Sciencexpress Brevia

Next-Generation Digital Information Storage in DNA

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

[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, Johns Hopkins University, Baltimore, MD 21205, USA.

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

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 ssDNA (*3*). Unlike most digital storage media, DNA storage is not restricted to a planar layer, and is often readable despite degradation in non-ideal conditions over millennia (*4*, *5*). Finally, 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). Here, we develop a strategy to encode arbitrary digital information using a novel encoding scheme that utilizes 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 1 JavaScript program into a 5.27 megabit bitstream (*3*). We then encoded these bits onto 54,898 159nt oligonucleotides (oligos) each encoding a 96-bit data block (96nt), a 19-bit address specifying the location of the data block in the bit stream (19nt), and flanking 22nt 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 PCR and then sequenced on a single lane of an Illumina HiSeq. We joined overlapping paired-end 100nt reads to reduce the effect of sequencing error (*9*). Then using 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 homo-polymer 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 synthesize, store, and sequence many copies of each individual oligo. Since errors in synthesis and sequencing are rarely coincident, each molecular copy corrects errors in the other copies. We use a purely in vitro approach that avoids cloning and stability issues of in vivo approaches. Finally, we leverage 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 100x 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*), while costs and times for writing and reading are currently impractical for all but century-scale archives (*3*). However, the cost 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*). Hand-held, 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*).

References and Notes

1. "Extracting Value from Chaos" (IDC, Framingham, MA, 2011), http://www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf.
2. C. Bancroft, T. Bowler, B. Bloom, C. T. Clelland, Long-term storage of information in DNA. *Science* **293**, 1763 (2001). doi:10.1126/science.293.5536.1763c Medline
3. Information on materials and methods is available on *Science* Online.
4. J. Bonnet *et al*., Chain and conformation stability of solid-state DNA: Implications for room temperature storage. *Nucleic Acids Res.* **38**, 1531 (2010). doi:10.1093/nar/gkp1060 Medline
5. S. Pääbo *et al*., Genetic analyses from ancient DNA. *Annu. Rev. Genet.* **38**, 645 (2004). doi:10.1146/annurev.genet.37.110801.143214 Medline
6. J. Davis, Microvenus. *Art J.* **55**, 70 (1996). doi:10.2307/777811
7. D. G. Gibson *et al*., Creation of a bacterial cell controlled by a chemically synthesized genome. *Science* **329**, 52 (2010). doi:10.1126/science.1190719 Medline
8. E. M. LeProust *et al*., Synthesis of high-quality libraries of long (150mer) oligonucleotides by a novel depurination controlled process. *Nucleic Acids Res.* **38**, 2522 (2010). doi:10.1093/nar/gkq163 Medline
9. J. St. John, *SeqPrep* https://github.com/jstjohn/SeqPrep (2011)
10. L. M. Adleman, Molecular computation of solutions to combinatorial problems. *Science* **266**, 1021 (1994). doi:10.1126/science.7973651 Medline
11. P. A. Carr, G. M. Church, Genome engineering. *Nat. Biotechnol.* **27**, 1151 (2009). doi:10.1038/nbt.1590 Medline
12. E. Pennisi, Search for pore-fection. *Science* **336**, 534 (2012). doi:10.1126/science.336.6081.534 Medline
13. S. Kosuri, A. M. Sismour, When it rains, it pores. *ACS Synth. Biol.* **1**, 109 (2012). doi:10.1021/sb300015f
14. S. A. Benner, Z. Yang, F. Chen, Synthetic biology, tinkering biology, and artificial biology. What are we learning? *C. R. Chim.* **14**, 372 (2011). doi:10.1016/j.crci.2010.06.013
15. B. Langmead, S. L. Salzberg, Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357 (2012). doi:10.1038/nmeth.1923 Medline
16. H. Li *et al*, The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078 (2009). doi:10.1093/bioinformatics/btp352 Medline
17. C. T. Clelland, V. Risca, C. Bancroft, Hiding messages in DNA microdots.

