NATURE | NEWS

Correction algorithms extend the reach of genome sequencing

Latest sequencers combined with older technology improve accuracy of genome assemblies.

Monya Baker

02 July 2012

No large genome-mapping project is ever really complete — genome assemblies usually have missing or rearranged sections, giving a partial picture of the genetic makeup of an organism. An algorithm published today in *Nature Biotechnology*[1] helps to reduce these gaps by combining data from different sequencing technologies, revealing more information than can be produced by any of these technologies alone.

Stephen Dalton/naturepl.com

The parrott genome was the first to be sequenced using a both second and third-generation sequencers working together.

‘Second-generation sequencers’ read genomes in pieces of 100–700 base pairs long, but those pieces are hard to stitch together in the correct order. ‘Third-generation’, or single-molecule sequencers such as the PacBio RS, made by Pacific Biosciences, produce reads as long as 23,000 bases, but make more errors than typical genome-analysis software can tolerate. Adam Phillippy, a bioinformatics researcher at the National Biodefense Analysis and Countermeasures Center in Frederick, Maryland, and his colleagues used short reads from second-generation Illumina or Roche 454 machines to correct errors in long single-molecule reads from the PacBio RS. They tested their correction algorithm on the genomes of *Escherichia coli* and yeast, as well as the collection of messenger RNA, or transcriptome, of maize, and found they could improve accuracy, from roughly 83% to as high as 99.9%. They also applied this hybrid correction strategy to a previously unsequenced genome of a parrot (*Melopsittacus undulatus*).

Pacific Biosciences could use some good news (see Pacific Biosciences gets sued – and that’s just business as usual). The company, based in Menlo Park, California, laid off more than a quarter of its workforce last fall in an increasingly competitive marketplace. But David Ferreiro, an analyst at Oppenheimer based in New York, thinks that an algorithm to improve accuracy will not do much to boost sales. “It's an incremental positive for [Pacific Biosciences], but it's not going to drive the market,” he says. Current sequencing technologies do a relatively good job of sequencing the protein-encoding portions of genes, which is what most researchers are interested in, and a technique that requires multiple instruments is too expensive and too complicated, he says.

But long corrected reads may have a place in more specialized applications. Elaine Mardis, co-director of the Genome Institute at Washington University in St Louis, Missouri, thinks that the algorithm might be useful for transcriptome analysis, because a single long read could encompass an entire messenger RNA. and so reveal different ways of assembling a protein. In the same issue of *Nature Biotechnology*, researchers from Pacific Biosciences describe another hybrid correction strategy for rapidly correcting bacterial genomes[2].

Phillippy says that the technique can work with long and short reads from many types of machines, and that he hopes this hybrid correction strategy will spur interest in precisely those non-coding regions of the genome that are otherwise ignored. “Normally people just get the genes out,” he says, “but you lose structural information. All of those regions have just been brushed under the rug because short reads don't interrogate them.” Some repetitive regions are still too big to be addressed with the technology available, he adds, and assembly programs will improve further as read lengths increase.

Study co-author Erich Jarvis, a neurobiologist at Duke University Medical Center in Durham, North Carolina, who studies vocal communication in birds, teamed up with Phillippy to get precisely this kind of information. Jarvis believes that some differences in vocal learning between species can be explained not by differences in the proteins certain genes encode but by differences in the amounts of protein the genes produce, which might be affected by non-coding regions within and around genes. “Without a good assembly of the regulatory regions, doing those experiments is just a fantasy.”

Nature | doi:10.1038/nature.2012.10924

References

1. Koren, S. *et al*. *Nature Biotechnol.* http://nature.com/doifinder/10.1038/nbt.2280 (2012).

2. Bashir, A. *et al*. *Nat Biotechnol.* http://www.nature.com/doifinder/10.1038/nbt.2288 (2012).