National Advisory Council for Human Genome Research

September 2021

Concept Clearance for RFA

Enhancing Diversity in Cloud-based Genomic Data Science Education (U24/R25)

Purpose:
The National Human Genome Research Institute (NHGRI) proposes to launch a new initiative that includes two Request for Applications (RFA) for Enhancing Diversity in Cloud-based Genomic Data Science Education. This initiative aims to support educational opportunities in genomic data science for institutions serving groups underrepresented in the biomedical sciences, as outlined in NIH's Notice of Interest in Diversity. This initiative also extends to institutions serving groups underrepresented in the data science workforce (e.g. women data scientists). The focus is on development of instructional content appropriate for the undergraduate and master's degree level that leverages genomic datasets, analysis tools and resources available through NIH cloud computing platforms.

Background:
The 2020 NHGRI Strategic Vision highlights championing a diverse workforce as one of the guiding principles and values for genomics. As detailed in NHGRI's 2021 "Building a Diverse Genomics Workforce: An NHGRI Action Agenda", a major implementation goal for achieving that vision is to develop and support training programs and networks that connect undergraduate and graduate education to careers in genomics. This includes ensuring that institutions that have a mission to serve groups that are currently underrepresented in the genomics enterprise are aware of and tightly connected to these networks.

Concurrently, the Strategic Vision highlights the importance of computational genomics and data science (CGDS) in sustaining and improving a robust foundation for genomics. CGDS is the cross-cutting area where statistics and computer science are used to understand, analyze, and interpret information from genome sequences. Genomics is inherently a "big data" field, and appropriate skills in data science and data stewardship are now prerequisites for becoming a genomics researcher. Therefore, the next generation of genomic scientists needs to be sufficiently trained in data science. However, not all undergraduate and master's level students who are interested in genomics have access to the infrastructure and specialized training needed to obtain those skills.

In recent years, cloud computing technology has revolutionized CGDS, by democratizing access to genomics data sets, tools, and computing resources, and by enabling efficient genomic data sharing and collaboration. Through platforms such as the NHGRI AnVIL, NHLBI BioData Catalyst, the *All of Us* Research Hub and the NCI Cancer Research Data Commons, NIH has been a major promoter of the use of cloud computing for genomics in biomedical research. NIH's Office of Data Science Strategy (ODSS), as part of its mission to implement the NIH Strategic Plan for Data Science, has been a key leader of NIH's cloud computing implementation.

This initiative will leverage the NHGRI-funded AnVIL, and other NIH cloud-based platforms, to facilitate hands-on exposure to CGDS at the undergraduate and master's degree levels. The program will support activities to increase awareness of CGDS and cloud computing concepts, and address barriers to effective CGDS education and training. The program will also support instructional content development at institutions that (a) have a historical and current mission to serve groups that are underrepresented in the biomedical sciences and data science workforce, and (b) have received no more than $7.5 million dollars per year (total costs) from NIH Research Project Grants in each of the preceding three fiscal years. Institutions in scope to be CGDS Educational Partner Sites (see details below) include

Historically Black Colleges and Universities (HBCUs), Hispanic-Serving Institutions (HSIs), Tribal Colleges and Universities (TCUs) and women's colleges (hereafter, collectively referred to as Eligible Institutions).

Proposed Scope and Objectives:
The overarching goal of this initiative is to increase the number of individuals from diverse backgrounds, including underrepresented groups, who obtain appropriate skills in CGDS at the undergraduate and master's degree level. By focusing on early career stages, this initiative will create opportunities for individuals to obtain a more robust foundation in genomic data science and facilitate the development of a diverse genomics workforce.

This goal will be achieved through three complementary activities:

a. identifying challenges in teaching cloud computing and CGDS concepts at Eligible Institutions, and testing strategies that address those challenges;
b. developing CGDS instructional content at Eligible Institutions that uses the NHGRI AnVIL and other NIH cloud platforms (with curricula including both classroom lectures and hands-on data analysis exercises); and
c. disseminating these curricula and associated materials to the broader educational community.

