



No Longer So Short: Bioinformaticians Scramble to Keep Pace with Short-read Sequencing

July 17, 2009

Newsletter: [BioInform](#)
[BioInform - July 17, 2009](#)

By [Vivien Marx](#)

Although rapid advances in sequencing technology are making it tough for bioinformatics developers to stay ahead of the curve, that isn't stopping them from trying.

At a recent meeting on algorithms for short-read analysis, several speakers noted how much the field has changed in the past year. Dubbed "short-SIG," the special interest group session held as part of the Intelligent Systems for Molecular Biology conference and European Conference on Computational Biology highlighted new developments in SNP and structural variation discovery, RNA sequencing, metagenomics, assembly, and statistics.

Some topics that were "completely hot" last year, such as alignment and assembly, are "mostly solved" now, the University of Toronto's Michael Brudno, a meeting organizer, told *BioInform*. In the case of short-read mapping, "in a year's time, we went from having almost no tools out there to having 12 or 13, of which almost all are good or very good," he said.

While last year's Short-SIG presentations focused on "basic algorithms" for read mapping and assembly, there were more talks this year on polymorphism detection, said Jens Stoye from the University of Bielefeld, another meeting organizer. This is a sign that "the field is maturing, away from technical problems that need to be solved, towards the biological and medical applications," he said.

Another change over the last 12 months is the increase in read length that second-generation sequencers can produce. "Short reads are no longer as short as they used to be," Stoye said.

The Illumina Genome Analyzer, for example, which was generating reads on the order of 30 to 40 base pairs last summer, is expected to reach the 100-base pair mark by the end of 2009 — an improvement that may require next year's meeting to be called "Mid-SIG," Brudno said.

But while read length growth is a positive development for end-users, Brudno noted that many bioinformaticians who have been focused on developing short-read algorithms over the last year are finding that they "don't just scale" to longer reads.

In the meantime, developers have been busy creating new tools for short-read analysis challenges. For example, several groups presented new algorithms for splice junction mapping, which "cannot be handled by 'classical' read mappers, where by 'classical' I mean those from 2008," Stoye said.

For example, TopHat is a new splice junction mapping algorithm for use with RNA-seq by Steven Salzberg's group at the University of Maryland's Center for Bioinformatics and Computational Biology. The scientists reported that it maps close to 2.2 million reads per CPU hour, a pace that will allow processing of an RNA-seq experiment in less than a day on a desktop computer.

Scientists from the University of Kentucky and the University of North Carolina at Chapel Hill presented another splice junction mapping tool called MapSplice, which they claim is faster than TopHat.

Another hot area of development is variation discovery. For example, SNVMix, a Bayesian mixture model developed by a research team at the British Columbia Cancer Agency, the BC Cancer Research Centre, and the University of British Columbia, was used to identify a possible "driver mutation" in a rare form of ovarian cancer, work that was published in the *New England Journal of Medicine* in June.

VARiD, a hidden Markov model-based framework developed in Brudno's lab for variation detection with Life Technologies' ABI SOLiD sequencer, demonstrated "similar performance" to ABI's Corona Lite pipeline on the same Sanger-validated data set.

Brudno said he and his team plan to release a user-friendly version of VARiD by the end of the summer. It is still in the prototyping stage, written in MatLab, so it is difficult to estimate the computational resources it will need. "We are going to re-write the whole thing in C," he said. "There is nothing in VARiD which is inherently computationally intensive," he said.

One field that has "moved by leaps and bounds" over the last year is assembly, Brudno said. Scientists from the Max Planck Institute for Developmental Biology presented their new assembly tool LOCAS, for Low Coverage Assembly Software, to handle low-coverage resequencing and to locally improve assembly by "rescuing" and incorporating unmapped and "left-over" reads to cover polymorphic regions. The software will be made available in September here, researcher Juliane Klein said in her presentation.

LOCAS is being developed for the 1001 Genomes *Arabidopsis* resequencing project as part of the ShoRE pipeline that is looking at *Arabidopsis* sequence variants and genome-wide polymorphisms. The project is a collaboration between researchers in Germany, the US, and UK.

The technique, as Klein explained, builds an alignment graph with the "overlap sequences," and the software tries to merge that subgraph with the local assembly. She presented data on a region of the first *Arabidopsis* chromosome and said that for sequencing with coverage up to 12x, LOCAS covered more of the original sequence than

Velvet.

A presentation of work by a Dutch-German team from Nijmegen Medical Center, Radboud University in Nijmegen, and the Max Planck Institute for Marine Microbiology outlined a technique for short-read metagenomic sequencing in which short reads from a population of microbial strains are mapped and assembled to create a genome that "captures the consensus" of the population's sequences.

Working the World

While sequencing vendors realize that they need to provide some informatics support with their machines, Brudno noted that manufacturers "do not consider informatics to be part of their core business," and are therefore open to collaborations with users and developers.

Neither Applied Biosystems nor Illumina have as much algorithm-development expertise and resources as "there exists in the world," Brudno said, adding that the scientific community will therefore continue to play a role in tool development.

Assim Siddiqui, director of Applied Biosystems' SOLiD bioinformatics at Life Technologies, told *BioInform* that making use of "what is out there" will help the firm provide "a consistent stable framework of software that we have validated."

This stability is particularly important for customers who may not have bioinformaticians right down the hall, he said.

ABI has a [SOLiD software development community site](#) with software tools from researchers as well as third-party vendors such as for SNP detection and copy number variation.

The company is also developing its own algorithms. For example, Jonathan Mangion, senior bioinformatics specialist for next-generation sequencing at Applied Biosystems in Oxford, UK, presented a spectral correction algorithm that he said has reduced the sequencing error rate from 4 percent to 1 percent in internal studies.

Siddiqui told *BioInform* that the firm is doing "early access testing" of the method, which is an adaptation of an algorithm developed by Pavel Pevzner of the University of California, San Diego, but will probably not release it before next year.

Also in development at ABI is a standalone software suite called BioScope that will include a plug-in architecture and a genome viewer to look at data from multiple sequencing runs, Mangion said.

BioScope, which will probably be released "later this year," will include applications for sequence and transcriptome analysis. The error-correction algorithm "will fold into this eventually," Siddiqui said.

"The idea is to have a framework that people can integrate their own pipelines into as well," Mangion said. The company will provide a set of tools, but researchers can use their own tools, a bit like the open source workflow framework Taverna.

Dirk Evers, director of computational biology at Illumina's Chesterford Research Park site,

said that Illumina scientists are working on making improvements to software tools for variant calling and "are going to probabilistic models," for such applications as SNP detection and indel and copy number variation detection.

Illumina's short-read alignment tool Eland version 2 and a new version of the Genome Analyzer's Consensus Assessment of Sequence and Variation, or CASAVA, software will both arrive "shortly," he said.

Like ABI, Illumina is also planning to create a "modular, open framework" that is expandable computationally from multi-core to grid-based application. He also mentioned plans to develop an application programming interface to allow third parties to hook into the framework.

Hogs Abound

Despite new and improved algorithms, computational challenges with short-read assembly remain, Brudno said, since tools such as Velvet are "memory hogs." Scientists tackle this drawback by buying more compute power.

While cloud computing appears to offer a solution, the challenge is that "there is insufficient bandwidth to get the data to the cloud," he said. Popping a disk into a FedEx box still is the best way to share large data sets, he said.

Genomeweb system

These settings are generally managed by the web site so you rarely need to consider them.

Issue Order: 2