Nature **399**, 533 (1999). doi:10.1038/21092 Medline
18. P. Wong, K. Wong, H. Foote, Organic data memory using the DNA approach. *Commun. ACM* **46**, 95 (2003). doi:10.1145/602421.602426
19. C. Gustafsson, For anyone who ever said there's no such thing as a poetic gene. *Nature* **458**, 703 (2009). doi:10.1038/458703a
20. N. Yachie, K. Sekiyama, J. Sugahara, Y. Ohashi, M. Tomita, Alignment-based approach for durable data storage into living organisms. *Biotechnol. Prog.* **23**, 501 (2007). doi:10.1021/bp060261y Medline
21. N. G. Portney, Y. Wu, L. K. Quezada, S. Lonardi, M. Ozkan, Length-based encoding of binary data in DNA. *Langmuir The Acs Journal Of Surfaces And Colloids* **24**, 1613 (2008). doi:10.1021/la703235y Medline
22. M. Ailenberg, O. Rotstein, An improved Huffman coding method for archiving text, images, and music characters in DNA. *Biotechniques* **47**, 747 (2009). doi:10.2144/000113218 Medline
23. Ecma International, *Data interchange on read-only 120mm optical data disks (CD-ROM)*, (ECMA Standard 130, Geneva, Switzerland 1996, http://www.ecma-international.org/publications/files/ECMA-ST/Ecma-130.pdf).
24. Ecma International, *120 mm DVD - Read-Only Disk*, (ECMA Standard 267, Geneva, Switzerland 2001, http://www.ecma-international.org/publications/files/ECMA-ST/Ecma-267.pdf).
25. Blu-Ray Disc Association, *White Paper – Blu-Ray Disc Format* (2nd Edition, Universal City, CA 2010, http://www.blu-raydisc.com/Assets/Downloadablefile/general_bluraydiscformat-15263.pdf).
26. Oracle, *StorageTek T10000 Family Tape* Cartridge (Oracle, Redwood Shores, CA 2010, http://www.oracle.com/us/products/servers-storage/storage/tape-storage/033617.pdf).
27. SanDisk, *SanDisk Develops Smallest 128Gb NAND Flash Memory Chip* (SanDisk, Milipitas, CA 2012, http://www.sandisk.com/about-sandisk/press-room/press-releases/2012/sandisk-develops-worlds-smallest-128gb-nand-flash-memory-chip).
28. Toshiba, *NAND Flash Memory in Multi Chip Package* (Toshiba, Tokyo, Japan, 2011, http://www.toshiba-components.com/memory/mcp.html).
29. Seagate, *Seagate Reaches 1 Terabit Per Square Inch Milestone In Hard Drive Storage With New Technology Demonstration* (Seagate, Cupertino, CA, 2012, http://www.seagate.com/about/newsroom/press-releases/terabit-milestone-storage-seagate-master-pr/).
30. S. Loth, S. Baumann, C. P. Lutz, D. M. Eigler, A. J. Heinrich, Bistability in atomic-scale antiferromagnets. *Science* **335**, 196 (2012). doi:10.1126/science.1214131
31. D. M. Eigler, E. K. Schweizer, Positioning single atoms with a scanning tunnelling microscope. *Nature* **344**, 524 (1990). doi:10.1038/344524a0
32. C. R. Moon, L. S. Mattos, B. K. Foster, G. Zeltzer, H. C. Manoharan, Quantum holographic encoding in a two-dimensional electron gas. *Nat. Nanotechnol.* **4**, 167 (2009). doi:10.1038/nnano.2008.415 Medline
33. T. Grotjohann *et al.*, Diffraction-unlimited all-optical imaging and writing with a photochromic GFP. *Nature* **478**, 204 (2011). doi:10.1038/nature10497 Medline
34. H. E. Kubitschek, Cell volume increase in *Escherichia coli* after shifts to richer media. *J. Bacteriol.* **172**, 94 (1990). Medline
35. "Screening Framework Guidance for Providers of Synthetic Double-Stranded DNA" *Federal Registrar* **75**, 62820-62832 (2010) FR Doc No: 2010-25728.

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 07 August 2012
Published online 16 August 2012
10.1126/science.1226355

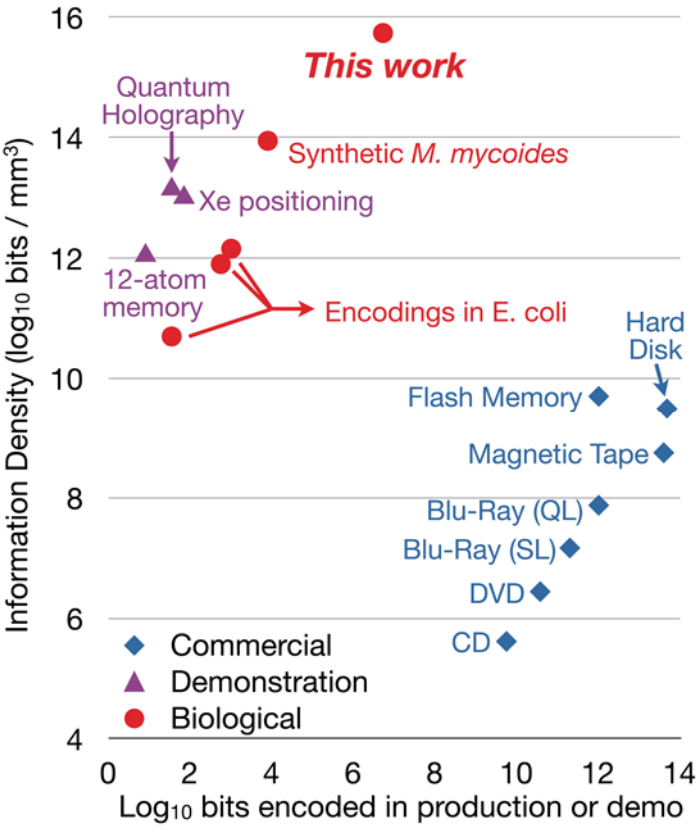

Fig. 1. Comparison to other measured by the $\log_{10}$ of bits encoded in the report or commercial technologies. We plotted information density ($\log_{10}$ of bits/mm^3) versus current scalability as unit (*3*).

www.sciencemag.org/cgi/content/full/science.1226355/DC1

Supplementary Materials for

Next-Generation Digital Information Storage in DNA

George M. Church, Yuan Gao, Sriram Kosuri*

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

Published16 August 2012 on *Science* Express
DOI: 10.1126/science.1226355

This PDF file includes:

Materials and Methods

Book Description

We wanted to go well beyond a limited encoding of uppercase text and considered several possible digital texts, including classics used in previous data storage milestones, like *A Tale of Two Cities*. We decided to use a draft HTML version of a book that one of us wrote called *Regenesis: How Synthetic Biology Will Reinvent Nature and Ourselves* (Church GM and Regis E) Basic Books (New York, NY; ISBN-13: 978-0465021758). We chose this in order to demonstrate modern formatting, images, and JavaScript. As with typical web pages, we used Universal Character Set Transformation Format, 8-bit (UTF-8), a variable-width encoding, which is backwards compatible with ASCII and UNICODE for special characters and fonts. There are 11 images that are black-and-white and JPEG encoded (typically a 10:1 data compression with little loss in quality). These are embedded "inline" (i.e. not separate files) in the html in base64 format. A consensus bit error in the middle of any of these JPEG segments would only affect data downstream within that segment. A bit error in the text will affect at most the 12 characters in that oligonucleotide containing the error. The JavaScript is a simple display of a 37-byte text string (mnemonic encoding of the genetic code) that can curve dynamically to follow the cursor position. The point being that DNA (like other digital media) can encode executable directives for digital machines.

