TABLE OF CONTENTS

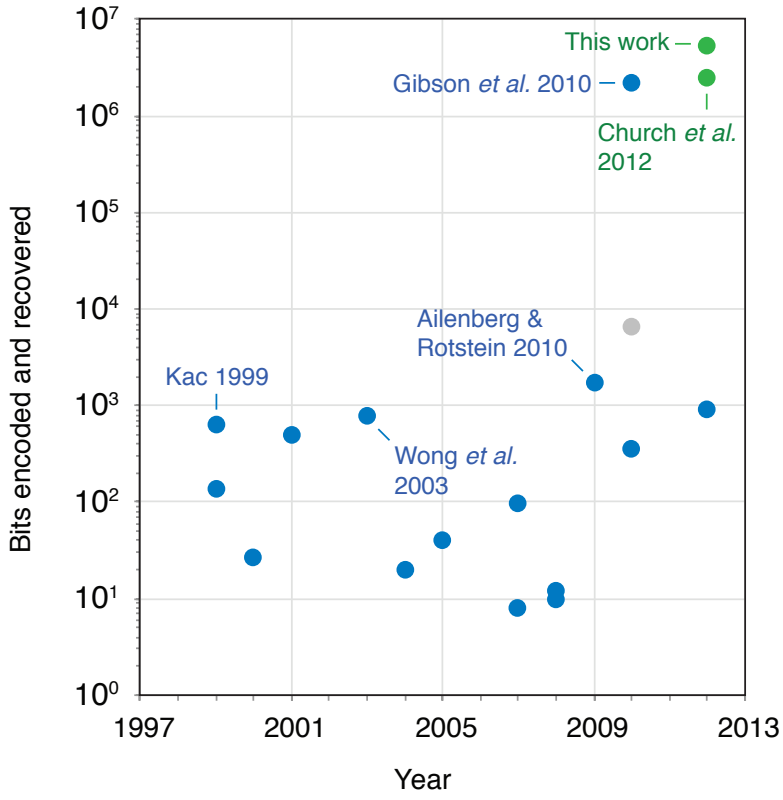
HUMAN-DESIGNED INFORMATION STORED IN DNA

Supplementary Table S1 and Supplementary Fig. S1 show the amounts of human-designed information stored in DNA and successfully recovered in this letter 16 previous studies. Supplementary Fig. S2 illustrates some of the information encoded in this study. The Shannon information content[10] of the designed messages was approximated by the minimum number of bits required to encode the message using any of the following methods:

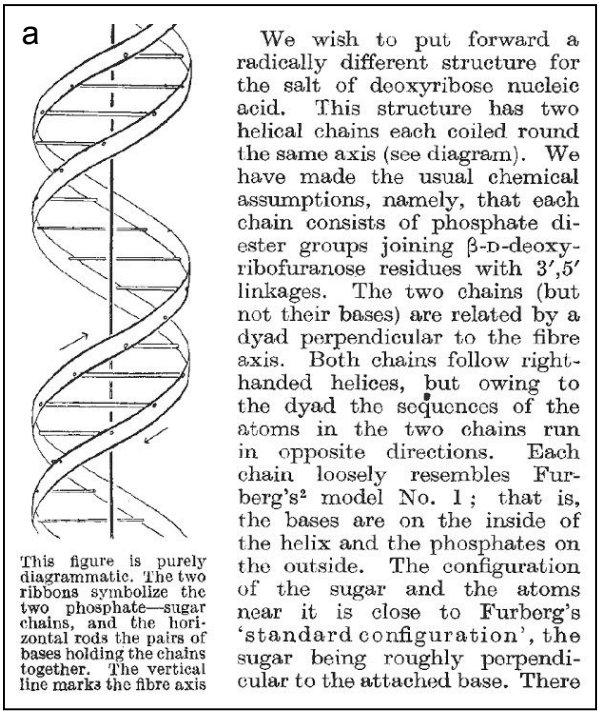
- compress ASCII file containing the message in natural form, using Unix command `gzip --best`
- compress ASCII file containing the message in natural form, using Unix command `bzip2 --best`
- for DNA sequence, 2 bits per base
- for simple English text, 5 bits per character (permits use of $2^5 = 32$ characters, e.g. 26 letters of the alphabet plus space and simple punctuation)
- for English/Latin text using reduced or extended alphabets, the number of bits per character is calculated similarly (e.g. 3 bits per character for an alphabet of $8 = 2^3$ characters, 6 bits/char for a 64-character alphabet)

Ref.	Authors	Year	Message type	Message length	Bases used	Shannon information (bits)	Notes
5	Clelland et al.	1999	English text	23 characters	69	138	
6	Kac	1999	English text	129 characters	360	645	Biblical quotation encoded in mutating *E. coli* genome as a work of art; decoded with 3 character errors, attributed to mutation
30	Leier et al.	2000	three 9-bit numbers	27 bits	810	27	
31	Bancroft et al.	2001	English text	106 characters	318	504	
32	Wong et al.	2003	English text (64 character alphabet)	185 characters*	560	800	*estimated
33	Arita & Ohashi	2004	English text	4 characters	24	20	
34	Kashimawura et al.	2005	DNA string	20 bases	65	40	
35	Skinner et al.	2007	four 2-bit numbers	8 bits	231	8	
36	Yachie et al.	2007	mathematical equation and date (256 character alphabet)	12 characters	250	96	
37	Heider & Barnekow	2008	English text	2 characters	5	10	
38	Portney et al.	2008	Latin text (using 8 character alphabet)	5 characters	80	12	final character of 5 (i.e. 3 bits) lost in decoding
7	Ailenberg & Rotstein	2009	English text, simple musical notation, simple line-drawing notation	349 characters	844	1715	
39	CUHK-iGEM	2010	English text	70 characters	438	350	
8	Gibson et al.	2010	bacterial genome with additional "watermark" sequences (see below)	1077947 bases	1077947	2155894	decoded with 8 base errors and two insertions of 768 and 85 bases, respectively
	of which:		"watermarks": English text plus programming symbols (64 character alphabet)	1280 characters	4658	6504	
9	Church et al.	2012	English text, JPEG images, computer code, all within HTML encoding	658776 characters (bytes)	6313270	2495760	decoded with 10 bit errors
40	Jarvis & NPC	2012	English text	180 characters	540	900	Article 1 of the Universal Declaration of Human Rights encoded in *E. coli* genome as work of art
	Goldman et al.	**this letter**	**Total**	**757051 bytes**	**17940195**	**5165800**	
	of which:		English text (all 154 Shakespeare sonnets)	107738 characters (bytes)	2533635	297856	file `wssnt10.txt` (from Project Gutenberg, `http://www.gutenberg.org/ebooks/1041`)
			PDF document (Watson and Crick, 1953)	280864 bytes	6659172	2119848	file `watsoncrick.pdf` (from the Nature website, `http://www.nature.com/nature/dna50/archive.html`, modified to achieve higher compression); see Supplementary Fig. S2a
			MP3 audio file (extract from Martin Luther King "I Have a Dream" speech)	168539 bytes	3997773	1227176	file `MLK_excerpt_VBR_45-85.mp3` (from `http://www.americanrhetoric.com/speeches/mlkihaveadream.htm`, modified to achieve higher compression: variable bit rate, typically 48–56 kbps; sampling frequency 44.1 kHz)
			JPEG 2000 image file (image of EBI, 640 x 480 pixels, 16.7M colours)	184264 bytes	4379076	1474000	file `EBI.jp2` (authors' own picture); see Supplementary Fig. S2b
			ASCII file (Huffman code used to convert bytes to base-3; human readable)	15646 bytes	370539	46920	file `View_huff3.cd.new`

