BD2K and Global Genomic Data Sharing

***David Haussler, University of California, Santa Cruz Genomics Institute**; Benedict Paten, University of California, Santa Cruz; Brian O'Connor, University of California, Santa Cruz; Kevin Osborn, University of California, Santa Cruz; David Steinberg, University of California, Santa Cruz; Melissa Cline, University of California, Santa Cruz; Mary Goldman, University of California, Santa Cruz; Mark Diekhans, University of California, Santa Cruz; Robert Currie, University of California, Santa Cruz*

It is clear to many BD2K centers that genomics has entered a critical phase where research is rapidly being translated into clinical practice, and thus genomics data must be standardized and joined to other clinical data types. At the same time, data is becoming increasingly global as we collectively recognize that it will take truly huge collections to map out the relationships we need to understand between genotype and phenotype.

Addressing the challenge, we at UCSC have forged a partnership with the Global Alliance for Genomics and Health (GA4GH), an organization dedicated to the global sharing of standardized genomics data. The GA4GH has grown to more than 400 institutional members from more than 40 countries. Translational Genomics (UCSC) and other BD2K centers have contributed to GA4GH, often through global demonstration projects like Beacon, BRCA Challenge, and the recently announced Cancer Gene Trust. Additionally, a set of BD2K centers, including UCSC, MD2K, PIC-SURE and bioCADDIE, have adapted (GA4GH) APIs for the secure exchange of genomic, mobile sensor and clinical data ("Count Everything"). These projects collectively demonstrate the power of globally shared data.

To make shared data analysis possible, orthogonal but vital to these projects is the general development of computational environments in the cloud where big data analysis can occur, inspired by the BD2K notion of a Data Commons. Groups within BD2K are harnessing the power of containerization tools like Docker and workflow languages like CWL to lay the groundwork for such an environment that can exist equivalently in many different clouds and data centers, and run modular, standardized workflows on standardized data accessed through widely accepted APIs. The GA4GH Containers and Workflows Task Team, with a BD2K co-leader, is one contributor to this vision, along with the KnowEng BD2K center. Achieving this vision would be transformational for biomedicine.