MedicalXpress

Research community comes together to provide new 'gold standard' for genomic data analysis

May 18 2015

Cancer research leaders at the Ontario Institute for Cancer Research, Oregon Health & Science University, Sage Bionetworks, the distributed DREAM (Dialog for Reverse Engineering Assessment and Methods) community and The University of California Santa Cruz published the first findings of the ICGC-TCGA-DREAM Somatic Mutation Calling (SMC) Challenge today in the journal *Nature Methods*. These results provide an important new benchmark for researchers, helping to define the most accurate methods for identifying somatic mutations in cancer genomes. The results could be the first step in creating a new global standard to determine how well cancer mutations are detected.

The Challenge, which was initiated in November 2013, was an open call to the research community to address the need for accurate methods to identify cancer-associated mutations from whole-genome sequencing data. Although genomic sequencing of tumour genomes is exploding, the mutations identified in a given genome can differ by up to 50 per cent just based on how the data is analyzed.

Research teams were asked to analyze three in silico (computer simulated) tumour samples and publicly share their methods. The 248 separate analyses were contributed by teams around the world and then analyzed and compared by Challenge organizers. When combined, the analyses provide a new ensemble algorithm that outperforms any single algorithm used in genomic data analysis to date.

The authors of the paper also report a computational method, BAMSurgeon (developed by co-lead author Adam Ewing, a postdoctoral fellow in the lab of Dr. David Haussler at UC Santa Cruz), capable of producing an accurate simulation of a tumour genome. In contrast to tumour genomes from real tissue samples, the Challenge organizers had complete knowledge of all mutations within the simulated tumour genomes, allowing comprehensive assessment of the mistakes made by all submitted methods, as well as their accuracy in identifying the known mutations.

The submitted methods displayed dramatic differences in accuracy, with many achieving less than 80 per cent accuracy and some methods achieving above 90 per cent. Perhaps more surprisingly, 25 per cent of teams were able to improve their performance by at least 20 per cent just by optimizing the parameters on their existing algorithms. This suggests that differences in how existing approaches are applied are critically important - perhaps more so than the choice of the method itself.

The group also demonstrated that false positives (mutations that were predicted but didn't actually exist) were not randomly distributed in the genome but instead they were in very specific locations, and, importantly, the errors actually closely resemble mutation patterns previously believed to represent real biological signals.

"Overall these findings demonstrated that the best way to analyze a human genome is to use a pool of multiple algorithms," said co-lead author Kathleen Houlahan, a Junior Bioinformatician at the Ontario Institute for Cancer Research working with the Challenge lead, Dr. Paul Boutros. "There is a lot of value to be gained in working together. People around the world are already using the tools we've created. These are just the first findings from the Challenge, so there are many more discoveries to share with the research community as we work through

the data and analyze the results."

"Science is now a team sport. As a research community we're all on the same team against a common opponent," said Dr. Adam Margolin, Director of Computational Biology at Oregon Health & Science University and co-organizer of the challenge. "The only way we'll win is to tackle the biggest, most challenging problems as a global community, and rapidly identify and build on the best innovations that arise from anywhere. All of the top innovators participated in this Challenge, and by working together for a year, I believe we've advanced our state of knowledge far beyond the sum of our isolated efforts."

"Paul and the whole team have done something truly exceptional with this Challenge. By leveraging the SMC Challenge to establish a living community benchmark, the Challenge organizers have made it run more like an "infinite game" where the goal is no longer one of winning the Challenge but instead of constantly addressing an ever-changing horizon," said Dr. Stephen Friend, President of Sage Bionetworks. "And given the complex heterogeneity of cancer genomes and the rapid rate with which next generation sequencing technologies keep changing and evolving, this seems like an ideal approach to accelerate progress for the entire field."

"We owe it to cancer patients to interpret tumour DNA information as accurately as we can. This study represents yet another great example of harnessing the power of the open, blinded competition to take a huge step forward in fulfilling that vision," said Josh Stuart, professor of biomolecular engineering at UC Santa Cruz and a main representative of The Cancer Genome Atlas project among the authors. "We still have important work ahead of us, but accurate mutation calls will give a solid foundation to build from."

More information: Combining tumor genome simulation with

crowdsourcing to benchmark somatic single-nucleotide-variant detection, DOI: 10.1038/nmeth.3407

The Challenge: https://www.synapse.org/#!Synapse:syn312572