By Kelly Rae Chi

LAB TOOLS

Benching Bases

How to do heavy computational lifting in genomes and transcriptomes

You've unpacked your next-generation sequencing system and popped in some DNA or RNA. Five days later, you've sequenced 50 million tiny strings of nucleotides. Then what?

Based on their sequences, you have to align all the fragments, called "reads," with the help of a reference genome—a fully assembled sequence from the same species. In the absence of a reference, you're left with assembling the genome based solely on the portions of the reads that overlap with each other. For both alignment and assembly, "computation becomes a big issue," says Steven Salzberg, director of University of Maryland's Center for Bioinformatics and Computational Biology. "That's a huge amount of data, and in fact even streaming the data off the machine onto other computers causes network bandwidth problems."

That's because most newer technologies generate shorter reads—roughly 25 to 50 nucleotides in length—than those generated using traditional Sanger sequencing. The newer methods create smaller and more numerous DNA or RNA fragments, so software must have more muscle to piece them together. But existing software can be too slow and intolerant of errors in the reads or mismatches in alignment. Reconstructing a long stretch of mRNA comes with an added challenge: Genomic DNA can be spliced in many alternative ways, which creates many possible versions of the mRNA transcript. Thus, there is no single reference transcriptome to help you piece together the mRNA short reads, so researchers must create special bioinformatics tools to align the mRNA pieces with genomic DNA.

The good news is that a new wave of alignment and assembly software solutions has caught up to next-generation sequencing. *The Scientist* talked to some of the developers. Here's what they said:

ASSEMBLE DIVERSITY

USER: René Warren, Bioinformatician, BC Cancer Research Centre, Vancouver, British Columbia, Canada

PROJECT: Developing an approach for sequencing all the types of T-cell receptor genes present in blood

PROBLEM: Warren's group needed to capture the portion of the T-cell receptor gene responsible for generating millions of receptor variations in a healthy individual. Because that hot spot is so diverse between individual T-cell receptors—it has more than 10^{15} theoretically possible sequences—there is no single reference genome. "Basically you're doing a de novo assembly," Warren says. He needed a way to assemble the diverse genetic region, 12 to 16 nucleotides, from scratch, using short-read data.

SOLUTION: Last summer, the group developed and tested iSSAKE, software that helps them assemble the genomic hot spot. It works by finding certain reads that are part of a known gene segment—the V-gene—that neighbors the hot spot. The strategy is like that used in assembling a section puzzle of a landscape, and picking out the pieces that include the transition from one image to another, like the border between the grass and sky. "We segregated all the reads that aligned to the V-genes but had unmatched bases at the end," Warren says. "Presumably these reads would actually capture part of the [neighboring spot of interest], maybe all of it."

The group tested the algorithm on a data simulation of T-cell receptors based on GenBank data and found that for read lengths of 36 nucleotides, the method is more than 90% sensitive and more than 99.9% accurate even for relatively rare T-cell receptor types (*Bioinformatics*, 25:458-64, 2009).

CONSIDERATIONS: The group is working with wet data and has found a 96% intersect between their computational reconstruction with iSSAKE and a small sample using traditional Sanger sequences. "The challenge that remains is that there are still some errors in short-read data—this might affect the quality of the outcome," Warren says.

Download: ftp://ftp.bcgsc.ca/supplementary/iSSAKE (free)