Supplementary Table S1 | Amounts of human-designed information stored in DNA and successfully recovered. Message length uses the natural measurement according to the Message type. Bases used indicates the number of DNA bases designed to contain a single copy of the encoded message and ignores the number of copies synthesised.



Supplementary Figure S1 | Amounts of human-designed information stored in DNA and successfully recovered. Information content is measured in bits; note the logarithmic scale on the *y*-axis. Blue points indicate studies not adapted to high-throughput data storage; green indicates high-throughput methods. The grey point indicates that part of the Gibson *et al.* (2010) experiment[8] that encoded information of non-biological origin.


a

This figure is purely diagrammatic. The two ribbons symbolize the two phosphate—sugar chains, and the horizontal rods the pairs of bases holding the chains together. The vertical line marks the fibre axis

We wish to put forward a radically different structure for the salt of deoxyribose nucleic acid. This structure has two helical chains each coiled round the same axis (see diagram). We have made the usual chemical assumptions, namely, that each chain consists of phosphate diester groups joining β-D-deoxyribofuranose residues with 3′,5′ linkages. The two chains (but not their bases) are related by a dyad perpendicular to the fibre axis. Both chains follow right-handed helices, but owing to the dyad the sequences of the atoms in the two chains run in opposite directions. Each chain loosely resembles Furberg's[2] model No. 1; that is, the bases are on the inside of the helix and the phosphates on the outside. The configuration of the sugar and the atoms near it is close to Furberg's 'standard configuration', the sugar being roughly perpendicular to the attached base. There


b

Supplementary Figure S2 | Digital information encoded in DNA. a, An excerpt from the Watson and Crick (1953) paper[18] (PDF format) and **b**, a digital photograph of the European Bioinformatics Institute (JPEG 2000 format) that were among the files encoded in DNA and successfully recovered in this study.

SUPPLEMENTARY METHODS

Digital information encoding. Five files of digital information stored on a hard disk drive were encoded using purpose-written computer software. Each byte of each file to be encoded was represented as a sequence of DNA bases via base-3 digits ('trits' 0, 1 and 2) using a purpose-designed Huffman code[10]. Each of the 256 possible bytes was represented by five or six trits; the Huffman code is given in Supplementary File `huffman.pdf`. Next, each trit was encoded as a DNA nucleotide selected from the three nucleotides different from the last one used, to exclude homopolymer runs. The resulting DNA sequence was converted to segments of length 100 bases, each overlapping the previous by 75 bases, to give strings of a length that was readily synthesised and to provide fourfold redundancy (each DNA base is included in four different segments). Alternate segments were reverse complemented. Indexing information, comprising two trits for file identification (permitting up to $3^2 = 9$ files to be distinguished, in this implementation), 12 trits for intra-file location information (permitting up to $3^{12} = 531{,}441$ locations per file, i.e. a total of up to $3^{14} = 4{,}782{,}969$ unique data locations) and one parity-check[10] trit, again encoded as non-repeating DNA nucleotides, was appended to the 100 information storage bases. Each indexed DNA segment had one further base added to each end, consistent with the 'no homopolymers' rule, that would indicate whether the entire fragment was reverse complemented during the 'reading' stage of the experiment. A full formal specification of the digital information encoding scheme is given in Supplementary File `file2features.pdf`. In total, the five files were represented by a total of 153,335 strings of DNA, each comprising 117 nt (= 1 + 100 + 2 + 12 + 1 + 1) to encode original digital information plus indexing information. The fourfold redundancy provides simple but effective error correction: as each base is encoded in four of the DNA segments, two of which are reverse complemented, any systematic or chance errors in synthesis or sequencing may be corrected by majority vote or more complex decoding schemes. We used simple majority voting (see below).

The data-encoding component of each string can contain Shannon information at 5.07 DNA bases per byte (i.e. 8/5.07 = 1.58 bits per base), close to the theoretical optimum capacity of 5.05 bits per DNA base (see `huffman.pdf`) for base-4 channels with runlength limited to 1 (i.e. no repeated nucleotides). After error-correction redundancy and addition of indexing and parity-check information, the data content of our encoding scheme was 4.94 bytes per string (= 757,051/153,335), or 0.0422 bytes per base (= 4.94/117) (i.e. 23.70 bases per byte). The 153,335 designed DNA strings are available online at http://www.ebi.ac.uk/goldman-srv/DNA-storage.

Our indexing scheme, with 14 nt per string available to record file identification and intra-file location, is easily extended by the addition of further indexing nucleotides. This is considered below, in our analysis of the scaling properties of our DNA-storage scheme. Increasing the number of indexing trits (and therefore bases) used to specify file and intra-file location by just two, to 16, gives $3^{16} = 43{,}046{,}721$ unique locations, in excess of the 16.8M that is the practical maximum for the Nested Primer Molecular Memory (NPMM) scheme[41,16]. While these indexing schemes share the aim of encoding which part of a larger total message any one string contains, ours is simpler and more-readily-extensible as it does not incorporate any system by which the indexing information is used to physically extract a subset of the information-bearing strings prior to decoding, as the NPMM scheme does.

DNA synthesis. The synthesis process was also used to incorporate 33 nt paired-end adapter sequences at the 5′ and 3′ ends of each oligonucleotide (oligo) to facilitate PCR amplification and sequencing on the Illumina platform:

- 5′ adapter: `5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCT-3'`
- 3′ adapter: `5'-AGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG-3'`

The 153,335 DNA oligo designs were synthesised in three distinct runs (with oligos randomly assigned to runs) using an updated version of Agilent Technologies' OLS (oligo library synthesis) process described previously[42,20]. This adapts the phosphoramidite chemistry developed previously[43] and employs inkjet printing and flowcell reactor technologies in the SurePrint *in situ* microarray synthesis platform. Inkjet printing within an anhydrous chamber allows the delivery of very small volumes of phosphoramidites to a confined coupling area on a 2D planar surface, resulting in the addition of hundreds of thousands of bases in parallel. Subsequent oxidation and detritylation are carried out in a flowcell reactor. Once DNA synthesis has been completed, the oligos are then cleaved from the surface and deprotected[44].

