# Synthetic double-helix faithfully stores Shakespeare's sonnets

**'Error-free' technique encodes large files in molecular form.**

**Ed Yong**

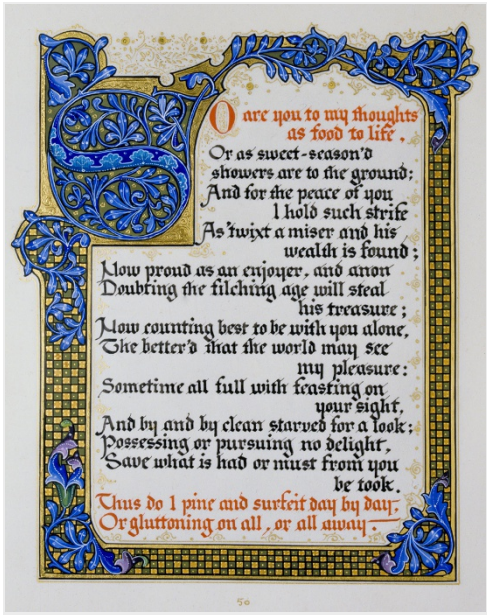
23 January 2013

A team of scientists has produced a truly concise anthology of verse by encoding all 154 of Shakespeare's sonnets in DNA. The researchers say that their technique could easily be scaled up to store all of the data in the world.

Along with the sonnets, the team encoded a 26-second audio clip from Martin Luther King's famous "I have a dream" speech, a copy of James Watson and Francis Crick's classic paper on the structure of DNA, a photo of the researchers' institute and a file that describes how the data were converted. The researchers report their results today on *Nature*'s website[1].

The project, led by Nick Goldman of the European Bioinformatics Institute (EBI) at Hinxton, UK, marks another step towards using nucleic acids as a practical way of storing information — one that is more compact and durable than current media such as hard disks or magnetic tape.

"I think it's a really important milestone," says George Church a molecular geneticist at Harvard Medical School in Boston, Massachusetts, who encoded a draft of his latest book in DNA last year[2]. "We have a real field now."




Nathan Benn / Alamy

Humanity's legacy, including Shakespeare's sonnets, may be best preserved in DNA databases.

DNA packs information into much less space than other media. For example, CERN, the European particle-physics lab near Geneva, currently stores around 90 petabytes of data on some 100 tape drives. Goldman's method could fit all of those data into 41 grams of DNA.

This information should last for millennia under cold, dry and dark conditions, says Goldman, as is evident from the recovery of readable DNA from long-extinct animals. "The experiment was done 60,000 years ago when a mammoth died and lay there in the ice," he says. "And those weren't even carefully prepared samples."

And whereas current media such as cassette tapes or compact discs become obsolete as soon as their respective players are replaced by new technology, scientists will always want to read and study DNA, Goldman says. Sequencers might change, but you can "stick the DNA in a cave in Norway for a thousand years and we'll still be able to read it". This creates enormous savings for archivists, who will not have to keep buying new equipment to rewrite their archives in the latest formats.

**Data capture**

Goldman's team encoded 5.2 million bits of information into DNA, roughly the same amount as Church's team did. But Church's team used a simple code, where the DNA bases adenine or cytosine represented zeroes, and guanine or thymine represented ones. This sometimes led to long stretches of the same letter, which is hard for sequencing machines to read and led to errors.

Goldman's group developed a more complex cipher in which every byte — a string of eight ones or zeroes — is represented by a word of five letters that are each A, C, G or T. To try to limit errors further, the team broke the DNA code into overlapping strings, each one 117 letters long with indexing information to show where it belongs in the overall code. The system encodes the data in partially overlapping strings, in such a way that any errors on one string can be cross-checked against three other strings.

Agilent Technologies in Santa Clara, California, synthesized the strings and shipped them back to the researchers, who were able to reconstruct all of the files with 100% accuracy.

The promise of extending DNA storage is largely hampered by the high cost of writing and reading DNA. The EBI team estimates that it

costs around $12,400 to encode every megabyte of data, and $220 to read it back. However, these costs are falling exponentially. The technique could soon be feasible for archives that need to be maintained long term, but that will rarely be accessed, such as CERN's data. If costs fall by 100-fold in ten years, the technique could be cost-effective if you want to store data for at least 50 years. And Church says that these estimates may be too pessimistic, as "the cost of reading and writing DNA has changed by a million-fold in the past nine years, which is unheard of even in electronics".

Goldman adds that DNA storage should be apocalypse-proof. After a hypothetical global disaster, future generations might eventually find the stores and be able to read them. "They'd quickly notice that this isn't DNA like anything they've seen," says Goldman. "There are no repeats, and everything is the same length. It's obviously not from a bacterium or a human. Maybe it's worth investigating."

*Nature* | doi:10.1038/nature.2013.12279

## References

1. Goldman, N. *et al*. *Nature* http://dx.doi.org/10.1038/nature11875 (2013).
2. Church, G. M., Gao, Y. & Kosuri, S. *Science* **337**, 1628 (2012).