving. Existing DNA storage technologies involve the synthesis and sequencing of multiple nominally identical molecules in parallel, resulting in information redundancy. We report the development of encoding and decoding methods that exploit this redundancy using composite DNA letters. A composite DNA letter is a representation of a position in a sequence that consists of a mixture of all four DNA nucleotides in a predetermined ratio. Our methods encode data using fewer synthesis cycles. We encode 6.4 MB into composite DNA, with distinguishable composition medians, using 20% fewer synthesis cycles per unit of data, as compared to previous reports. We also simulate encoding with larger composite alphabets, with distinguishable composition deciles, to show that 75% fewer synthesis cycles are potentially sufficient. We describe applicable error-correcting codes and inference methods, and investigate error patterns in the context of composite DNA letters.

DNA-based data storage systems are particularly appealing owing to the high information capacity, in terms of physical volume, of DNA as compared to current state of the art storage media. Storing digital information on DNA involves encoding the information into a sequence over the DNA alphabet (that is, A, C, G and T), producing synthetic DNA molecules with the desired sequence and storing the synthetic biological material. Reading the stored information requires sequencing of the DNA and decoding to obtain the original digital information.

DNA-based storage systems[1–8] involve several technological and design challenges. Biochemical and technical constraints require the use of custom coding schemes to accommodate possible dropouts and common DNA synthesis and sequencing errors[4,7,9]. Random access at reduced sequencing overhead requires efficient design of large pools of mutually compatible PCR primers[5,6,8,10]. Recently, innovative synthesis approaches have been introduced[11,12], which may lead to more cost-effective DNA-based data storage. Other molecular biology techniques can also be used for DNA-based storage[13]. DNA synthesis technology, which is based on phosphoramidite chemistry[14,15], yields high numbers of molecules for each of the designed DNA sequences[16]. Oligonucleotide multiplicity, an important inherent property of current DNA synthesis and sequencing technologies, has not yet been exploited in DNA-storage technologies based on synthesis.

The efficiency of DNA-based storage systems can be evaluated using several quantitative metrics. One is the physical density of the storage medium, as measured by data unit per gram of DNA (for example, gigabytes per gram). A recent study demonstrated a DNA-based storage system with a physical density of 215 PB g^{-1}. This density, when converted to volumetric density, represents roughly six orders of magnitude improvement over current storage media[7]. A second performance metric is the number of synthesis cycles required for a unit of data. This is termed logical density and is the main focus of this current work (as well as of another recent study that was made available during the late stages of this project[17]).

We introduce the use of composite DNA letters to increase the logical density of DNA storage above the strict, single-molecule, theoretical limit of 2 bits per synthesis cycle. A composite DNA letter is a representation of a position in a sequence that constitutes a mixture of all four standard DNA nucleotides in a specified predetermined ratio. We use composite DNA letters to form the basis of a DNA synthesis approach that trades sequence multiplicity for increased complexity of the synthesized DNA. This increased complexity effectively extends the available alphabet and therefore allows higher data content per synthesis cycle. We demonstrate an implementation of a complete large-scale, DNA-based storage system using composite DNA letters, develop related methods including error-correction codes and investigate trade-offs and performance metrics.

Results

Composite DNA letters extend the DNA alphabet. A composite DNA letter is a representation of a position in a sequence that constitutes a mixture of all four standard DNA nucleotides in a specified predetermined ratio $\sigma=(\sigma_A,\sigma_C,\sigma_G,\sigma_T)$ where $k=\sigma_A+\sigma_C+\sigma_G+\sigma_T$ is defined as the resolution parameter of the composite letter (Methods). For example, $\sigma=(1,1,2,0)$ represents a position in a composite DNA sequence of resolution $k=4$ in which there is a 25%, 25%, 50% and 0% chance of seeing A, C, G and T, respectively. Writing a composite DNA letter at a given position of a DNA sequence is equivalent to producing (synthesizing) multiple copies (oligonucleotides) of the sequence, so that in this given position the different DNA nucleotides are distributed across the synthesized copies according to the specification of σ. Reading a composite letter requires the sequencing of multiple independent molecules representing the same composite sequence and inferring the original ratio or composition from the observed base frequencies (Fig. 1). Introducing composite letters extends the available alphabet and thus allows the coding of longer messages within a fixed synthesized molecule length. A composite DNA alphabet is a set of composite DNA letters, usually, but not necessarily, sharing a common

[1]Computer Science Department, Technion – Israel Institute of Technology, Haifa, Israel. [2]Faculty of Biotechnology and Food Engineering, Technion – Israel Institute of Technology, Haifa, Israel. [3]School of Computer Science, Herzliya Interdisciplinary Center, Herzliya, Israel. *e-mail: leon.anavy@gmail.com; zohar.yakhini@gmail.com

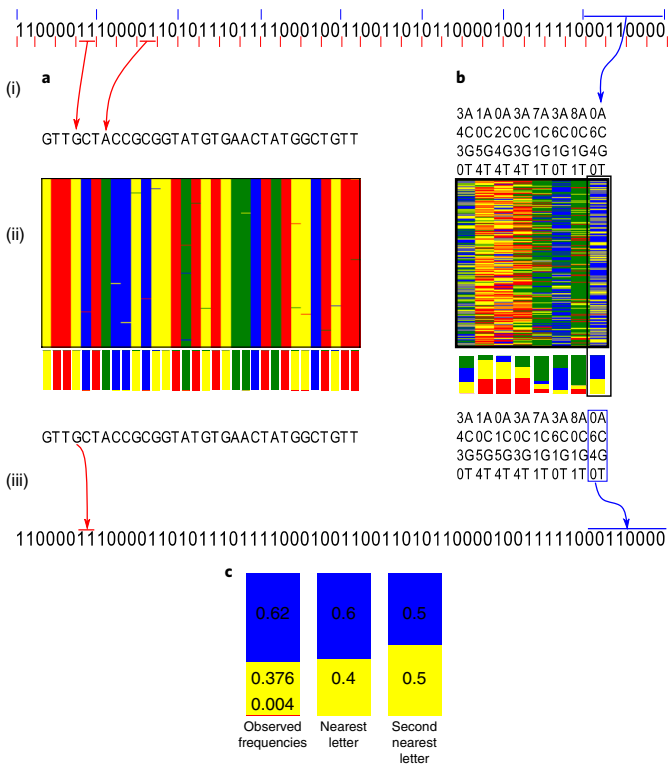


Fig. 1 | Encoding a binary message using standard and composite DNA. A binary message, depicted on top, is encoded into DNA. **a**, Standard DNA-based storage scheme[7]. The binary message is encoded to DNA by mapping every 2 bits (represented by the short red separating lines) to a DNA letter or synthesis cycle (i), the designed DNA sequence will then be synthesized and sequenced by a noisy procedure that introduces some errors (ii). The sequencing output is then used to infer the DNA composition at every position (iii). Decoding of the original message is done assuming the use of an error-correcting code. **b**, The same message is encoded using a composite DNA alphabet of resolution $k=10$ by mapping every 8 bits (represented by the blue separating lines) of the binary message to a single composite DNA position (a single synthesis cycle when using appropriate hardware). Sufficiently deep sequencing allows correct identification of the original composite letters (the right most position, in a black frame, is shown in **c**) and decoding of the message. The decoding also uses an error-correction mechanism; our implementation uses Reed–Solomon over the appropriate finite field. **c**, An example of the inference step at a single synthesized composite position. The observed frequencies are used to infer the source, $\sigma=(0,6,4,0)$, as the closest composite letter, using Kullback–Leibler divergence (Methods). Note that the inference at any fixed position is affected by the sequencing depth obtained there, as well as by sequencing and synthesis errors.

resolution k. The full composite alphabet of resolution k, denoted Φ_k, is the set of all $\sigma=(\sigma_A,\sigma_C,\sigma_G,\sigma_T)$ so that $\sum_{i\in\{A,C,G,T\}}\sigma_i=k$. Note that $|\Phi_k|=\binom{k+3}{3}$, thus the composite alphabet size grows with the resolution parameter and so does the potential logical density, as measured by data units per synthesis cycle (Supplementary Fig. 1).

To read a message coded using composite DNA letters correctly, we must infer, from the observed reads, the original composite letters in sufficiently many positions of the total message. The sequencing readout (that is, the observed sequencing reads) is the product of a complex process, consisting of DNA synthesis, long-term storage[2,18], sampling and DNA sequencing. The distribution of counts, for every letter in {A,C,G,T}, resulting from σ at depth N can be described by a single model in which the readout counts are multinomial:

$$X^{(N)}(\sigma,P_{syn},P_{deg},P_{seq}) \sim \text{Multinomial}(N,(p_A(\sigma),p_C(\sigma),p_T(\sigma),p_G(\sigma))) \quad (1)$$

The parameters of the distribution are the designed input letter σ, the sequencing depth N and the errors introduced in the synthesis, storage and sequencing steps of the process, P_{syn}, P_{deg} and P_{seq} (Methods). While each step introduces different errors and biases, the most important parameters that affect the readout are the sampling of molecules to be sequenced and the sequencing depth.