Up to ~99.8% coupling efficiency is achieved by using thousands-fold excess of phosphoramidite and activator solution. Similarly, millions-fold excess of detritylation agent drives the removal of the 5′-hydroxyl protecting group to near-completion. A novel controlled process in the flowcell reactor significantly reduces depurination, the most prevalent side reaction[20]. With the latest platform, up to 244,000 unique sequences are synthesised in parallel and delivered as ~1–10 pmol pools of oligos. This is equivalent to ~2.5–25 × 10^6 oligos for each designed sequence (= 1–10 × 10^{-12} × 6.02 × 10^{23}/244,000). In our experiment, three runs were used to synthesise 153,335 designs, leading to the higher figure of ~12–120 × 10^6 (= 3–30 × 10^{-12} × 6.02 × 10^{23}/153,335).

Error rates in the Agilent OLS process are approximately 1 per 500 bases synthesised[42] and synthesis errors are believed to occur independently in different oligos (SC and EML, unpublished data). Combined with our data encoding scheme, this gives further error tolerance. The probability that a given oligo is synthesised entirely correctly is ~0.79 (= $(1 - 1/500)^{117}$), giving a large pool of correct oligos; oligos with a small number of errors in their 100 nt data region may also contribute to correct decoding, with the majority of positions contributing correct information and a small number of errors being outweighed by contributions from other reads (see below).

Library preparation and sequencing. The three samples of lyophilised oligos were resuspended in Tris buffer to a concentration of 5 ng/ml. Samples were then purified from residual synthesis by-products on Ampure XP paramagnetic beads (Beckman Coulter). The reconstituted oligo library was amplified in a total of 22 cycles using thermocycler conditions selected for even A/T vs. G/C processing[45]. PCR was performed with high-fidelity AccuPrime reagents (Invitrogen), a combination of Taq and *pyrococcus* polymerases with a thermostable accessory protein, and paired-end PCR primers (Illumina) complementary to the synthesised adapter sequences flanking each DNA-storage oligo to incorporate additional sequences necessary for flowcell attachment. PCR amplification enabled enrichment for full-length oligos with both 5′ and 3′ adapters correctly synthesised, and allowed us to achieve appropriate concentration for sequencing while simultaneously incorporating the additional sequences necessary for flowcell attachment and cluster formation. The amplified library products were bead-purified and quantified on the Agilent 2100 Bioanalyzer (concentration

determined to be 15.1 ng/µl, i.e. 86 nM given a peak construct size measured at 270 bp and approximating 650 pg/pmol per bp), diluted to a concentration of 16 pM for flowcell loading and sequenced in paired-end mode on the Illumina HiSeq 2000. The sequencing reaction consumed ~0.1% of the DNA in the initial library: 337 pg of DNA (120µl at 16 pM) from a starting value of 302 ng (20µl at 86 nM). Further details of the sample preparation are given in Supplementary Table S2.

Sample stage	Vol (µl)	Conc (ng/µl)	Conc (nM)	Amount of DNA (pg)	Description
A	20	15.1	86	302000	PCR products, quantified on the Agilent 2100 Bioanalyzer
B	2	15.1	86	30200	2 µl of (A) extracted for sequencing; remaining 18 µl reserved for future analysis
C	17.2	1.76	10	30200	(B) then diluted in Tris to 10 nM concentration
D	2	1.76	10	3512	2 µl of (C) retained (remainder discarded in this experiment)
E	20	0.176	1	3512	(D) then denatured in 100mM NaOH, giving dilution to 1 nM
F	16	0.176	1	2809	16 µl of (E) retained (remainder discarded in this experiment)
G	1000	0.00281	0.016	2809	(F) then diluted to 1000 µl
H	120	0.00281	0.016	337	120 µl of (G) sent for clustering (remainder discarded in this experiment)

Supplementary Table S2 | Sample preparation details. Products at stage A (blue) were measured; other values were computed from these.

Base calls were computed from observed intensities using the AYB software[21], producing 79.6M read-pairs of 104 bases in length. (Illumina's base calling software Bustard produced 65.9M read-pairs, 17.2% fewer than AYB, but led to qualitatively identical decoding results.) Quality control of the reads was performed using FastQC (version 0.10.1; ref. 46). Overall the QC report (available as Supplementary File `FastQC.pdf`) indicated a high-quality sequencing run. Per-cycle quality scores were as expected for an Illumina HiSeq run. The mean quality (Q) score was 36.7, with 95% of quality scores ≥ Q30. The GC content of the sequenced reads and the *k*-mer frequencies along the reads were consistent with the structure of the designed DNA strings. The read duplication levels were high, in concordance with the design of the library providing many reads covering any single string.

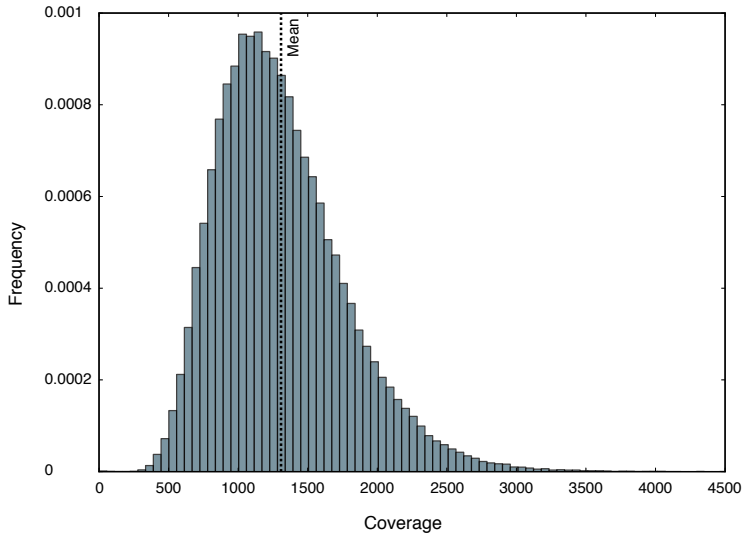
As further quality assessment, but not used for subsequent digital information decoding, the reads were aligned to the designed DNA strings using BWA version 0.6.1-r104 in paired-end mode[47]. Per-cycle error rates were calculated from the resulting alignment using the ErrorRatePerCycle functionality of the GATK package (version 2.1-8-g5efb575; ref. 48). Per-cycle error rates were as expected for an Illumina HiSeq run (see Supplementary File `GATK.txt`) and the mean error rate of 0.001774 after 12.81% unmappable reads were discarded is in line with the combination of current estimates of synthesis error (1 base in 500, above) and sequencing error (1 base in 1,000; ref. 49).