Encoding into DNA

We encoded the book by converting the draft to html format (with embedded jpg images). Once made, we read the book in bit form and converted individual bits to A or C for 0 and T or G for 1. Bases were chosen randomly while disallowing homopolymer runs greater than three. Addresses of the bitstream were 19 bits long and numbered consecutively, starting from 0000000000000000001. The script *Bits2DNA.pl* (see code section) is the program we wrote for encoding the html file into DNA segments.

Synthesis and Amplification

We synthesized 54,898 oligonucleotides on Agilent's Oligo Library Synthesis microarray platform. DNA was eluted by Agilent to give an ~1 picomole pool of oligonucleotides in 100µL TE (10mM Tris-Cl pH 7.5, 0.1mM EDTA).

We amplified the libraries as follows. We used 1µL (~10 femtomole expected) of library in a 50µL PCR amplification reaction using 200nM each of primers MD-Test-1F and MD-Test-1R for 6 cycles using Sybr Fast Master Mix (Kapa Biosystems) in a BioRad CFX96 Real-Time PCR machine and monitored the Sybr Green channel during amplification.

1. 95˚C for 3 min
2. 95˚C for 10 sec
3. 60˚C for 30 sec
4. Read Sybr Green Channel
5. Goto Step 2 for a total of 10 cycles
6. 68˚C for 30 sec
7. Hold at 4˚C

The resulting PCR product was purified using Qiagen MinElute PCR cleanup column according to manufacturer's instructions into 10 µL of Buffer EB (10mM Tris-Cl, pH 8.5). The eluted DNA gave a concentration of 36.8 ng/µL (A260/A280 = 1.85) as measured by a NanoDrop 2000c spectrophotometer.

We then amplified two tubes of 1 µL of 1:11 diluted (in water) amplification reaction for nine cycles using the same conditions but this time using 200nM of PE-PCR Primer 1.0 – F and PE-PCR Primer 2.0 – R. PCR reactions were cleaned up using Ampure beads per manufacturer's suggestion (Agencourt) to remove residual primers and resuspended in 50µL of TE. The final product was ~22 ng/µL as quantified both through NanoDrop and agarose gel imaging.

Finally, we use 1% of the initial library for this experiment. From the resulting amplifications, we made enough template for millions of potential sequencing runs. However, if pulling such information from a repository occurs without replacement, using the protocols described herein, we could read the library ~100 disparate times.

Primers Used
** Denotes phosphorothioate linkage*
>MD Test 1–F
ACACTCTTTCCCTACACGACGCTCTTCCGATC*T
>MD Test 1–R
CTCGGCATTCCTGCTGAACCGCTCTTCCGATC*T
>PE PCR Primer 1.0 – F
AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATC*T
>PE PCR Primer 2.0 – R
CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGAT
C*T

Sequencing and Processing

We sequenced the amplified library by loading 1mL of 14pM library (14 fmoles; ~1:1000 of the amplified library) on a single lane of a HiSeq 2000 using paired end 100 reads. From the lane we got 346,151,426 million paired reads with 87.14% >= Q30 and mean Q score of 34.16. Since we were sequencing a 115bp construct with paired 100bp reads, we used SeqPrep (*9*) to combine overlapping reads into a single contig using the following command (for a single tile of the HiSeq lane):

```
SeqPrep -f MTMC2_NoIndex_L006_R1_002.fastq.gz -r MTMC2_NoIndex_L006_R2_002.fastq.gz -
1 tile2r1.fastq.gz -2 tile2r2.fastq.gz -s tile2-merged.fastq.gz -E tile2-align.txt.gz -o 50 -m 0.1 -n 0.8
```

After SeqPrep, 292,371,030 contigs were formed. We aligned the contigs to the reference using Bowtie2 version 2.0.0-beta5 (*15*) and SamTools version 0.1.18 (*16*) using the following command:

```
zcat *merged* | bowtie2 -p 10 --end-to-end -x ../../../agilentlib -U - | samtools view -bS -  > alltiles-
merged.bam
```

After alignment, 267,993,389 (92%) aligned to one member of the synthetic library giving average coverage of 4882 ± 1261 (±1 standard deviation). We then filtered SeqPrepped contigs that give the full-length 115bp contig, resulting in 190,284,472 reads and 3419 ± 998 average coverage. On average for each member of the library, ~69.5% ± 0.4 of the reads were of full length. The construct with the fewest number of reads was an oligo md-37545, which had 94 and 9 reads before and after 115bp filtering; the resulting consensus was still correct for this oligo.

We took all 115bp contigs from SeqPrep, and filtered for perfect barcode sequences, which resulted in 164,831,518 reads (~87% of all 115bp contigs). Because of our degenerate encoding scheme, choosing perfect barcoded bases has a 2^{19}-fold advantage over choosing perfect barcoded bits in terms of likelihood of confusing two barcodes. We sorted these reads, and counted unique sequences. Then for each barcode, we made a consensus sequence that determined the payload sequence using a custom python script (*buildConsensus.py*) to construct a consensus library that could be compared to the designed library.

Errors

From the consensus library we found discrepancies between designed and read sequences that are summarized in Supplementary Table 1. As shown, we found 22 discrepancies, 10 of which resulted in bit errors (bolded). Most of the errors (20/22) were located within the last 15 bases of the sequence where we only had single coverage during sequencing. In addition, most of the errors (18/22) resulted in runs of at least 3 consecutive repeated nucleotides. Screening out homopolymer reads of 4 or more repeated nucleotides (greyed boxes), which we purposefully avoided in the original design, would result in 12 discrepancies, 7 of which are bit errors.

