Next-Generation Digital Information Storage in DNA

George M. Church,[1,2] Yuan Gao,[3] Sriram Kosuri[1,2]*

As digital information continues to accumulate, higher density and longer-term storage solutions are necessary (*1*). DNA has many potential advantages as a medium for immutable, high latency information storage needs (*2*). For example, DNA storage is very dense. At theoretical maximum, DNA can encode two bits per nucleotide (nt) or 455 exabytes per gram of single-stranded DNA (*3*). Unlike most digital storage media, DNA storage is not restricted to a planar layer and is often readable despite degradation in nonideal conditions over millennia (*4*, *5*). Lastly, DNA's essential biological role provides access to natural reading and writing enzymes and ensures that DNA will remain a readable standard for the foreseeable future.

Storing messages in DNA was first demonstrated in 1988 (*6*), and the largest project to date encoded 7920 bits (*7*). The small scale of previous work stems from the difficulty of writing and reading long perfect DNA sequences and has limited broader applications (table S1). We developed a strategy to encode arbitrary digital information by using a novel encoding scheme that uses next-generation DNA synthesis and sequencing technologies (fig. S1). We converted an html-coded draft of a book that included 53,426 words, 11 JPG images, and one JavaScript program into a 5.27-megabit bitstream (*3*). We then encoded these bits onto 54,898 159-nt oligonucleotides (oligos) each encoding a 96-bit data block (96 nt), a 19-bit address specifying the location of the data block in the bit stream (19 nt), and flanking 22-nt common sequences for amplification and sequencing. The oligo library was synthesized by ink-jet printed, high-fidelity DNA microchips (*8*). To read the encoded book, we amplified the library by limited-cycle polymerase chain reaction and then sequenced on a single lane of an Illumina HiSeq. We joined overlapping paired-end 100-nt reads to reduce the effect of sequencing error (*9*). Then with only reads that gave the expected 115-nt length and perfect barcode sequences, we generated consensus at each base of each data block at an average of ~3000-fold coverage (fig S2). All data blocks were recovered with a total of 10 bit errors out of 5.27 million (table S2), which were predominantly located within homopolymer runs at the end of the oligo, where we only had single sequence coverage (*3*).

Our method has at least five advantages over past DNA storage approaches. We encode one bit per base (A or C for zero, G or T for one), instead of two. This allows us to encode messages many ways in order to avoid sequences that are difficult to read or write such as extreme GC content, repeats, or secondary structure. By splitting the bit stream into addressed data blocks, we eliminate the need for long DNA constructs that are difficult to assemble at this scale. To avoid cloning and sequence verifying constructs, we synthesized, stored, and sequenced many copies of each individual oligo. Because errors in synthesis and sequencing are rarely coincident, each molecular copy corrects errors in the other copies. We used a purely in vitro approach that avoids cloning and stability issues of in vivo approaches. Lastly, we leveraged next-generation technologies in both DNA synthesis and sequencing to allow for encoding and decoding of large amounts of information for ~100,000-fold less cost than first-generation encodings.

The density (5.5 petabits/mm^3 at 100× synthetic coverage) and scale (5.27 megabits) of this work compare favorably to other experimental storage technologies while only using commercially available materials and instruments (Fig. 1 and table S3). DNA is particularly suitable for immutable, high-latency, sequential access applications such as archival storage. Density, stability, and energy efficiency are all potential advantages of DNA storage (*10*), although costs and times for writing and reading are currently impractical for all but century-scale archives (*3*). However, the costs of DNA synthesis and sequencing have been dropping at exponential rates of 5- and 12-fold per year, respectively—much faster than electronic media at 1.6-fold per year (*11*). Handheld, single-molecule DNA sequencers are becoming available and would vastly simplify reading DNA-encoded information (*12*). Our general approach of using addressed data blocks combined with library synthesis and consensus sequencing should be compatible with future DNA sequencing and synthesis technologies. Reciprocally, large-scale use of DNA such as for information storage could accelerate development of synthesis and sequencing technologies (*13*). Future work could use compression, redundant encodings, parity checks, and error correction to improve density, error rate, and safety. Other polymers or DNA modifications can also be considered to maximize reading, writing, and storage capabilities (*14*).