Digital information decoding. As the central 91 bases of each oligo were sequenced from both ends, rapid computation of full-length (117 base) oligos and removal of reads

inconsistent with our designs was straightforward. Sequencing reads were decoded using purpose-written software that exactly reverses the encoding process. The numbers of reads used in different stages of the information decoding process are given in Supplementary Table S3. At the final stage of decoding, the five files were reconstructed from 50.1M strings, giving a mean sequencing depth of 1,308× coverage (standard deviation 459). Supplementary Fig. S3 shows the distribution of sequencing depths over encoded data locations (bases of the files' DNA representations). Virtually every location within each decoded file was detected in hundreds or thousands of different sequenced DNA oligos.

Analysis stage	Number of reads	% of total	% of previous stage	Notes	Possible reasons for losses relative to previous analysis stage
A	79564267	100		read-pairs from AYB base caller	
B	55047046	69.19	69.19	117 nt fragment reads recovered from combining 104 nt paired-end reads with 91 nt overlap as expected, with at most 6 mismatches within the overlap region	synthesis error, sequencing error, contamination
C	50145113	63.02	91.10	reads with indexing information indicating they belong to one of the five files encoded in this experiment	synthesis error or sequencing error leading to dinucleotide repeat, invalid file identification in indexing information or parity-check failure
of which:	18270252	22.96	33.19	file `watsoncrick.pdf`	
	8064484	10.14	14.65	file `wssnt10.txt`	
	11966357	15.04	21.74	file `EBI.jp2`	
	802908	1.01	1.46	file `View_huff3.cd.new`	
	11041112	13.88	20.06	file `MLK_excerpt_VBR_45-85.mp3`	
D	50141326	63.02	99.99	reads contributing 'votes' to final decoding of file	synthesis error or sequencing error in indexing information leading to invalid location in file
of which:	18269250	22.96	99.99	file `watsoncrick.pdf`	
	8063761	10.13	99.99	file `wssnt10.txt`	
	11965715	15.04	99.99	file `EBI.jp2`	
	802414	1.01	99.94	file `View_huff3.cd.new`	
	11040186	13.88	99.99	file `MLK_excerpt_VBR_45-85.mp3`	

Supplementary Table S3 | Numbers of reads used during decoding.



Supplementary Figure S3 | Distribution of sequencing depths over encoded locations. The mean is at a coverage of 1,308; the standard deviation is 459.

Majority voting was used to resolve any discrepancies caused by DNA synthesis or sequencing errors. The error rate amongst the 'votes' used to reconstruct the five files was 0.004004 (20.08M errors in 5,014M bases counted). This is higher than the combined synthesis and sequencing error rate reported above because of cases where one or a small number of errors in indexing information led to a read being misplaced in its correct file, or placed in an incorrect file, generating on average 75 incorrect votes (100 misplaced votes, each with probability ~0.75 of being incorrect).

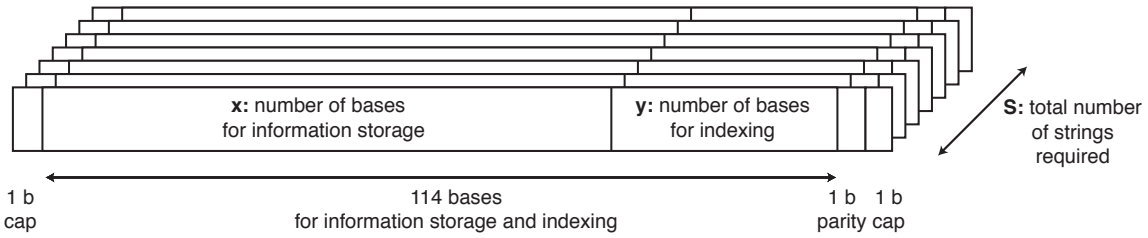
On completion of this procedure, four of the five original files were reconstructed perfectly. The fifth file required manual intervention to correct two regions each of 25 bases that were not recovered from any sequenced read, as described below.

Scaling properties of the DNA-storage scheme. In the following sections we demonstrate that, even constrained to today's technology, the scheme presented here scales nearly linearly, well beyond any realistically needed range, i.e. beyond 20 orders of magnitude larger than the estimated global amount of digital data of 3 ZB (3×10^{21} bytes).

The global data volume was estimated by adding the 1.8 ZB estimate of data produced in 2011 (ref. 22) to the estimated production in 2010, calculated assuming a doubling time of two years[22], to give 3 ZB (= $1.8+1.8/\sqrt{2}$).

As each 117 base string can be synthesised and sequenced independently, it is reasonable to assume that the costs associated with these processes is linear in the number of strings. Therefore, to demonstrate the scaling behaviour of our scheme, we show that the number of strings required to reliably store the data increases nearly linearly with the amount of data. First, we focus on the relationship between the number of strings required and the amount of data to be stored. Second, we show that increase the amount of data to be stored does not lead to higher error rates. This is achieved by both theoretical and empirical estimates of the error rate as a function of amount of data and sequencing coverage.

Scaling of the total number of strings required. Let I be the information to be stored, in bytes. We now show that the number of 117 nt strings required to encode these data scales nearly linearly with respect to I. Recall that in each 117 nt string, 114 bases can be used to store data and indexing information. As the amount of data to be stored increases, we may need to use more than the current 14 bases for indexing. Supplementary Fig. S4 shows how, in general, the strings may be partitioned into x data bases and y indexing bases. Note that y depends on the total number of strings to be used, S. Optimally, $y = \lceil \log_3(S) \rceil$ and therefore $x = 114 - \lceil \log_3(S) \rceil$.



Supplementary Figure S4 | Schematic representation of information encoding in DNA. Multiple strings are used, each comprising 117 bases of which x may be used for storing information and y for indexing, with $x + y = 114$.

Intuitively, we can anticipate that the relationship between S and I is not linear: as the amount of information increases, the number of strings required also increases; this in turn requires more indexing information (and thus greater y), which leaves fewer bases to store information in each string (thus smaller x).

