nature biotechnology

Prepare for the deluge

The gobs of data produced by next-generation sequencing are a key problem limiting wider adoption.

Nearly three decades since Fred Sanger and Wally Gilbert received the Nobel Prize for chemistry, sequencing technology is getting a major overhaul. Cyclic array sequencing platforms technologies currently on the market dispense with the Sanger biochemistry and capillary electrophoresis-based readouts of yesteryear in favor of multi-parallel, arrayed formats that call bases via cutting-edge fluorescence imaging technology. The sheer throughput and production of base calls from each of these instruments is unprecedented, opening up new horizons in biology and new lines of experimental investigation. As illustrated by this focus—produced with the generous support of Roche—the research community's appetite for this technology continues to grow apace. The downside is that the data glut is creating a considerable market barrier for wider uptake of the technology.

Gone are the days when a researcher, using a typical Applied Biosystems' (ABI) 3730XL capillary electrophoresis sequencer, generated at most ~60 megabases per year. When it was launched in 2005, even the first iteration of Roche/454's pyrosequencer could generate as much data as more than 50 ABI capillary sequencers. And since then, the data indigestion problem has been getting worse, with other manufacturers' instruments coming on line. One single run on ABI's next-generation SOLiD (supported oligonucleotide ligation and detection) platform, for example, can produce as much as 6 gigabases of sequence.

This type of output translates into 12–15 gigabytes per run for a Roche/454 sequencer all the way up to 10 terabytes for a two-hour run on an Illumina Genome Analyzer (GAII) system. Although some manufacturers, like Applied Biosystems, provide 11.25 TB of storage with their machines, for most laboratory information management systems, the overwhelming amounts of data being produced are the equivalent of taking a drink from a fire hose.

A side effect of this is that users cannot routinely archive raw image data but must instead let company software digest them to reads, which can then be saved. This ditching of raw data after every experiment does not rest easy with many researchers. But the alternative of paying tens of thousands of dollars for a tape backup of each sequence run means simply running the sequencing experiment again might be a simpler, cheaper option. Instrument manufacturers make the argument that nothing new can be learned from looking at raw image files later. But at least, in the case of the 454 sequencer, users have been able to get better sequence from archived raw data when employing upgraded software that improves base calling and reduces error.

Beyond the data management issues, significant computational infrastructure is also required to reassemble the short reads from many of these instruments into genomic scaffolds or exons. The problem is that the instrument manufacturers are only providing the software to analyze what comes off the sequencer in a limited set of applications, such as targeted resequencing, gene expression analysis, chromatin immunoprecipitation or *de novo* genome sequencing; they don't provide any type of support for other types of downstream bioinformatics analysis. And as the research community is becoming more familiar with the potential of these platforms to interrogate biology, it is looking to ask new questions and run new types of experiments beyond the current bioinformatics.

In this respect, the open SOLiD software development community (http://solidsoftwaretools.com/gf/), which in July recruited two new software companies, is a step in the right direction. The group makes software analysis tools available under an open source license with the intention of getting the bioinformatics community to work on the code and tailor algorithms for new applications.

For users, it would also be beneficial if data formats and the statistics for comparing performance of the different instruments could be standardized. This is particularly important as the commercial environment surrounding these sequencing platforms is highly competitive at present and each manufacturer is working hard to present its own data in the best light. A related endeavor that would help better benchmark the different next-generation sequencing technologies would be to carry out an initiative similar to the Microarray Quality Control consortium where different platforms would be compared against one another as well as against DNA microarrays or quantitative PCR.

What all of this means is that for the foreseeable future, next-generation sequencing platforms may remain out of the hands of labs lacking the deep pockets needed for bioinformatics support. For many laboratories, DNA microarrays are likely to remain a cheaper alternative, even if deep sequencing provides more bang for more bucks: in transcription profiling, for example, one deep sequencing read provides not only gene-abundance information (with greater dynamic range) but also splice-variant information and SNP information (all of which would require separate DNA microarrays).

Are there practical initiatives that might help? For a start, grant-awarding bodies should start focusing on the back-end bioinformatics as much as the sequencing technology itself. And as the bioinformatics bottleneck threatens to limit instrument sales, manufacturers as a group have a massive incentive to unblock it. They might find willing allies among researchers and potential co-funders at grant-awarding bodies. But it would be important that they commit to producing community-based, rather than proprietary-based, bioinformatics solutions.

Thus, if the next-generation platforms are to truly democratize sequencing—bringing genomics out of the main sequencing centers and into the laboratories of single investigators or small academic consortia—much more effort needs to be expended in developing cost-effective software and data management solutions. At the moment, the cost of bioinformatics support threatens to keep the technology in the hands of the few. And this powerful technology clearly needs to be in the hands of many. If the data problem is not addressed, ABI's SOLiD, 454's GS FLX, Illumina's GAII or any of the other deep sequencing platforms will be destined to sit in their air-conditioned rooms like a Stradivarius without a bow.