Code

Bits2DNA.pl

```
# cd "\Perl\gmc\Bin_DNA"
# \Perl\bin\perl Bits2DNA.pl    GMC Jul-2011 & 27-May-2012
# docstore.mik.ua/orelly/perl/cookbook/ch02_05.htm (bin) ch01_05.htm (char)
# http://perldoc.perl.org/functions/pack.html  rand.html
# Each oligo is L(19)+8N(12)= 115 bp, long flanked by 22-mer amplification primers.
# DNA Encoded Artifacts Registry (DEAR) to coordinate global standards.

open IN,"in.html";  open OUT,">Bits2DNA.txt"; binmode IN;
$t{"0"}="a"; $t{"1"}="G"; # lowercase a,c = zero bit.
$t{"a"}="c"; $t{"G"}="T"; $t{"c"}="a"; $t{"T"}="G";
$u1=""; $u2=""; $u3="";  # Initialize; keep homopolymer runs < 4
$N=12; # Length of segment in bytes (not including segment number)
$L=19; # 2^19 = 524,288 = max number of oligos   L=00010011
$seed=2; srand($seed); # remove this line to get a random seed
print int2bp(262144)," ",int2bp(262145);
$f="CTACACGACGCTCTTCCGATCT";   # forward 'universal' sequencing & amplification
primer
$r="AGATCGGAAGAGCGGTTCAGCA";   # reverse 22-mer primer
```

```
$n=0; print OUT $f,int2bp(0),"";    ###
while (read (IN, $text, 65536)) {
  @ascii_num = unpack("C*", $text);
  foreach $val (@ascii_num) {
    print OUT byt2bp($val);           ###
    $n++;
    if($n%$N==0){
      print OUT $r,"\n",$f,int2bp($n/$N),""; ###
    }   # N bases per output line
  }   # each byte
}   # 65 Kbytes
for ($k=$n%$N; $k<$N; $k++){
  print OUT byt2bp(int(rand(256)));   ###
}   # pad last data line to keep all oligos same size.
  print OUT "$r\n";      ###

sub byt2bp {   # convert rightmost 8 bits (MSB first byte) to 8 bp
  my $b = unpack("B32", pack("N", shift));
  $p="";
  for ($i=24; $i<=31; $i++){
    $x=substr($b,$i,1); # bits 24 to 31 inclusive
    $u=$t{$x};
    if(rand(2)<1){$u=$t{$u};} # pick synonym a=c; G=T
    if(($u eq $u1) && ($u eq $u2) && ($u eq $u3)){$u=$t{$u};}
    $u1=$u2; $u2=$u3; $u3=$u; # Shift previous base string
    $p = $p.$u;
  }
  return $p;
}
sub int2bp { # convert rightmost $L bits of 32 bit integers to $L bp
  my $b = unpack("B32", pack("N", shift));
  $p="";
  for ($i=31; $i>=32-$L; $i--){
    $x=substr($b,$i,1); # bits 31 to $L
    $u=$t{$x};
    if(rand(2)<1){$u=$t{$u};} # pick synonym a=c; G=T
    if(($u eq $u1) && ($u eq $u2) && ($u eq $u3)){$u=$t{$u};}
    $u1=$u2; $u2=$u3; $u3=$u; # Shift previous base string
    $p = $p.$u;
  }
  return $p;
}
```

buildConsensus.py

```
import sys

#builds consensus sequence from individual base counts
def getConsensus(finalbuckets):
    sequence = ''
    for i in range(len(finalbuckets)):
        letterindex = finalbuckets[i].index(max(finalbuckets[i]))
        if letterindex == 0:
            sequence += 'A'
        elif letterindex == 1:
            sequence += 'C'
```

```
        elif letterindex == 2:
            sequence += 'G'
        elif letterindex == 3:
            sequence += 'T'
    return sequence


oligolength = 115
currentbarcode = ''
#initialize vector to building consensus
buckets = [[0 for col in range(4)] for row in range(oligolength)]

for line in sys.stdin:
    splitline = line.split()
    count = int(splitline[0])
    barcode = splitline[1]
    sequence = splitline[2]
    if not barcode == currentbarcode:
        if not currentbarcode == '':
            print getConsensus(buckets)

        buckets = [[0 for col in range(4)] for row in range(oligolength)]
        currentbarcode = barcode
    for i in range(oligolength):
        if sequence[i] == 'A':
            buckets[i][0] += count
        elif sequence[i] == 'C':
            buckets[i][1] += count
        elif sequence[i] == 'G':
            buckets[i][2] += count
        elif sequence[i] == 'T':
            buckets[i][3] += count

#print final consensus
print getConsensus(buckets)
```

Supplementary Text

Previous DNA Storage Works

There have been many previous implementations of DNA information storage. We list previous attempts where information (digital or otherwise) was stored in DNA and retrieved in table S1. We did not include other work on associative memory and DNA computing because they are designed for different purposes for DNA information storage needs, and are all at much lower scales than what is found in our listing.

Calculations on Data Density

Theoretical DNA density was calculated by using 2 bits per nucleotide of single stranded DNA. The molecular weight of DNA we used was based on an average of 330.95 g/mol/nucleotide of anhydrous weight for the sodium salt of an ATGC balanced

library. This results in a weight density of 1 bit per 2.75 x 10^{-22} g, and thus 4.5 x 10^{20} bytes per gram. Of course, practical maximums would be several orders of magnitude less dense depending the types of redundancy, barcoding, and encoding schemes desired. This theoretical maximum calculation is not used in Figure 1B.

For data plotted on fig. 1, we made best faith efforts to do comparisons between very different technologies. In cases of planar density calculations where thickness was not reported, we chose 100µm as depth (this is ~10x smaller than a hard drive platter, and 33% smaller than current Flash memory stacking). For our work, we assumed current information encoding density (96 bits per 159bp), and 100x synthesized coverage of the DNA in storage. We also assumed an approximate volume of 1 g/cm^3, the density of pure water, which is probably a slight underestimate for dry DNA. See Table S2 for more detailed information on individual calculations.

