Research overcomes key obstacles to scaling up DNA data storage

June 3 2019, by Matt Shipman

Credit: DataBase Center for Life Science

Researchers from North Carolina State University have developed new techniques for labeling and retrieving data files in DNA-based information storage systems, addressing two of the key obstacles to widespread adoption of DNA data storage technologies.

"DNA systems are attractive because of their potential information

storage density; they could theoretically store a billion times the amount of data stored in a conventional electronic device of comparable size," says James Tuck, co-corresponding author of a paper on the work and an associate professor of electrical and computer engineering at NC State.

"But two of the big challenges here are, how do you identify the strands of DNA that contain the file you are looking for? And once you identify those strands, how do you remove them so that they can be read—and do so without destroying the strands?"

"Previous work had come up with a system that appends short, 20-monomer long sequences of DNA called primer-binding sequences to the ends of DNA strands that are storing information," says Albert Keung, co-corresponding author of the paper and an assistant professor of chemical and biomolecular engineering at NC State. "You could use a small DNA primer that matches the corresponding primer-binding sequence to identify the appropriate strands that comprise your desired file. However, there are only an estimated 30,000 of these binding sequences available, which is insufficient for practical use. We wanted to find a way to overcome this limitation."

To address these problems, the researchers developed two techniques that, taken together, they call DNA Enrichment and Nested Separation, or DENSe.

The researchers tackled the file identification challenge by using two, nested primer-binding sequences. The system first identifies all of the strands containing the initial binder sequence. It then conducts a second "search" of that subset of strands to single out those strands that contain the second binder sequence.

"This increases the number of estimated file names from approximately 30,000 to approximately 900 million," Tuck says.

Once identified, the file still needs to be extracted. Existing techniques use polymerase chain reaction (PCR) to make lots (and lots) of copies of the relevant DNA strands, then sequence the entire sample. Because there are so many copies of the targeted DNA strands, their signal overwhelms the rest of the strands in the sample, making it possible to identify the targeted DNA sequence and read the file.

"That technique is not efficient, and it doesn't work if you are trying to retrieve data from a high-capacity database—there's just too much other DNA in the system," says Kyle Tomek, a Ph.D. student at NC State and co-lead author of the paper.

So the researchers took a different approach to data retrieval, attaching any of several small molecular tags to the primers being used to identify targeted DNA strands. When the primer finds the targeted DNA, it uses PCR to make a copy of the relevant DNA—and the copy is attached to the molecular tag.

The researchers also utilized magnetic microbeads coated with molecules that bind specifically to a given tag. These functionalized microbeads "grab" the tags of targeted DNA strands. The microbeads can then be retrieved with a magnet, bringing the targeted DNA with them.

"This system allows us to retrieve the DNA strands associated with a specific file without having to make many copies of each strand, while also preserving the original DNA strands in the database," Keung says.

"We've implemented the DENSe system experimentally using sample files, and have demonstrated that it can be used to store and retrieve text and image files," Keung adds.

"These techniques, when used in tandem, open the door to developing

DNA-based data storage systems with modern capacities and file-access capabilities," Tomek says.

"Next steps include scaling this up and testing the DENSe approach with larger databases," Tuck says. "A big challenge there is cost."

The paper, "Driving the Scalability of DNA-Based Information Storage Systems," is published in the journal *ACS Synthetic Biology*.

More information: Kyle J. Tomek et al. Driving the Scalability of DNA-Based Information Storage Systems, *ACS Synthetic Biology* (2019). DOI: 10.1021/acssynbio.9b00100

Provided by North Carolina State University