LETTER

doi:10.1038/nature11875

Towards practical, high-capacity, low-maintenance information storage in synthesized DNA

Nick Goldman[1], Paul Bertone[1], Siyuan Chen[2], Christophe Dessimoz[1], Emily M. LeProust[2], Botond Sipos[1] & Ewan Birney[1]

Digital production, transmission and storage have revolutionized how we access and use information but have also made archiving an increasingly complex task that requires active, continuing maintenance of digital media. This challenge has focused some interest on DNA as an attractive target for information storage[1] because of its capacity for high-density information encoding, longevity under easily achieved conditions[2–4] and proven track record as an information bearer. Previous DNA-based information storage approaches have encoded only trivial amounts of information[5–7] or were not amenable to scaling-up[8], and used no robust error-correction and lacked examination of their cost-efficiency for large-scale information archival[9]. Here we describe a scalable method that can reliably store more information than has been handled before. We encoded computer files totalling 739 kilobytes of hard-disk storage and with an estimated Shannon information[10] of 5.2×10^6 bits into a DNA code, synthesized this DNA, sequenced it and reconstructed the original files with 100% accuracy. Theoretical analysis indicates that our DNA-based storage scheme could be scaled far beyond current global information volumes and offers a realistic technology for large-scale, long-term and infrequently accessed digital archiving. In fact, current trends in technological advances are reducing DNA synthesis costs at a pace that should make our scheme cost-effective for sub-50-year archiving within a decade.

Although techniques for manipulating, storing and copying large amounts of existing DNA have been established for many years[11–13], one of the main challenges for practical DNA-based information storage is the difficulty of synthesizing long sequences of DNA *de novo* to an exactly specified design. As in the approach of ref. 9, we represent the information being stored as a hypothetical long DNA molecule and encode this *in vitro* using shorter DNA fragments. This offers the benefits that isolated DNA fragments are easily manipulated *in vitro*[11,13], and that the routine recovery of intact fragments from samples that are tens of thousands of years old[14,15] indicates that well-prepared synthetic DNA should have an exceptionally long lifespan in low-maintenance environments[3,4]. In contrast, approaches using living vectors[6–8] are not as reliable, scalable or cost-efficient owing to disadvantages such as constraints on the genomic elements and locations that can be manipulated without affecting viability, the fact that mutation will cause the fidelity of stored and decoded information to reduce over time, and possibly the requirement for storage conditions to be carefully regulated. Existing schemes used for DNA computing in principle permit large-scale memory[1,16], but data encoding in DNA computing is inextricably linked to the specific application or algorithm[17] and no practical storage schemes have been realized.

As a proof of concept for practical DNA-based storage, we selected and encoded a range of common computer file formats to emphasize the ability to store arbitrary digital information. The five files comprised all 154 of Shakespeare's sonnets (ASCII text), a classic scientific paper[18] (PDF format), a medium-resolution colour photograph of the European Bioinformatics Institute (JPEG 2000 format), a 26-s excerpt from Martin Luther King's 1963 'I have a dream' speech (MP3 format) and a Huffman code[10] used in this study to convert bytes to base-3 digits (ASCII text), giving a total of 757,051 bytes or a Shannon information[10] of 5.2×10^6 bits (see Supplementary Information and Supplementary Table 1 for full details).

The bytes comprising each file were represented as single DNA sequences with no homopolymers (runs of $\geq$2 identical bases, which are associated with higher error rates in existing high-throughput sequencing technologies[19] and led to errors in a recent DNA-storage experiment[9]). Each DNA sequence was split into overlapping segments, generating fourfold redundancy, and alternate segments were converted to their reverse complement (see Fig. 1 and Supplementary Information). These measures reduce the probability of systematic failure for any particular string, which could lead to uncorrectable errors and data loss. Each segment was then augmented with indexing information that permitted determination of the file from which it originated and its location within that file, and simple parity-check error-detection[10]. In all, the five files were represented by a total of 153,335 strings of DNA, each comprising 117 nucleotides (nt). The perfectly uniform fragment lengths and absence of homopolymers make it obvious that the synthesized DNA does not have a natural (biological) origin, and so imply the presence of deliberate design and encoded information[2].

