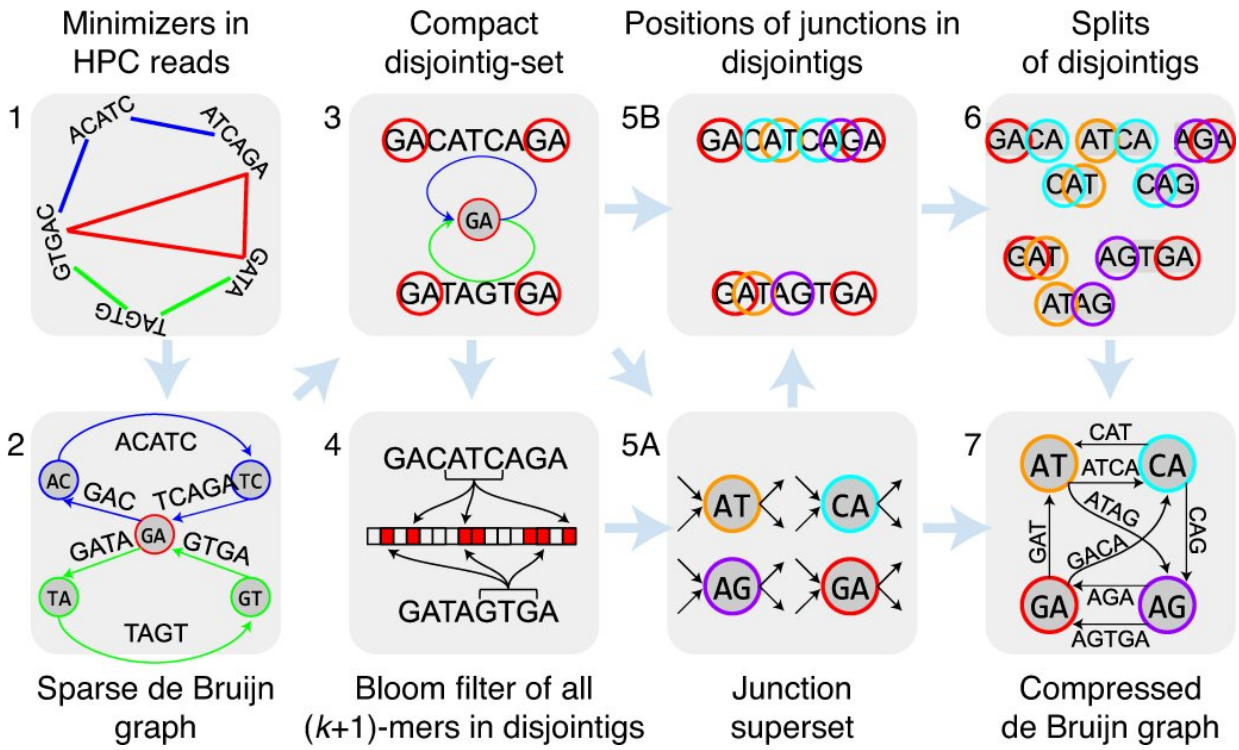
New, highly accurate algorithm scales ability to assemble complete genomes

March 11 2022



JumboDBG pipeline. Credit: *Nature Biotechnology* (2022). DOI: 10.1038/s41587-022-01220-6

An international team led by researchers at University of California San Diego's Department of Computer Science and Engineering has shown that a new genome assembly algorithm, called the La Jolla Assembler (LJA), vastly improves large genomes reconstruction, the process by

which DNA snippets are arranged into complete genomes, which is an essential aspect of genomic sequencing.

In addition, LJA significantly reduces error rates and boosts the ability to scale complete human genome assembly. This will make it easier to conduct large populations studies, in which thousands or millions of people are sequenced and their genomes compared to better understand the genetic factors that contribute to disease. The study was published this week in the journal *Nature Biotechnology*.

"We used LJA to completely reconstruct almost half of the chromosomes in the human genome in a completely automatic fashion," said Pavel Pevzner, the Ronald R. Taylor Distinguished Professor of Computer Science and senior author on the paper. "This reduced assembly errors five-fold compared to other assembly algorithms utilizing long, high-fidelity (HiFi) reads. The accuracy of this approach will bring important benefits, particularly for large population studies of complex and poorly studied regions of the human genome, such as centromeres or antibody-generating loci."

Genome assemblers are computational tools that reconstruct genomes based on a collection of shorter sequences (reads). For many years, researchers relied almost exclusively on short read technologies, which generate reads up to 300 nucleotides. These provided crucial genomic information but left gaps in genomic sequences—many in biomedically important regions. As a result, the Human Genome Project, completed two decades ago, left thousands of unassembled regions—unexplored DNA that could have clinical and scientific significance.

"This incomplete human genome assembly produced a revolution in biology and medicine 20 years ago," said Anton Bankevich, a postdoctoral researcher in the Department of Computer Science and Engineering and first author on the paper. "However, the missing pieces

of the genome may hold many more secrets."

More recently, scientists have begun adopting long, HiFi reads (greater than 10,000 nucleotides), which have helped them sequence complete human and other genomes. The first complete human genome, generated by the Telomere-to-Telomere (T2T) consortium last year, was an important milestone. However, this feat required intensive manual work and would be virtually impossible to scale to hundreds, let alone millions, of genomes.

To automate the process and increase speed and accuracy, Pevzner's team adopted a computational approach called de Bruijn graphs, which helped them assemble millions of reads into complete genomes. Originally an obscure mathematical approach invented by Dutch mathematician Nicolaas de Bruijn, this technique has become a sequencing workhorse, modeling a genome as a complex road network that connects various cities (short genomic fragments) and finding ways to traverse the network while using each road. In a way, this was history repeating itself. More than 20 years ago, Pevzner and colleagues used de Bruijn graphs to make sense of short reads.

"Although it looks like simply applying this 20-year old technique to HiFi reads would lead to excellent human genome assemblies, all previously developed algorithmic ideas fall apart when faced with constructing the enormously complex de Bruijn graph of the human genome," said Andrey Bzikadze, a graduate student in the Bioinformatics and Systems Biology Program at UC San Diego and co-author on the paper. "Reusing old methods would require a prohibitive amount of computer memory, making them impossible to implement."

LJA solves this problem, reducing the data footprint as well as assembly errors. It sets the stage for improved speed and accuracy in large population studies, in which scientists will need to assemble millions of

genomes to identify the gene sequences that confer good health or cause disease.

"Assembling a single genome is not enough to drive biological discovery," said Pevzner. "It is by comparing different genomes that scientists can understand their functions and associations with diseases. That is why we need to scale genome assembly efforts and create algorithms that produce the same quality of genome assembly as the T2T human genome but can do it automatically."

More information: Anton Bankevich et al, Multiplex de Bruijn graphs enable genome assembly from long, high-fidelity reads, *Nature Biotechnology* (2022). DOI: 10.1038/s41587-022-01220-6

Provided by University of California - San Diego