A new tool to correct DNA sequencing errors using consensus and context

September 3 2014, by Paul Greenfield

The rapid development of next-generation DNA sequencing has revolutionized biological and ecological research in the last few years. The cost of DNA sequencing has fallen dramatically, and sequencing machines are becoming a standard piece of lab equipment. Low-cost sequencing is enabling researchers to uncover the gene differences that make some people more susceptible to diseases; to explore the genetic

makeup microbial communities from the human gut or the bottom of the ocean; and to rapidly identify the organism responsible for a life-threatening infection.

But while the costs of sequencing have plummeted, the accuracy of the data produced has improved only slowly: about 1 percent of the bases generated are still called incorrectly. The bioinformatics community has responded to this problem by building specialized error correction tools that use the inherent redundancy in sequence data to find and repair miscalls and other sequencing errors. Tests have shown that incorporating the best of these error-correction tools into standard bioinformatics analytical pipelines can result in much better quality genomes and more accurately called gene variants.

However, accurately correcting errors turns out to be a difficult problem, largely because of the repetitive and ambiguous nature of genomes. It is easy to correct simple substitution errors, such as when 50 sequence reads say that a given base is an A, and only the read being corrected says it's a G. Such simple errors are well handled by downstream tools such as assemblers and aligners. The challenge is making the right correction when there are multiple plausible corrections—such as when 50 reads say A, 49 say G, and the read being corrected says T—as happens whenever reads fall across the end of a repeated region within a genome. Just to make things more challenging, this correction has to be done without any knowledge of the genomes being sequenced, and the only clues about which corrections are '"right" comes from the sequence data itself.

My colleagues and I at the Commonwealth Scientific and Industrial Research Organisation (CSIRO) have just released a new error correction tool we've developed for use by the research community. We call it "Blue." Blue is a high-performance C# application that runs natively on Windows systems, and under Mono on Linux and OS X. As

we reported in a paper published in *Bioinformatics*, test results show that Blue is significantly faster than other available tools—especially on Windows—and is also more accurate as it recursively evaluates possible alternative corrections in the context of the read being corrected.

Another uncommon feature of Blue is that it can correct all three types of possible errors (substitutions, deletions, and insertions), making it suitable for use of data produced by the Roche 454 and Life Technologies Ion Torrent systems. Blue also allows for the correction of one set of reads with a consensus derived from another set of reads, and this capability has been used to correct small numbers of long (and expensive) Roche 454 reads with a consensus derived from a large file of cheaper (but shorter) Illumina reads. This "cross-correction" method has been used very effectively to improve the quality of several reference assemblies, ranging in size from bacteria to moths and grasses.

More information: Paul Greenfield, Konsta Duesing, Alexie Papanicolaou, and Denis C. Bauer. "Blue: correcting sequencing errors using consensus and context." Bioinformatics first published online June 11, 2014 DOI: 10.1093/bioinformatics/btu368

Blue and its associated tools can be downloaded from CSIRO Bioinformatics: www.bioinformatics.csiro.au/blue/

Provided by Microsoft