Title	Collaboration and interoperable, multi-platform workflow development
Authors	*Brad Chapman*, Rory Kirchner, Lorena Pantano, Shannan Ho Sui, Oliver Hofmann
Affiliations	Harvard Chan School, Bioinformatics Core (http://bioinformatics.sph.harvard.edu/), The University of Melbourne Centre for Cancer Research (http://www.gla.ac.uk/researchinstitutes/cancersciences/ics/facilities/wwcrc/)
Contact	bchapman@hsph.harvard.edu
Availability	https://github.com/chapmanb/bcbio-nextgen
Documentation	https://bcbio-nextgen.readthedocs.org/en/latest/contents/cwl.html
License	MIT

bcbio (https://github.com/chapmanb/bcbio-nextgen) is an open, community effort to develop validated and scalable variant calling, RNA-seq and small RNA analyses. Last year at BOSC, we discussed our work to port bcbio's internal workflow representation to use the community developed Common Workflow Language (CWL: http://www.commonwl.org/) standard. This transition removed barriers that prevented bcbio interoperability.

The practical benefit of changing to standardized workflow definitions is that bcbio works on multiple heterogeneous platforms. Using CWL, bcbio runs on Curoverse's Arvados (https://arvados.org/), UCSC's Toil (http://toil.ucsc-cgl.org/) and Seven Bridges' rabix bunny (http://rabix.io/). In addition, in progress work converting to the Workflow Description Language (WDL: https://software.broadinstitute.org/wdl/) provides support for Broad's Cromwell (https://github.com/broadinstitute/cromwell) and DNAnexus's APIs (https://github.com/dnanexus-rnd/dxWDL). There is also ongoing work with other communities actively developing CWL integration, including Nextflow (https://github.com/nextflow-io/cwl2nxf) and Galaxy (https://github.com/common-workflow-language/galaxy).

Widespread bcbio interoperability allows running in many computational environments without the overhead of maintaining bcbio specific integrations. Users can still run locally or on high performance computing clusters with schedulers like SLURM, SGE and PBSPro. In addition, CWL enabled runners work across the three major cloud providers: Amazon Web Services, Google Compute Engine and Microsoft Azure. Commercial platforms like Curoverse, Seven Bridges and DNAnexus enable clinical labs to run in controlled environments. The key component of this diverse support is collaboration through the CWL standard. This demonstrates the importance of community standard development, especially in research environments where it is typically difficult to fund maintenance of large scale infastructure development.

The talk will discuss the practicalities of adjusting bcbio to use CWL and WDL. We have to balance infrastructure work for the transition to CWL with continued improvement of workflows and community support. Testing and documentation of bcbio becomes more complex since we validate analyses, like germline and somatic variant calling with both small and large variants, across many environments. This requires coordination between groups with different focus and directions as platforms, analyses and standards develop. This type of high level coordination will become increasingly important as we do more complex science, and

we'll describe the role of the open bioinformatics community in enabling it.