2022 ASHG Abstract

Empowering Discovery in Childhood Cancer: Genomic Harmonization at the Kids First Data Resource Center

The Gabriella Miller Kids First Pediatric Research Program (GMKF) is an NIH Common Fund initiative focused on providing large-scale clinically annotated genomic data for pediatric cancer and structural birth defect cohorts.
Kids First is challenged with taking genomic data from a wide range of sources and providing researchers and physicians with harmonized genetic data. Preparing large-scale harmonized datasets presents unique challenges in scalability, reproducibility, and transparency. The Kids First Data Resource Center (DRC) tackles these challenges using open-source, community-standard workflows that are deployed in cloud-based HPC environments. Our workflows, written in Common Workflow Language (CWL), are modeled after established best practices workflows, optimized and verified through internal benchmarking, and, when possible, made in collaboration with our network of researchers. These workflows are made freely available both as open-source code on GitHub and as public apps via CAVATICA, an Amazon Web Services (AWS) based cloud computing platform associated with the Kids First DRC Portal co-developed by Seven Bridges Genomics, where workflows feature scatter-gather parallelization, conditional execution, and AWS resource optimization.
Today, the DRC has six production level workflows producing harmonized datasets at scale for whole genome sequencing, exome sequencing, and RNA-seq technologies. Additionally, we have over a dozen non-production workflows for everything from germline to tumor-only to long reads applications. These workflows have been run across the 24 Kids First studies and 20,000 participants already released on the Kids First Data Resource Portal. In total, we have more than 1.0 PB of data, with more being released yearly. Here we present our process for creating, validating, and distributing these workflows.