Title	Towards traceable, scriptable, and efficient data distribution for next-generation genomics
Author	*John Bradley*, Dan Leehr, Erich S. Huang, Jonathan Turner, Hilmar Lapp
Affiliation	Duke University, Center for Genomic and Computational Biology
Contact	john.bradley@duke.edu
URLs	https://github.com/Duke-GCB/DukeDSClient https://github.com/Duke-GCB/DukeDSHandoverService
License	MIT

At many research institutions, next-generation genomics data starts their lifecycle at a core facility. Depending on the type of core this may be as primary data generated by instruments, or as secondary data generated by analysis or other processing. From there, data will be handed over to a principal investigator, who then derives scientific conclusions underpinned by the data and publishes them in scholarly journals. To be consistent with open and reproducible science principles, this lifecycle creates a number of challenges, many of which can be traced back to the shortcomings of distributing and sharing data via traditional, yet still very common, networked block storage. In particular, (1) lifecycle progression, including key steps such as handing over data from core to investigator, and depositing data for permanent archival in a repository, is difficult or impossible to formalize as a documented and programmable transaction; (2) the metadata of data, which importantly includes tracing a data object's entire provenance chain, are difficult to track and tie to the data; (3) versioning of data, including identifying in a globally unique way the data used for each analysis step, is highly cumbersome at best. These challenges are further exacerbated by the large volume of data generated by ever-evolving next-generation genomics technologies.

To enable these challenges to be addressed in a principled way, a collaboration spearheaded by an interdisciplinary team at Duke University is building an open-source informatics infrastructure for managing and tracking data at scale through its lifecycle, called the Duke Data Service (DDS)[1]. DDS is inspired by years of experience developing and using SAGE Bionetworks' Synapse[2], comes with an extensive API deployed on the Heroku cloud, and for storing, versioning, and identifying data federates across cloud-based and on-premise object store service APIs, including AWS S3 and OpenStack Swift.

Here we report on our work using this infrastrucure to build scalable command line and user-interface tools for digital genomics data during the lifecycle period from generation in a core facility to hand-off to investigator. We highlight how the tools efficiently and traceably ingest and register data and their provenance; how investigators can receive the data on high-performance computing environments; and how the object store API enables parallelization of data streaming to better saturate available I/O bandwidth for high-volume data. Finally, we will show how a web-based user-interface implemented on top of this infrastructrure can formalize the hand-over of data from core to investigator as a documented transaction that signifies the passing of data stewardship from one party to another.

References:

1. https://github.com/Duke-Translational-Bioinformatics/duke-data-service
2. http://sagebase.org/synapse