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]

[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.

Nature 494, 77–80 (07 February 2013) doi:10.1038/nature11875
Received 15 May 2012; Accepted 12 December 2012; Published online 23 January 2013

Outline

- Motivation
- Introduction
- Materials and Methods
- Results and discussion
- Summary

Motivation

- Archiving is an increasingly complex task
- DNA as storage
  - Capacity for high-density
  - Longevity under easily achieved conditions
  - Track record as an information bearer
- Previous DNA-based storage approaches
  - Trivial amounts of information
  - Not easily scalable
  - No robust error-correction

New approach

- Scalable and reliable method
- Successfully encoded computer files totaling 739 kilobytes
- Synthesized the DNA, sequenced it and reconstructed original files with **100% accuracy**
- Can theoretically be scaled beyond current global data volumes and could be cost-effective within a decade

Introduction

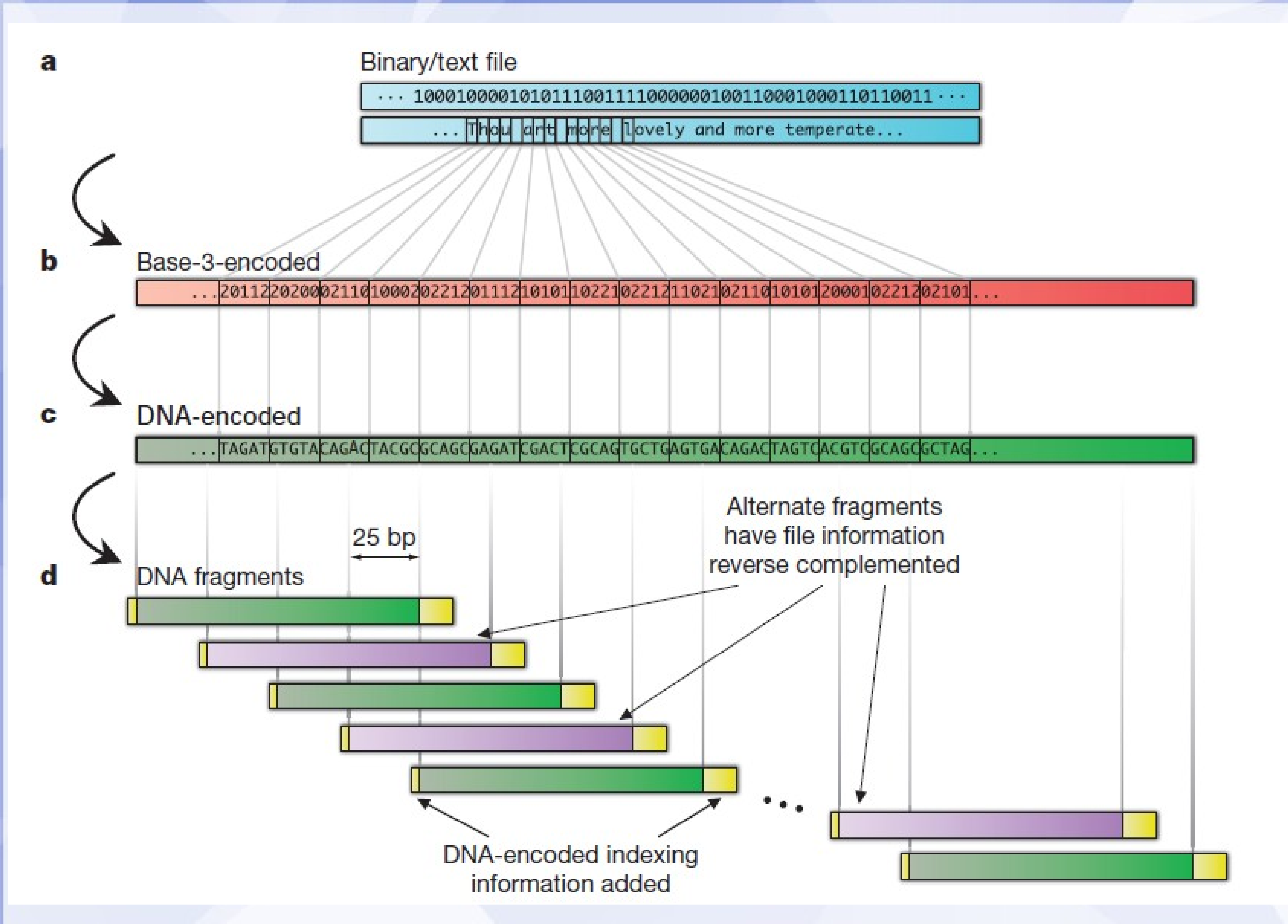
- Main challenge of existing DNA storage techniques is the “difficulty of synthesizing long sequences of DNA *de novo* to an exactly specified design.”

