Sentieon Seeks to Build Business Around Faster, More Accurate Genomic Analysis Software

Jul 21, 2016

Uduak Grace Thomas

Premium

NEW YORK (GenomeWeb) – Mountain View, Calif.-based Sentieon is hoping to earn a livelihood from improved or redesigned versions of widely used bioinformatics programs.

The company intends to "enable precision genomics data for precision medicine" by offering more efficient implementations of existing tools for secondary genomic data analysis so that they produce comparable or more accurate results in a fraction of the time, according to Sentieon CEO Jun Ye.

Sentieon, which officially opened its doors in July 2014, has so far launched two products for calling germline mutations and for identifying somatic mutations in sequenced tumor-normal samples that use the same methodologies as the Broad Institute's Genome Analysis Toolkit and MuTect software but are much faster and in some cases more accurate than the original tools.

Its DNAseq product, for example, launched last November, offers capabilities for calling variants from FASTQ and BAM files. It uses the same mathematics as the Genome Analysis Toolkit best practice workflow using HaplotypeCaller, has a similar user interface, and produces identical results but is a much faster implementation of the methodology, according to Sentieon. "We rewrote the entire software from scratch but we implemented [it] using [a] better computer algorithm and [a] better software implantation," Ye said.

According to benchmarks from Sentieon, in terms of number of core hours needed to process data, DNAseq offers 10-fold faster variant calling from FASTQ files and a 20- to 50-fold increase in processing speed from BAM to variant call files over the standard BWA-GATK pipeline, without a corresponding increase in hardware requirements. In fact, these speedups result in a 10-fold improvement in productivity of existing compute infrastructure if the software is run on a local cluster, as well as a 10-fold reduction in computing costs if the system is run on a cloud infrastructure, according to Sentieon's calculations.

For example, DNAseq is able to process a 30x whole genome from FASTQ file to variant calls in about six hours on a single 32-core Amazon server. On the precisionFDA system, which runs the DNAnexus platform, DNAseq completes the data processing through to variant calls in less than 10 hours on a single server. If a customer wants a shorter turnaround time, DNAseq supports distributed processing, meaning users can run their analysis on multiple servers, which further reduces the FASTQ to VCF time from six hours to an hour. The company has published details of these and other benchmarks in a preprint paper that came out in *PeerJ Preprints* in January this year.

Furthermore, unlike the GATK, DNAseq does not downsample in high coverage regions, has no thread-dependencies, and its performance is consistent from one run to the next. These features make

it suitable for clinical applications, which require repeatability, and for use in applications involving samples sequenced at high coverage, Sentieon said. DNAseq also eliminates the randomness and noise inherent in the GATK implementation of the underlying mathematics, which contributes to its improved accuracy compared to the Broad's iteration. The software can also perform joint calling of a large number of samples – up to 100,000 – without requiring an intermediate file merging step, according to the company.

In April, Sentieon also began marketing TNAseq, its software solution for detecting somatic variants in tumor-normal pairs. It is similar to the Broad'sMuTect and MuTect2 software solutions, which identify somatic mutations in NGS cancer data, but is more than 10 times faster in terms of core hours, according to Sentieon. Like DNAseq, TNseq performs consistently from one run to the next with no downsampling in high coverage regions, which makes it suitable for high coverage applications like liquid biopsy analysis, the company said.

Also like DNAseq, TNseq is multi-threaded and supports distributed processing. Compared to MuTect2, which also uses multi-threaded processing, TNseq processes data over 10-fold faster. When run on a single-thread server, TNseq is able to process data more than 20 times faster than MuTect.

Sentieon charges a license fee for its software, which can be installed on local clusters, cloud platforms, or personal computers. The starting price is $1,000 per core year, meaning that if, for example, a customer wants to use the software on 50 cores over the course of a year, they will pay a list price of $50,000. That price point, Ye explained, is based on Sentieon's estimation of how many genomic jobs its software can process in a single core year, which he claims will be cheaper than running GATK on the same amount of data and will require fewer compute resources. However, the exact costs that users pay will likely vary from one customer to the next, depending on what additional capabilities and features that users want, Ye noted. The company also offers a 50 percent discount to academic users who want to license its software.

The company will update both DNAseq and TNseq as new features are added to GATK and MuTect. After new versions of both GATK and MuTect come out, it will wait a few weeks until all the initial bugs have been fixed before it will begin updating its solutions to match the changes in the new releases, Ye said, adding that the expected lag time for customers will usually be one to two months.

The company says that its products have been installed at over 100 sites worldwide where researchers have evaluated their performance on both whole-genome and whole-exome datasets.

One such user is Wing Wong, a Stanford University professor of statistics and biomedical data science. His lab develops statistical methods for analyzing large quantities of next-generation sequencing data from gene regulation and human disease studies.