The sequencing readout frequencies will most likely not exactly match any letter from the original alphabet. Inference of the original letter is performed by converting the readout to a vector of base frequencies and comparing it to the base frequencies of the candidate letters in the composite alphabet. The comparison can be done, for example, using the Kullback–Leibler divergence or the L^1 norm. To assess the performance of this inference step, we developed a simulation model and analyzed inference rates for various composite alphabets (Supplementary Fig. 2). The Kullback–Leibler divergence, which corresponds to a maximum-likelihood estimator (Supplementary Note), was found to perform much better and was thus used in the remainder of this study, including the molecular implementation (Methods).

Large-scale composite DNA-based data storage. To show the feasibility of the composite DNA alphabet concept and to demonstrate its potential for improving DNA-based data archiving systems, we performed a large-scale molecular implementation of a storage system based on a six-letter composite alphabet. The system consisted of using our composite letter encoding approach together with an error-correction system, which was based on a combination of Reed–Solomon[19] and fountain[7,20] schemes (Methods), to produce a composite DNA encoding pipeline (Fig. 2). We first used our system to store and successfully retrieve a 2.12 MB data file from Erlich and Zielinski[7]. Our encoded DNA pool consisted of 58,000 six-letter composite 152-nucleotide oligonucleotides, as compared to 72,000 oligonucleotides of the same length that were required using standard DNA, demonstrating a ~24% increase in logical density, as measured by bits per synthesis cycle (Table 1 and Supplementary Table 4). The six-letter composite alphabet used here was $\Sigma_6=\{A,C,G,T,M,K\}$, where $M=(1,1,0,0)$ and $K=(0,0,1,1)$. Note that $\Sigma_6\subset\Phi_2$ ($|\Sigma_6|=6,|\Phi_2|=10$). Our error-correcting scheme uses a Reed–Solomon code at the composite DNA level (using the appropriate Galois field) and not at the binary bits level, thereby improving the robustness of the system (Fig. 2; Methods). We further demonstrated the increased logical density of composite DNA by encoding a bilingual interactive version of the Bible, compressed to a 6.42 MB file, using three different composite alphabets. The above six-letter alphabet Σ_6 required 174,000 oligonucleotides, while a five-letter alphabet $\Sigma_5=\{A,C,G,T,M\}\subset\Phi_2$ required 193,000 oligonucleotides and a standard four-letter alphabet $\Sigma_4=\Phi_1$ required 217,000 synthetic oligonucleotides, all of the same length of 152 nucleotides (Table 1 and Supplementary Table 4). All the composite DNA oligonucleotides mentioned above were synthesized by Twist Bioscience, using standard DNA-writing hardware and an optimized synthesis process to obtain the desired nucleotide ratios for the letters K and M. We thus demonstrated large-scale composite DNA synthesis. Using the data acquired, we further investigated the characteristics of this approach to composite DNA synthesis.

The synthesized DNA was amplified using two different primer pairs as technical repeats. We then sequenced the resulting synthetic DNA sample (100-nucleotide paired-end reads, Illumina HiSeq at the Technion Genome Center). Our library and reaction

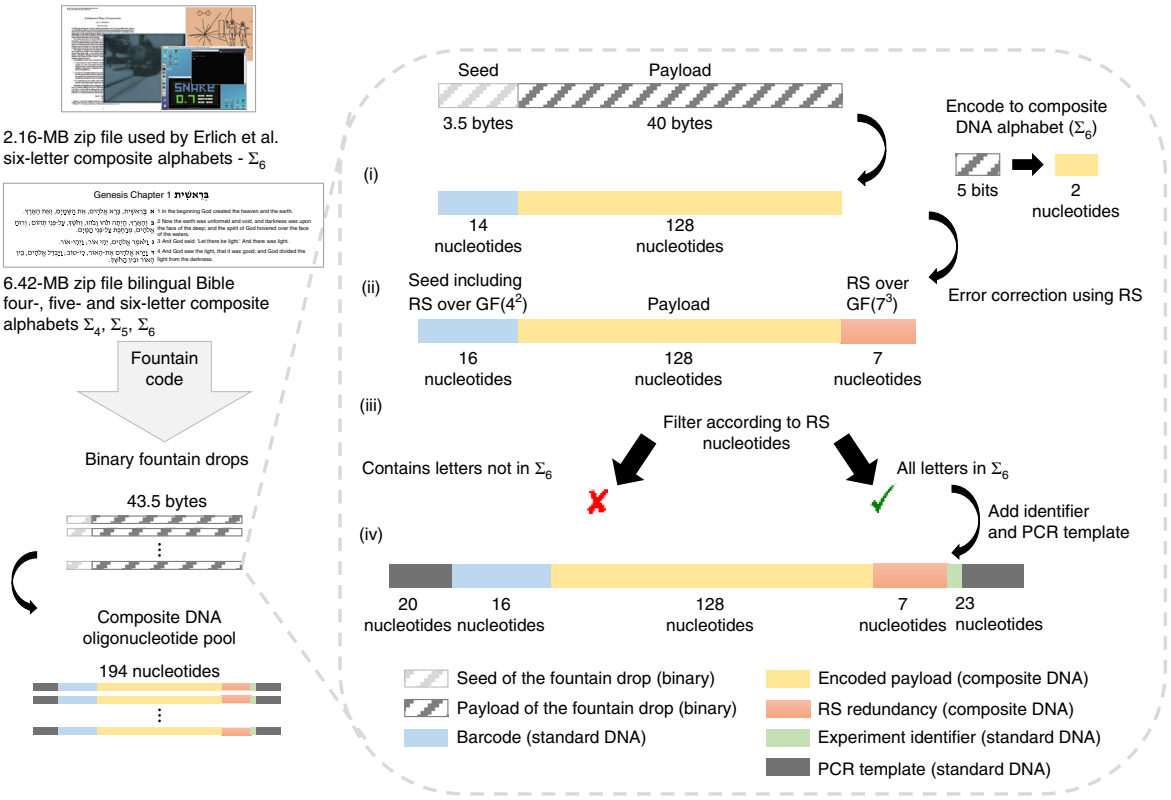


Fig. 2 | Encoding pipeline of a large-scale composite DNA-based data storage. A compressed input file was processed by the fountain code to produce binary droplets. A composite DNA encoding workflow was then applied for each droplet (Methods). (i) The binary message was translated into a composite DNA sequence. The seed sequence was translated to standard DNA sequence, which served as a barcode for the decoding process. The payload was translated to a six-letter composite DNA alphabet (Σ_6) in 5-bit chunks. (ii) Error-correction nucleotides were added to the DNA sequence by using a systematic Reed–Solomon (RS) encoding. The barcode was encoded using Reed–Solomon over $GF(2^4)$ and the payload was padded and encoded using Reed–Solomon over $GF(7^3)$. (iii) Each encoded message was then filtered to verify that the Reed–Solomon redundancy letters are all from Σ_6 (note that the Reed–Solomon code used here is systematic). (iv) Experiment identifier and amplification template sequences were appended to each valid sequence. Similar coding schemes were used for four- and five-letter alphabets (Methods; Supplementary Figs. 3–6).

design allowed for separately decoding each one of the four test messages, as described above, with each of the primer pairs. In Fig. 3, we describe the process of decoding the results of a sequencing reaction, performed on a synthesized composite DNA library originating from one message (in this case, the 6.4 MB Bible encoded into Σ_6 with a single pair of primers), and of inferring the underlying binary message (Methods). In brief, we first preprocessed the raw reads by assembling paired-end reads, filtering by length and grouping by putative barcode sequence (prefixes). Next, we filtered out prefixes with less than 20 associated reads generating a set of putative barcodes each associated with a group of reads. We inferred the full composite oligonucleotide for each putative barcode using Kullback–Leibler inference. The resulting composite oligonucleotides were Reed–Solomon verified (over $GF(7^3)$ in the case of Σ_6) and the valid oligonucleotides were converted into binary drops. Finally, we applied a binary fountain code decoding to obtain the original message, if successful. We note (Fig. 3c) that the average observed multiplicity, for each one of the inferred putative barcodes was 96 reads. Inference of the full composite oligonucleotide was only done for putative barcodes with more than 20 reads. Figure 3d,e depict frequencies and Kullback–Leibler inference decision boundaries (red dashed lines) for positions that were originally designed as composite. Note that individual synthesized positions were correctly inferred at an error rate of less than 10^{-5} for these filtered putative barcodes. These errors were either corrected by the Reed–Solomon code (over $GF(7^3)$ in the case of Σ_6) or rejected by the same mechanism. Eventually, sufficiently many accepted drops (172,608 for the Bible encoded into Σ_6) made it to the last decoding step, which used an adaptation of the fountain code mechanism proposed by Erlich and Zielinski[7] (Fig. 3f).

