GENOMICS

Genome builders face the competition

Three independent projects seek to contrast approaches in preparation for routine analysis of genetic data.

BY ERIKA CHECK HAYDEN

Sequencing DNA on an industrial scale is no longer difficult: the challenge is in assembling a full genome from the multitude of short, overlapping snippets that second-generation sequencing machines churn out. Researchers can call on any of two dozen computer programs to do the job, but all have their flaws. With genome sequencing fast becoming standard practice across the life sciences, researchers want to know how to choose.

The answer may come from three separate genome-assembly projects, each of which aims to test different algorithms on batches of raw sequence data and to compare the results.

There won't be any single 'winner', researchers stress. There is no consensus way to determine the absolute quality of a genome, and different assemblers might do a better job of handling different types of data.

"My dream is that a few years from now, a person who is about to do a genome project will be able to say, 'This is our budget, these are the characteristics of our genome; what is the combination of sequencing technologies and genome-assembly program that best fits our project?'" says Ian Korf at the University of California, Davis, who helped to organize the Assemblathon, one of the three genome-assembly evaluation projects.

Last December, the Assemblathon released a computer-generated human genome data set. Scientists were invited to use their assembler of choice to stitch the data into a genome. Seventeen teams from seven countries took up the challenge. Korf's team then evaluated the assemblies on the basis of commonly used criteria for the quality of genome assemblies — such as the portion of the genome that is assembled into large chunks of DNA, or contigs — as well as less-common measurements, such as how many genes each assembly is able to capture.

At a meeting last week at the University of California, Santa Cruz, three winners emerged: ALLPATHS-LG, developed by the Broad

Institute in Cambridge, Massachusetts; ABySS, developed at the British Columbia Cancer Agency's Genome Sciences Centre in Vancouver, Canada; and SOAPdenovo, developed by the Beijing Genomics Institute. But, Korf notes, "it's not just the software, it's how people are running it" that determines the quality of each assembly.

A similar genome-assembly project called dnGASP has been organized by the National Center for Genome Analysis in Barcelona, Spain. Its results are set to be discussed at a workshop on 4–7 April.

A third project, led by Steven Salzberg of the University of Maryland, College Park, is evaluating just five assemblers, among them ALLPATHS-LG and SOAPdenovo. Salzberg's group will perform and evaluate all the assemblies. In

addition, the researchers will use real genome data from four species, including the Argentine ant and the common eastern bumblebee. "With purely simulated data, you don't get a realistic picture of how these assemblers perform," says Salzberg.

Later this year, the Assemblathon will launch another round of evaluation, comparing efforts to

assemble two previously unreleased genomes, that of a parrot and a cichlid fish. And although the three current efforts are focused on data generated by the popular Illumina sequencers, new sequencing methods could become commercially available as early as next year.

Their output will differ from that of the Illumina machines; the single molecule, real-time (SMRT) technology developed by Pacific Biosciences of Menlo Park, California, for instance, produces longer reads but has higher error rates (see *Nature* **470**, 155; 2011). This creates a new challenge, says Gene Robinson, an entomologist at the University of Illinois at Urbana-Champaign, whose bee sequence data are being used by the University of Maryland project. "Biologists really want assembly algorithms that can make use of multiple forms of reads and build the best possible assembly," Robinson says.

The contest is just beginning. ■

Test case: researchers will evaluate programs assembling this bee's genome.

D. WROBEL/VISUALS UNLIMITED/CORBIS