National Human Genome Research Institute

Genomic Data Science Working Group Report

September 1, 2020

The NHGRI Genomic Data Science Working Group (GDSWG) is a subcommittee of the National Advisory Council for Human Genome Research (NACHGR) established in 2017 to facilitate a deeper engagement of the NACHGR in the numerous and increasingly complex issues at the interface between genomics and data science. The working group has been charged with providing input to the NHGRI Director about relevant trans-NIH issues related to data science, for which NHGRI is frequently called upon for leadership and expertise. The general remit is deliberately broad, with the intent for the working group to be poised to consider the full spectrum of data science challenges (from data management to data analyses, including policy-related elements associated to data sharing, security, etc.) as they relate to all areas of genomics, from basic science to genomic medicine implementation.

The current GDSWG roster include (as of August 1, 2020):

- Gill Bejerano, Stanford University
- Michael Boehnke, University of Michigan School of Public Health
- Eric Boerwinkle, UTHealth School of Public Health
- Lon Cardon, BioMarin
- Nancy Cox, Vanderbilt University
- Mark Craven, University of Wisconsin-Madison
- George Hripcsak, Columbia University
- Trey Ideker, University of California, San Diego
- Gail Jarvik, University of Washington Medical Center
- Mark Johnston, University of Colorado School of Medicine
- Anthony Philippakis, The Broad Institute of MIT and Harvard

NHGRI staff:

- Eric Green, NHGRI Director
- Carolyn Hutter, Director, Division of Genome Sciences
- Valentina Di Francesco, Program Director, Division of Genome Sciences
- Natalie Kucher, Executive Secretary

Since the last GDSWG report to the NHGRI Council in February 2019, the GDSWG discussed the following topics:

a. NHGRI 2020 Strategic Plan Related Activities

NHGRI has been keeping the GDSWG appraised about the Strategic Plan related events organized by the NHGRI Data Science Focus Group, and shared with the group documents and other materials that were used as the foundation for the soon to be released Strategic Plan. As

the various themes in the data science landscape took shape, the group provided input on NHGRI's opportunities to lead in genomic data science. A number of priority areas for NHGRI in genomic data science were identified, such as algorithms and computational tools development, approaches to facilitate genomic data sharing and genomic medicine, and interdisciplinary training. The GDSWG also encouraged NHGRI to consider addressing genomic data science questions ahead of or in parallel with data generating research. Of particular interest to the GDSWG was how technologies like machine learning could be either adopted or driven forward in genomics by NHGRI.

b. Machine Learning and Artificial Intelligence in Genomics

Dr. Kundaje Anshul (Stanford University) is a member of the NIH Advisory Council to the Director (ACD) Artificial Intelligence Working Group (AIWG) and was invited by the GDSWG to present the recently released AIWG Report. Based in part on the report recommendations, members of the group urged NHGRI to identify a set of benchmarks for machine learning methods and to drive forward solutions for making datasets "AI-ready".

The GDSWG outlined the barriers and needs that must be filled to apply these technologies to genomics. They recognized the importance of education at the intersection of machine learning, biology, and medicine to grow the field and the difficulty in attracting and retaining individuals with this mix of expertise.

The group identified opportunities to use machine learning methods in existing NHGRI research areas, such as the analysis of the non-coding regions of the genome, prediction of genome functionality, and the association of genomic and phenotypic data from EHRs. The group emphasized that the evaluation and reduction of biases in the training sets for these methods will have to be carefully taken into account for the development of machine learning for genomics.

The group also discussed the concept for a new NIH Common Fund initiative in artificial intelligence that was approved by the NIH Council of Councils in May 2020. NHGRI is one of the NIH Institute Leads, together with NLM and NIBIB. The group recommended that the NHGRI carefully evaluate whether the genomic research needs for machine learning will be adequately addressed under the Common Fund initiative, or whether the Institute should consider an independent initiative for machine learning in genomics.

c. Machine Learning for Genomics Workshop

These discussions led the idea of bringing the genomics community together for a workshop. The goal of the workshop is to identify key genomic focus areas for NHGRI to invest in the development of machine learning techniques. The workshop is expected to identify scientific, technical, and ELSI areas, to stimulate and define the training and educational needs in these areas. To bring together the relevant genomic and data science expertise, the workshop will

include participants from a wide array of career levels and fields. The GDSWG is serving as the organizing committee and is targeting a virtual workshop in spring 2021.

There was also excitement about hosting a hands-on hackathon-style workshop to engage participants with genomic research questions that could be addressed with machine learning techniques using large datasets. Such an event could excite the community and directly involve younger computational scientists. However, this idea was not pursued further by the GDSWG given the restrictions for in-person meetings during the SARS-CoV-2 pandemic.

d. Other topics

Throughout the year, the GDSWG briefly discussed other genomic data science activities at NHGRI, including the increasing need for NHGRI funded programs and initiatives to more systematically and widely share metadata and phenotypic information in addition to sequence data; updates about the organizational plan for the NHGRI-funded Model Organism Databases and the Alliance of Genome Resources; and a report from the training jamboree for Genome Sequencing Program consortium investigators hosted by the NHGRI AnVIL platform.

Plans for the upcoming year

To reduce the burden of participation on the GDSWG for the current members, and bring into the group new expertise and ideas, in the fall of 2020 a few GDSWG members will rotate out and other investigators will be added to the roster.

In addition to organizing the Machine Learning for Genomics Workshop, we anticipate that the following key topics will be brought up for discussion: the implementation of genomic data science activities in accordance with the vision expressed in the NHGRI 2020 Strategic Plan; how the AnVIL program can improve genomic data sharing and analysis; training and education programs for data scientists in genomic research, or for genomicists in data science; and how to engage the broad genomic data science community.

The GDSWG often addresses the complexity of how genomic data science efforts fit across NHGRI programs and across the NIH more broadly. The group will continue to advise on the role for NHGRI to lead and break barriers across the NIH and the genomics field more generally.

NHGRI staff authors: Valentina Di Francesco and Natalie Kucher