Dilution and physical density of composite DNA storage. To assess the physical density achieved by our composite DNA-based storage system we performed a dilution experiment. The encoding DNA was sequentially diluted and then amplified, sequenced and processed through our decoding pipeline. Our results include four different encodings in four dilution experiments. In a physical density of 6 PB g^{-1} we managed to successfully decode the Bible message encoded using Σ_6 (See Supplementary Table 1 for physical density calculations). We recovered 167,093 of the 174,000 original composite oligonucleotides (Fig. 3b). Further dilution of the DNA yielded only partial recovery of the composite oligonucleotides, below the redundancy level that is recoverable by the fountain code. From the message from Erlich and Zielinski[7], encoded using Σ_6 and representing a 30 PB g^{-1} density, we successfully recovered 92% of the original oligonucleotides. This was slightly lower than required by the fountain code to decode the message. We observed that even in the standard DNA (Σ_4) experiment we achieved a lower physical density than that reported by Erlich and Ziellinski. This could be due to the modified synthesis process or to the larger scale of the experiment. From the complete set of dilution results, we estimated that using a six-letter composite alphabet we can achieve a physical density of 20–30 PB g^{-1}.

Table 1 | Comparison of published DNA-based data storage schemes

Study	Experiment	Error correction	Robustness to dropouts	Input data (MB)	Total oligonucleotide library length (nucleotides)	Logical density (bits per synthesis cycle)
Church et al.[3]		–	–	0.66	6,313,270	0.83
Goldman et al.[4]		+	Repetition (4)	0.76	17,940,195	0.34
Grass et al.[18]		Reed–Solomon	Reed–Solomon	0.08	583,947	1.14
Bornholt et al.[5]		–	Repetition (1.5)	0.05	18,120,000	0.02
Erlich and Zielinski[7]		Byte-level Reed–Solomon	Fountain (1.07)	2.12	10,944,000	1.57
Organick et al.[8]		Byte-level Reed–Solomon	Reed–Solomon	200.2	~2,000,000,000	1.1
This work (message from Erlich and Zielinski[7])	Composite Σ_6	DNA-level Reed–Solomon	Fountain (1.1)	2.12	8,758,000	1.93
This work (Bible)	Standard Σ_4	DNA-level Reed–Solomon	Fountain (1.08)	6.42	32,767,000	1.57
	Composite Σ_5	DNA-level Reed–Solomon	Fountain (1.08)	6.42	29,143,000	1.76
	Composite Σ_6	DNA-level Reed–Solomon	Fountain (1.08)	6.42	26,274,000	1.96
This work (small scale)	Composite Φ_3	–	–	22.5 bytes after binary Huffman	42	4.29

The schemes are ordered chronologically. Information for previous studies is taken from Erlich and Zielinski[7]. Information for Organick et al.[8] is taken from their report. Logical density is calculated by dividing total binary input data, measured in bits, by the total number of synthesis cycles (Supplementary Table 2). Fountain code redundancy level is specified in parentheses.

Our dilution experiment evidently involved PCR amplification of the diluted material. By investigating the distribution of the K and M compositions in the different dilution steps, we therefore also examined the potential composition biases introduced by PCR together with those originating from the dilution itself. The analysis is presented in Supplementary Fig. 7. We concluded that the distribution of base frequencies had higher variance but there was only a minimal shift in the mean frequencies (0.3% for K and 0.05% for M after three additional cycles of PCR).

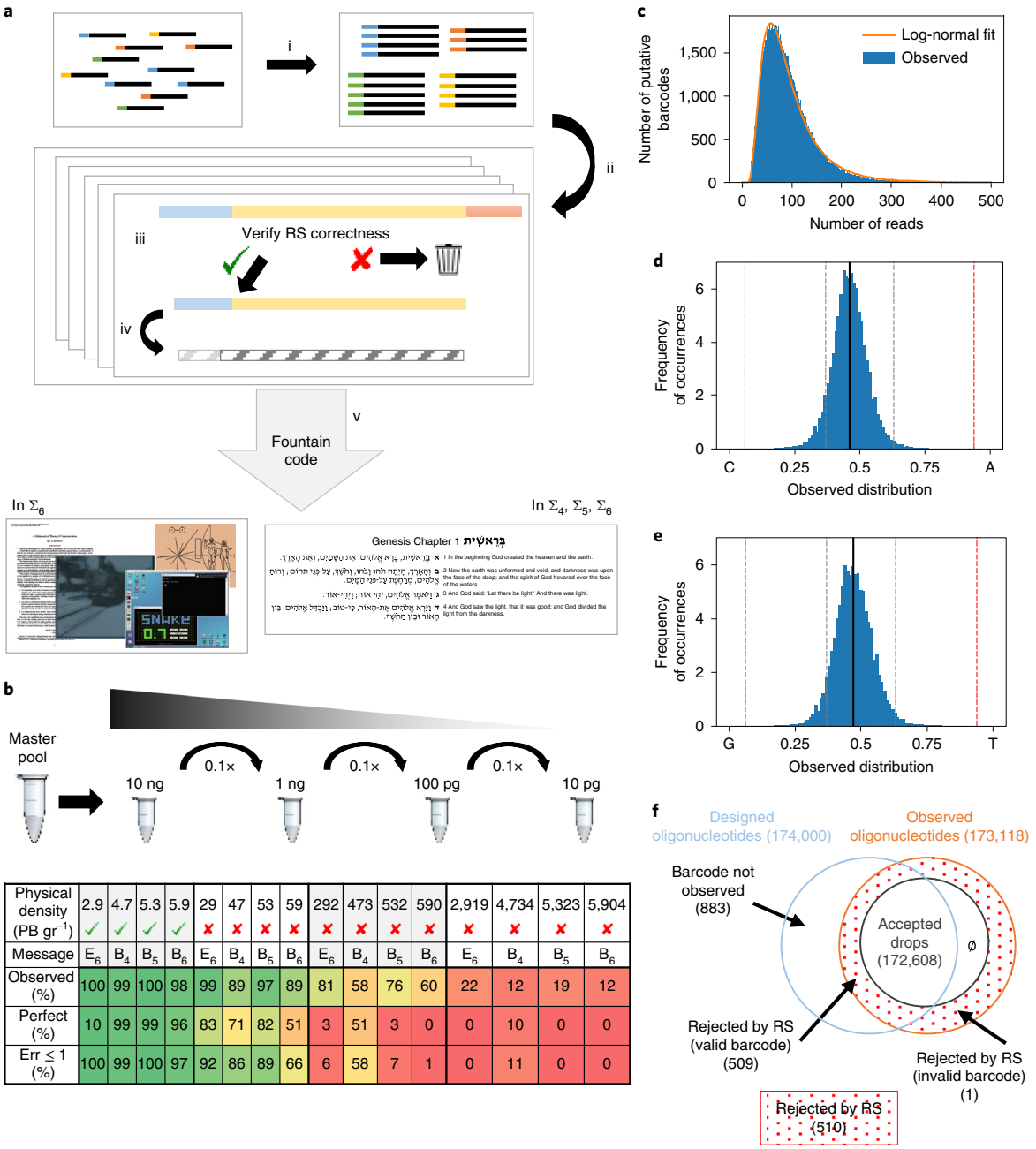
Higher resolutions, compositions and sequencing depth. As $\Phi_2 \subset \Phi_k$ for every even value of k, we can use the two composite letters from Φ_2 (that is, K and M) to calculate, on the basis of our experimental data, correct inference rates for these two letters in the context of larger composite alphabets. In Fig. 3d,e we further indicate the decision boundaries that would have been used under Φ_4 to distinguish, for example, $K=(0,0,2,2)$ from $\sigma=(0,0,3,1)$. Using these decision boundaries we would have had up to 7% of the positions designed as K (or M) potentially leaked to be interpreted as one of the two neighboring composite letters in Φ_4 (at the current sequencing depth). We further analyzed the effect of sequencing depth and the implication of extending the composite alphabet in Fig. 4. We observed that the mean base frequency of the composite letters K and M was slightly shifted toward G and C, respectively (Fig. 4b and Supplementary Fig. 8). As an immediate result, the leakage into neighboring letters was mainly toward G and C when considering Φ_2 decision boundaries. As expected, the leakage rate was anticorrelated with sequencing depth.

