OVERVIEW

Reports from the fifth edition of CAGI: The Critical Assessment of Genome Interpretation

Gaia Andreoletti1 | Lipika R. Pal2 | John Moult2,3 | Steven E. Brenner1,4

1Department of Plant and Microbial Biology, University of California, Berkeley, California
2Institute for Bioscience and Biotechnology Research, University of Maryland, Rockville, Maryland
3Department of Cell Biology and Molecular Genetics, University of Maryland, College Park, Maryland
4Center for Computational Biology, University of California, Berkeley, California

Correspondence
John Moult, Institute for Bioscience and Biotechnology Research, University of Maryland, 9600 Gudelsky Drive, Rockville, MD 20850.
Email: jmoult@umd.edu

Steven E. Brenner, Department of Plant and Microbial Biology, University of California, Berkeley, CA 94720.
Email: brenner@compbio.berkeley.edu

Abstract

Interpretation of genomic variation plays an essential role in the analysis of cancer and monogenic disease, and increasingly also in complex trait disease, with applications ranging from basic research to clinical decisions. Many computational impact prediction methods have been developed, yet the field lacks a clear consensus on their appropriate use and interpretation. The Critical Assessment of Genome Interpretation (CAGI, /ˈkɑːdʒɪ/) is a community experiment to objectively assess computational methods for predicting the phenotypic impacts of genomic variation. CAGI participants are provided genetic variants and make blind predictions of resulting phenotype. Independent assessors evaluate the predictions by comparing with experimental and clinical data.

CAGI has completed five editions with the goals of establishing the state of art in genome interpretation and of encouraging new methodological developments. This special issue (https://onlinelibrary.wiley.com/toc/10981004/2019/40/9) comprises reports from CAGI, focusing on the fifth edition that culminated in a conference that took place 5 to 7 July 2018. CAGI5 was comprised of 14 challenges and engaged hundreds of participants from a dozen countries. This edition had a notable increase in splicing and expression regulatory variant challenges, while also continuing challenges on clinical genomics, as well as complex disease datasets and missense variants in diseases ranging from cancer to Pompe disease to schizophrenia. Full information about CAGI is at https://genomeinterpretation.org.

KEYWORDS
CAGI, Critical Assessment of Genome Interpretation, genomics, genetic variation, cancer genetics, variant impact predictors, SNP

1 INTRODUCTION

Interpretation of genome sequence variation plays a major and increasing role in both basic research and in clinical medicine. These applications necessitate robust and reliable computational approaches to aid in determining the phenotypic impact of variants. Many such methods have been developed (see Hu et al., 2019 in this issue). However, the appropriate use and accuracy of most methods have not been objectively determined.

The Critical Assessment of Genome Interpretation (CAGI, pronounced /ˈkɑːdʒɪ/) addresses this need by providing an objective evaluation of the state-of-the-art in relating human genetic variation to phenotype, particularly health. Each edition of CAGI provides about a dozen challenges to understand performance of prediction methods in a given scenario. Participants in a challenge are provided genetic variants and make predictions of resulting molecular, cellular, or organismal phenotypes. Data sets for the challenges include germline and somatic cancer variation, rare disease, common disease, and pharmacogenomics. The scale of challenges ranges from single...
nucleotides to whole genomes, as well as complementary multomics and environmental information. Variant types include those affecting expression, splicing, and amino acid sequence, and may be single base changes, insertions or deletions, and structural variation.

The CAGI cycle commences with the organizers soliciting data sets from researchers and clinicians, whose studies demonstrate relationships between genotype and phenotype. Such data sets must be fully developed (“ripe”) but not publicly available (“spoiled”) until after the CAGI prediction season. Data providers and organizers work together to develop these data sets into prediction challenges with well-defined data sets and goals. The CAGI Ethics Forum evaluates and adjusts these challenges for suitability and sensitivity. In parallel, predictors are enrolled, bound by the CAGI Data Use Agreement [https://genomeinterpretation.org/data-use-agreement], and vetted for access to the CAGI experiment with tiers reflecting the sensitivity of the data. The prediction season launches with release of the challenges and concludes with the submission of predictions from predictors. For the fifth edition of CAGI (CAGI5), the prediction season extended until May 2018. The submitted predictions are evaluated by independent assessors. Each CAGI edition culminates in a conference to discuss the outcome; the CAGI5 conference was held from 5 to 7 July 2018, for which videos and slide sets are publicly available to registered CAGI members. [https://genomeinterpretation.org/content/5-conference]

Results from the previous CAGI edition (CAGI4) were published in a special issue of this journal (Hoskins et al., 2017). The present issue contains assessment and participant papers primarily from the most recent experiment, CAGI5, which included 14 challenges and attracted participants from 12 countries.

