**CAGI SickKids challenges: Assessment of bioinformatic strategies for predicting patients’ phenotypes and identifying potential pathogenic variants and disease genes from their genomes**

M. Meyn 1,2; L. Kasak 3,4; J.M. Hunter 5; R. Udani 5; Z. Hu 3; C. Bakolitsa 3; S.E. Brenner 3; The CAGI SickKids challenge predictors

**Affiliations:**  
1) Center for Human Genomics and Precision Medicine, University of Wisconsin School of Medicine and Public Health, Madison, Wisconsin.;

2) Department of Paediatrics, The Hospital for Sick Children, Toronto, Canada;

3) Department of Plant and Microbial Biology, Univ of California, Berkeley, California;

4) Institute of Biomedicine and Translational Medicine, University of Tartu, Tartu, Estonia;

5) Department of Paediatrics and Wisconsin State State Laboratory of Hygiene, University of Wisconsin, Madison, Wisconsin

Most patients who undergo whole genome sequencing (WGS) lack a molecular diagnosis, largely due to the large number of undiscovered disease genes and inability to assess the pathogenicity of most genomic variants. The CAGI SickKids challenges were public competitions designed to address this knowledge gap by assessing state-of-the-art methods for predicting clinical phenotypes from genomes.  
  
CAGI4 and CAGI5 participants were provided with WGS data and clinical descriptions for 25 and 24 patients respectively who had undergone WGS but remained undiagnosed after evaluation by the SickKids Genome Clinic project (NPJ Genom Med 2016 13:12016 and Genet Med 20:435 2018). Predictors were asked to identify primary and secondary causal variants. Additionally, for CAGI5, groups attempted to match each genome to one of three classes of disease (neurologic, ophthalmologic and connective), and separately to each patient. A team of clinical and molecular geneticists assessed group submissions.  
  
The four groups participating in CAGI4 proposed 191 diagnostic variants for 25 patients. Most variants were nominated by only one group and did not fully explain the patient’s phenotype. However, two of the ten variants proposed by two groups were deemed to be diagnostic.  
  
The CAGI5 challenge was more difficult, as genomes were not linked to patients’ clinical descriptions. The eight groups participating in CAGI5 all prioritized how well a candidate gene explained a phenotype over the pathogenicity of its variants. They did no better than chance in matching genomes to the three disease categories but two groups performed significantly better than chance in matching genomes to specific patients. Importantly, patients with ophthalmologic disorders were the most likely to be matched to their genomes and had the strongest correlation between the informational content of the clinical description and their chance of matching. Despite partial success in matching genomes to patients, no candidate variants in CAGI5 were nominated by more than one group and none were deemed to be clinically diagnostic. However, several nominated variants of unknown significance are candidates for phenotype expansion.  
  
Despite their clinical naïveté, the bioinformaticians solved two cases and identified multiple plausible causal variants in candidate disease genes that bear further evaluation. We will discuss implications for improving in silico assessm-ent of genomic variants and identifying new disease genes