Subsampling of reads. We performed a subsampling experiment in which we repeatedly sampled different portions of the reads and assessed the read subsets using our decoding pipeline. For the message from Erlich and Zielinski[7], we show that using as little as 29 reads per oligonucleotide on average (30% sampling) was still sufficient to successfully decode the message with ~97% of the oligonucleotides successfully recovered (Fig. 4a).

Subsampling of the reads resulted in a wider distribution of base frequencies (Fig. 4b), while examining only oligonucleotides with higher coverage generated a narrow distribution (Fig. 4c and Supplementary Fig. 9). In particular, inference of both K and M under hypothetical use of Φ_2 or of Φ_4 was perfect, even at 160 reads per barcode (Fig. 4d). When considering Φ_6 and Φ_8, we obtained reasonable performance at the higher depths. It is important to note that some errors in inference can be tolerated as we use a Reed–Solomon error correction on the complete composite oligonucleotide at the composite alphabet level.

Composite alphabets increase logical density. To assess and establish the potential of large composite alphabets we combined simulations of large-scale composite DNA systems and a smaller-scale experimental proof of concept.

First, we calculated the potential logical density of storage systems based on large composite alphabets (Supplementary Table 2). A system using Φ_{10}, which consists of 286 letters, potentially achieves logical density of 6.4 bits per synthesis cycle, which is a fourfold increase over the standard DNA system (Supplementary Table 2). For Φ_{10} we further performed a full simulation study, working with experimentally motivated error rates, to understand the potential under non-perfect conditions. Planted errors include deletions, mismatches and insertions as derived from our data (Methods). We encoded the message from Erlich and Zielinski[7] using 17,585 composite oligonucleotides and simulated the synthesis and sequencing using different error rates and sequencing depths (Supplementary Tables 5 and 6). At an average sequencing depth of 2,000 reads and an overall error rate of 1:500 bases or less we correctly infer more than 99.95% of the composite oligonucleotides allowing for correct decoding of the message (Fig. 5a). Using Φ_5, an alphabet with 56 letters, we achieve a logical density of 4.5 bits per synthesis cycle (a 2.8-fold increase) and encoded the same message using 24,848 composite oligonucleotides (Supplementary Tables 5 and 6). With an average sequencing depth of 2,000 reads we successfully decoded the message even with an error rate of 1:50 bases while with



Physical density (PB gr^{-1})	2.9 ✓	4.7 ✓	5.3 ✓	5.9 ✓	29 ✗	47 ✗	53 ✗	59 ✗	292 ✗	473 ✗	532 ✗	590 ✗	2,919 ✗	4,734 ✗	5,323 ✗	5,904 ✗
Message	E_6	B_4	B_5	B_6	E_6	B_4	B_5	B_6	E_6	B_4	B_5	B_6	E_6	B_4	B_5	B_6
Observed (%)	100	99	100	98	99	89	97	89	81	58	76	60	22	12	19	12
Perfect (%)	10	99	99	96	83	71	82	51	3	51	3	0	0	10	0	0
Err ≤ 1 (%)	100	99	100	97	92	86	89	66	6	58	7	1	0	11	0	0

Fig. 3 | Performance of a large-scale composite DNA-based storage system. Decoding a composite library to infer the original encoded message. **a**, The steps of the decoding process (Methods): (i) Preprocessing and grouping by prefix sequences; (ii) generation of a set of putative oligonucleotides; (iii) inference of composite oligonucleotides using Kullback–Leibler inference and Reed–Solomon error correction; (iv) conversion into binary drops; and (v) binary fountain code decoding to obtain the original message, if successful. **b**, A dilution experiment testing the physical density achieved by the composite DNA approach. DNA was sequentially diluted, amplified, sequenced and tested for decoding. For each dilution, physical density is presented for all four encodings. (The message from Erlich and Zielinski[7] is marked E6 and the Bible is marked B_i.) The percentage of observed barcodes is presented together with the composite oligonucleotide inference rates and the rate of composite oligonucleotides inferred with up to one error (Err ≤ 1). **c**–**f**, Descriptive statistics related to the decoding process. Numbers indicated are for the 6.4-MB Bible message encoded into Σ_6 composite DNA. **c**, The number of reads associated to each 16-nucleotide prefix (putative barcode). The distribution follows a log-normal shape with a median of 81 reads and a mean of 96 reads. **d**,**e**, The distribution of base frequencies per synthesized position. For this counting we consider the positions that were designed to be composite—either K or M. Kullback–Leibler decision boundaries are also depicted. **f**, Acceptance statistics for the designed composite oligonucleotides.

500 reads we decoded the message with an error rate of 1:500 bases or less (Supplementary Fig. 10).

Using composite DNA has the potential to reduce the costs of DNA-based storage. This reduction is due to the increased logical density leading to reduced DNA synthesis cost, which is related to the total number of synthesis cycles. We analyzed the effect of using a large composite alphabet on the overall cost of a DNA-based storage system, taking into account the reduction in synthesis cost together with the increase in sequencing costs (Methods). We performed the analysis using different assumptions on the synthesis cost to sequencing cost ratio (C_{syn}:C_{seq}). With a moderate cost ratio of 1,000:1 we observe that using Φ_5 (56 letters) is optimal, with an

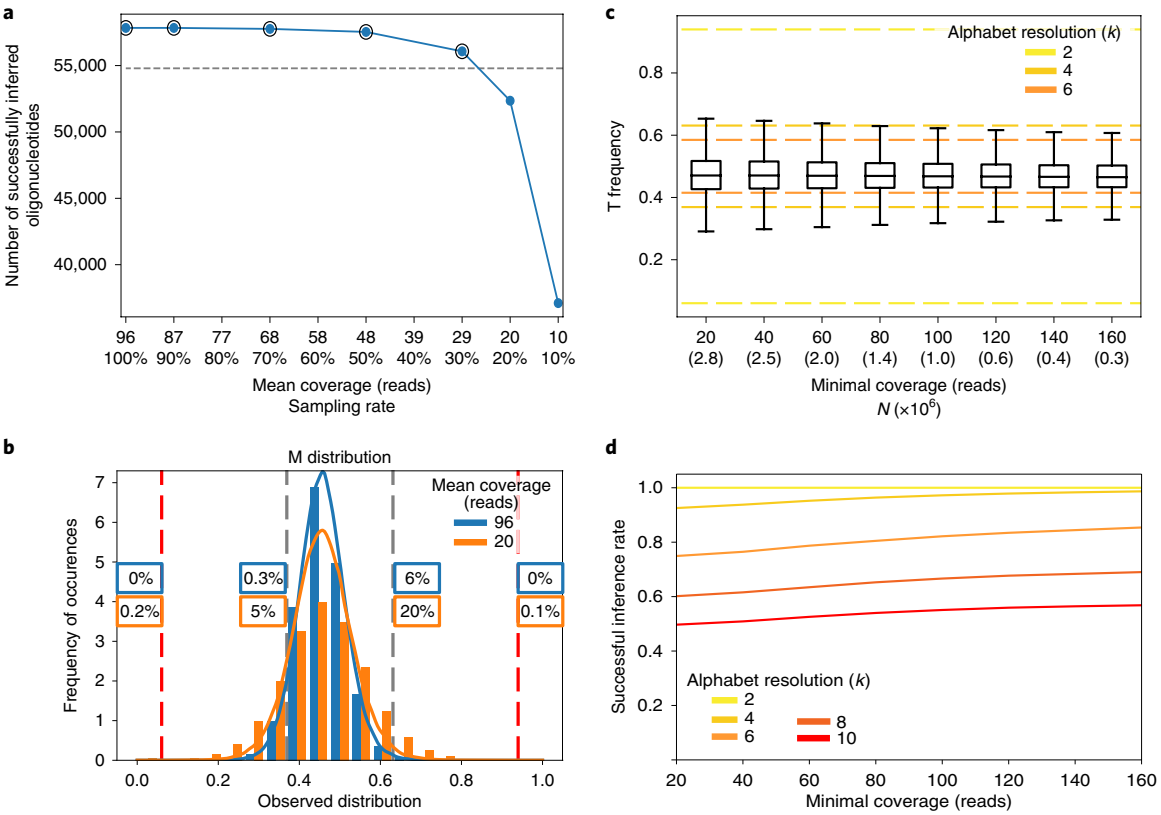


Fig. 4 | Analysis of higher-resolution composite alphabets using large-scale experiments. The effect of sequencing depth on the inference of composite letters. **a**, Composite oligonucleotide inference rate as a function of sequencing sampling rate based on the message from Erlich and Zielinski[7]. Successful decoding is marked with black circles. The theoretical limit of the fountain code (with 0.001 failure probability) is shown as a dashed gray line. **b**, The distribution of the A frequencies for the composite letter M in different sequencing sampling rates. The dashed lines depict the decision boundaries for Φ_2 and Φ_4. Rates of miss-inferred positions are depicted in the colored boxes. **c**, Performance of the K composite letter as a function of minimal sequencing coverage. The box plots depict the distribution of observed T frequencies by presenting the median, quartiles and whiskers that represent 1.5× the interquartile range. The number of analyzed positions in each box plot is depicted below the minimal coverage. The dashed lines depict the decision boundaries for Φ_2, Φ_4 and Φ_6. **d**, Successful inference rates of the composite letter K over different composite alphabets as a function of minimal sequencing cover.

overall expected cost reduction of 52% (Fig. 5b). With these cost assumptions, extending the alphabet up to Φ_{10} leads to a 30% cost reduction. Repeating the analysis with different cost ratios yields similar results with the optimal composite alphabet ranging between Φ_4 (500:1) and Φ_9 (5,000:1) (Supplementary Fig. 11).

