Building the First Integrated DNA Storage and Compute Platform
“The biological example of writing information on a small scale has inspired me to think of something that should be possible. Biology is not simply writing information; it is doing something about it”*

-- Richard Feynman, There’s Plenty of Room at the Bottom

*“And that something is compute”
History of data storage in DNA

(Mikhail Neiman, 1964)

“The biophysical information systems and processes open favorable prospects in the direction on microminiaturisation of information storage and processing devices. These processes are, in particular, in the recording of the hereditary information in single-chain polymer molecules of DNA.”

First artificial data stored: (Joe Davis, 1988) Designed and synthesized an 18-bp message and transformed into E. coli

The first one to do it, as far as I can find, is Joe Davis, an artist collaborating with molecular biologist Dana Boyd in Jon Beckwith’s lab at Harvard Medical School. In 1988 he designed and synthesized an 18-base-pair message encoding the image of the ancient Germanic rune representing life and the female earth.

The Microvenus message was then pasted into a vector and transformed into E. coli, creating a living work of art. A little interesting tidbit about this work is that it was inspired by the Arecibo telescope message, and there is this interesting article that talks about the artist’s intentions if you care to visit it.
Our writing technology is similar movable type, we are re-using DNA oligos in a combinatorial assembly process to synthesize a large number of unique DNA molecules. This approach dramatically reduces cost and increases speed of our DNA synthesis.
Moveable type encoding unlocks write speed and compute

Components are pre-synthesized in bulk

The position of each bit is encoded in an identifier, and the presence or absence of the identifier represents the value of that bit

Identifiers are pooled to represent a data set

Component

Identifier

Identifier Library
We’ve developed a custom instrument that allows us to do this. There are 4 modules to the system. Chassis provides the substrate, a slightly hydrophobic polypropylene webbing, that traverses the entire instrument. This substrate is where DNA is dispensed and reaction drop/spots are formed. The Print Engine is an array of industrial inkjet printheads that dispense pL size volumes of each DNA component in specific locations that result in the creation of ligation reactions that assemble identifiers. The final printhead is reserved for an enzyme that catalyzes the assembly reaction. The instrument creates about 500K rxns/second, and we’ve done studies showing that each reaction can assemble up to 32 identifiers. For appropriate environmental conditions, an incubator is used. In the incubator, the reactions located on the webbing or substrate, are threaded through a series of rollers that extend the time for reactions to remain in this environment and eventually empty into a basin that pools all identifiers. This is the instrument that we’ve previously mentioned being shipped from the UK and is currently being re-built. It brought us from off-the-shelf instrumentation which was much less than Kb/s write speeds to Mb/s write speeds which is where we are currently at, and
Computing on DNA Encoded Data

- **Two Critical Aspects of DNA for Computing**
  - Random Access
  - Massive Parallelism (courtesy of easy replication)

- **Value Based Search**
  - Lower cost of retrieval (time invariant with data volume)

- **Other Types of Chemical Instructions**
  - New algorithms
  - Storage Becomes “Active”
  - Compute and storage merged into single platform
Limitless Storage

**Hyper Dense:** 1,000,000x denser than SSD (solid state drive)

**Massive redundancy:** DNA is easily replicable into multitudes of copies for simultaneous computing/query

**Ultra-persistent:** Stable for 1000s of years – once archived will last forever

Limitless Compute

**In-Storage:** Compute directly on stored data without costly movement between memory tiers

**Scale-Free:** The time and cost required to process a GB of data in DNA is the same as that required to process a PB of data

**DNA-Native:** Rely on structure / physical properties inherent in DNA to perform unique computing operations