bio[INF0]RM
The integrated informatics news source

As Next-Gen Sequencers Push IT Limits, Labs Respond With a Range of Informatics Answers

March 21, 2008
By Bernadette Toner

As more and more labs adopt next-generation sequencing technology, they are finding that the new instruments are straining their current informatics infrastructures — a challenge the groups are trying to meet with a wide range of different solutions.

Platforms like Roche/454 Life Science's GS FLX, Illumina's Genome Analyzer, and Applied Biosystems' SOLiD system all generate unprecedented quantities of sequence data, which poses IT challenges for small labs and large genome centers alike.

For smaller research groups, next-generation sequencing technology offers an affordable way to instantly gain the sequencing capacity of a genome center, but labs with only a handful of researchers are finding that they lack the bioinformatics resources necessary to properly support these systems. On the other end of the spectrum, larger groups with well-established informatics teams are finding that these new technologies are forcing them to alter how they allocate their resources.

"The bottom line is that things are turning upside down," David Dooling, assistant director of the Genome Sequencing Center at Washington University, told *BioInform*. While Sanger sequencing was "production heavy … but fairly informatics light," next-gen instruments are "the opposite," he said: "You can generate a lot of sequences with a single box, but you need a lot of informatics, you need IT, you need computation, you need storage, you need bioinformatics people, systems people, [and] programmers to support that."

Dooling said that while the GSC's headcount is pretty much the same now as it was around the time of the Human Genome Project — around 300 people — its informatics staff has increased from around 20 at that time to close to 100 now, "and will probably continue to increase as we purchase more and more of these instruments."

Meanwhile, smaller labs that don't have the option to hire more informatics staff are turning to consulting firms like the BioTeam, which has identified the next-gen sequencing market as a key growth area.

"The general trend that we're seeing is that these instruments are being provisioned to research labs that might be new to large-scale storage and large-scale computing," said Bill Van Etten, director of services at the BioTeam.

Another challenge these labs face, he said, is that "it's a new technology, and there isn't a great deal of information out there in terms of what are the requirements for a particular use case, so there's nobody to copy out in the field."

The firm is currently working with a handful of clients specifically on IT infrastructure and data-management issues linked to next-gen sequencing, including Cold Spring Harbor Lab, the WiCell Research Institute, the MIT Center for Cancer Research, the Navy Medical Research Center, Cornell University, and Emory University.

The company has identified a particular niche among labs that have found the clusters that instrument vendors provide with their next-gen sequencers are inadequate for their needs.

"The instrument manufacturers know what they need to analyze whatever files come off a sequencer, and the clusters they sell with their instruments are designed specifically for that purpose alone," said Stan Gloss, managing director of the BioTeam. "The instrument manufacturers are not in the business of helping people build a bioinformatics compute core, but some of customers, when you get out in the field, are saying, 'I'm looking downstream in the analysis and I want to do more research and I want to be able to have the instrument do this plus something else.'"

Chris Dagdigian, director of technology for the BioTeam, noted that it's important for smaller labs to understand the full IT costs associated with a typical next-gen sequencer because instrument vendors are unlikely to consider a customer's downstream computing needs when recommending a system. A lack of awareness of the total costs — including computational infrastructure, storage, and backup — could have "really bad consequences for institutions correctly budgeting," he said. "If you're off by six figures, you might buy your instrument and have no money to actually support it ongoing."

> *"You can generate a lot of sequences with a single box, but you need a lot of informatics, you need IT, you need computation, you need storage, you need bioinformatics people, systems people, [and] programmers to support that."*

Most sequencing labs — big and small — are finding that they need additional compute clusters to fully analyze the data from these instruments. Christopher Bauser, director of bioinformatics for genomics services firm GATC Biotech, said that it's not possible to rely "entirely" on the clusters that come with the systems, "but you can rely on them easily for what they're designed to do — the image analysis, the initial base calling, the initial alignment to the reference genome."