Current synthesis technology also supports the use of DNA mixtures that represent higher-resolution composite alphabets, albeit on a small scale. To further explore the properties of large alphabets we encoded a short message (38 bytes in ASCI, 22.5 bytes after a binary Huffman compression) using composite alphabets of four different types, resulting in logical density of up to 4.29 bits per synthesis cycle (Table 1 and Supplementary Table 2; Methods).

The four different alphabets used are the standard DNA alphabet Φ_1, the full composite alphabets Φ_2 and Φ_3, and an alphabet containing the 15 IUPAC letters. The input English phrase, 'DNA STORAGE ROCKS!', was encoded to each of these alphabets using Huffman coding with the appropriate alphabet. The four resulting composite oligonucleotides were synthesized by IDT and sequenced by the Technion Genome Center (Methods; Supplementary Fig. 12 and Supplementary Table 3).

First, we examined the minimal sequencing depth required to decode the message correctly for each of the four composite alphabets. As expected, extending the alphabet by using higher resolutions requires deeper sequencing. In all four alphabets that were tested, a fully successful decoding was observed in as little as 100 reads (Fig. 5c) while a near-perfect decoding was obtained with as little as 50 reads (Supplementary Fig. 13). These results are better than those inferred from the aforementioned simulations, providing support for the cost estimations. In accordance with the theoretical analysis, Kullback–Leibler inference also performed much better than L^1 norm inference on the experimental data (Supplementary Figs. 14 and 15).

As predicted by the statistical model, some composite letters are harder to identify than others (Fig. 5d). However, contrary to the model prediction, when examining different letters from the same composite archetype (that is, letters that are different permutations of the same probability vector) we observe significant differences ($P < 10^{-10}$; Z test for proportion difference for the letters GGA and GGC) (Fig. 5d). These higher-resolution results also suggest that the position of the letter in the synthesized oligonucleotide affects the identification rate. To further explore the differences between different composite letters, we designed another synthetic DNA oligonucleotide containing all the equimolar letters (represented by the 15-letter IUPAC alphabet), with multiple copies of each composite letter distributed along the designed sequence (Methods; Supplementary Fig. 16). We examined the inference rate at a depth of 15 reads and reported the results as a function of the letter and the position in the oligonucleotide (Supplementary Fig. 17). We observed a small but persistent decrease in inference rates as a function of the position on the synthesized oligonucleotide, starting from the 5′ end.

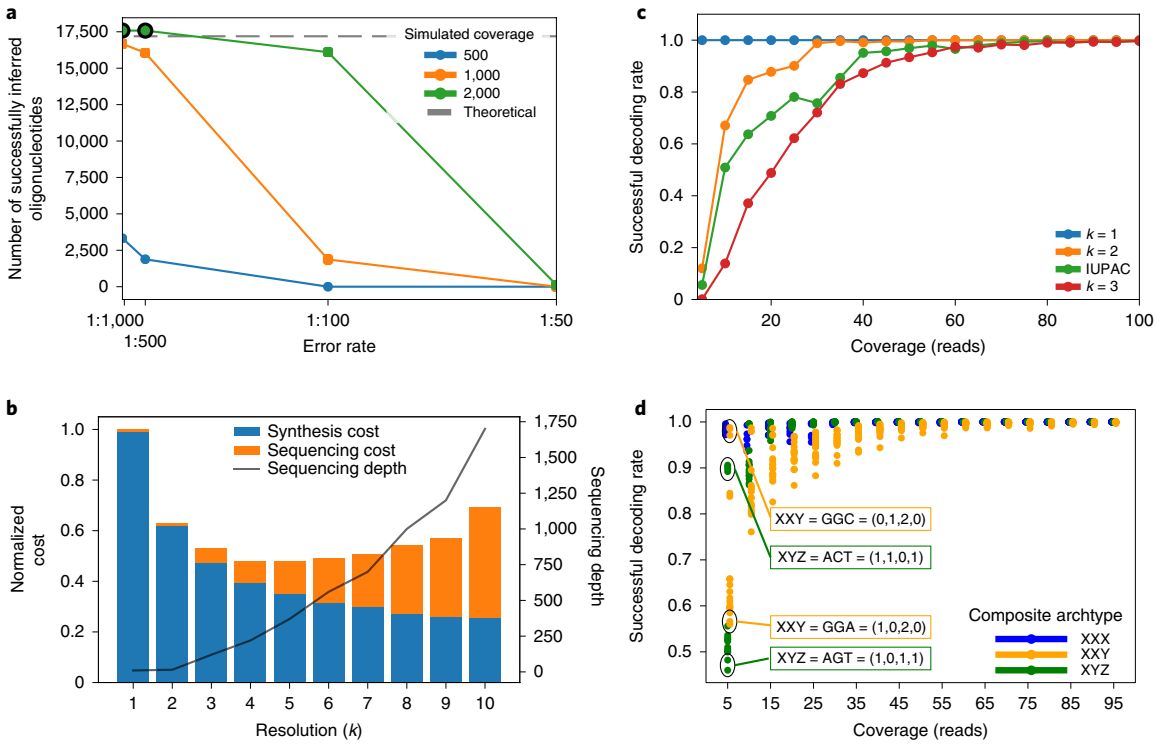


Fig. 5 | Data storage systems based on large composite alphabets. a, Successful inference of a Φ_{10} composite storage system storing the message from Erlich and Zielinski[7] on the basis of simulations ($N = 5$). The number of correctly inferred oligonucleotides is shown as a function of the simulated error rates for three sequencing depths (Methods). The theoretical limit of the fountain code (with 0.001 failure probability) is shown as a dashed gray line. Instances for which a successful decoding was achieved are marked with a black circle. **b**, Cost analysis for a composite DNA-based storage system using different alphabets. Cost components are normalized to the total cost of a standard DNA-based system. **c**, Successful decoding rate experimentally obtained in the molecular implementation using Kullback–Leibler inference as a function of sequencing depth for the four composite DNA alphabets. **d**, Inference rates for letters in Φ_3 (multiple occurences along the oligonucleotide) as a function of sequencing depth. The letters are colored by their composite archetype.

Discussion

We applied composite DNA letters to enable DNA-based data storage using fewer DNA synthesis cycles. Composite DNA schemes could be combined with other approaches such as orthogonal base pair systems[21], efficient coding[6,7,9,22] and random access approaches[5,6,10,23,24] to increase capacity and fidelity of DNA-based storage systems. However, the logical density advantage of using composite DNA is traded off with several performance metrics as discussed below.

Incorporating composite DNA into future DNA-based storage systems will require further investment in several directions. First and foremost, any large-scale implementation will require scaling up the currently limited commercial hardware for synthesis of composite DNA. The current implementation of a six-letter alphabet did not increase the cost per synthesized oligonucleotide. Progress to even higher resolutions will require slight modifications in the design of synthesis hardware or the adaptation of other synthesis approaches. In a recent study, the authors describe the construction of a flexible laboratory-size synthesis system[25]. This system can be configured to accommodate higher-resolution alphabets. Second, using highly multiplexed composite DNA sequences will require better understanding of the effect of composite DNA on different chemical processes involved in DNA manipulation. Previous studies dealt with the chemical limitations of these processes either by employing strict encoding schemes[3–6,9] or by using coding methodologies like DNA fountains to handle sequence dropout[7]. Employing composite DNA inherently generates balanced DNA molecules, resulting from the combinatorial space associated with every designed composite sequence. While unwanted sequences will unavoidably be part of the ensemble of synthesized molecules, the inherent independence of the different positions renders them negligible, representing an extra benefit of the composite DNA approach. Third, the design principles for composite DNA sequences, or of related coding approaches, as well as the decoding pipeline, can be further tuned for optimal results. Mixed composite alphabets can be generated to minimize inference errors without compromising the alphabet size, by only selecting subsets of the full alphabets Φ_k. Technical calibration of the actual base frequencies, on the basis of further experimental investigation thereof (such as in Fig. 4b), can be added to the decoding pipeline allowing the correction of systematic synthesis biases.

