Published in final edited form as:
Nat Genet. 2014 April ; 46(4): 318–319. doi:10.1038/ng.2932.

Global optimization of somatic variant identification in cancer genomes with a global community challenge

Paul C Boutros[#1,2,3], **Adam D Ewing**[4], **Kyle Ellrott**[4], **Thea C Norman**[5], **Kristen K Dang**[5], **Yin Hu**[5], **Michael R Kellen**[5], **Christine Suver**[5], **J Christopher Bare**[5], **Lincoln D Stein**[1,6], **Paul T Spellman**[7], **Gustavo Stolovitzky**[8], **Stephen H Friend**[5], **Adam A Margolin**[#5], and **Joshua M Stuart**[#4]

[1]Informatics and Biocomputing Program, Ontario Institute for Cancer Research, Toronto, Ontario, Canada.

[2]Department of Medical Biophysics, University of Toronto, Toronto, Ontario, Canada.

[3]Department of Pharmacology and Toxicology, University of Toronto, Toronto, Ontario, Canada.

[4]Department of Biomolecular Engineering, University of California, Santa Cruz, Santa Cruz, California, USA.

[5]Sage Bionetworks, Seattle, Washington, USA.

[6]Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada.

[7]Department of Molecular and Medical Genetics, Knight Cancer Institute, Oregon Health & Science University, Portland, Oregon, USA.

[8]IBM Computational Biology Center, T.J. Watson Research Center, Yorktown Heights, New York, USA.

[#] These authors contributed equally to this work.

To the Editor:

Cancer is a family of diseases caused by somatic genetic mutations. Fundamental questions remain about the causes of these mutations and their roles in shaping cellular phenotypes[1,2]. For example, are there generalizable mutational profiles shared across tumor types? How do mutation rates vary with sequence and cellular context? How many and which mutations in non-exomic DNA drive tumor progression and resistance to therapy? Can a patient's genomic information be used to guide treatment? High-throughput sequencing projects, such as the recent Pan-Cancer studies[3], are discovering complex, intertwined mutagenic and selective processes.

In the face of these fundamental questions, the fields of cancer genomics and precision medicine must confront the reality that identifying somatic variants is extremely

paul.boutros@oicr.on.ca, jstuart@ucsc.edu or margolin@sagebase.org

COMPETING FINANCIAL INTERESTS The authors declare no competing financial interests.

challenging. Cancer samples are a complex mixture of normal cells of different types and multiple tumor subclones, which are combined in ways that vary spatially within individual tumors[4]. Further, the requirements of clinical care often lead to degraded or non-representative samples being used for genome sequencing. Specialized analysis techniques different from those used for germline analysis are therefore needed to dissect cancer signals from these complex and noisy data. As noted by a recent editorial in these pages[5], the accuracy and robustness of pipelines for somatic variation analysis vary dramatically[6–8].

Thus far, over 20 software solutions for somatic variation calling have been published. However, a lack of accepted benchmarks has slowed the adoption of community standards and has hindered the evolution of best-in-class methods through collaborative efforts. The two largest international cancer genomics efforts—The Cancer Genome Atlas (TCGA) and the International Cancer Genomics Consortium (ICGC)—have recently joined forces to launch the ICGC-TCGA DREAM Somatic Mutation Calling Challenge, a crowd-sourcing effort to identify the best pipelines for the detection of mutations in high-throughput sequencing reads for cancer genomes (https://www.synapse.org/#!Synapse:syn312572). The Challenge is being organized as part of the DREAM series of open challenges in computational biology[9,10] and is being run on the Sage Bionetworks Synapse platform for open computational science (http://synapse.org/). Data are hosted on a storage system donated by Hitachi and available for download via Annai Systems' GeneTorrent software. Further, to open the door for scientists without ready access to large local computer clusters, Google has made their Google Cloud Platform available to approved Challenge participants, including cost-free access to contest data and credits for Google Compute Engine. The Challenge opened for participation on 7 November 2013, and contestants will have until July 2014 to optimize their predictive models.

The Challenge will include two components. First, to help bring in researchers from diverse fields, a series of synthetic tumors of increasing complexity will be simulated and made available to any team in the world, with a live leaderboard showing top results. Second, a set of ten tumor-normal pairs from actual patients will be made available to any team, after approval by the ICGC Data Access Compliance Office. Importantly, methods will be evaluated by experimentally verifying calls on the same patient DNA used for the original sequencing. Validation will be conducted for thousands of predictions (5,000–10,000) via deep sequencing using an independent technology. Somatic single-nucleotide and structural variation prediction accuracy will be benchmarked on both synthetic and patient-derived data, providing a global picture of mutation detection accuracy.

The best-performing methods will be applied retrospectively to over 10,000 cancer genomes stored in CGHub, and the results will be distributed to the research community. Moreover, the top-scoring methods will be made available as open source tools, allowing users around the world to process their own data with the same pipelines validated and used by the ICGC and TCGA. Nature Publishing Group has stepped up to coordinate publication models stemming from the Somatic Mutation Calling Challenge. Challenge-assisted peer review and early editorial feedback will help identify publishable themes that cut across multiple approaches. The involvement of major journals introduces the possibility of reaching a

broad audience and raises the impact and exposure of contestant contributions, thereby increasing incentives and overall morale.

This Challenge will create a ‘living benchmark’ for mutation detection pipelines with the potential to continually evaluate best methods to accelerate the adoption of standards. The general platform leveraged is extensible to addressing other key problems in cancer genome analysis such as reconstructing tumor phylogeny, detecting fusion transcripts in RNA sequencing data and distinguishing driver from passenger mutations, among others. Indeed, if the Challenge framework continues its successful run, community-evolved solutions could contribute foundation stones for a wide range of precision medicine applications.

References

1. Alexandrov LB, et al. Nature. 2013; 500:415–421. [PubMed: 23945592]
2. Lawrence MS, et al. Nature. 2013; 499:214–218. [PubMed: 23770567]
3. Weinstein JN, et al. Nat. Genet. 2013; 45:1113–1120. [PubMed: 24071849]
4. Ding L, et al. Nature. 2012; 481:506–510. [PubMed: 22237025]
5. Anonymous. Nat. Genet. 2013; 45:1263. [PubMed: 24165723]
6. Alkan C, Coe BP, Eichler EE. Nat. Rev. Genet. 2011; 12:363–376. [PubMed: 21358748]
7. Kim SY, Speed TP. BMC Bioinformatics. 2013; 14:189. [PubMed: 23758877]
8. O’Rawe J, et al. Genome Med. 2013; 5:28. [PubMed: 23537139]
9. Prill RJ, et al. PLoS ONE. 2010; 5:e9202. [PubMed: 20186320]
10. Margolin AA, et al. Sci. Transl. Med. 2013; 5:181re1.