Large-scale Storage Considerations

At some point, storing DNA as a single large mass with extremely large barcodes is both unrealistic and cumbersome no matter the future sequencing and synthesis technologies. To understand where this trade-off lies, we hypothetically imagine such a data store without constraining ourselves on sequencing/synthesis costs.

A larger 48-bit address (2.8e14 unique addresses) block with a 128bit data block would require 216nt length oligo synthesis, which is already available. Such a scheme would give 1.85e-19 g/bit, which would give 1.48 mg of DNA for storage of 1 petabyte of information (at 100x coverage). This DNA, stored in 1536 well plate, would give ~1.5 exabytes with dimensions of 128mm x 86mm x 13mm.

Reading each well (petabyte) would require ~1 exabase of sequencing, or 1.8e6 HiSeq runs (600e9 bases per run). Thus, we would need ~6 orders of magnitude improvement in sequencing technologies for routinely reading petabytes of DNA information. For synthesis, current Agilent arrays top out at 1e6 features (we need 6.25e13 features for a petabyte); so ~7-8 orders of magnitude improvement in synthesis technologies is required. However, reading and writing costs do not dominate in very long-term storage applications (e.g., century or longer archival storage), and could currently be cost competitive due to the expected lower maintenance costs.

Comment on Safety

We consider a few points of our DNA information scheme with respect to safety:

(A) DNA could be a form of cryptography, and hence includes legal restrictions and import/export regulations. The DNA could encode computer viruses (we've embedded a mouse tracking program to symbolize a spyware threat) or the computer code could contain human viruses (see item C below).

(B) NIH Guidelines: "If the synthetic DNA segment is not expressed in vivo as a biologically active polynucleotide or polypeptide product, it is exempt from the NIH Guidelines." These 159 bp long fragments are unlikely to replicate on their own or encode anything biologically active (longest ORF is 36 amino acids). However, if placed in the wild, they could get incorporated into a living organism. EPA, FDA, OSHA, USDA, DHS guidelines might then apply.

(C) Certain apparently innocuous digital documents (e.g. images) once converted to DNA by the methods described here could result in infectious DNAs, so rules governing DNA surveillance should be applied *(35)*. This book's html, when converted to DNA form, could (but does not) encode parts of one or more "select agents", which would have set off alarms at the facility manufacturing the DNA (unless prior explicit justification is on file). However, the current guidelines are only recommendations and only apply to DNA longer than 200 base pairs.

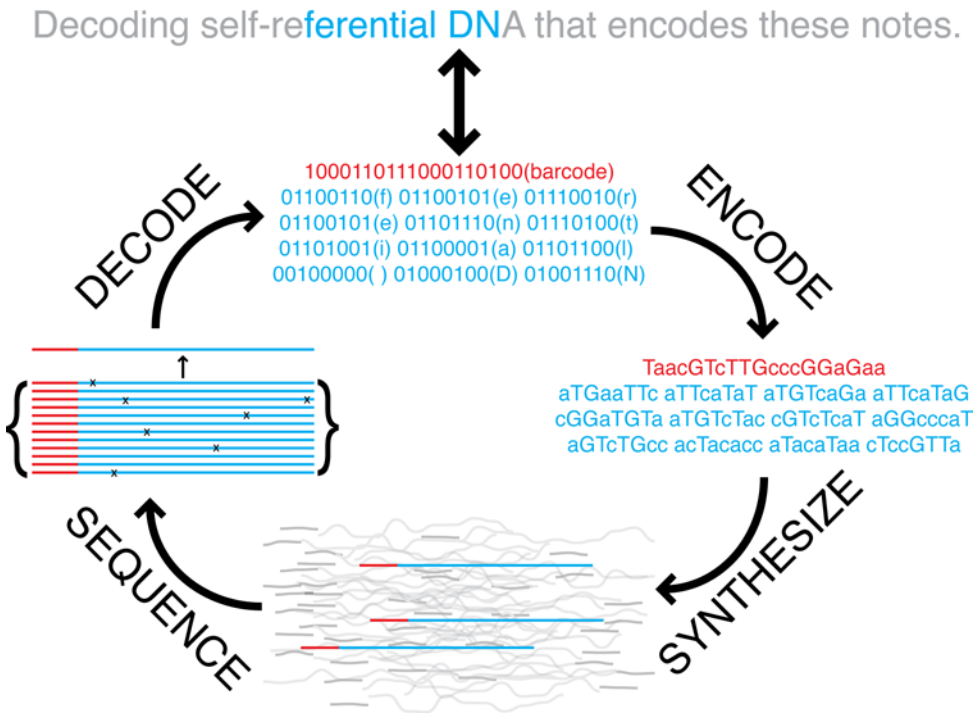

Fig. S1. Schematic of DNA information storage.

A 12-byte portion of a sentence within the encoded html book is converted to bits (blue) with a 19-bit barcode (red) that determines the location of the encoded bits within the overall book. The bit sequence is then encoded to DNA using a 1 bit per base encoding (a,c = 0; T,G = 1), while also avoiding 4 or more nucleotide repeats and balancing GC content. The entire 5.27 megabit html book used 54,898 oligonucleotides and was synthesized and eluted from a DNA microchip. After amplification (common primer sequences to all oligonucleotides are not shown), the oligonucleotide library was sequenced using next-generation sequencing. Individual reads with the correct barcode and length were screened for consensus, and then reconverted to bits obtaining the original book. In total, the writing, amplification, and reading resulted in 10 bit errors out of 5.27 megabits.