The use of composite DNA affects required sequencing depth and physical density, as our data show. Using current technologies, synthesis cost per position is approximately four orders of magnitude larger than sequencing cost per base, yielding a potential overall reduction in cost when using composite DNA. This holds despite the increase in sequencing costs entailed by the required depth. We further analyzed the effect of both factors on the overall cost, as described in the text and in Fig. 5b and Supplementary Fig. 11. The physical density reported herein is about a single order of magnitude less than the best previously reported.

It is important to note, in relation to the work presented here, that increased alphabet size for data storage can also be achieved, to a limited extent, by introducing synthetic orthogonal nucleotide pairs[21,26]. In the early days of DNA sequencing by hybridization, degenerate and semidegenerate bases were proposed as

wildcards for increasing the fidelity of the system[27–30]. Recently, a DNA sequencing approach that uses mixtures of nucleotides and associated error correction was described[31]. Transcription-factor binding and other regulatory cellular functions are often based on partially redundant recognition[32,33]. DNA synthesis is used to study and mimic regulatory systems[34–36], for tagging and tracking[37] and in many other applications.

The current study and suggested methodology adds DNA-based data storage to the potential applications of composite DNA nucleotides and will hopefully contribute to further interest and the development of efficient composite DNA synthesis, which could be used in all relevant applications.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of code and data availability and associated accession codes are available at https://doi.org/10.1038/s41587-019-0240-x.

Received: 26 September 2018; Accepted: 25 July 2019;
Published online: 9 September 2019

References

1. Cox, J. P. Long-term data storage in DNA. *Trends Biotechnol.* **19**, 247–250 (2001).
2. Zhirnov, V., Zadegan, R. M., Sandhu, G. S., Church, G. M. & Hughes, W. L. Nucleic acid memory. *Nat. Mater.* **15**, 366–370 (2016).
3. Church, G. M., Gao, Y. & Kosuri, S. Next-generation digital information storage in DNA. *Science* **337**, 1628 (2012).
4. Goldman, N. et al. Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. *Nature* **494**, 77–80 (2013).
5. Bornholt, J. et al. Toward a DNA-based archival storage system. *IEEE Micro* **37**, 98–104 (2017).
6. Tabatabaei Yazdi, S. M. H. et al. A rewritable, random-access DNA-based storage system. *Sci. Rep.* **5**, 14138 (2015).
7. Erlich, Y. & Zielinski, D. DNA Fountain enables a robust and efficient storage architecture. *Science* **355**, 950–954 (2017).
8. Organick, L. et al. Random access in large-scale DNA data storage. *Nat. Biotechnol.* **36**, 242–248 (2018).
9. Gabrys, R., Kiah, H. M. & Milenkovic, O. Asymmetric lee distance codes for DNA-based storage. In *Proc. 2015 IEEE International Symposium on Information Theory (ISIT)* 909–913 (IEEE, 2015)..
10. Levy, M. & Yaakobi, E. Mutually uncorrelated codes for DNA storage. In *Proc. 2017 IEEE International Symposium on Information Theory (ISIT)* 3115–3119 (IEEE, 2017).
11. 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).
12. Palluk, S. et al. De novo DNA synthesis using polymerase–nucleotide conjugates. *Nat. Biotechnol.* **36**, 645–650 (2018).
13. Roquet, N., Park, H. & Bhatia, S. P. Nucleic acid-based data storage. US patent 20180137418 (2017).
14. LeProust, E. M. et al. Synthesis of high-quality libraries of long (150mer) oligonucleotides by a novel depurination controlled process. *Nucleic Acids Res.* **38**, 2522–2540 (2010).
15. Barrett, M. T. et al. Comparative genomic hybridization using oligonucleotide microarrays and total genomic DNA. *Proc. Natl Acad. Sci. USA* **101**, 17765–17770 (2004).
16. Kosuri, S. & Church, G. M. Large-scale de novo DNA synthesis: technologies and applications. *Nat. Methods* **11**, 499–507 (2014).
17. Choi, Y. et al. High information capacity DNA-based data storage with augmented encoding characters using degenerate bases. *Sci. Rep.* **9**, 6582 (2019).
18. 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. Engl.* **54**, 2552–2555 (2015).
19. Reed, I. S. & Solomon, G. Polynomial codes over certain finite fields. *J. Soc. Ind. Appl. Math.* **8**, 300–304 (1960).
20. MacKay, D. J. C. Fountain codes. *IEE Proc. Comm.* **152**, 1062 (2005).
21. Jiménez-Sánchez, A. DNA computer code based on expanded genetic alphabet. *Eur. J. Comput. Sci. Inf. Technol.* **2**, 8–20 (2014).
22. Tabatabaei Yazdi, S. M. H. et al. DNA-based storage: trends and methods. *IEEE Trans. Mol. Biol. Multiscale Commun.* **1**, 230–248 (2015).
23. Raviv, N., Schwartz, M. & Yaakobi, E. Rank modulation codes for DNA storage. In *Proc. 2017 IEEE International Symposium on Information Theory (ISIT)* 3125–3129 (IEEE, 2017).
24. Yazdi, S. M. H. T., Kiah, H. M., Gabrys, R. & Milenkovic, O. Mutually uncorrelated primers for DNA-based data storage. Preprint at https://arxiv.org/abs/1709.05214 (2017).
25. 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).
26. Hoshika, S. et al. Hachimoji DNA and RNA: a genetic system with eight building blocks. *Science* **363**, 884–887 (2019).
27. Bains, W. Hybridization methods for DNA sequencing. *Genomics* **11**, 94–301 (1991).
28. Pevzner, P. A. Rearrangements of DNA sequences and SBH. *Comput. Chem.* **18**, 221–223 (1994).
29. Preparata, F. P. & Oliver, J. S. DNA sequencing by hybridization using semi-degenerate bases. *J. Comput. Biol.* **11**, 753–765 (2004).
30. Snir, S., Yeger-Lotem, E., Chor, B., and Yakhini, Z. Using restriction enzymes to improve sequencing by hybridization. Technical report CS-2002-14 (Technion, 2002).
31. Chen, Z. et al. Highly accurate fluorogenic DNA sequencing with information theory-based error correction. *Nat. Biotechnol.* **35**, 1170–1178 (2017).
32. Davidson, E. H. *The Regulatory Genome: Gene Regulatory Networks in Development and Evolution* (Academic, 2006).
33. Sandelin, A., Alkema, W., Engström, P., Wasserman, W. W. & Lenhard, B. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.* **32**, D91–D94 (2004).
34. Levy, L. et al. A synthetic oligo library and sequencing approach reveals an insulation mechanism encoded within bacterial σ54 promoters. *Cell Rep.* **21**, 845–858 (2017).
35. Sharon, E. et al. Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nat. Biotechnol.* **30**, 521–530 (2012).
36. Gilbert, L. A. et al. CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. *Cell* **154**, 442–451 (2013).
37. Mikutis, G. et al. Silica-encapsulated DNA-based tracers for aquifer characterization. *Environ. Sci. Technol.* **52**, 12142–12152 (2018).

Acknowledgements

We thank T. Katz-Ezov and T. Hashimshony from the Technion Genome Center for advice and assistance with oligonucleotide design and sequencing experiments. We also thank P. Weiss from Twist Bioscience for technical support and assistance with DNA synthesis. Finally, we thank the Yakhini and Amit research groups for valuable comments and discussions. L. Anavy is supported by the Adams Fellowships Program of the Israel Academy of Sciences and Humanities. This project received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement 664918 (MRG-Grammar).

Author contributions

L.A. and Z.Y. initiated and designed the coding and algorithmic approach. L.A. developed the software and performed data analysis. I.V. and O.A. performed the experiments. L.A., R.A. and Z.Y. wrote the manuscript. R.A. and Z.Y. supervised the study.

Competing interests

L.A, Z.Y and R.A are the inventors of a patent application for the method described in this article. The initial filing was assigned United States Provisional Patent Application No. 62/674,114. The remaining authors declare no competing financial interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41587-019-0240-x.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to L.A. or Z.Y.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Definition of composite DNA letters. A composite DNA alphabet is defined as

$$\Phi_k = \left\{ (\sigma_A, \sigma_C, \sigma_G, \sigma_T) : \sigma_{i\in\{A,C,G,T\}} \in \mathbb{Z}_{\geq 0}, \sum_{i\in\{A,C,G,T\}} \sigma_i = k \right\} \quad (2)$$

In this notation, $\sigma = (\sigma_A,\sigma_C,\sigma_G,\sigma_T) \in \Phi_k$ represents a composite letter and k is a tunable parameter that represents the resolution of the composite alphabet.