Notable in this CAGI edition is the increasing representation of regulatory and noncoding variant challenges. Previously, CAGI has had just one small splicing challenge [https://genomeinterpretation.org/content/Splicing-2012]. CAGI5 included two full-scale splicing challenges (Mount et al., 2019) and these have resulted in five papers from participants (Chen, Lu, Zhao, & Yang, 2019; Cheng, Çelik, Nguyen, Aveec, & Gagneur, 2019; Gotoa, Margolin, & Elnitski, 2019; Naito, 2019; Wang, Wang, & Hu, 2019). The issue also contains an overview paper from one of the splicing data providers (Rhine et al., 2019). There had also been only one previous expression regulatory variant challenge, in CAGI4 (Kreimer et al., 2017). CAGI5 has a new expression regulatory challenge (Shigaki et al., 2019), and there are also two participant papers (Dong & Boyle, 2019; Kreimer, Yan, Ahituv, & Yosef, 2019).

CAGI5 continued the emphasis on the interpretation of clinically relevant large-scale sequence data, with a challenge on the risk of thrombosis in African-American cohort given whole exome sequence (Mclnnes et al., 2019; Wang & Bromberg, 2019); the identification of variants contributing to intellectual disability phenotypes given gene panel sequence (Aspromonte et al., 2019; Carraro et al., 2019; Chen, 2019); and a challenge of matching whole genome sequences to clinical profiles for patients at Toronto’s Hospital for Sick Children (SickKids) and identifying causal variants (Kasak, Hunter et al., 2019; Pal, Kundu, Yin, & Moul, 2019). The latter challenge is related to the CAGI4 SickKids challenge, also described in the assessment paper here.

Over all CAGI editions, the plurality of challenges have been on the interpretation of isolated missense variants, and CAGI5 continues that trend. There are assessment, data provider, and participant papers for the prediction of the destabilizing effect of missense mutations in a cancer-relevant protein (Frataxin, with biophysical measurements of protein stability; Petrosino et al., 2019; Savojardo, Petrosino et al., 2019; Strokach, Corbi-Verge, & Kim, 2019); on the effect of missense changes in a human calmodulin, assayed using a high-throughput yeast complementation assay (Zhang et al., 2019); the effect of missense mutations related to schizophrenia in human Pericentriolar Material 1 (PCM1), using a zebrafish development model (Miller, Wang, & Bromberg, 2019; Monzon et al., 2019); the effect of missense mutations in two cancer-related proteins, PTEN and TPMT, on intracellular protein levels, measured in a high-throughput assay (Pejaver et al., 2019); and the effect of missense changes in a monogenic disease related protein, acid alpha-glucosidase (GAA), with measurements of total intracellular enzyme activity (Adhikari, 2019). Three participant papers describe results on all the missense challenges (Garg & Pal, 2019; Katsonis & Lichtarge, 2019; Savojardo, Babbì et al., 2019). The issue also contains assessment articles from two earlier missense challenges on monogenic disease related proteins: N-acetyl-glucosaminidase (NAGLU; Clark et al., 2019), with total intracellular enzyme activity measured; and cystathionine beta-synthase (CBS), using the metric of yeast growth in a complication assay (Kasak, Bakolitsa et al., 2019).

In addition to the other cancer-related challenges outlined above, there are two that required prediction of the pathogenicity of germline variants in cancer-related proteins: one for breast cancer risk from variants in BRCA1 and BRCA2 as characterized by the ENIGMA consortium (Cao et al., 2019; Cline et al., 2019; Padilla et al., 2019; Parsons et al., 2019), and the other for cancer risk of variants in CHEK2 in Latina breast cancer cases and ancestry matched controls (Voskanian et al., 2019).

CAGI5 will continue to bear fruit. This edition introduced a time-based challenge in association with dbNSFP, whereby CAGI accepted predictions for all possible missense variants in the genome. The results are to be vetted periodically in the future as the impact of some of these variants are experimentally or clinically established. Additional papers on the challenges and other aspects of CAGI are presently in development and will be added to the CAGI5 papers collection at Human Mutation [https://onlinelibrary.wiley.com/toc/10981004/2019/40/9].

Full information about CAGI5 and earlier editions is at [https://genomeinterpretation.org].

