Title	Supporting dynamic community developed biological pipelines
Author	*Brad Chapman*, Rory Kirchner, Oliver Hofmann, Winston Hide
Affiliation	Harvard School of Public Health
Contact	bchapman@hsph.harvard.edu
URL	https://github.com/chapmanb/bcbio-nextgen
License	MIT

bcbio-nextgen is a community developed set of validated, scalable pipelines for running variant calling and RNA-seq analyses. It creates an infrastructure from open source tools that is easy to install and run. The goal is to implement best-practice approaches that scale across multiple architectures ranging from single machines to large clusters, and combine this with automated validation of results for correctness against reference standards.

For example, the practical goal of the variant calling pipeline within bcbio-nextgen is to let biologists work with best-practice variant calls, instead of struggling with processing raw next-generation sequencing reads. To do this, we integrate aligners like bwa-mem and variant callers like GATK and FreeBayes alongside other BAM and variant manipulation tools, and then validate variants against reference callsets from the Genome in a Bottle consortium. The outcome is a push button analysis framework that parallelizes a complex set of tools to produce high quality variants as ready to analyze outputs.

The challenge associated with supporting bcbio-nextgen is that it relies on many open-source tools and works across a wide range of heterogeneous platforms. The result is a large amount of community time spent on installation issues, rather than answering biological questions. We produced an automated installer and updater using CloudBioLinux which solved many adoption issues but also requires work to maintain, extend and test.

At BOSC, we'll discuss two approaches designed to improve ease of use:

- Isolating dependencies within lightweight Docker containers. This provides a standard distribution environment containing all third-party code and tools. This avoids the need to compile and install these on a wide variety of systems. Additionally it provides a reproducible analysis environment for export, archival and sharing.
- Providing a Amazon Web Services implementation that is resilient to failure and makes using of spot instances. This helps overcome two major hurdles to cloud adoption: difficulty scaling on less reliable commodity hardware and justifying spending on external compute. We'll share our experiences redesigning bcbio-nextgen to run on non-shared filesystems and handle higher failure rates found in cloud environments.

These improvements move towards the goal of having shared community developed pipelines usable by researchers, clinical labs and the general public. By removing the separation between up to date research grade tools and validated clinical grade tools, we enable contributions from multiple communities and standardization around stable reliable tools for the overlapping needs of the diverse translational research community.