The size of the composite DNA alphabet grows with resolution as follows:

$$|\Phi_k| = \binom{k+3}{k} \quad (3)$$

Using a naive mapping from $\{0,1\}^*$ to Φ_k^*, we see that the length of the encoded message decreases with the resolution of the composite alphabet

$$L_k = \frac{L_b}{\log_2 |\Phi_k|} \quad (4)$$

where L_k is the length of the message encoded using composite alphabet of resolution k and L_b is the length of the original binary message.

A multinomial model for the composite DNA letters. The distribution of the observed count vectors is governed by the sampling process of N independent molecules that are sequenced and counted to generate the count vector. Given the original composite letter $\sigma = (\sigma_A,\sigma_C,\sigma_G,\sigma_T)$, assuming a sequencing depth N (that is, the number of independent copies sequenced), and ignoring other factors beside the sampling, the observed counts constitute a random variable with a multinomial distribution

$$\begin{aligned} X_{seq}^{(N)}(\sigma) &= (X_A(\sigma), X_C(\sigma), X_G(\sigma), X_T(\sigma)) \\ &\sim \text{Multinomial}(N, (\sigma_A/k, \sigma_C/k, \sigma_G/k, \sigma_T/k)) \end{aligned} \quad (5)$$

and so

$$X_{seq}^{(N)} \in \mathcal{X}^{(N)} = \left\{ (X_A, X_C, X_G, X_T) : X_{i\in\{A,C,G,T\}} \in \mathbb{Z}_{\geq 0}, \sum_{i\in\{A,C,G,T\}} X_i = N \right\} \quad (6)$$

The observed read counts are, in actuality, also affected by the following parameters: synthesis error rate, represented by $\{P_{syn}\}_{i\to j} = P(j \text{ synthesized}|i \text{ designed})$; degradation rates, represented $\{P_{deg}\}_{i\to j} = P(j \text{ present after storage}|i \text{ synthesized})$; and sequencing error rate, represented by $\{P_{seq}\}_{i\to j} = P(j \text{ read}|i \text{ present})$.

Deletion and insertion events are a special class of errors in DNA synthesis and sequencing. These affect the read counts for all positions following the event position.

Assuming independence of the different error sources, we can incorporate all errors into a generalized multinomial model with slightly altered probabilities

$$X^{(N)}(\sigma, P_{syn}, P_{deg}, P_{seq}) \sim \text{Multinomial}(N, p(\sigma)), X^{(N)} \in \mathcal{X}^{(N)} \quad (7)$$

where $\mathbf{p}(\sigma) = (p_A(\sigma),p_C(\sigma),p_G(\sigma),p_T(\sigma))$ is a corrected probability vector.

Inference of composite DNA letters. To correctly read a message coded using composite DNA letter alphabet we must infer the original composite letter σ from the observed $\mathbf{x}^{(N)} \in \chi^{(N)}$. Namely, we need to define a decoding map:

$$f: \mathcal{X}^{(N)} \to \Phi_k \quad (8)$$

We infer the original letter from the observed vector $\mathbf{x}^{(N)}$ by first calculating a proportion vector $\boldsymbol{\pi}(x)$

$$\begin{aligned} \boldsymbol{\pi}(x) &= (\pi_A(x), \pi_C(x), \pi_G(x), \pi_T(x)) \\ &= (x_A/N_x, x_C/N_x, x_G/N_x, x_T/N_x), N_x = \sum_{i\in\{A,C,G,T\}} x_i \end{aligned} \quad (9)$$

Next we use one of following mapping approaches.

L^p norm:

$$f(x^{(N)}) = \text{argmin}_{\sigma\in\Phi_k}(\|\pi(x^{(N)}) - \pi(\sigma)\|_p) \quad (10)$$

Kullback–Leibler:

$$f(x^{(N)}) = \text{argmin}_{\sigma\in\Phi_k}(\text{KL}(\pi(x^{(N)}), \pi(\sigma))) \quad (11)$$

Where KL(P,Q) stands for the Kullback–Leibler divergence:

$$\text{KL}(P, Q) = \sum_i P_i \log\left(\frac{P_i}{Q_i}\right) \quad (12)$$

and i runs over the four letters in {A,C,G,T}.

When using the error-aware multinomial model, the Kullback–Leibler approach is equivalent to a maximum-likelihood mapping (Supplementary Note). As the Kullback–Leibler measure is highly sensitive for letters on the edges of the simplex, we implemented this approach using a variation of the composite alphabet in which zero entries in the probability vectors are replaced with some small value ϵ.

Simulations of composite DNA letter inference. The probability of correctly identifying the original letter from the observed count vector is defined as

$$C(\sigma) = \text{Prob}(f(X^{(N)}(\sigma)) = \sigma) \quad (13)$$

where $X^{(N)}$ is distributed in accordance with the process parameters.

We simulated the process of DNA synthesis, storage and sequencing to examine the properties of the composite DNA inference mechanisms. For simplicity, we used a single error rate parameter. We then used the inference mechanisms described above to infer the original composite letter. For a given composite alphabet we repeated the process for $R = 1{,}000$ times for each composite letter to estimate the inference success rate for every letter σ (Supplementary Fig. 2).

Encoding in the composite DNA-based storage system. We designed a composite DNA-based storage system consisting of the following components:

1. DNA fountain encoding with no Reed–Solomon error correction to an extended output alphabet. We altered the previously described DNA fountain code[7] to support composite DNA sequences. The seed value of the DNA fountain was limited to fit in 3.5 bytes. The conversion of the binary droplet to a DNA sequence was altered so that the droplet seed, which is encoded in the first 3.5 bytes, was converted to a 14-nucleotide standard DNA sequence acting as a barcode, and the rest of the binary sequence was converted to the desired composite DNA alphabet. For Σ_4 (standard DNA) every two bits were converted to a single DNA letter. For Σ_5 every 9 bits were converted to a four-letter composite word. For Σ_6 every 5 bits were converted to a two-letter composite word.
2. Addition of Reed–Solomon error correction directly to composite DNA sequence. We implemented Reed–Solomon codes over finite fields of various orders. Two bases of standard DNA were added to the barcode sequence by using a systematic (7, 8) Reed–Solomon code over $GF(2^4)$. The remaining 128-nucleotide composite sequence was padded to be $129 = 43 \times 3$ nucleotides and then encoded using a (43, 45) Reed–Solomon code over $GF(4^3)$, $GF(5^3)$ and $GF(7^3)$ for the composite alphabets Σ_4, Σ_5 and Σ_6, respectively. This generated 151-nucleotide composite sequences. To overcome the mismatch between the six-letter alphabet Σ_6 and the seven-letter finite field, an additional filtration step was used in which encoded oligonucleotides were included in the final set of oligonucleotides only if all the Reed–Solomon redundancy bases were in Σ_6. This entailed a 13% overhead in the generation time of the final set of oligonucleotides (Supplementary Figs. 3–6).

The resulting set of composite oligonucleotides was incorporated into a constant DNA backbone containing 20-nucleotide amplification primers on each side (we used two different sets of primers as technical repeats) and a 3-nucleotide barcode marking the experiment ID (input dataset and output alphabet). This resulted in a set of 194-nucleotide composite oligonucleotides. Combining the sets from all four experiments and two sets of primers resulted in 1.4 million composite oligonucleotides (Supplementary Table 4) that were synthesized by Twist Bioscience.

Synthesis of composite DNA oligonucleotides. Composite DNA oligonucleotides for all libraries were synthesized by Twist Bioscience using standard DNA-writing hardware and an optimized synthesis process to obtain the desired nucleotide ratios for the letters K and M. In addition to the standard A, C, G and T phosphoramidite solutions, calibrated mixtures of G/T to obtain K and A/C to obtain M were prepared and then added to two additional discreet print heads on the standard Twist Bioscience DNA writer. The process was then run and took 24 h to complete.

Sequencing of the composite DNA storage library. The synthetic DNA library was amplified using 14 cycles of PCR and sequenced by the Technion Genome Center. Sequencing was done on two lanes of an Illumina HiSeq machine and resulted in ~230 million reads for both primer sets.

Decoding in the composite DNA-based storage system. The decoding of the message was split into the following two steps:

1. Generation of a set of composite oligonucleotides:
   a. Preprocessing of the reads including assembly of paired-end reads using PEAR[38], filtration on the basis of length and the existence of the primer sequence, and generation of eight separate sets for the different experiment.
   b. Grouping of the reads according to the 16-nucleotide prefix to generate a set of putative barcodes each with an associated set of reads.
   c. Filtration of the putative barcode set to include only sequences with at least 20 reads associated to them.

d. Inference of the composite sequence using the Kullback–Leibler inference method.
e. Decoding of the composite sequence using the appropriate Reed–Solomon decoder. Only error detection was performed.