We synthesized oligonucleotides (oligos) corresponding to our designed DNA strings using an updated version of Agilent Technologies' OLS (oligo library synthesis) process[20], creating $\sim 1.2 \times 10^7$ copies of each DNA string. Errors occur only rarely ($\sim$1 error per 500 bases) and independently in the different copies of each string, again enhancing our method's error tolerance. We shipped the synthesized DNA in lyophilized form that is expected to have excellent long-term preservation characteristics[3,4], at ambient temperature and without specialized packaging, from the USA to Germany via the UK. After resuspension, amplification and purification, we sequenced a sample of the resulting library products at the EMBL Genomics Core Facility in paired-end mode on the Illumina HiSeq 2000. We transferred the remainder of the library to multiple aliquots and re-lyophilized these for long-term storage.

Our base calling using AYB[21] yielded 79.6×10^6 read-pairs of 104 bases in length, from which we reconstructed full-length (117-nt) DNA strings *in silico*. Strings with uncertainties due to synthesis or sequencing errors were discarded and the remainder decoded using the reverse of the encoding procedure, with the error-detection bases and properties of the coding scheme allowing us to discard further strings containing errors. Although many discarded strings will have contained information that could have been recovered with more sophisticated decoding, the high level of redundancy and sequencing coverage rendered this unnecessary in our experiment. Full-length DNA sequences representing the original encoded files were then reconstructed *in silico*. The decoding process used no additional information derived from knowledge of the experimental design. Full details of the encoding, sequencing and decoding processes are given in Supplementary Information.

Four of the five resulting DNA sequences could be fully decoded without intervention. The fifth however contained two gaps, each a run

[1]European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton CB10 1SD, UK. [2]Agilent Technologies, Genomics–LSSU, 5301 Stevens Creek Boulevard, Santa Clara, California 95051, USA.

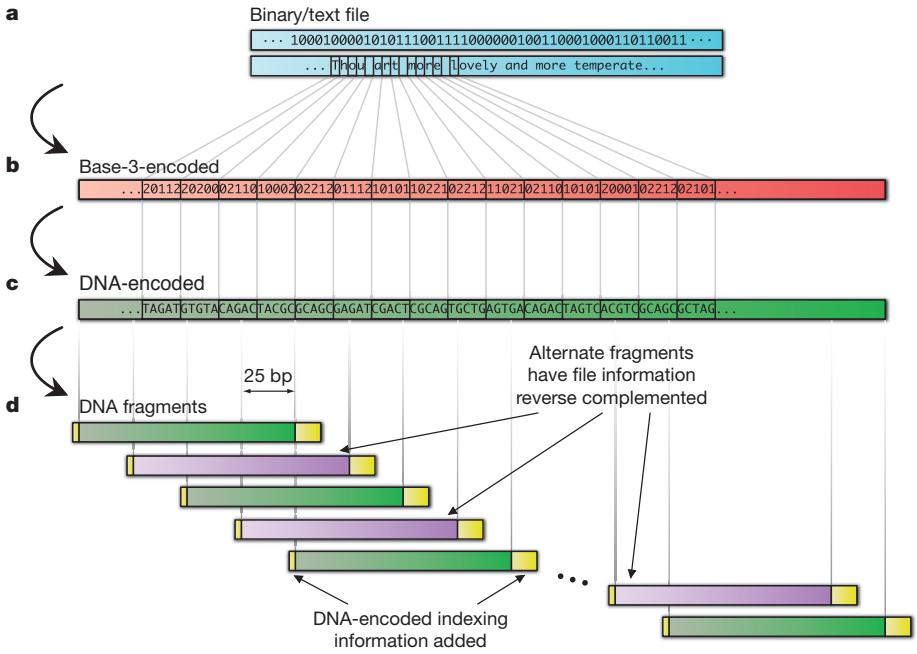

Figure 1 | Digital information encoding in DNA. Digital information (**a**, in blue), here binary digits holding the ASCII codes for part of Shakespeare's sonnet 18, was converted to base-3 (**b**, red) using a Huffman code that replaces each byte with five or six base-3 digits (trits). This in turn was converted *in silico* to our DNA code (**c**, green) by replacement of each trit with one of the three nucleotides different from the previous one used, ensuring no homopolymers were generated. This formed the basis for a large number of overlapping segments of length 100 bases with overlap of 75 bases, creating fourfold redundancy (**d**, green and, with alternate segments reverse complemented for added data security, violet). Indexing DNA codes were added (yellow), also encoded as non-repeating DNA nucleotides. See Supplementary Information for further details.