"We tested Sentieon's software because we have been looking for ways to accelerate our WGS analysis pipeline, [which uses] mostly BWA and GATK," he told GenomeWeb in an email. "We are very happy with the Sentieon results as it gave us almost identical results as GATK but with six- to 15-fold speedup in various stages of the pipeline running on our cluster." Wong is now a scientific advisor to Sentieon, although he was not one at the time his lab first tested the software.

Another researcher who has used the software is Jong Bhak, a professor of biomedical engineering at the Ulsan National Institute of Science and Technology in South Korea. Bhak is also CEO and founder of two companies —Korea-based Geromics and San Diego, Calif-based Theragenomics – as well as director and founder of the Genome Research Foundation, a nonprofit research foundation in Korea.

Bhak has worked on several reference genome projects, including the first tiger and whale reference genomes and the first Korean human reference genome. He is now involved in the Korean Genome Project, which is currently sequencing about 10,000 genomes, as well as the Korean arm of the Personal Genome Project.

"Sentieon is the most reliable and fastest genetic variation detection algorithm," Bhak told GenomeWeb in an email. He also highlighted the software's reduced compute resource footprint as a key benefit. He further stated that researchers involved in the Korean Genome Project plan to use the software to analyze data from the initiative. His team will also use Sentieon's solution for tumor-normal analysis internally. "We are very happy with its performance and accuracy," he said. "It is going to be the preferred commercial tool, in my opinion."

Xiaole Shirley Liu, a professor of biostatistics and computational biology at Harvard University and director of Dana-Farber's Center for Functional Cancer Epigenetics, similarly highlighted the efficacy of the software for tumor analysis. "We are involved in a project to compare the mutational landscape of different tumor subtypes and different outcomes," she said in an email. "Sentieon's method seems to give robust mutation calling results from the tumor sequencing data but can run much faster than other widely used academic solutions."

Hongyu Zhao, chair of Yale University's biostatistics department, pointed out the speed of Sentieon's products and their accuracy as well as their consistency. His lab at Yale Medical School uses the software to analyze both whole-exome and whole-genome data. Zhao told GenomeWeb that the company's tools lethis team analyze data more quickly, freeing the team members to focus on the biological questions they want to answer.

Jeremy Edwards, a professor of molecular genetics and microbiology at the University of New Mexico School of Medicine, expressed similar sentiments. "Sentieon software is enabling us to ask questions that weren't possible with GATK due to computational bottlenecks," he told GenomeWeb. "Also, I am very excited by the new developments that provide improved cancer genome performance," he added.

Sentieon's products benefit from the depth of experience that the company's employees bring to the bioinformatics space from other industries, which have already found ways to deal with big data computational challenges, Ye told GenomeWeb. Genomics is a more recent entrant into the big data world with the recent reduction in sequencing costs but in the semiconductor space, for example, where Ye worked before moving into bioinformatics, that was a problem 10 to20 years ago, he said. Working in these areas, "we developed many technologies and techniques … to solve a math problem more efficiently [and] rigorously [and] we bring those experiences and skills into this field."

Companies like Intel have also recently begun offering optimized versions of bioinformatics tools, such as the GATK and Blast. But "our improvement is on the entire pipeline algorithm, not just on a single module, and not just by parallelization," Ye noted, adding that the company would be interested in partnering with Intel in the future.

In recent months, Sentieon has participated in various bioinformatics competitions that pit its software against other algorithms and programs and offer an opportunity for the company to showcase the benefits and features of its solutions to potential customers. In at least three challenges the company recently participated in, its solutions have outperformed more established software solutions, including commercial offerings from DNAnexus.

For example, Sentieon's products won multiple awards in each of the two challenges run by the developers of the precisionFDA initiative. In the Consistency challenge, which ran from February 25 to April 25, Sentieon's DNAseq was named the top overall performer of all the participating solutions. DNAseq also had the highest reproducibility of all the software used for the challenge, beating tools from the Broad Institute and DNAnexus, among other contenders, according to results posted on the precisionFDA site. Over the course of this challenge, Sentieon developed a new algorithm for DNAseq called the Process Matching Model, which is designed to compensate for process-specific biases during analysis.

Sentieon also won multiple awards in the second precisionFDA challenge, which ran from April 26 to May 26. For the so-called Truth challenge, participants tested their pipelines on data from an uncharacterized sample and their results were evaluated against a "truth" dataset. Here, the company's software had the best SNP recall and the highest indel precision compared to other participating solutions.

The company is now participating in one of the ICGC-TCGA DREAM mutation calling challenges focused on tumor-normal somatic variant calling. The challenge was supposed to close on April 22 but was extended to August to allow greater participation. Currently, Sentieon software is in first place in all three variant calling categories – SNV, indels, and structural variation – according to the leaderboard for the challenge.

Sentieon plans to release additional products, including tools for RNA analysis, Ye said. The company is already preparing to release a new variant detection algorithm, called Cloud9, that it is currently using for the DREAM mutation calling challenge. The solution, which will be available in September or October, will include some newly developed capabilities and functionalities beyond what is currently available in its existing products, for example a tool for structural variant detection.