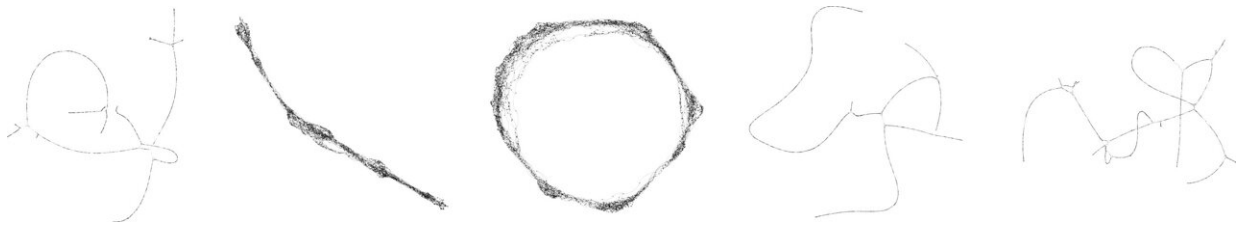
Scientists can now assemble entire genomes on their personal computers in minutes

September 14 2021



This image shows parts of the pangenome graph of 661,405 bacterial genomes, with each component representing the common structure of multiple genomes from closely-related species. Credit: Ekim et al./*Cell Systems*

Scientists at the Massachusetts Institute of Technology (MIT) and the Institut Pasteur in France have developed a technique for reconstructing whole genomes, including the human genome, on a personal computer. This technique is about a hundred times faster than current state-of-the-art approaches and uses one-fifth the resources. The study, published September 14 in the journal *Cell Systems*, allows for a more compact representation of genome data inspired by the way in which words, rather than letters, offer condensed building blocks for language models.

"We can quickly assemble entire genomes and metagenomes, including microbial genomes, on a modest laptop computer," says Bonnie Berger, the Simons Professor of Mathematics at the Computer Science and AI Lab at MIT and an author of the study. "This ability is essential in

assessing changes in the gut microbiome linked to disease and bacterial infections, such as sepsis, so that we can more rapidly treat them and save lives."

Genome assembly projects have come a long way since the Human Genome Project, which finished assembling the first complete human genome in 2003 for the cost of about $2.7 billion and more than a decade of international collaboration. But while human genome assembly projects no longer take years, they still require several days and massive computer power. Third-generation sequencing technologies offer terabytes of high-quality genomic sequences with tens of thousands of base pairs, yet genome assembly using such an immense quantity of data has proved challenging.

To approach genome assembly more efficiently than current techniques, which involve making pairwise comparisons between all possible pairs of reads, Berger and colleagues turned to language models. Building from the concept of a de Bruijn graph, a simple, efficient data structure used for genome assembly, the researchers developed a minimizer-space de Bruin graph (mdBG), which uses short sequences of nucleotides called minimizers instead of single nucleotides.

"Our minimizer-space de Bruijn graphs store only a small fraction of the total nucleotides, while preserving the overall genome structure, enabling them to be orders of magnitude more efficient than classical de Bruijn graphs," says Berger.

The researchers applied their method to assemble real HiFi data (which has almost perfect single-molecule read accuracy) for *Drosophila melanogaster* fruit flies, as well as human genome data provided by Pacific Biosciences (PacBio). When they evaluated the resulting genomes, Berger and colleagues found that their mdBG-based software required about 33 times less time and 8 times less random-access

memory (RAM) computing hardware than other genome assemblers. Their software performed genome assembly for the HiFi human data 81 times faster with 18 times less memory usage than the Peregrine assembler and 338 times faster with 19 times less memory usage than the hifiasm assembler.

Next, Berger and colleagues used their method to construct an index for a collection of 661,406 bacterial genomes, the largest collection of its kind to date. They found that the novel technique could search the entire collection for antimicrobial resistance genes in 13 minutes—a process that took 7 hours using standard sequence alignment.

"We knew our representation was efficient but did not know it would scale so well on real data, after further optimizations of the code," says Berger.

"The overall idea just works and does not require some of the usually expensive pre-processing steps, like error correction, done by most other genome assembly methods," says Rayan Chikhi, a researcher and group leader at Institut Pasteur and an author of the study.

"We can also handle sequencing data with up to 4% error rates," adds Berger. "With long-read sequencers with differing error rates rapidly dropping in price, this ability opens the door to the democratization of sequencing data analysis."

Berger notes that while the method currently performs best when processing PacBio HiFi reads, which fall well below a 1% error rate, it may soon be compatible with ultra-long reads from Oxford Nanopore, which currently has 5-12% error rates but may soon offer reads at 4%.

"We envision reaching out to field scientists to help them develop fast genomic testing sites, going beyond PCR and marker arrays which might

miss important differences between genomes," says Berger.

More information: *Cell Systems*, Ekim et al.: "Minimizer-space de Bruijn graphs: Whole-genome assembly of long reads in minutes on a personal computer" www.cell.com/cell-systems/full … 2405-4712(21)00332-X , DOI: 10.1016/j.cels.2021.08.009

Provided by Cell Press