of 25 bases, for which no segment was detected corresponding to the original DNA. Each of these gaps was caused by the failure to sequence any oligo representing any of four consecutive overlapping segments. Inspection of the neighbouring regions of the reconstructed sequence permitted us to hypothesize what the missing nucleotides should have been (see Supplementary Information) and we manually inserted those 50 bases accordingly. This sequence could also then be decoded. Inspection confirmed that our original computer files had been reconstructed with 100% accuracy.

An important issue for long-term digital archiving is how DNA-based storage scales to larger applications. The number of bases of synthesized DNA needed to encode information grows linearly with the amount of information to be stored, but we must also consider the indexing information required to reconstruct full-length files from short fragments. As indexing information grows only as the logarithm of the number of fragments to be indexed, the total amount of synthesized DNA required grows sub-linearly. Increasingly large parts of each fragment are needed for indexing however and, although it is reasonable to expect synthesis of longer strings to be possible in future, we modelled the behaviour of our scheme under the conservative constraint of a constant 114 nt available for both data and indexing information (see Supplementary Information). As the total amount of information increases, the encoding efficiency decreases only slowly (Fig. 2a). In our experiment (megabyte scale) the encoding scheme is 88% efficient; Fig. 2a indicates that efficiency remains >70% for data storage on petabyte (PB, 10^{15} bytes) scales and >65% on exabyte (EB, 10^{18} bytes) scales, and that DNA-based storage remains feasible on scales many orders of magnitude greater than current global data volumes[22]. Figure 2a also shows that costs (per unit information stored) rise only slowly as data volumes increase over many orders of magnitude. Efficiency and costs scale even more favourably if we consider the synthesized fragment lengths available using the latest technology (Supplementary Fig. 5).

As the amount of information stored increases, decoding requires more strings to be sequenced. A fixed decoding expenditure per byte of encoded information would mean that each base is read fewer times and so is more likely to suffer decoding error. But extension of our scaling analysis to model the influence of reduced sequencing coverage on the per-decoded-base error rate (see Supplementary Information) revealed that error rates increase only very slowly as the amount of information encoded increases to a global data scale and beyond (Supplementary Table 4). This also suggests that our mean sequencing coverage of 1,308 times was considerably in excess of that needed for reliable decoding. We confirmed this by subsampling from the 79.6×10^6 read-pairs to simulate experiments with lower coverage. Figure 2b indicates that reducing the coverage by a factor of 10 (or even more) would have led to unaltered decoding characteristics, which further illustrates the robustness of our DNA-storage method.

DNA-based storage might already be economically viable for long-horizon archives with a low expectation of extensive access, such as government and historical records[23,24]. An example in a scientific context is CERN's CASTOR system[25], which stores a total of 80 PB of Large Hadron Collider data and grows at 15 PB yr^{-1}. Only 10% is maintained on disk, and CASTOR migrates regularly between magnetic tape formats. Archives of older data are needed for potential future verification of events, but access rates decrease considerably 2–3 years after collection. Further examples are found in astronomy, medicine and interplanetary exploration[26]. With negligible computational costs and optimized use of the technologies we employed, we estimate current costs to be \$12,400 MB^{-1} for information storage in DNA and \$220 MB^{-1} for information decoding. Modelling relative long-term costs of archiving using DNA-based storage or magnetic tape shows that the key parameters are the ratio of the one-time cost of synthesizing the DNA to the recurrent fixed cost of transferring data between tape technologies or media, which we estimate to be 125–500 currently, and the frequency of tape transition events (Supplementary Information and Supplementary Fig. 7). We find that with current technology and our encoding scheme, DNA-based storage may be cost-effective for archives of several megabytes with a ~600–5,000-yr horizon (Fig. 2c). One order of magnitude reduction in synthesis costs reduces this to ~50–500 yr; with two orders