To accomplish these objectives, NHGRI proposes two Request for Applications (RFAs):

a. A CGDS Educational Hub ("the Hub"): The Hub will be funded as a five-year award to carry out foundational activities to increase awareness of cloud computing and CGDS concepts among faculty and students at Eligible Institutions and to facilitate the development of educational material to be used at those institutions. The scope of activities to be performed by the Hub includes (a) organizing well-structured workshops and hands-on training sessions on cloud computing and CGDS concepts (b) collecting input on challenges faced by faculty in teaching CGDS and working with appropriate stakeholders and collaborators to develop and test strategies to address these challenges (c) disseminating educational materials developed through the other RFA in this initiative (see below), (d) collecting necessary data and evaluating effectiveness of the Hub activities, (e) providing logistical and coordinating support for awardees of the other RFA in this initiative (see "shared activities" below) and (f) coordinating with the NIH staff and Principal Investigators for the various NIH cloud computing platforms that will be used in this initiative.

b. The CGDS Educational Partner Sites ("the Sites"): The Sites will be funded as three-year awards with a primary focus on developing educational materials designed to enhance training of undergraduate and masters level students in cloud-based CGDS. The awardees must be from Eligible Institutions (as described above). These Sites will be solicited at the beginning of the 2nd year of the Hub and will be funded starting in the 3rd year of the Hub. The scope of activities to be performed by the Sites include: (a) development of hands-on cloud computing educational content that uses the AnVIL and other NIH cloud-based platforms for genomics (b) implementing the educational content in real-world settings with their students (c) collecting necessary data and evaluating effectiveness of the educational content, and (d) making the educational material available publicly available.

Shared activities: To connect the Hub and the Sites, this proposed initiative will also have the following shared activities (administered through the Hub):

a. A competitive opportunity for the Sites to receive $50,000 of support for hands-on research projects. These projects will engage students at the Sites in computational genomics research, allowing Sites to model how to use the cloud for more advanced

curricula or research experiences. This will be available to the Sites in their 2nd or 3rd year of funding (4th and 5th years of the Hub).

b. Annual in-person meetings (organized by the Hub) for members of the Hub, the Sites, representatives of participating NIH cloud platforms and NIH program staff. These meetings will focus on sharing best practices and lessons learned among the sites, and on how to disseminate the curricula and teaching aids developed through this initiative with institutions beyond the immediate awardees.

c. Train-the-trainer content for faculty members beyond the funded Sites. This will include holding teaching workshops adopting the curricula developed at the Sites as part of this initiative and will enable more faculty to take these curricula into their classrooms.

NHGRI expects that all educational or training materials, including curricula and syllabi, developed by the Hub and the Sites will be made publicly available by the beginning of the final year of each award. The materials should be provided in a shareable format that is widely available at no cost.

Relationship to Ongoing Activities:
This concept is complementary to ongoing activities at NIH designed to use cloud computing platforms as a tool for CGDS workforce development, such as the workshops and data jamborees conducted by the AnVIL project and the *All of Us* Researcher Workbench training activities. Those activities focus on short-term workshops and outreach activity, rather than on the development of curricula and educational tools. NHGRI has engaged faculty at Eligible Institutions to address educational use of the cloud at diverse institutions through the Genomic Data Science Community Network (GDSCN). GDSCN is a one-year project scheduled to end before this initiative begins. The output from the GDSCN will be available as a starting point for the Hub in this initiative. ODSS is engaged in early planning of complementary data science initiatives for workforce diversity and cloud computing (see NOT-OD-21-079 and NOT-OD-21-158, respectively). NHGRI will coordinate with ODSS to examine potential synergies between this initiative and those activities as they develop.

Mechanism of Support and Funds Anticipated:
Two RFAs are proposed for this concept:

The CGDS Educational Hub will be solicited as a single five-year U24 award (U24 Resource-Related Research Projects - Cooperative Agreements). Funding is anticipated to start in FY23 and be up to $1.5M total costs in FY23-25, increasing to $1.8M total costs for FY26-FY28 when more advanced research projects are funded at the Sites. ODSS has been approached as a partner for funding this RFA.

The CGDS Educational Partner Sites will be solicited as multiple three-year R25 awards (R25 Research Education Program). The Funding Opportunity Announcement will have one receipt date per year for two years, with awards starting in FY25 and FY26. Funding is anticipated to be up to $150K total costs per award per year. One set of awards will be made for FY25-27, with another set funded for FY26-28. In each set, NHGRI anticipates funding 5-8 awards, for an aggregate total of $750K-$1.2M per year. NHLBI, NIMHD, and the All of Us Research Program have been approached as possible partners to co-fund R25 applicants aligned with their scientific or diversity/outreach missions.

NHGRI will work with the NIH STRIDES program to provide cost-effective access to cloud resources for the Hub and the Sites, including using models for providing cloud credits as appropriate.