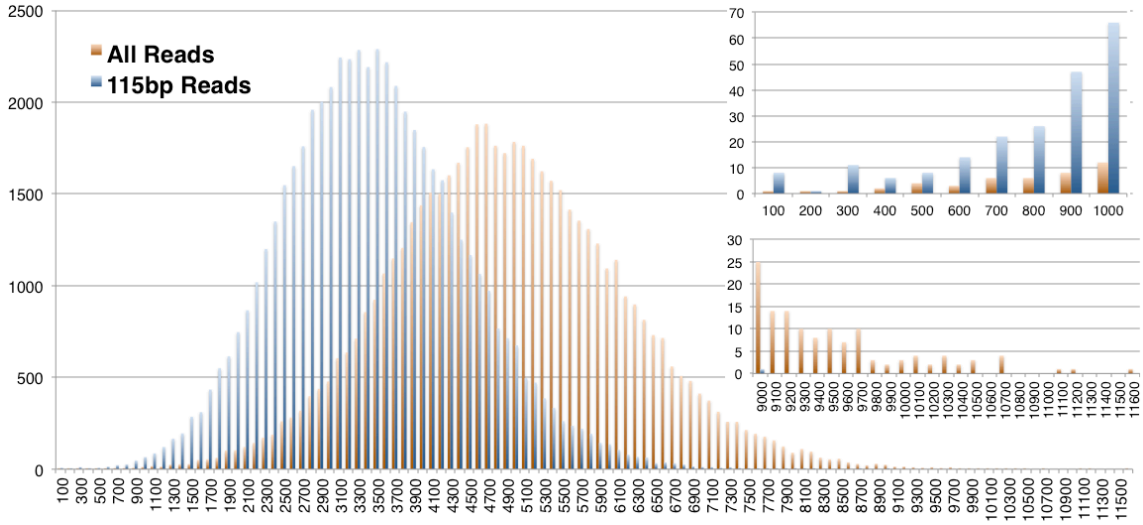

Fig. S2. Histogram of number of observations for each oligo of the designed library

All reads that formed contigs from SeqPrep (i.e., had overlaps between reads) were aligned against the synthesized library using Bowtie2, binned and plotted (red). The same information is displayed in green for only contigs 115 bp in length. Insets show zoomed in views of the distribution tails.

Table S1. Previous DNA Storage works. Bolded lines are included in Figure 1 as representative sequences. Other DNA information works are included here for reference. We do not include works related to DNA computation or associative memory where bits are encoded for purposes other than storage considerations (these works usually encode only modest amounts of information).

Year	Bits Encoded	Ref	Storage Mechanism	Described Usage	Notes
1988	**35**	**(*6*)**	**Plasmid/E. coli**	**Art**	**Encoded image; Microvenus**
1998	718	see note	Plasmid/E. coli	Art	Text from Bible (Genesis) http://www.ekac.org/geninfo.html
1999	138	(*17*)	DNA Microdot	Encrypted Message	"JUNE 6 INVASION:NORMANDY"
2001	**561**	**(*2*)**	**Plasmid/E.coli**	**Archival Storage**	**Lines from Dickens**
2003	1106	(*18*)	Plasmid/E.coli	Archival Storage	Parts of "It's a Small World"
2005	**1007**	**(*19*)**	**Plasmid/E.coli**	**Art**	**Encoded poem "Tomten"**
2007	120	(*20*)	E.coli genome	Archival Storage	"E=mc^2 1905!"; use multiple encodings to correct errors
2008	12	(*21*)	plasmid	Archival Storage	Use restriction fragment length to encode data
2009	1688	(*22*)	Plasmid/E.coli	Archival Storage	Text/Music/Image
2010	**7920**	**(*7*)**	**Mycoplasma genome**	**Watermark**	**Watermarking of synthetic genome**

Table S2. Data Density Calculations Used in Figure 1

In order to compare vastly different technologies for information encoding, we converted all data density information into volumetric data densities by making various assumptions. For commercial technologies, we used information about thickness where available. In the case of flash memory, we combined best in class data density with chip-stacking thickness from different manufacturers. For demonstration data storage technologies, substrate thicknesses were not reported. Therefore we assumed 100µm thickness, which is 1/3rd the current thickness of stacked flash storage technologies. We approximated the density of dried DNA to water's density. For other biological demonstrations using cloned DNA, we used volumes of individual cells as volume. Finally, greyed rows are not shown in Figure 1 as they were obscured by other data points, but are included here for completeness.

Type	Label	Date	Ref	bits	bits/mm^3	Comments
Commercial	CD	1982	*(23)*	5.6e9	4.13e5	1.2 mm thick CD; 120mm diameter
Commercial	DVD-SL	1996	*(24)*	3.76e10	2.77e6	1.2 mm thick DVD-SS-SL; 120mm diameter
Commercial	DVD-QL	2000	*(24)*	1.37e11	1.01e7	1.2 mm thick DVD-DS-DL; 120mm diameter
Commercial	Blu-Ray (SL)	2002	*(25)*	2.00e11	1.47e7	1.2 mm thick Blu-ray disk (1 layer)
Commercial	Blu-Ray (QL)	2010	*(25)*	1.02e12	7.52e7	1.2 mm thick Blu-ray disk (XL 4 layer)
Commercial	Magnetic Tape	2010	*(26)*	4.00e13	5.59e8	Oracle StorageTek T10000 T2 - 5TB 5.2µm thickness, 1147m length, ~12 mm wide
Commercial	Flash Memory	2012	*(27,28)*	1.02e12	5.02e9	NAND Flash; Sandisk for density of a single chip *(22)* 128Gbits in 170mm2; 150µm depth taken from Toshiba chip stacking *(23)*
Commercial	Hard Disk	2012	*(29)*	*4.80e13*	3.10e9	Hard Drive -> Seagate 1 Terabit /inch2 = 1.55e9 bits /mm2 =1.55e9 bits /mm2 assuming ds1mm platter
Demonstration	12-atom memory	2012	*(30)*	8	1.11e12	9 nm^2 / bit (assuming *100 µm thickness*) low temperature non-volatile memory
Demonstration	Xe positioning	*1991*	*(31)*	*70*	*1E13*	Spelled IBM with Xe atoms spaced 1nm apart on a 14 x 5 nm^2 lattice; 1 bit/nm^2; assuming 100µm thickness
Demonstration	Quantum Holography	2008	*(32)*	3.5E+1	1.38e13	35 bit image pair, 17x17 nm2 overhead atoms and 4x5 read space = ((4x5)/(17x17)) * 20bits/nm2 =1.38 bits/nm2 = 1.38e12 bits/mm2; assuming 100µm thickness
Biological	Super-resolution GFP	2011	*(33)*	27	4.0E10	9 3x3 bit fields (81 bits); 250nm center-to-center spacing; 1 bit/250nm^2; assuming 100µm thickness
Biological	DNA in E. coli	1988	*(6,34)*	35	5.0e10	E.coli, 0.7 µm^3 from *(34)*
Biological	DNA in E. coli	2001	*(2,34)*	561	8.01e11	E. coli 0.7 µm^3 – 118 characters (27 possibilities) = 27^{118} = 2^x; x = 561
Biological	DNA in E. coli	2005	*(19,34)*	1007	1.44e12	E.coli, 0.7 µm^3 – 233 characters (20 possibilities) = 20^{233} = 2^x; x=1007
Biological	Mycoplasma	2010	(*7*)	7920	8.80E+13	Mycoplasma, volume of ~0.09 µm^3;
Biological	This Work	2012		5.27e6	5.49e15	Assuming 1e-3g/mm^3; 330.95 g/mol/nucleotide; 96 bits per 159bp; 100x fold coverage; 330.95*2*159 = 105242.1 g/mol = 1.748e-19 g/molecule = 1.748e-16 g per 100 molecules = 96 bits / 1.748e-13 mm3