2. DNA fountain decoding of the resulting set of composite oligonucleotides using the altered DNA fountain decoder.

Sampling experiments for investigating sequencing coverage analysis. Reads for the message from Erlich and Zielinski[7] were sampled by taking different subsets of the reads covering 10, 20, 30, 50, 60, 70, 80 and 90% of the reads. Each subset was then processed using the same decoding pipeline. Each sampling rate was repeated twice.

Simulation of large-scale composite DNA-storage systems. The message from Erlich and Zielinski[7] was encoded using Φ_5 and Φ_{10} and error correction was performed using the Reed–Solomon method over the appropriate finite field. For Φ_5, every 23 bits were converted to a four-letter composite DNA word. The resulting composite sequence was then encoded using a (65, 68) Reed–Solomon code over $GF(59^2)$ (Supplementary Table 5). As Φ_5 contain only 56 letters, a similar approach to the encoding of Σ_6 was used, and messages containing letters 57–59 in the Reed–Solomon redundancy were dropped. For Φ_{10}, every 8 bits were converted to a single composite DNA letter and only 256 of the possible 286 composite letters were used. The resulting composite sequence was then encoded using a (130, 136) Reed–Solomon code over $GF(2^8)$ (Supplementary Table 5). The resulting 17,585 composite oligonucleotides for Φ_{10} (24,848 for Φ_5) were then simulated using the following procedure with the following parameters: number of oligonucleotides (N); error rate (r); and mean sequencing depth (D). For each oligonucleotide a read count is sampled as a random variable $X \sim \mathrm{Bin}\left(n = N \times D, p = \frac{1}{N}\right)$. Then, for each position in the oligonucleotide with a composite letter σ, the base frequency is sampled using a multinomial random variable $Y \sim \mathrm{Multinomial}(n = x, \pi(\sigma))$ and instantiated as a random permutation over A, C, G and T. Errors are then introduced by randomly selecting $\frac{r}{2}$ of the positions and replacing their value to one of the three other letters. Finally $\frac{r}{2}$ of the positions are deleted. The simulation was repeated five times for each alphabet. The simulated read sets were then processed using the same decoding pipeline (Supplementary Table 6).

Cost analysis. The overall cost of composite DNA-based storage systems was calculated by using the encoding presented in Supplementary Table 2 and calculating the synthesis cost directly. Required sequencing depth was determined using simulations by finding the depth in which the worst-case letter of the used alphabet had an error rate of less than 10^{-4}. The cost components were normalized by dividing each component by the total cost of a standard DNA system.

Experiments with large composite alphabets. We encoded a short input message ('DNA STORAGE ROCKS!') using an encoding pipeline consisting of the following steps:

- Mapping of the message to a binary sequence using the standard ASCI code for the English language.
- Huffman coding the binary sequence into a sequence of composite DNA letters of resolution k using the complete Shakespeare corpus[39] to generate the Huffman coding scheme[40].
- To achieve equal sequence length for all designed oligonucleotides (of different resolutions k), we repeated the encoded message to fit a predetermined length of 42 bases.

This process was performed for four different resolutions ($k = 1$, 2 and 3, and a special case in which the composite alphabet consists of only equimolar combinations of bases (representing the 15 different IUPAC codes)).

To calculate the logical density of these encodings we used similar Huffman coding of the exact same message into a binary sequence and calculated

$$\text{Logical density} = \frac{\text{Length of binary message after Huffman coding}}{\text{Length of composite message after Huffman coding}} \quad (14)$$

For each of the above four configurations, we inserted the encoded composite DNA sequence into a synthetic construct containing amplification primer templates, a unique molecular identifier and a barcode to obtain a total oligonucleotide length of 99 nucleotides (Supplementary Fig. 12 and Supplementary Table 3).

The four designed oligonucleotides were then commercially synthesized (IDT), amplified using PCR primers from the Illumina small RNA sequencing kit and sequenced using an Illumina Mi-Seq at the Technion Genome Center.

We obtained 5,421,556 50-base-pair paired-end reads of the four different samples. We merged the read pairs using PEAR[38] to generate 4,855,676 reads, 95% of which were of the designed length of 52 bases. We then split the reads into four different samples using the barcode value, placing ~25% of the reads in each sample.

Next, we decoded the original message using a decoding pipeline consisting of the following steps:

- Reading of the sample reads.
- Filtering of the reads on the basis of read length and removing reads containing undetermined bases ('N' output in the sequencing) and reads of lengths other than 52 bases.
- Inference of the composite sequence using the inference mechanisms described above.
- Decoding of the original messages using the same Huffman coding used for encoding.

For each sample we tested the ability to decode the entire message (including the repetition introduced to equalize oligonucleotide length), as well as only the first occurrence of the original encoded message text.

To test for the required sequencing depth for each sample representing a specific resolution, we sampled different numbers of reads for each resolution and repeated the decoding process for each such subsample data. We repeated the sampling process $R = 100$ times for each sampling rate and recorded the inference rates and the overall decoding outcome for each sample.

Error analysis for composite DNA letters. We designed a synthetic composite DNA oligonucleotide using the same overall design with the following alterations:

- The barcode and unique molecular identifier were removed as they were unnecessary for this analysis.
- The length of the composite DNA sequence was 145 nucleotides yielding a total oligonucleotide length of 192 nucleotides.
- The 145 composite nucleotides consisted of all the possible pairs of composite letters. This oligonucleotide design was constructed as a de Bruijn sequence. A balanced circular de Bruijn sequence over an alphabet of 12 letters composed of the eleven composite letters (15 IUPAC letters minus the four standard bases) plus one extra letter was constructed. The occurrences of the extra letter were then replaced by the standard DNA bases in a cyclic manner (Supplementary Fig. 16 and Supplementary Table 3).
- This 192-nucleotide oligonucleotide (de Bruijn sequence and primers) was then synthesized, processed and sequenced using similar procedures to the above with the following differences: the oligonucleotide was synthesized using IDT Ultramer synthesis technology for long synthetic DNA oligonucleotides; and sequencing was performed using the Nano Mi-Seq kit yielding 150-base-pair paired-end reads.

We obtained 1,086,991 150-base-pair paired-end reads. We merged the read pairs using PEAR[38] to generate 1,017,813 reads, 90% of which were of the designed length of 145 bases. We used a similar pipeline to the one described above to calculate inference rates for each position in the sequence and to investigate the properties of the error rates.

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

Data availability

All raw sequencing data are available from the European Nucleotide Archive (ENA) under accession PRJEB32427. This includes sequencing of the large-scale experiment described in Figs. 2–4, sequencing of the experiment with large alphabets described in Fig. 5 and sequencing of the error analysis experiment described in Fig. 5. All other data are available within the article or its supplementary information.

Code availability

All original software code included in this study is available online. Alteration of the previously published DNA fountain code to support composite DNA is available from https://github.com/leon-anavy/dna-fountain. Code used for Reed–Solomon error correction (altered from previously published code) is available from https://github.com/leon-anavy/Reed-Solomon. Custom code used for the analyses presented in this study is available from https://github.com/leon-anavy/composite-DNA.

References

38. Zhang, J., Kobert, K., Flouri, T. & Stamatakis, A. PEAR: a fast and accurate Illumina paired-end read merger. *Bioinformatics* **30**, 614–620 (2014).
39. Shakespeare, W. *The Complete Works of William Shakespeare* http://www.gutenberg.org/ebooks/100 (1994)
40. Huffman, D. A. A method for the construction of minimum-redundancy codes. *Proc. IRE* **40**, 1098–1101 (1952).

nature research

Corresponding author(s): Leon Anavy, Zohar Yakhini

Last updated by author(s): May 19, 2019

Reporting Summary

Nature Research 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 Research policies, see Authors & Referees 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	- We altered the previously published DNA fountain code. Available under: https://github.com/leon-anavy/dna-fountain - Custom python code was used for the Reed-Solomon related encode and decode. Available under: https://github.com/leon-anavy/Reed-Solomon
Data analysis	- For assembling the paired end NGS reads we used PEAR v0.9.10 - Custom python code was used for the analyses presented in the paper. Available under: https://github.com/leon-anavy/composite-DNA

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/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research 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 list of figures that have associated raw data
- A description of any restrictions on data availability

All raw sequencing data is available in the European Nucleotide Archive (ENA) with accession PRJEB32427.

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	"Sample size" was determined by the technical properties of DNA synthesis and sequencing technologies.
Data exclusions	Raw sequencing data was filtered and excluded from downstream analysis as described in the online methods section.
Replication	The large scale experiments were repeated twice. The encoded message was successfully decoded in both technical repeats.
Randomization	this is not relevant to the study as they were no participants
Blinding	this is not relevant to the study as they were no participants

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
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

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