ACKNOWLEDGMENTS

The primary contributors to CAGI are predictors, and without their willingness to test their methods in this way, the experiments would not be possible. Participants over the five CAGI editions are: Allison


We offer great thanks to the CAGI assessors who have devoted extensive time to analysis and evaluation of the results: Aashish Adhikari, Michael A. Beer, Yana Bromberg, Emidio Capriotti, Marco Carraro, John-Marc Chandria, Rui Chen, Luigi Chicco, Wyatt T. Clark, Melissa Cline, Roxana Daneshjou, Roland Dunbrack, Iddo Friedberg, Mabel Furutuki, Gad Getz, Manuel Giolli, Nick Grishin, Maricel Kann, Rachel Karchin, Anat Kreimer, Greg McInnes, M. Stephen Meyn, Sean D. Mooney, Alexander A. Morgan, John Moult, Steve Mount, Robert L. Nussbaum, Ayoti Patra, Francesco Reggiani, Jeremy Sanford, David B. Sears, Dustin Shigaki, Artem Sokolov, Josh Stuart, Shamal Sunayev, Sean Tavtigian, Silvio C. E. Tosatto, Alin Voskanian, QiFang Xu, and Nir Yosef.

Our advisory board and scientific council members have been the source of much wisdom and assistance. The board has comprised: Russ Altman, Serafim Batzoglou, George Church, Tim Hubbard, Scott Kahn, Rachel Karchin, Sean D. Mooney, Pauline Ng, Robert L. Nussbaum, Susanna Repo, John Shon, and Michael Snyder; our Scientific Council
additionally has included: Patricia Babbitt, Yana Bromberg, Atul Butte, Garry R. Cutting, Laura Elnitski, Reece Hart, Ryan Hernandez, Michael Snyder, Shamil Sunyaev, Joris Veltman, Liping Wei, and Justin Zook. The ethicists and patient advocates of the CAGI Ethics Forum have worked with organizers, undertaking a critical role in ensuring CAGI operates on a firm ethical foundation and informing its activities: Wylie Burke, Larry Carr, Flavia Chen, Steve Grossman, Julie Harris-Wai, Kirsten Isgro, Barbara A. Koenig, Selena Martinez, Robert L. Nussbaum, Jodi Paik, and Mark Yarborough.

We also thank those who have helped organize the CAGI experiments and who assisted with the associated technology: Daniel Barsky, John-Marc Chandonia, Ajithavalli Chellappran, Flavia Chen, Navya Dabbiru, Reece Hart (who coined the term “CAGI”), Roger Hoskins, Naveen Kolippara, Melissa K. Ly, Andrew J. Neumann, Gaurav Pandey, Sadhna Rana, Susanna Repo, Rajgopal Srinivasan, Stephen Yee, Sri Jyothsna Yeleswarapu, and Maya Zuhl. We specially thank those who helped in writing several papers for CAGI5 special issue: Constantina Bakolitsa, Zhiqiang Hu, and Laura Kasak.

We are truly grateful to NHGRI and NCI for the support of the CAGI experiment and conference. We especially recognize Tata Consultancy Services (TCS), which has been a collaborator in organizing all the CAGI experiments. We also thank the members of the Brenner and Moul laboratories who have contributed to CAGI.

For this special issue we greatly appreciate the efforts of the anonymous peer reviewers, the Lead Guest Editor Rachel Karchin, and the Consulting Guest Editors who oversaw the peer-review process: Tim Hubbard, Sean D. Mooney, Robert L. Nussbaum, Predrag Radivojac, Shamil Sunyaev, and Joris Veltman. John Moult and Steven E. Brenner were the Issue Editors, and Lipika R. Pal was the Organizing Editor for this special issue of Human Mutation. We thank Mohammad Israr, Christine Murray, Mark Paalman, and Allain Selga for their work in coordinating the editorial and production operations at Wiley.

Finally, we also wish to acknowledge our great debt to the many individuals who shared their private genetic and phenotypic or clinical information as participants in the research studies and clinical data sets included in the CAGI challenges.

CONFLICT OF INTERESTS

There is no conflict of interest involved in this study.

FUNDING

The CAGI experiment coordination is supported by NIH / NHGRI and NCI U41 HG007346 and the CAGI conference by R13 HG006650. Support has also been provided via a research agreement with Tata Consultancy Services (TCS).

ORCID

Gaia Andreoletti https://orcid.org/0000-0002-0452-0009
Lipika R. Pal https://orcid.org/0000-0002-3390-110X

John Moult http://orcid.org/0000-0002-3012-2282
Steven E. Brenner http://orcid.org/0000-0001-7559-6185

REFERENCES

Cline, M. S., Babbì, G., Bonache, S., Cao, Y., Casadio, R., ... de la Cruz, X. ENIGMA Consortium (2019). Assessment of blind predictions of the clinical significance of BRCA1 and BRCA2 variants. Human Mutation, https://doi.org/10.1002/humu.23861