Table S3. Discrepancies between designed and read library.

Each error is one row, displaying the barcode the error is associated with, the position in the oligo (out of 115), the error type, whether or not the error resulted in a bit change, the original context, and the new context (error position is in the middle of dashes), and finally whether or not the error resulted in a run of 4 bases that could have been filtered out (but we did not to be conservative). Lines that resulted in bit errors are bolded, and lines that could have been filtered based on runs of 4 consecutive bases are shaded.

Barcode	Error Position	Error Type	Bit Error	Reference Context	Read Context	Homo-polymer
AACTGTCGCTATTCACTCA	**115**	**A->G**	**yes**	**CAC-A-**	**CAC-G-**	**no**
ACTAACGCACCTGGAATCA	106	A->C	no	CTT-A-CCT	CTT-C-CCT	no
ACTTTCGGCATGAGTACCC	103	A->C	no	CGC-A-CCC	CGC-C-CCC	yes
AGCGGCTCGTCGGTGTCCC	**40**	**T->C**	**yes**	**GGT-T-CCG**	**GGT-C-CCG**	**no**
ATACGGCTCATTACAAACC	**105**	**G->C**	**yes**	**TCT-G-CCC**	**TCT-C-CCC**	**yes**
ATGCGGGCAAATCACAGCA	106	A->C	no	AAC-A-CCT	AAC-C-CCT	yes
ATGGCCGTAATGGAGAAAC	102	C->A	no	TAG-C-AAG	TAG-A-AAG	no
ATGTTCTGAATTAGCGCCC	**108**	**C->G**	**yes**	**CAA-C-GAG**	**CAA-G-GAG**	**no**
CAATGTAGATCCTCGAAAC	106	A->C	no	CAG-A-CCC	CAG-C-CCC	yes
CCGGCCTAAACGGCACGCC	106	A->C	no	CTC-A-CCT	CTC-C-CCT	yes
CGATATTCGGGAACACCCA	**102**	**G->C**	**yes**	**AAC-G-CCC**	**AAC-C-CCC**	**yes**
CGATATTCGGGAACACCCA	106	A->C	no	CCC-A-CCT	CCC-C-CCT	yes
CGGCGGAGCGGAGACGCCA	106	C->A	no	AGG-C-AAG	AGG-A-AAG	no
CTGCTCTTCAACCGCTACA	115	T->G	no	CGA-T-	CGA-G-	no
GGTAATTTCTAGTACAGCC	105	A->C	no	GCA-A-CCC	GCA-C-CCC	yes
GGTCGCATAAACTTGACCC	**105**	**A->G**	**yes**	**CGC-A-CGA**	**CGC-G-GGA**	**no**
GGTCGCATAAACTTGACCC	**106**	**C->G**	**yes**	**GCA-C-GAG**	**GCG-G-GAG**	**no**
GTGCCAATAAAGTGGTCCC	**102**	**T->C**	**yes**	**TCG-T-CCG**	**TCG-C-CCG**	**no**
GTGCCAATAAAGTGGTCCC	106	C->A	no	CCG-C-AAG	CCG-A-AAG	no
GTGTCCCACCCACCCACCC	**83**	**A->G**	**yes**	**ACA-A-CTG**	**ACA-G-CTG**	**no**
TCCCAGGCAGCTACCCGCA	**102**	**T->C**	**yes**	**GCG-T-CCC**	**GCG-C-CCC**	**yes**
TGACGCGCCGGTTGGGCCC	106	A->C	no	ACC-A-CCT	ACC-C-CCT	yes

References and Notes

1. “Extracting Value from Chaos” (IDC, Framingham, MA, 2011), http://www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf.