GATC, which currently has instruments from 454, Illumina, and Solexa, uses the vendor-provided computers for primary data analysis and a separate cluster for secondary and tertiary analyses, Bauser said.

Other groups are circumventing the manufacturers' computers entirely. "Currently the 454 just has a single computer that ships with the instrument, and if you want to do the analysis on the instrument, then that's time that you're not running the instrument, so we take that and we move it off the instrument and we do the analysis," Wash U's Dooling said.

He noted that 454 plans to ship a cluster with the next version of its sequencer, but it's unlikely the GSC would use that either "because it doesn't buy us any benefit. It's just another system that we'll have to manage and control and integrate with, as opposed to our current cluster, where we already have the systems in place to do that."

On the other hand, he said that Illumina's plan to ship a cluster with the next version of its sequencer for image analysis [*BioInform* 02-22-08] might be useful if it enables real-time analysis for quality control checking during the course of the run.

"If there are advantages that this on-instrument analysis provides, we would go with that," he said.

Smaller labs may not have the resources to offload all their analysis to a cluster, however. Charlie Whittaker of the bioinformatics and computing core facility at the MIT Center for Cancer Research said that his group — which includes himself and one other person — is using a 30-processor Apple Xserve cluster to process sequence date from its Illumina sequencer, but that is currently eating up about a third of the cluster's time because it takes two days to process the data.

Working with the BioTeam, Whittaker decided to move the downstream informatics work for the sequencer to a Mac Pro dual-processor quad core machine.

"The idea is a single, fully dedicated machine to run the analysis pipeline," he said. "The way we're doing it now is working fine, but it's just that [the cluster] has other bioinformatics jobs besides the sequencer, so it sometimes gets a little too busy to be ideal."

Intel Inside Your Sequencer?

A project underway at Intel might help assuage the next-gen data overload. The company is eyeing the sector as a key application area for its QuickAssist initiative, which aims to make it easier to deploy accelerated chips like field-programmable gate arrays and application-specific integrated circuits on Intel platforms.

The goal of QuickAssist is to more tightly integrate FPGAs and ASICs with the Intel platform through a direct connection to the front side bus of Intel's computers. Wilfred Pinfold, general manager of integrated analytic solutions at Intel, said that this approach is expected to offer several advantages over plugging an FPGA into a computer's PCI slot.

"If your FPGA is out on a PCI express card, the latency — the time it takes to get to the FPGA, get that problem run, and get that data back — can be significant because you have to move all the data," he said. "That chunk of time can completely drown out any benefit you get from performance."

The alternative, he said, is to move the whole program to the FPGA, but “the problem now is that a lot of things that would be a lot easier to do with the tool suite and infrastructure you have on a general purpose processor, you have to redo … on this complicated FPGA.”

With QuickAssist, “what we’ve done is moved them so close together that they share memory space,” said Pinfold. “Now all I need to do is send a ping off to the FPGA, it knows how to get to the memory space, and it will do that and then it will get out of the way again and let the GPU go on.”

Pinfold said that Intel sees an opportunity in working with sequencing vendors to speed on-instrument processing, though he stressed that Intel has no formal partnerships with any manufacturers at this time.

“Most of the sequences today are image based, and there are some interesting things you can do in image analysis that, if you could do it quickly enough, you could improve the quality of the reads,” he said. “In association with that, there’s all the assembly and alignment work, and as you do that with these shorter-read systems, could you do some of that as the machine reads, and could you improve the quality of the output by doing more of that?”

Pinfold said that Intel is working with several life-science firms in order to ensure that QuickAssist meets the demands of the genomics market. “We’re bringing in the ability to add accelerators to the platform, but we don’t want to just complicate customers’ lives — we want to simplify them greatly,” he said. “So we’re working with people like BioTeam and some of the [software vendors] in the community to see if there’s anything we can do to improve the usability of these systems in the ecosystem.”