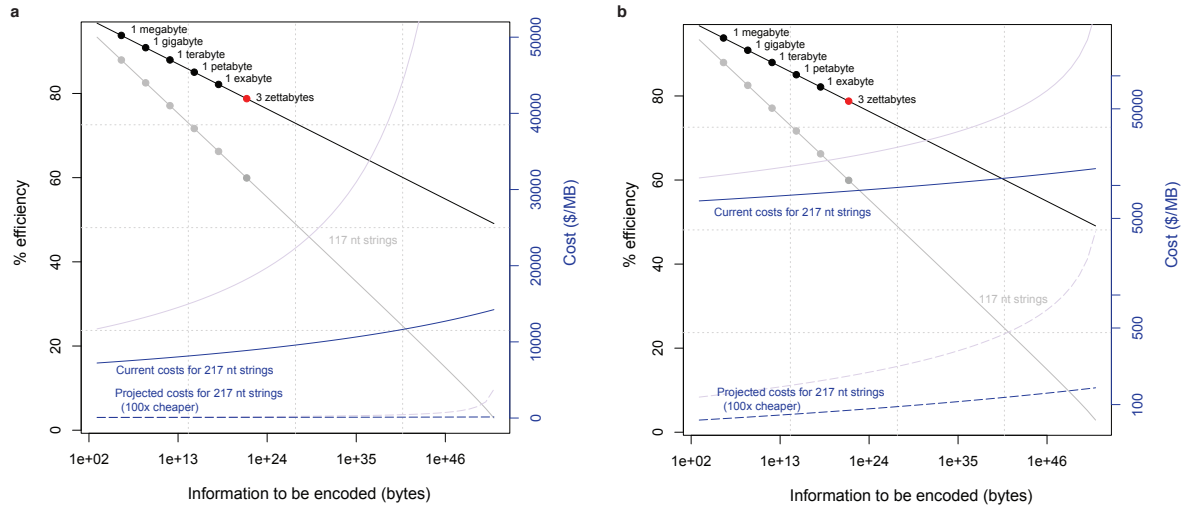
Expressed in bases, the information to be stored is $B = 5.07I$, as we can encode 5.07 bases/byte (see above). Since the encoding strings are 75% overlapping, the relationship between B and S is:

$$B = Sx / 4 = S(114 - \lceil \log_3(S) \rceil) / 4 \qquad (1)$$

This can be solved for S numerically. Fig. 2a (Main Text) illustrates the relationship of information stored (I) and the efficiency of encoding, measured as the proportion of bases synthesised (data and indexing) that are used to hold data ($eff = B / 117S$). Two features become apparent. First, the proportion of the total DNA available for encoding data decreases slowly, and is reasonable across the entire relevant data size range. Second, even constrained to 117 base-long strings, the current encoding scheme makes it possible to encode > 20 orders of magnitude more data than is currently practically relevant. (The theoretical limit comes when every base is required for indexing, and none remains to store information.)

DNA-storage costs are affected by the efficiency achieved for different information volumes. Current costs are about $12,400/MB (below), based on our efficiency of $B / 117S = 0.88$. Costs scale as the inverse of efficiency, and Fig. 2a (Main Text) also shows the cost function $12{,}400(0.88 / eff)$.

We repeated the above calculations based on longer synthesised strings. The Agilent OLS process can already produce 300-base oligos, with 244,000 designed strings costing approximately $30,000. Assuming that this would provide 217 nt strings for DNA-storage, an increase of 100 nt, we get twice as many available bases for 30/25 times the price. This would give a current cost of about $7,440/MB ($= 12{,}400 \times (30{,}000 / 25{,}000) / 2$) based on an efficiency of 0.94 (the encoding efficiency that would be achieved repeating our experiment with 217 nt strings) and consequently a cost function of $7{,}440(0.94 / eff)$. Supplementary Fig. S5 repeats the information of Fig. 2a (Main Text), and adds the corresponding results for these longer strings. This shows that with achievable improvement in DNA synthesis technology the scalability of DNA-storage is substantially improved, with higher efficiency, lower cost and slower decline in efficiency and increase in cost for larger data volumes.



Supplementary Figure S5 | Scaling properties of DNA-storage. The graphs show how encoding efficiency and costs change as the amount of stored information increases, for longer strings with 217 nt available for data and indexing. The *x*-axis (logarithmic scale) represents the total amount of information to be encoded. Common data scales are indicated, including the 3 ZB global data estimate. The black line (*y*-axis scale to left) indicates encoding efficiency, measured as the proportion of synthesised 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). The pale grey and pale blue lines give the corresponding results for 117 nt strings, for ease of comparison with Fig. 2a (Main Text). **a**, Linear cost scale; **b**, logarithmic cost scale.

Scaling of the decoded data error rate. Assuming that all synthesised strings have the same probability of being sequenced, the mean error rate per base of encoded data depends on three variables:

- ε: the mean error rate per base at the level of a sequencing read (due to synthesis and sequencing error), set to 0.004 as determined for our experiment (above)
- S: the total number of designed strings
- c_B: the base coverage (mean number of times each base of encoded information is sequenced)

Recall that the decoding scheme calls each base of encoded information based on a majority vote of all the read bases corresponding to its position. Because each base of encoded information is represented in four strings (due to the 75% overlap in encoded data between neighbouring strings), the mean string coverage is $c_S = c_B / 4$. Thus, in total, there are $Sc_S = Sc_B / 4$ reads. The probability that any read covers base i of encoded information is $4 / S$. Thus, the number x_i of base reads for encoded base i follows a binomial distribution with number of trials $Sc_B / 4$ and probability of success $4 / S$, which we write as $B(Sc_B / 4, 4 / S)$.

Next, consider that for encoded base i to be correctly called, the majority of the x_i read bases (votes) need to be correct. The distribution of correct bases $x_{i,\text{correct}}$ is $B(x_i, 1 - \varepsilon)$. The majority vote regarding encoded base i is wrong if $x_{i,\text{correct}} < x_i / 2$. (This is the worst case scenario that all incorrect votes are for the same incorrect base. Our results are not significantly altered when other, more-favourable, scenarios are considered.) Because of this

dependency, the expected encoded base error rate is not straightforward to compute analytically, but it is easily estimated by Monte Carlo simulation. Supplementary Fig. S6 shows an R function that performs this estimation.

```
# R function to estimate the encoded base error rate:
estimateBaseErrorRateMC = function(eps,S,cS,nsamples) {
  x = rbinom(nsamples, cS*S ,4/S);
  s = sapply(x,function(t) rbinom(1,t,1-eps));
  e = 1 - sum (s > x/2) / nsamples;
  return(e);
}
```

Supplementary Figure S6 | R function to estimate encoded base error rate.

Supplementary Table S4 provides estimates of the encoded base error rate, as a function of data size and sequencing effort (per encoded byte) relative to our experiment, based on 10^5 Monte Carlo samples (`nsamples` in Supplementary Fig. S6) per cell. This shows that, keeping sequencing effort per encoded byte constant, the error rate increases only very slowly with the increase in amount of data.