The Method

- Hypothetical long molecule encoded *in vitro* using shorter DNA fragments
- Long lifespan in low-maintenance environment
- Synthetic is better than living vectors
- Shorter fragments are easier to encode and manipulate

The Encoded Data

- Five files: 757,051 bytes total
  - All 154 of Shakespeare's sonnets (ASCII)
  - Classic scientific paper (PDF)
  - Medium resolution color photo (JPEG)
  - 26-s excerpt from Martin Luther King's 1963 'I have a dream' speech (MP3)
  - Huffman code used to convert bytes to base-3 digits (ASCII)



a
Binary/text file
··· 1000100001010111001111000000100110001000110110011 ···
... Thou art more lovely and more temperate...
b
Base-3-encoded
...20112202000211010002022120111210101102210221211021021101010120001022120210l...
c
DNA-encoded
...TAGATGTGTACAGACTACGCGCAGCGAGATCGACTCGCAGTGCTGAGTGACAGACTAGTCACGTCGCAGCGCTAG...
d
DNA fragments
25 bp
Alternate fragments have file information reverse complemented
DNA-encoded indexing information added

Encoding Process

- Each byte = single DNA sequence with no homopolymers (low error rates)
- Overlapping segments with fourfold redundancy
- All 5 files were represented by 153,335 strings of DNA, each 117 nucleotides
- Non-biological origin → deliberate design

Synthesis and Storage

- Agilent Technologies OLS (oligo library synthesis process)
- Synthesized DNA shipped to Germany
  - Lyophilized form
  - Ambient temperature
  - No specialized package
- Resuspended, amplified and purified
- Sequenced on the Illumina HiSeq 2000

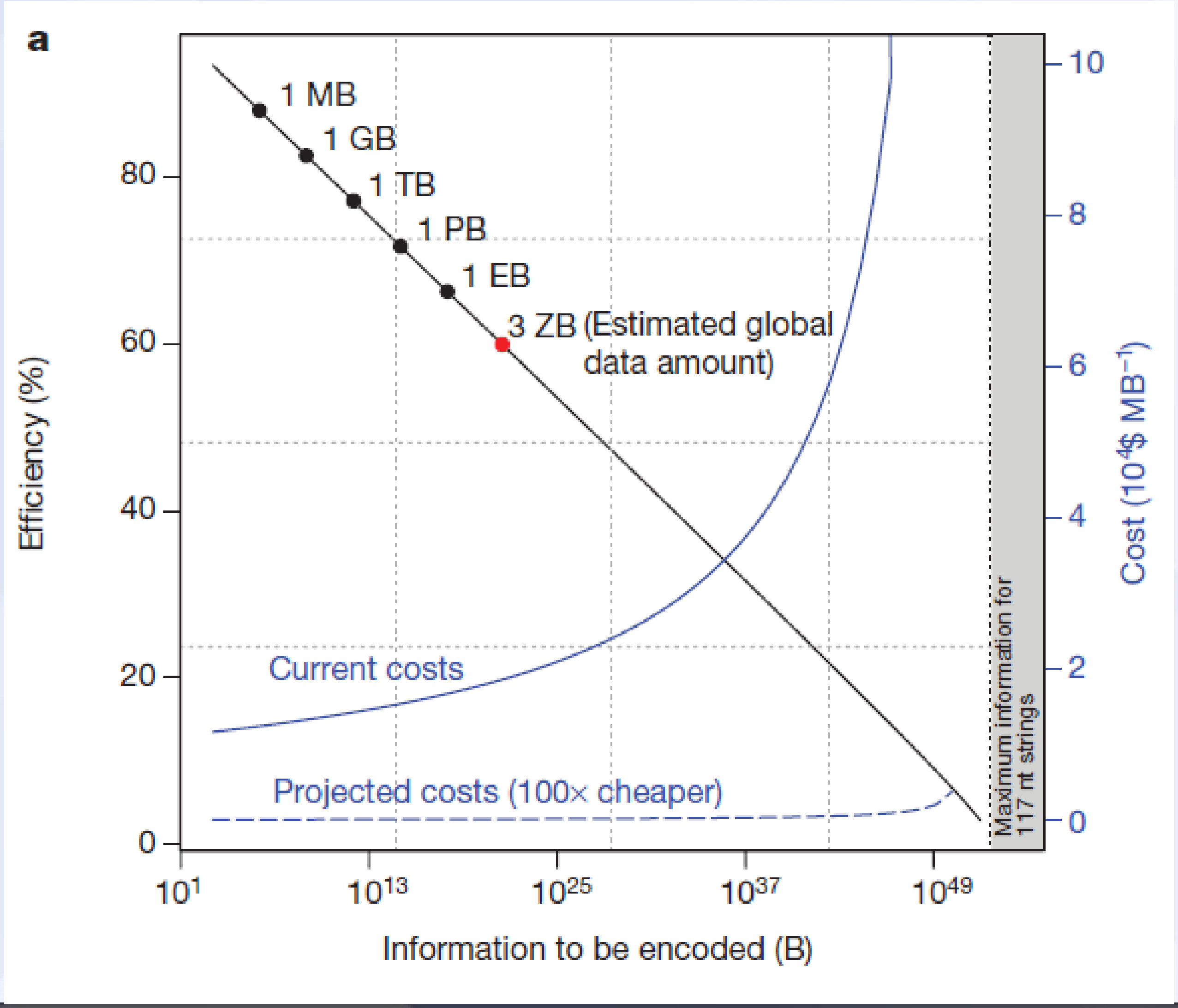
Decoding Process

- 79.6 x 10^6 read-pairs of 104 bases in length
- Full length (117 nt) DNA strings reconstructed *in silico*
- Reverse of the encoding procedure systematically discarding strings containing errors due to high level of redundancy
- The DNA sequences representing all of the files were then reconstructed *in silico*.

