Mind the technology gap

New technologies are often inherently more complex than the technologies they supplant. Users must be aware of the impact this has on data interpretation.

Over the next few months, many scientists will unwrap new sequencing machines. The February 2012 meeting of Advances in Genome Biology and Technology was abuzz with news from Oxford Nanopore Technologies that its new instrument could sequence single continuous molecules tens of kilobases long. More new types of machines and improved versions of older machines will follow. Both Illumina and Life Technologies recently announced new offerings.

Of course, initial claims in product announcements are not always borne out. Researchers will need to wait for actual data to know how these machines perform. In any case, improvements in sequencing machines will not guarantee that genome assemblies and other analyses are correct.

In fact, the question of whether or not a technology is accurate is a bit like asking whether flour tastes good. Taste is determined by how all the raw ingredients are combined and cooked. Already, the complexities of technologies such as high-throughput sequencing, mass spectrometry and super-resolution imaging can make assessment of data quality difficult, and the inherent complexity of some techniques is increasing. How a sample is prepared and data are processed and analyzed matter as much or more than the instruments themselves.

As the number of variables in implementing a method and processing the data increases, so does the complexity of the results. Unfortunately, this complexity is typically hidden behind familiar data representations that have been used for much less complex data. Assembled sequences are still long sequences of As, Cs, Gs and Ts. Super-resolution microscopy produces a familiar-looking image even if it is built up from localizing huge numbers of individual molecules, each with their own uncertainties. Scientists may understandably fail to appreciate that the figures appearing in the results section are less straightforward than they used to be.

More care needs to be taken so that the complexity behind the displayed results is understood. Even now, experts say, too few biologists pay attention to the fact that not all sequence assemblies are created equal; some are much more useful for addressing particular biological questions. Researchers probing a newly assembled genome to design follow-up experiments should remember that data quality varies. They may need explicit guidance about how to consider the data on which to draw conclusions, such as how many reads support the presence of a variant and how well all those reads agree.

Technologists can do much to help biologists out. When possible, analyzing data sets for which the answers are already established can show what combination of sample preparation and software produces reliable, reproducible outcomes. Better awareness of flaws will make for better results. Next-generation sequencing technology improved vastly once people began learning the error characteristics of each machine.

To be sure, it is not always straightforward to test new techniques on well-trodden ground. Gold standards against which to measure sequencing and analysis pipelines need to be defined for various applications. After all, old sequencing data may have been obtained on an older machine using now-defunct library-preparation protocols. The raw data may not be available, and the algorithms used in previous analyses may be outdated or performed without clearly recorded parameters. But if the genome or its variants are known and there is a renewable source of DNA that all investigators can access, scientists could run a control experiment. For this to happen, however, researchers need financial support and other incentives that make resequencing or reassembly projects attractive.

Projects aimed at development of user-friendly controls should be lauded and encouraged. Consortia are coming up with standards for structural variation, for example. And researchers have launched large-scale comparative work and community experiments. Over the last few months, groups of researchers have begun gathering to compare performance of various genome-assembly algorithms on common datasets, the better to understand strengths and weaknesses of various programs and programming parameters.

Some of these efforts are translatable to other fields and others are not, but regardless, the technologists who understand the caveats and pitfalls of the methodologies would be well served to make sure that biologists understand the complexities. Proper use of a technology with transparent data treatment is the best promotion to biologists and prevents mistakes that could damage adoption.

Journals, funding agencies and thought leaders should all support evaluation and improvement of protocols as well as the creation of new protocols. It is only as a community gains experience with a new technology that it learns to recognize and address errors. The best way to do this is by testing new techniques against established ones and, if possible, an answer key. With that compass in hand, researchers can feel more secure venturing into the wilds of using new and improved technologies.