Error rate as function of data size and sequencing effort

% sequencing effort (per encoded byte) relative to experiment	Information size in byte																
	1000	1MB	1GB	1TB	1PB	1EB	1ZB	1.00E+24	1.00E+27	1.00E+30	1.00E+33	1.00E+36	1.00E+39	1.00E+42	1.00E+45	1.00E+48
1000%	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
316%	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
100%	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
31%	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10%	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3%	0	0	0	0	0	0	0	0	0	0	0	0	0	1.00E-05	0.00038	0.00479
1%	0	0	1.00E-05	4.00E-05	3.00E-05	7.00E-05	9.00E-05	0.00016	0.00054	0.00118	0.00317	0.00657	0.0145	0.03268	0.07506	0.17551
0.3%	0.01221	0.01764	0.02209	0.02725	0.03568	0.04084	0.05129	0.06733	0.08702	0.11427	0.14752	0.19205	0.24854	0.32076	0.42418	0.5758
0.1%	0.24756	0.2693	0.29461	0.31652	0.34522	0.34376	0.37356	0.40983	0.44767	0.50255	0.54485	0.59052	0.63929	0.69121	0.7514	0.81493
0.03%	0.63294	0.66131	0.67665	0.69713	0.71112	0.70758	0.72623	0.74469	0.76418	0.78336	0.80056	0.8224	0.84375	0.86439	0.88685	0.90861
0.01%	0.859	0.87701	0.88289	0.89178	0.90067	0.87071	0.8787	0.88407	0.89109	0.89681	0.90474	0.91015	0.91856	0.9251	0.93259	0.94003
0.003%	0.9575	0.95954	0.96263	0.96321	0.96711	0.92772	0.9297	0.9313	0.93422	0.9357	0.93776	0.94094	0.9441	0.94557	0.94769	0.94916
0.001%	0.97811	0.98709	0.98762	0.9881	0.98941	0.99047	0.94677	0.94583	0.9464	0.94967	0.94959	0.94987	0.95178	0.95045	0.95207	0.95358

Supplementary Table S4 | Error rate as a function of data size and sequencing effort. Percentage sequencing effort is measured relative to our experiment; the highlighted row corresponds to the same sequencing effort per encoded byte as realised in our experiment.

In contrast, the error rate depends strongly on the coverage. Supplementary Table S4 suggests that the effective coverage of our actual experiment (1,308×; see above) could be substantially lowered without impacting on the error rate. To confirm this theoretical analysis using our empirical data, we subsampled the 79.6M read-pairs at varying fractions and attempted to reconstruct the five encoded files using our original protocol. Fig. 2b (Main Text) presents results on the per-encoded-base error rate for recovery of the file `watsoncrick.pdf`, which always has at least 50 base errors due to encoded bases that were not recovered from any sequenced read (see below), for the recovery of the other four encoded files combined and for our theoretical predictions based on the analysis above. The plot shows the error rate (*y*-axis, as a percentage) as a function of subsampling percentage (*x*-axis, logarithmic scale). This indicates good agreement of our theoretical and empirical

results. The difference between the `watsoncrick.pdf` results and the other four files is explained by the unrecoverable 50 bases described above. In this case, the minimum possible error rate is 0.0036% (10 bytes not recovered, out of 280,864). The discrepancy between the theoretical and empirical curves for relatively high subsampling fractions is probably due to model violation. For example, unlike in our model, the true sampling probability of strings is almost certainly not uniform due to unequal DNA synthesis and sequencing efficiencies. Note however that the discrepancy corresponds to a difference of only a few per cent in terms of reads used, and the model remains a good approximation. The plots confirm that we could reduce sequencing coverage by a factor of 10 or even 100 without significantly impacting on our ability to recover the encoded information.

Modelling cost-effectiveness of DNA-storage. We modelled the costs of DNA-storage over time according to:

$$C_{\mathrm{D}}(t) \;=\; D_0 + F t \tag{2}$$

where $C_{\mathrm{D}}(t)$ is the cost to archive 1 MB of information for a period of t years. D_0 is the initial cost to write this information to DNA-storage — this will decrease over time as DNA synthesis technology improves — and F is the cost per year to maintain a DNA-storage facility (per MB of information stored). Storage on magnetic tape was modelled according to:

$$C_{\mathrm{T}}(t) \;=\; T_0 + F t + \sum_{i=1}^{f\,t} \left[R_{\mathrm{fix}} + \frac{T_0 + R_{\mathrm{dim}}}{\left(2^{f^{-1}/2.5}\right)^{i}} \right] \tag{3}$$

where $C_{\mathrm{T}}(t)$ is the cost to archive 1 MB. T_0 is the initial cost to write this information to tape and F is the cost per year to maintain a tape storage facility (e.g. data centre), assumed equal to the corresponding cost for a DNA-storage archive. This assumption is likely to strongly favour tape over DNA due to the costs of power and of recurrent replacement of computing and tape hardware. The parameter f is the frequency of 'tape transfer events', i.e. how often it is necessary to read and re-write the information using the current technology as the previous one becomes obsolete. Industry standards suggest f is likely to be approximately $^1/_5$–$^1/_{10}$ yr^{-1}, i.e. data must be read and re-written to new technology every 5–10 years[50,51]. The summation represents the costs (per MB) of the ft transfer events occurring in t years, each comprising fixed cost R_{fix} (e.g. finite labour cost of retrieval of existing tape archive, set-up of copying process, storage of new archive material) and diminishing costs proportional to T_0 (for new storage media) and R_{dim} (other expenses, e.g. costs proportional to time spent reading and re-writing information), both of which are assumed to halve every 2.5 years due to technological improvements[52].

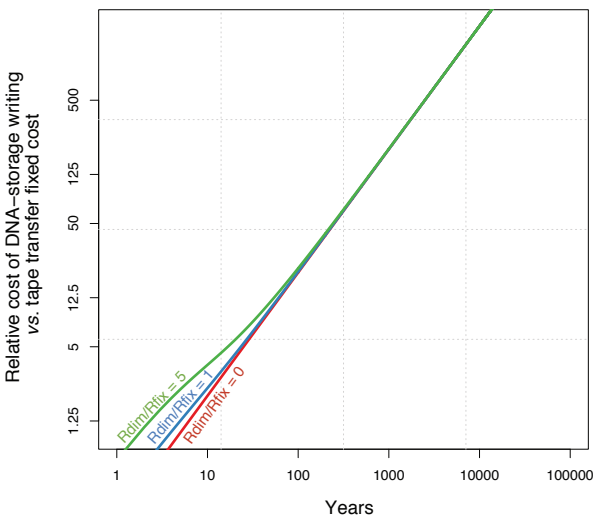
Tape costs are already very low, and so we set $T_0 = 0$. The break-even point when DNA-storage achieves the same cost as tape comes when $C_{\mathrm{D}}(t) = C_{\mathrm{T}}(t)$; equating equations (2) and (3) allows us to write:

$$\frac{D_0}{R_{\mathrm{fix}}} \;=\; f t + \frac{R_{\mathrm{dim}}}{R_{\mathrm{fix}}} \left(\frac{1 - 2^{-t/2.5}}{2^{f^{-1}/2.5} - 1} \right) \tag{4}$$