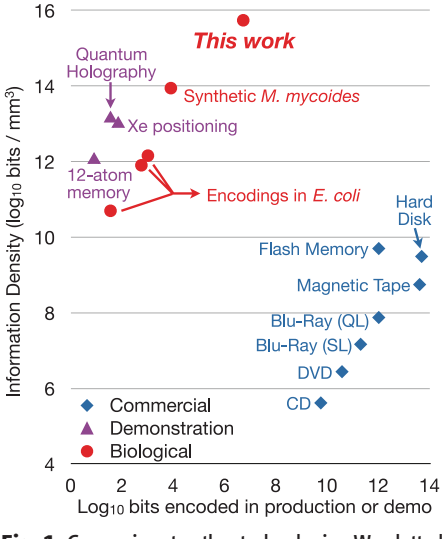

Fig. 1. Comparison to other technologies. We plotted information density ($\log_{10}$ of bits/mm^3) versus current scalability as measured by the $\log_{10}$ of bits encoded in the report or commercial unit (*3*).

References and Notes

1. J. Gantz, D. Reinsel, "Extracting value from chaos" [International Data Corporation (IDC), Framingham, MA, 2011], www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf.
2. C. Bancroft, T. Bowler, B. Bloom, C. T. Clelland, *Science* **293**, 1763 (2001).
3. Information on materials and methods is available on *Science* Online.
4. J. Bonnet *et al.*, *Nucleic Acids Res.* **38**, 1531 (2010).
5. S. Pääbo *et al.*, *Annu. Rev. Genet.* **38**, 645 (2004).
6. J. Davis, *Art J.* **55**, 70 (1996).
7. D. G. Gibson *et al.*, *Science* **329**, 52 (2010); 10.1126/science.1190719.
8. E. M. LeProust *et al.*, *Nucleic Acids Res.* **38**, 2522 (2010).
9. J. St. John, SeqPrep (2011), https://github.com/jstjohn/SeqPrep.
10. L. M. Adleman, *Science* **266**, 1021 (1994).
11. P. A. Carr, G. M. Church, *Nat. Biotechnol.* **27**, 1151 (2009).
12. E. Pennisi, *Science* **336**, 534 (2012).
13. S. Kosuri, A. M. Sismour, *ACS Synth. Biol.* **1**, 109 (2012).
14. S. A. Benner, Z. Yang, F. Chen, *C. R. Chim.* **14**, 372 (2011).

Acknowledgments: This work was supported by U.S. Office of Naval Research (N000141010144), Agilent Technologies, and the Wyss Institute. Agilent Technologies is a commercial provider for DNA microchips. G.M.C. and S.K. designed and performed all experiments and analyses and wrote the manuscript; Y.G. performed the sequencing. We thank J. Aach, C. Fracchia, S. Raman, H. H. Wang, A. W. Briggs, J. Lee, T. Wu, and D. B. Goodman for helpful suggestions on the manuscript.

Supplementary Materials

www.sciencemag.org/cgi/content/full/science.1226355/DC1
Materials and Methods
Supplementary Text
Figs. S1 and S2
Tables S1 to S3
References (*15*–*35*)

20 June 2012; accepted 7 August 2012
Published online 16 August 2012;
10.1126/science.1226355

[1]Department of Genetics, Harvard Medical School, Boston, MA 02115, USA. [2]Wyss Institute for Biologically Inspired Engineering, Boston, MA 02115, USA. [3]Department of Biomedical Engineering, Neuroregeneration and Stem Cell Biology Program, Institute for Cell Engineering, Lieber Institute for Brain Development, Johns Hopkins Medical Campus, Johns Hopkins University, Baltimore, MD 21205, USA.

*To whom correspondence should be addressed. E-mail: sri.kosuri@wyss.harvard.edu