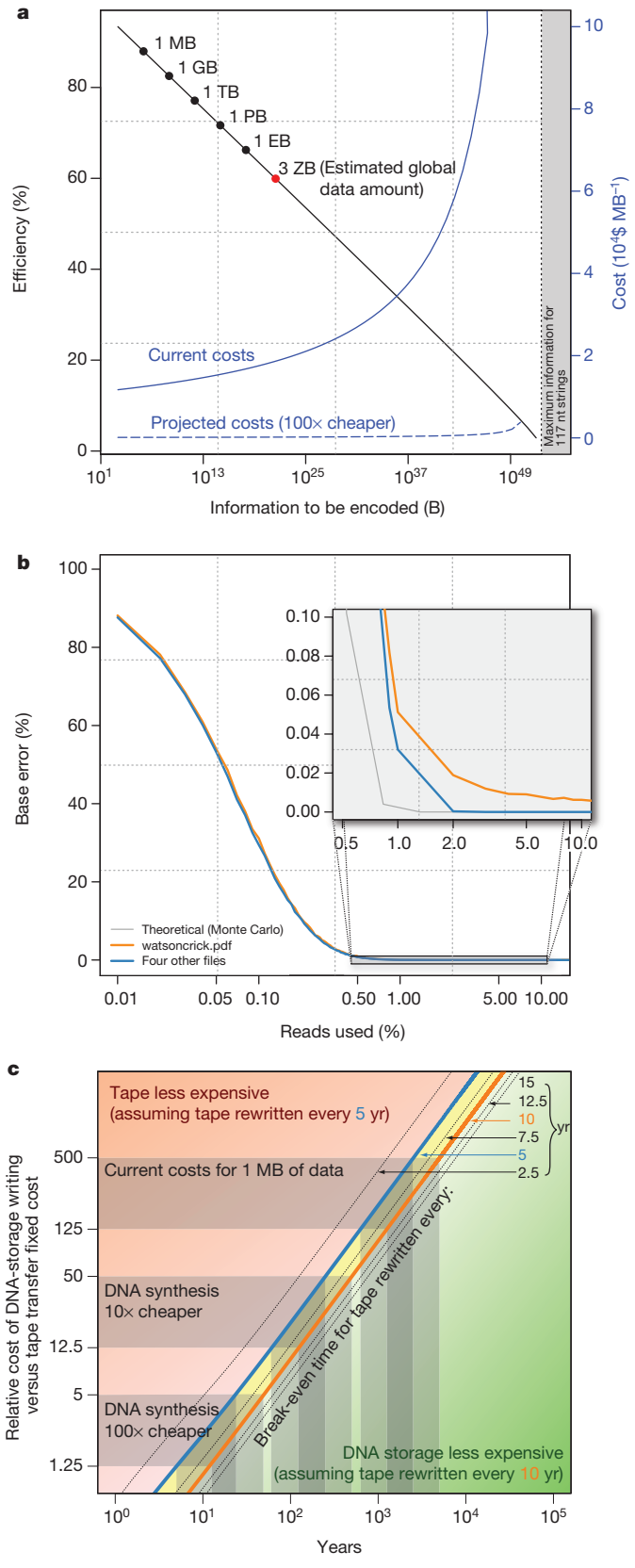

Figure 2 | Scaling properties and robustness of DNA-based storage. **a**, Encoding efficiency and costs change as the amount of stored information increases. The *x* axis (logarithmic scale) represents the total amount of information to be encoded. Common data scales are indicated, including the three zettabyte (3 ZB, 3×10^{21} bytes) global data estimate, shown red. The black line (*y*-axis scale to left) indicates encoding efficiency, measured as the proportion of synthesized bases available for data encoding. The blue curves (*y*-axis scale to right) indicate the corresponding effect on encoding costs, both at current synthesis cost levels (solid line) and in the case of a two-order-of-magnitude reduction (dashed line). **b**, Per-recovered-base error rate (*y* axis) as a function of sequencing coverage, represented by the percentage of the original 79.6×10^6 read-pairs sampled (*x* axis; logarithmic scale). The blue curve represents the four files recovered without human intervention: the error is zero when ≥2% of the original reads are used. The grey curve is obtained by Monte Carlo simulation from our theoretical error rate model. The orange curve represents the file (watsoncrick.pdf) that required manual correction: the minimum possible error rate is 0.0036%. The boxed area is shown magnified in the inset. **c**, Timescales for which DNA-based storage is cost-effective. The blue curve indicates the relationship between break-even time beyond which DNA storage is less expensive than magnetic tape (*x* axis) and relative cost of DNA-storage synthesis and tape transfer fixed costs (*y* axis), assuming the tape archive has to be read and rewritten every 5 yr. The orange curve corresponds to tape transfers every 10 yr; broken curves correspond to other transfer periods as indicated. In the green-shaded region, DNA storage is cost-effective when transfers occur more frequently than every 10 yr; in the yellow-shaded region, DNA storage is cost-effective when transfers occur every 5–10 yr; in the red-shaded region tape is less expensive when transfers occur less frequently than every 5 yr. Grey-shaded ranges of relative costs of DNA synthesis to tape transfer are 125–500 (current costs for 1 MB of data), 12.5–50 (achieved if DNA synthesis costs are reduced by one order of magnitude) and 1.25–5 (costs reduced by two orders of magnitude). Note the logarithmic scales on both axes. See Supplementary Information for further details.