D_0/R_{fix} is the relative cost of writing DNA-storage compared to the fixed costs of tape transfer events. Equation (4) indicates the balance between values of D_0, f, t, R_{fix} and R_{dim} that leads to break-even point of DNA-storage; smaller values of D_0/R_{fix} or greater values of t correspond to conditions where DNA-storage is more cost-effective than tape; conversely, larger D_0/R_{fix} or smaller t make tape favourable.

The current commercial cost of the Agilent OLS process is approximately $25,000 for 244,000 designed oligos of length 200 bases (approximately $0.05/100 bases). We encoded 739 kB in 153,335 DNA strings of length 117 bases, leading to a value for D_0 of approximately $12,400/MB (= 25,000/(0.739 × (244,000/153,335) × (200/117))). For archives of a few megabytes, we estimate that the cost in personnel, labour and management of a corresponding tape technology transition might be of the order of $25–100, leading to a current estimate of D_0/R_{fix} in the range 125–500. Other current DNA synthesising methods, e.g. maskless photolithography, can be used to produce shorter oligos with potentially higher error rates[53], but are less expensive per base synthesised. It is possible that these could be used to reduce the cost per MB (D_0).

For realistic values of f (above), the second term on the right-hand side of equation (4) rapidly becomes small. Supplementary Fig. S7 shows the relationship of D_0/R_{fix} and break-even points (timescale on which DNA-storage and tape storage costs equate) when $R_{dim}/R_{fix} = 0$, 1 and 5 and $f = {}^1/_5$. It is clear that even for $R_{dim}/R_{fix} = 5$, which is unrealistically large, the effect of R_{dim}/R_{fix} is negligible for values of D_0/R_{fix} that are likely to be achieved in the near future (see Main Text). The same is true for $f = {}^1/_{10}$ (not shown). Consequently, we have assumed $R_{dim}/R_{fix} = 1$ for illustrative purposes. Fig. 2c (Main Text) plots D_0/R_{fix} against time t in this case, highlighting the break-even points for $f = {}^1/_5$ and $f = {}^1/_{10}$.



Supplementary Figure S7 | Effect of R_{dim}/R_{fix} on DNA-storage break-even timescale. The x-axis is the break-even time beyond which DNA-storage is less expensive than magnetic tape, assuming the tape archive has to be read and re-written every 5 years ($f = {}^1/_5$); the y-axis is the relative cost of DNA-storage synthesis and tape transfer fixed costs. Lines plotted are for $R_{dim}/R_{fix} = 0$ (red), 1 (blue) and 5 (green). Note the logarithmic scales on both axes.

Information decoding costs. In our experiment, we decoded 739 kB of information using one lane of the Illumina HiSeq 2000, at a sequencing cost of approximately $1,600. This gives a decoding cost of ~$2,200/MB (= 1,600/0.739). As shown above, we could have sequenced 10 times as much encoded information in the same run and still recovered our data. This suggests that ~$220/MB (= 2,200/10) is a reasonable approximation for the decoding costs for optimised use of existing technologies.

SUPPLEMENTARY DISCUSSION

Repair of file with missing reads. During decoding, one file (ultimately determined to be `watsoncrick.pdf`) reconstructed *in silico* at the level of DNA (prior to decoding, via base-3, to bytes) contained two regions, each of 25 bases in length, that were not recovered from any sequenced read. Given the overlapping segment structure of our encoding, each such region indicates the failure of any oligo representing any of four consecutive segments to be synthesised or sequenced successfully, as any one of four consecutive overlapping segments would have contained the bases corresponding to this location. Inspection of the two regions indicated that the non-detected bases fell within long repeats of the following 20-base motif:

5′-GAGCATCTGCAGATGCTCAT-3′

(colours used to highlight motif repeats; see below). We noticed that repeats of this motif have a self-reverse complementary pattern (Supplementary Fig. S8) and we hypothesised that long, self-reverse complementary DNA fragments might not be readily sequenced using the Illumina process. In terms of DNA synthesis, we know no reason to expect self-reverse complementary fragments to be problematic in the Agilent OLS system. The PCR conditions used for library construction involved denaturing the template oligos at 98 °C, initially for a period of 3 min and for 80 s per cycle thereafter. These conditions are more than sufficient to denature and amplify any self-annealing oligos. A further denaturing step separates the double-stranded products prior to dilution and flowcell loading, and bridge amplification should also proceed nominally to form clusters of clonally amplified library constructs.

Supplementary Figure S8 | Self-reverse complementary nature of the 20-base motif.

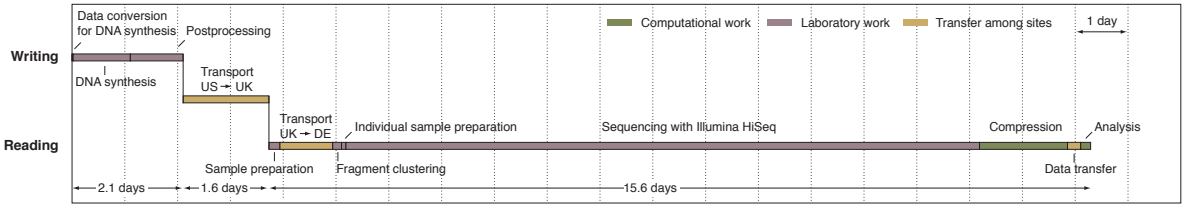
However, the subsequent sequencing conditions are intended to promote DNA hybridisation/annealing, and so in principle internal secondary structures might form and remain stable throughout the process. We therefore reason that each of the two self-complementary regions produced by our encoding scheme led to the formation of stem-loop structures within the target sequences, inhibiting the sequencing-by-synthesis reaction over these nonlinear stretches. The associated clusters would be omitted from the imaging readout, thereby resulting in a gap in the sequencing data when the files were reconstructed.

Examination of our coding methods indicated that such long repeats of the 20-base motif would arise when the original computer file contained repeats of byte value 255 (hexadecimal FF). Consequently, we modified the *in silico* reconstructed DNA sequence to repair the repeating motif pattern and subjected this to subsequent decoding steps. No further problems were encountered, and the final decoded file matched perfectly the file `watsoncrick.pdf`.