2. C. Bancroft, T. Bowler, B. Bloom, C. T. Clelland, Long-term storage of information in DNA. *Science* **293**, 1763 (2001). doi:10.1126/science.293.5536.1763c Medline

3. Information on materials and methods is available on *Science* Online.

4. J. Bonnet *et al*., Chain and conformation stability of solid-state DNA: Implications for room temperature storage. *Nucleic Acids Res.* **38**, 1531 (2010). doi:10.1093/nar/gkp1060 Medline

5. S. Pääbo *et al*., Genetic analyses from ancient DNA. *Annu. Rev. Genet.* **38**, 645 (2004). doi:10.1146/annurev.genet.37.110801.143214 Medline

6. J. Davis, Microvenus. *Art J.* **55**, 70 (1996). doi:10.2307/777811

7. D. G. Gibson *et al*., Creation of a bacterial cell controlled by a chemically synthesized genome. *Science* **329**, 52 (2010). doi:10.1126/science.1190719 Medline

8. E. M. LeProust *et al*., Synthesis of high-quality libraries of long (150mer) oligonucleotides by a novel depurination controlled process. *Nucleic Acids Res.* **38**, 2522 (2010). doi:10.1093/nar/gkq163 Medline

9. J. St. John, *SeqPrep* https://github.com/jstjohn/SeqPrep (2011)

10. L. M. Adleman, Molecular computation of solutions to combinatorial problems. *Science* **266**, 1021 (1994). doi:10.1126/science.7973651 Medline

11. P. A. Carr, G. M. Church, Genome engineering. *Nat. Biotechnol.* **27**, 1151 (2009). doi:10.1038/nbt.1590 Medline

12. E. Pennisi, Search for pore-fection. *Science* **336**, 534 (2012). doi:10.1126/science.336.6081.534 Medline

13. S. Kosuri, A. M. Sismour, When it rains, it pores. *ACS Synth. Biol.* **1**, 109 (2012). doi:10.1021/sb300015f

14. S. A. Benner, Z. Yang, F. Chen, Synthetic biology, tinkering biology, and artificial biology. What are we learning? *C. R. Chim.* **14**, 372 (2011). doi:10.1016/j.crci.2010.06.013

15. B. Langmead, S. L. Salzberg, Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357 (2012). doi:10.1038/nmeth.1923 Medline

16. H. Li *et al*, The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078 (2009). doi:10.1093/bioinformatics/btp352 Medline

17. C. T. Clelland, V. Risca, C. Bancroft, Hiding messages in DNA microdots. *Nature* **399**, 533 (1999). doi:10.1038/21092 Medline

18. P. Wong, K. Wong, H. Foote, Organic data memory using the DNA approach. *Commun. ACM* **46**, 95 (2003). doi:10.1145/602421.602426

19. C. Gustafsson, For anyone who ever said there's no such thing as a poetic gene. *Nature* **458**, 703 (2009). doi:10.1038/458703a

20. N. Yachie, K. Sekiyama, J. Sugahara, Y. Ohashi, M. Tomita, Alignment-based approach for durable data storage into living organisms. *Biotechnol. Prog.* **23**, 501 (2007). doi:10.1021/bp060261y Medline

21. N. G. Portney, Y. Wu, L. K. Quezada, S. Lonardi, M. Ozkan, Length-based encoding of binary data in DNA. *Langmuir The Acs Journal Of Surfaces And Colloids* **24**, 1613 (2008). doi:10.1021/la703235y Medline

22. M. Ailenberg, O. Rotstein, An improved Huffman coding method for archiving text, images, and music characters in DNA. *Biotechniques* **47**, 747 (2009). doi:10.2144/000113218 Medline

23. Ecma International, *Data interchange on read-only 120mm optical data disks (CD-ROM)*, (ECMA Standard 130, Geneva, Switzerland 1996, http://www.ecma-international.org/publications/files/ECMA-ST/Ecma-130.pdf).

24. Ecma International, *120 mm DVD - Read-Only Disk*, (ECMA Standard 267, Geneva, Switzerland 2001, http://www.ecma-international.org/publications/files/ECMA-ST/Ecma-267.pdf).

25. Blu-Ray Disc Association, *White Paper – Blu-Ray Disc Format* (2nd Edition, Universal City, CA 2010, http://www.blu-raydisc.com/Assets/Downloadablefile/general_bluraydiscformat-15263.pdf).

26. Oracle, *StorageTek T10000 Family Tape* Cartridge (Oracle, Redwood Shores, CA 2010, http://www.oracle.com/us/products/servers-storage/storage/tape-storage/033617.pdf).

27. SanDisk, *SanDisk Develops Smallest 128Gb NAND Flash Memory Chip* (SanDisk, Milipitas, CA 2012, http://www.sandisk.com/about-sandisk/press-room/press-releases/2012/sandisk-develops-worlds-smallest-128gb-nand-flash-memory-chip).

28. Toshiba, *NAND Flash Memory in Multi Chip Package* (Toshiba, Tokyo, Japan, 2011, http://www.toshiba-components.com/memory/mcp.html).

29. Seagate, *Seagate Reaches 1 Terabit Per Square Inch Milestone In Hard Drive Storage With New Technology Demonstration* (Seagate, Cupertino, CA, 2012, http://www.seagate.com/about/newsroom/press-releases/terabit-milestone-storage-seagate-master-pr/).

30. S. Loth, S. Baumann, C. P. Lutz, D. M. Eigler, A. J. Heinrich, Bistability in atomic-scale antiferromagnets. *Science* **335**, 196 (2012). doi:10.1126/science.1214131

31. D. M. Eigler, E. K. Schweizer, Positioning single atoms with a scanning tunnelling microscope. *Nature* **344**, 524 (1990). doi:10.1038/344524a0

32. C. R. Moon, L. S. Mattos, B. K. Foster, G. Zeltzer, H. C. Manoharan, Quantum holographic encoding in a two-dimensional electron gas. *Nat. Nanotechnol.* **4**, 167 (2009). doi:10.1038/nnano.2008.415 Medline

33. T. Grotjohann *et al*., Diffraction-unlimited all-optical imaging and writing with a photochromic GFP. *Nature* **478**, 204 (2011). doi:10.1038/nature10497 Medline

34. H. E. Kubitschek, Cell volume increase in *Escherichia coli* after shifts to richer media. *J. Bacteriol.* **172**, 94 (1990). Medline

35. "Screening Framework Guidance for Providers of Synthetic Double-Stranded DNA" *Federal Registrar* **75**, 62820-62832 (2010) FR Doc No: 2010-25728.