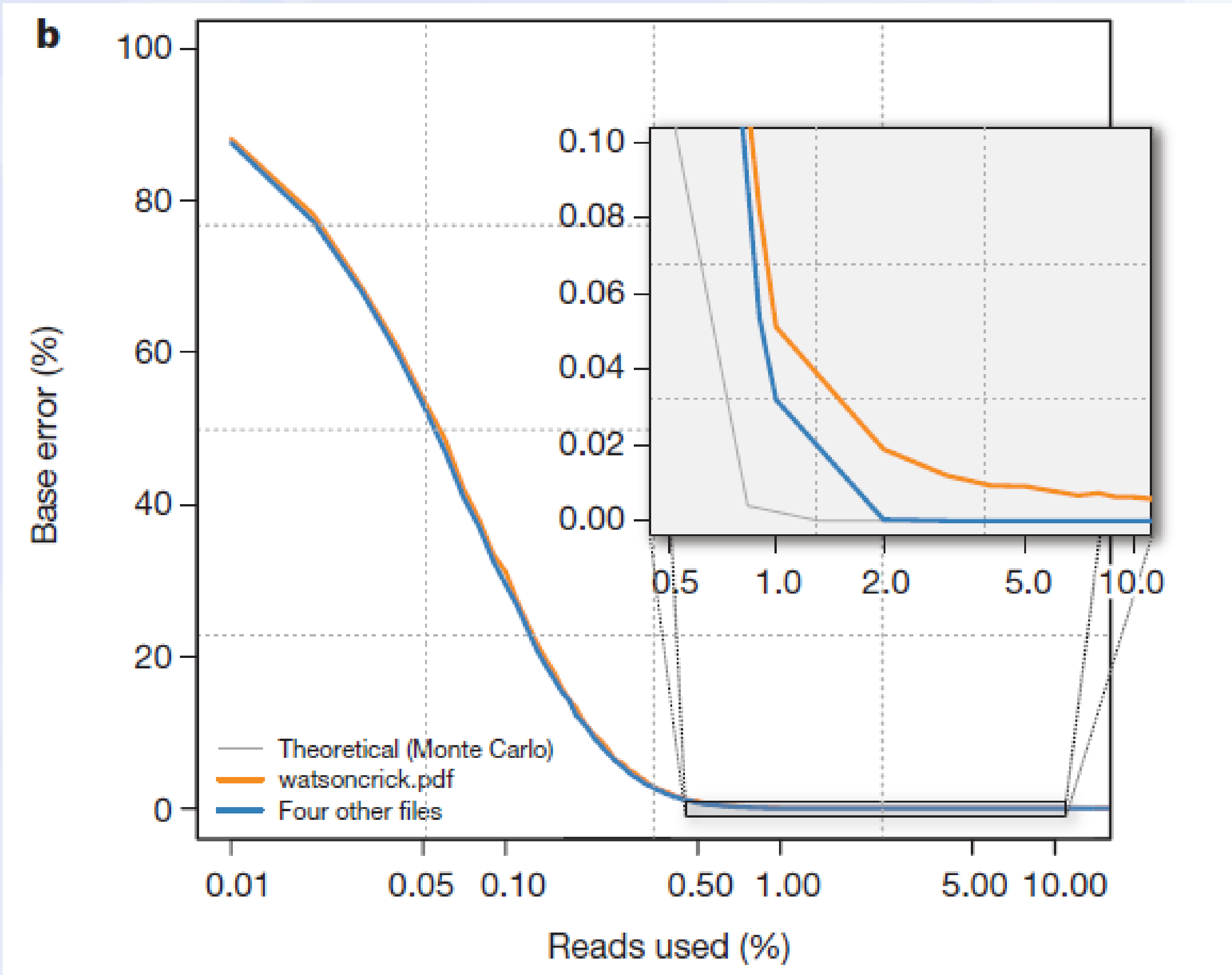
Results

- 4 out of 5 files were fully decoded without intervention
- Fifth contained 2 gaps, 25 bases each
- Inspection of neighboring regions allowed to hypothesize missing fragments
- After manually inserting 50 bases, the original file had been reconstructed with 100% accuracy

Scalability



Reliability



Summary

- DNA-based storage is a high-capacity and low-maintenance alternative to magnetic storage devices
- The new approach allows for storage density of 2.2 PB/g while consuming just 10% of the library produced from the synthesized DNA
- New methods will improve efficiency, cost and reliability, as well as storage time

The End!

- “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.”