was able to store only about 4 megabytes on a device that weighed more than a ton.

To enable large-scale selective access, Organick *et al.*[1] scale up a simple and elegant solution for retrieving an arbitrary file. Specifically, they attach distinct primers to each set of DNA molecules carrying information about a file. This allows them to retrieve a given file by selectively amplifying and sequencing only the molecules with the primer marking the desired file. To test their scheme, they design a primer library that allows them to uniquely tag data stored in DNA (**Fig. 1**). They encode 35 digital files into 13,448,372 DNA sequences, each 150-nucleotides long. Redundant information using error detection codes is also included to increase robustness to missing sequences and errors. The complete library of sequences is then synthesized and stored.

To improve recovery of the information, Organick *et al.*[1] develop a clustering and consensus algorithm that aligns and filters reads before error correction. In contrast to previous work[3–6], this algorithm also takes into account reads that differ from the correct length. When nanopore sequencing is used for reading information from the DNA, the clustering approach notably improves information recovery, and enables recovery even at low sequencing coverages. If, however, high-throughput sequencing is used for recovery—as it is for the most part in this work[1] as well as in previous studies[3–6]—then the gains are marginal because most reads have the correct length.

Many challenges must be solved before DNA drives run in data centers. The data longevity and information density of current DNA data storage systems already surpass those of traditional storage systems, but the cost and the read and write speeds do not. Storing one megabyte of data in DNA with existing technology[1,3–7] costs hundreds of dollars, compared with less than $0.0001 per year using tape, the standard for archival data storage. The price of DNA storage will undoubtedly drop substantially as the costs of DNA synthesis and sequencing fall. The more pressing challenge is that DNA synthesis and sequencing are inherently slow. While the millisecond access times of hard drives are not necessary for archival data storage, and DNA synthesis and sequencing DNA can be extensively parallelized, their slow speeds limit the amount of data that can be written and read in a given time interval. The bottleneck for both cost and speed is synthesis, and improved synthesis methods are key for moving forward.

Finally, current pipelines for storing data in DNA[1,3–7] illustrate the difficulty of automating all the steps required to encode and decode information in this medium. A fully automated DNA drive would include synthesis and sequencing technology, components to store and handle the DNA, as well as a supply of chemicals. In contrast, a hard drive requires only a magnet and a spinning disk.

These caveats aside, DNA data storage is an exciting field for researchers at the intersection of biology and computation. As the foundational technologies improve, the molecule encoding all biological information may one day become a robust, compact, and reliable means of digital archiving.

COMPETING FINANCIAL INTERESTS

The author declares no competing financial interests.

1. Organick, L. *et al. Nat. Biotechnol.* **36**, 242–248 (2018).
2. Wiener, N. *US News & World Report*, 02–24, 84–86 (1964).
3. Church, G.M., Gao, Y. & Kosuri, S. *Science* **337**, 1628 (2012).
4. Goldman, N. *et al. Nature* **494**, 77–80 (2013).
5. Grass, H. *et al. Chem. Int. Ed.* **54**, 2552–2555 (2015).
6. Erlich, Y. & Zielinski, D. *Science* **355**, 950–954 (2017).
7. Yazdi, S.M.H.T. *et al. Sci. Rep.* **1**, 230-248 (2015).

Research Highlights

Papers from the literature selected by the Nature Biotechnology editors. (Follow us on Twitter, @NatureBiotech #nbtHighlight)

Miniaturized neural system for chronic, local intracerebral drug delivery
Dagdeviren, C. *et al. Sci. Transl. Med.* **10**, eaan2742 (2018).

Aging and neurodegeneration are associated with increased mutations in single human neurons
Lodato, M.A. *et al. Science* **359**, 555–559 (2018).

Partial DNA-guided Cas9 enables genome editing with reduced off-target activity
Yin, H. *et al. Nat. Chem. Biol.* doi:10.1038/nchembio.2559 (2018).

Unification of protein abundance datasets yields a quantitative *Saccharomyces cerevisiae* proteome
Ho, B., Baryshnikova, A. & Brown, G.W. *Cell Syst.* doi:10.1016/j.cels.2017.12.004 (2018).

Multiplexed gene synthesis in emulsions for exploring protein functional landscapes
Plesa, C., Sidore, A.M., Lubock, N.B., Zhang, D. & Kosuri, S. *Science* **359**, 343–347 (2018).

Cloning of macaque monkeys by somatic cell nuclear transfer
Liu, Z. *et al. Cell* doi:10.1016/j.cell.2018.01.020 (2018).