With hindsight, we should have devised a code that ensured that no long self-complementary regions existed in any of our designed DNA segments. One way to achieve this would be to pre-process the files to be encoded using a one-time pad or other stream

cipher with a standard or known key stream, leading to the DNA segments having random properties[54].

Timescale of dna-storage experiment. Supplementary Fig. S9 shows the time taken for each stage of our DNA-storage experiment. The experiment was not optimised for speed. All encoding and decoding computations were performed on one core of an Intel i5-2540M processor running at 2.60 GHz, except for the reconstruction of full-length (117 base) oligos from paired-end (104 base) reads which was performed using one core of an Intel Xeon X5650 at 2.67GHz. In a large-scale DNA-storage archive, transfer periods could be eliminated by having encoding, storage and decoding taking place at one site. Both computer software and laboratory procedures could readily be optimised and parallelised; laboratory procedures could also be automated with liquid-handling robotics for high-throughput applications[55]. Both computational and laboratory equipment are subject to continual innovation, improving their speed.



Supplementary Figure S9 | Timeline of DNA-storage experiment. We report only periods of active work on the experiment. We have omitted time taken to devise repairs for the file with two information gaps (above).

Information storage density. We recovered 757,051 bytes of information from 337 pg of DNA (above), giving an information storage density of ~2.2 PB/g (= 757,051/337 × 10^{-12}). We note that this information density is enough to store the US National Archives and Records Administration's Electronic Records Archives' 2011 total of ~100 TB (ref. 56) in < 0.05 g of DNA, the Internet Archive Wayback Machines's 2 PB archive of web sites[57] in ~1 g of DNA, and CERN's 80 PB CASTOR system for LHC data[25] in ~35 g of DNA.

REFERENCES FOR SUPPLEMENTARY INFORMATION

30. Leier, A., Richter, C., Banzhaf, W. & Rauhe, R. Cryptography with DNA binary strands. *Biosystems* **57**, 13–22 (2000)
31. Bancroft, C., Bowler, T., Bloom, B. & Clelland, C. T. Long-term storage of information in DNA. *Science* **293**, 1763–1765 (2001)
32. Wong, P. C., Wong, K.-K. & Foote, H. Organic data memory. Using the DNA approach. *Comm. ACM* **46**, 95–98 (2003)
33. Arita, M. & Ohashi, Y. Secret signatures inside genomic DNA. *Biotechnol. Prog.* **20**, 1605–1607 (2004)
34. Kashiwamura, S., Yamamoto, M., Kameda, A., Shiba, T. & Ohuchi, A. Potential for enlarging DNA memory: the validity of experimental operations of scaled-up nested primer molecular memory. *Biosystems* **80**, 99–112 (2005)

35. Skinner, G. M., Visscher, K. & Mansuripur, M. Biocompatible writing of data into DNA. *J. Bionanoscience* **1**, 17–21 (2007)
36. Yachie, N., Sekiyama, K., Sugahara, J., Ohashi, Y. & Tomita, M. Alignment-based approach for durable data storage into living organisms. *Biotechnol. Prog.* **23**, 501–505 (2007)
37. Heider, D. & Barnekow, A. DNA watermarks: a proof of concept. *BMC Mol. Biol.* **9**, 40 (2008)
38. Portney, N. G., Wu, Y., Quezada, L. K., Lonardi, S. & Ozkan, M. Length-based encoding of binary data in DNA. *Langmuir* **24**, 1613–1616 (2008)
39. CUHK iGEM 2010. Bacterial-based storage and encryption device (2010) http://2010.igem.org/Team:Hong_Kong-CUHK accessed online, 10 May 2012
40. Jarvis, C. & Netherlands Proteomics Centre. Blighted by Kenning (2012) http://www.artforeating.co.uk/restaurant/index.php?/blighted-by-ken/project-overview/ accessed online, 8 November 2012
41. Yamamoto, M., Kashiwamura, S., Ohuchi, A. & Furukawa, M. Large-scale DNA memory based on the nested PCR. *Natural Computing* **7**, 335–346 (2008)
42. Kosuri, S. *et al.* A scalable gene synthesis platform using high-fidelity DNA microchips. *Nature Biotech.* **28**, 1295–1299 (2010)
43. Beaucage, S. L. & Caruthers, M. H. Deoxynucleoside phosphoramidites — a new class of key intermediates for deoxypolynucleotide synthesis. *Tetrahedron Lett.* **22**, 1859–1862 (1981)
44. Cleary, M. A. *et al.* Production of complex nucleic acid libraries using highly parallel *in situ* oligonucleotide synthesis. *Nature Methods* **1**, 241–248 (2004)
45. Aird, D. *et al.* Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol.* **12**, R18 (2011)
46. Andrews, S. FastQC. A quality control tool for high throughput sequence data (2012) http://www.bioinformatics.babraham.ac.uk/projects/fastqc accessed online, 5 September 2012
47. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009)
48. McKenna, A. *et al*. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. **20**, 1297–1303 (2010)
49. Minoche, A. E., Dohm, J. C. & Himmelbauer, H. Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and Genome Analyzer systems. *Genome Biol*. **12**, R112 (2011)
50. Vican, I. & Stančić, H. Long-term inactive data retention through tape storage technology. *In* Stančić, H. *et al.* (eds.) *INFuture2009: the Future of Information Sciences — Digital Resources and Knowledge Sharing*. pp. 105–114. (Department of Information Sciences, University of Zagreb, 2009)
51. Klingler, S. L. Information storage technologies. *In* Bates, M. J. (ed.) *Understanding Information Retrieval Systems: Management, Types and Standards.* pp. 245–258. (CRC Press, 2012)
52. Fujifilm. Fujifilm to manufacture 5TB tape cartridge for Oracle's StorageTek T10000C drive (2011) http://www.fujifilm.com/news/n110201.html accessed online, 10 May 2012
53. Agbavwe, C. *et al.* Efficiency, error and yield in light-directed maskless synthesis of DNA microarrays. *J. Nanobiotechnology* **9**, 57 (2011)
54. Paar, C. & Pelzl, J. *Understanding Cryptography. A Textbook for Students and Practitioners.* (Springer, 2010)

55. Quail, M. A., Swerdlow, H. & Turner, D. J. Improved protocols for the Illumina genome analyzer sequencing system. *Curr. Protoc. Hum. Genet.* **62**, 18.2.1–18.2.27 (2009)
56. US National Archives & Records Administration (2012) http://www.archives.gov/era/status.html accessed online, 10 May 2012
57. Internet Archive. The Wayback Machine (2012) http://archive.org/about/faqs.php accessed online, 7 September 2012