May 6, 2019

The European Bioinformatics Institute

Wellcome Genome Campus

Cambridgeshire, CB10 1SD

To Whom It May Concern:

I am applying for the position of Project Lead with the European Genome-phenome Archive (EBI01399) to provide project coordination between the EGA and the UK Biobank to archive and distribute whole genome and exome data. I am excited for this opportunity to take on a leadership and coordination role, and I believe that I possess the qualities and experience required to be a productive and effective Project Lead with the EGA.

Coordinating the integration of UK Biobank genome and exome data with the EGA is especially attractive to me as it aligns with my belief in the potential for big data to have a profoundly positive impact on the quality of human life. This potential can only be realized, however, if data are findable, accessible, interoperable, and re-usable. As a graduate student and post-doctoral research fellow, I largely depended on published next-generation sequencing data to achieve my research aims. These datasets were often accessible (deposited in standard archives) and interoperable (used standard formats and vocabularies), but not necessarily findable or re-usable without significant effort. A growing desire to improve the FAIR-ness of biological data led me to join the Human Cell Atlas (HCA) Data Coordination Platform (DCP) project where I work to define metadata standards for cellular resolution data to adhere to the FAIR data principles. As a Project Lead for the EGA, I hope to continue working towards maximizing the positive impact of big data on the scientific community and, ultimately, on human life.

While part of the Galaxy Project and the HCA DCP, I have coordinated work within multiple teams to achieve project goals. As a Galaxy Training Network trainer, I worked with other trainers to identify what bioinformatic analysis topics our users wanted training on and to improve the quality of the training materials by incorporating feedback from workshop attendees.Within the HCA DCP I have multiple roles which require me to be highly organized in order to address changing priorities. As the senior data wrangler, I plan biweekly sprints to coordinate work with the five other data wranglers from both the EBI and the University of California at Santa Cruz (UCSC). Early communication struggles between the data wranglers led me to implement and chair weekly data wrangler meetings where we report on, plan, and prioritize our work. Planning this work requires me to translate DCP-wide priorities decided by the project leads into the specific data wrangling tasks that need to be accomplished. As part of the metadata team, I chair monthly metadata meetings to coordinate work between the EBI and UCSC metadata team members and to discuss metadata requirements from other DCP teams at Broad Institute and the Chan Zuckerberg Initiative. Additionally, I often deputize for our project coordinator and project lead to chair project coordination calls or attend project management meetings when they are unavailable. The project coordination skills I’ve learned from these experiences will be valuable for collaborating with both the EGA and UK Biobank teams to translate the project’s goals into high-quality deliverables.

The Galaxy Project and the HCA DCP are both projects whose aims are to provide a resource for scientists that supports FAIR data. The success of these projects relies on not only internal coordination of teams from multiple institutions, but also on collaborations with other scientists and organizations. I strongly believe that big data projects must be able to integrate with external resources to take advantage of others’ expertise and improve FAIR-ness of data. Recently, I had the opportunity to coordinate work between the HCA DCP team at the EBI, the Unified Submission Interface team, and EBI archive teams (ENA, BioSamples, BioStudies) in order to establish an automated pipeline for accessioning HCA datasets in EBI archives. Through prioritization of features, assessment of solutions to eliminate blockers, and facilitation of communication between relevant developers, I was able to ensure delivery of a workable pipeline and successfully archive the first two HCA datasets. This work required me to have technical discussions with developers and priority discussions with project managers, skills that are highly relevant for a Project Lead role with the EGA. In addition to communicating with external scientists, I believe that it’s important to engage in discussions about big data with the public to foster a sense of shared stewardship and build trust. To this end, I’ve discussed the HCA DCP with a wide range of audiences at public engagement events. For example, last summer I spoke to members of the Cambridge Society for the Application of Research about the HCA, answering their questions about data privacy and describing how the HCA is contributing to improving human health. These examples of my interactions with external groups emphasize a track record of successful and productive collaborative efforts with a variety of teams and stakeholders. I am confident that these skills will be valuable to achieving the EGA’s mission of making UK Biobank genome and exome sequencing data FAIR for the scientific community.

As a Project Lead coordinating the archive and distribution of UK Biobank genome and exome sequencing data, I believe I can pursue my scientific and career goals as well as contribute to the overall mission of the EGA. I welcome the opportunity to further discuss my application, and I look forward to hearing from you soon.

Yours sincerely,

Mallory Freeberg, Ph.D.