of magnitude reduction, as can be expected in less than a decade if current trends continue (ref. 13, and http://www.synthesis.cc/2011/06/new-cost-curves.html), DNA-based storage becomes practical for archives with a horizon of less than 50 yr. The speed of DNA-storage writing and reading are not competitive with current technology, but both processes can be accelerated through parallelization (Supplementary Information).

The DNA-based storage medium has different properties from traditional tape- or disk-based storage. As DNA is the basis of life on Earth, methods for manipulating, storing and reading it will remain the subject of continual technological innovation. As with any storage system, a large-scale DNA archive would need stable DNA management[27] and physical indexing of depositions. But whereas current digital schemes for archiving require active and continuing maintenance and regular transferring between storage media, the DNA-based storage medium requires no active maintenance other than a cold, dry and dark environment[3,4] (such as the Global Crop Diversity Trust's Svalbard Global Seed Vault, which has no permanent on-site staff[28]) yet remains viable for thousands of years even by conservative estimates. We achieved an information storage density of ~2.2 PB g^{-1} (Supplementary Information). Our sequencing protocol consumed just 10% of the library produced from the synthesized DNA (Supplementary Table 2), already leaving enough for multiple equivalent copies. Existing technologies for copying DNA are highly efficient[11,13], meaning that DNA is an excellent medium for the creation of copies of any archive for transportation, sharing or security. Overall, DNA-based storage has potential as a practical solution to the digital archiving problem and may become a cost-effective solution for rarely accessed archives.

Received 15 May; accepted 12 December 2012.
Published online 23 January 2013.

1. Baum, E. B. Building an associative memory vastly larger than the brain. *Science* **268,** 583–585 (1995).
2. Cox, J. P. L. Long-term data storage in DNA. *Trends Biotechnol.* **19,** 247–250 (2001).
3. Anchordoquy, T. J. & Molina, M. C. Preservation of DNA. *Cell Preserv. Technol.* **5,** 180–188 (2007).
4. Bonnet, J. *et al.* Chain and conformation stability of solid-state DNA: implications for room temperature storage. *Nucleic Acids Res.* **38,** 1531–1546 (2010).
5. Clelland, C. T., Risca, V. & Bancroft, C. Hiding messages in DNA microdots. *Nature* **399,** 533–534 (1999).
6. Kac, E. *Genesis* (1999); available at http://www.ekac.org/geninfo.html (accessed 10 May 2012).