To Store or Delete?

But a speedup in computational power won’t solve one of the biggest challenges for labs working with next-gen sequencing data: storage.

BioTeam’s Dagdigian noted that the advent of multicore hardware is enabling smaller and smaller compute clusters, which promises to do away with machine rooms altogether for some computing applications. However, even though “the footprint of the compute power you need to analyze and process this stuff is rapidly shrinking, you still have significant storage space requirements.”

Forty terabytes may only take up five inches of space, he said, “but when you’re adding other things like backup and disaster recovery, you sort of get pushed back to the traditional data center.”

MIT’s Whittaker said his lab’s 12.5 TB system is “filled to the gills” after running about one run per week on the Illumina system since it was installed in late November. “I have to start deleting image files,” he said. “I’ve just been putting it off as long as possible.”

The question of whether to delete the massive image files associated with these instruments remains unanswered for many in the community. Some labs are finding that the cost of storing data on RAID systems is approaching — and in some cases surpassing — the cost of running the experiment again. As disk space fills up, deleting the image files appears to be a viable option, but runs counter to many researchers’ instincts.

“It seems like a strange thing to do — to spend so much time and energy on the data set, and then deleting it,” Whittaker said. “But you’re really not deleting it. It’s just the raw data — it’s an intermediate step that is just too big to sit around. It’s close to 700 GB per run worth of pictures. And once you process them, you shouldn’t need to ever look at them again.”

Whittaker said that when he first received the Illumina instrument, “I didn’t have a lot of experience and was worried that maybe there was something I could do differently with those first runs to try to improve the image processing and base calling and things like that,” so he decided it was best to save the image files.

Now, however, “I’m gaining confidence and I’m going to have to start deleting soon. Some researchers want their image files and in that case I’m asking them to provide me with external drives,” he said.

Others aren’t struggling with the decision so much. “As a service provider, we do not yet have a problem with data storage,” GATC’s Bauser said.

“We don’t need to store all the data longer than a few months. Scientists come to us and say, do these experiments, generate the data, and send us the results, and when our customers are confident that they’ve got the results that they need, we have the luxury of being able to delete all of that from our system,” he said.

“Biology is a cheaper way to store biological information than a disk, but in some settings that’s not appropriate,” said BioTeam’s Van Etten. “Maybe the reagents are $5,000 to run it once, but it might be $10,000 to store it, so it might make more sense to hold onto the biology,” he said. “But if it took weeks or months for several PhD scientists to generate that sample, or if the sample is not easily replaceable, then you don’t have the opportunity to get that back. So it depends on what you’re doing.”

Wash U’s Dooling noted that while deleting images is probably the easiest way to reduce storage, “another way is to more tightly control what’s being stored where and for how long.”

He said that his group at the GSC is building its own storage infrastructure “to not only drive down costs, but also to get a little bit more control over the disk environment, the storage environment, and ensure that it is more tightly integrated with our LIMS.”

With next-generation technologies, he said, “it’s really important to tightly integrate not just the generation of the data and the tracking of those steps, but also the tracking of the analysis, and tracking the disk space and how that’s going to be used, and tracking the backups and the archives and things like that.”

That approach, he said, “will help us reduce the burdens on our backup infrastructure and reduce the needs of our storage infrastructure to grow as much as it would have if we didn’t take these steps.”

Dooling said that the GSC will be purchasing “hundreds of terabytes” for the new data center it is building [*BioInform* 11-09-07], which it plans to occupy on May 1, though he noted that the exact storage capacity for the center still hasn’t been determined.

“I have no doubt that however much we purchase, we’ll be able to use it,” he said, “because while we certainly are able to be a little bit smarter about what we store and what we don’t store as far as the raw data, the analysis on these sorts of platforms is still very much a work in progress, and those [analytical steps] can easily generate orders of magnitude more data than the raw data itself.”