7. Ailenberg, M. & Rotstein, O. D. An improved Huffman coding method for archiving text, images, and music characters in DNA. *Biotechniques* **47,** 747–754 (2009).
8. Gibson, D. G. *et al.* Creation of a bacterial cell controlled by a chemically synthesized genome. *Science* **329,** 52–56 (2010).
9. Church, G. M., Gao, Y. & Kosuri, S. Next-generation digital information storage in DNA. *Science* **337,** 1628 (2012).
10. MacKay, D. J. C. *Information Theory, Inference, and Learning Algorithms* (Cambridge Univ. Press, 2003).
11. Erlich, H. A., Gelfand, D. & Sninsky, J. J. Recent advances in the polymerase chain reaction. *Science* **252,** 1643–1651 (1991).
12. Monaco, A. P. & Larin, Z. YACs, BACs, PACs and MACs: artificial chromosomes as research tools. *Trends Biotechnol.* **12,** 280–286 (1994).
13. Carr, P. A. & Church, G. M. Genome engineering. *Nature Biotechnol.* **27,** 1151–1162 (2009).
14. Willerslev, E. *et al.* Ancient biomolecules from deep ice cores reveal a forested southern Greenland. *Science* **317,** 111–114 (2007).
15. Green, R. E. *et al.* A draft sequence of the Neandertal genome. *Science* **328,** 710–722 (2010).
16. Kari, L. & Mahalingam, K. in *Algorithms and Theory of Computation Handbook* Vol. 2, 2nd edn (eds Atallah, M. J. & Blanton, M.) 31-1–31-24 (Chapman & Hall, 2009).
17. Păun, G., Rozenberg, G. & Salomaa, A. *DNA Computing: New Computing Paradigms* (Springer, 1998).
18. Watson, J. D. & Crick, F. H. C. Molecular structure of nucleic acids. *Nature* **171,** 737–738 (1953).
19. Niedringhaus, T. P., Milanova, D., Kerby, M. B., Snyder, M. P. & Barron, A. E. Landscape of next-generation sequencing technologies. *Anal. Chem.* **83,** 4327–4341 (2011).
20. 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).
21. Massingham, T. & Goldman, N. All Your Base: a fast and accurate probabilistic approach to base calling. *Genome Biol.* **13,** R13 (2012).
22. Gantz, J. & Reinsel, D. *Extracting Value from Chaos* (IDC, 2011).
23. Brand, S. *The Clock of the Long Now* (Basic Books, 1999).
24. Digital archiving. History flushed. *Economist* **403,** 56–57 (28 April 2012); available at http://www.economist.com/node/21553410 (2012).
25. Bessone, N., Cancio, G., Murray, S. & Taurelli, G. Increasing the efficiency of tape-based storage backends. *J. Phys. Conf. Ser.* **219,** 062038 (2010).
26. Baker, M. *et al.* in *Proc. 1st ACM SIGOPS/EuroSys European Conf. on Computer Systems* (eds Berbers, Y. & Zwaenepoel, W.) 221–234 (ACM, 2006).
27. Yuille, M. *et al.* The UK DNA banking network: a ''fair access'' biobank. *Cell Tissue Bank.* **11,** 241–251 (2010).
28. Global Crop Diversity Trust. Svalbard Global Seed Vault. (2012); available at http://www.croptrust.org/main/content/svalbard-global-seed-vault (accessed 10 May 2012).

Supplementary Information is available in the online version of the paper.

Acknowledgements At the University of Cambridge: D. MacKay and G. Mitchison for advice on codes for run-length-limited channels. At CERN: B. Jones for discussions on data archival. At EBI: A. Löytynoja for custom multiple sequence alignment software, H. Marsden for computing base calls and for detecting an error in the original parity-check encoding, T. Massingham for computing base calls and advice on code theory and K. Gori, D. Henk, R. Loos, S. Parks and R. Schwarz for assistance with revisions to the manuscript. In the Genomics Core Facility at EMBL Heidelberg: V. Benes for advice on Next-Generation Sequencing protocols, D. Pavlinić for sequencing and J. Blake for data handling. C.D. is supported by a fellowship from the Swiss National Science Foundation (grant 136461). B.S. is supported by an EMBL Interdisciplinary Postdoctoral Fellowship under Marie Curie Actions (COFUND).

Author Contributions N.G. and E.B. conceived and planned the project and devised the information-encoding methods. P.B. advised on oligo design and Next-Generation Sequencing protocols, prepared the DNA library and managed the sequencing process. S.C. and E.M.L. provided custom oligonucleotides. N.G. wrote the software for encoding and decoding information into/from DNA and analysed the data. N.G., E.B., C.D. and B.S. modelled the scaling properties of DNA storage. N.G. wrote the paper with discussions and contributions from all other authors. N.G. and C.D. produced the figures.

Author Information Data are available at http://www.ebi.ac.uk/goldman-srv/DNA-storage and in the Sequence Read Archive (SRA) with accession number ERP002040. Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to N.G. (goldman@ebi.ac.uk).