# Abstract

Genomics researchers increasingly turn to cloud computing as a means of accomplishing large-scale analyses efficiently and cost-effectively. Successful operation in the cloud requires careful instrumentation and management to avoid common pitfalls, such as resource bottlenecks and low utilization that can both drive up costs and extend the timeline of a scientific project.

We developed the Butler framework for large-scale scientific workflow management in the cloud to meet these challenges. The cornerstones of Butler design are: ability to support multiple clouds, declarative infrastructure configuration management, scalable, fault-tolerant operation, comprehensive resource monitoring, and automated error detection and recovery. Butler relies on industry-strength open-source components in order to deliver a framework that is robust and scalable to thousands of compute cores and millions of workflow executions. Butler’s error detection and self-healing capabilities are unique among scientific workflow frameworks and ensure that analyses are carried out with minimal human intervention.

Butler has been used to analyze over 725TB of DNA sequencing data on the cloud, using 1500 CPU cores, and 6TB of RAM, delivering results with 43% increased efficiency compared to other tools. The flexible design of this framework allows easy adoption within other fields of Life Sciences and ensures that it will scale together with the demand for scientific analysis in the cloud for years to come.

Because many bioinformatics tools have been developed in the context of small sample sizes they often struggle to keep up with the demands for large-scale data processing required for modern research and clinical sequencing projects due to the limitations in their design. The Rheos software system is designed specifically with these large data sets in mind. Utilizing the elastic compute capacity of modern academic and commercial cloud computing centers, Rheos takes a service-oriented containerized approach to the implementation of modern bioinformatics algorithms which allows the software to achieve the scalability and ease-of-use required to succeed under increased operational load of massive data sets generated by projects like ICGC Argo and the All of Us initiative.

Rheos algorithms are based on an innovative stream-based approach for processing genomic data which enables Rheos to make faster decisions about the presence of genomic mutations that drive diseases such as cancer, thereby improving the tool’s efficacy and relevance to clinical sequencing applications. Our testing of the novel germline SNP and deletion variant calling algorithms developed within Rheos indicates that Rheos achieves ~98% accuracy in SNP calling and ~85% accuracy in deletion calling, which is comparable with other leading tools such as GATK, freebayes, and Delly.

The two frameworks that we developed provide important contributions to solve the ever-growing need for large scale genomic data analysis on the cloud, by enabling more effective use of existing tools, in the case of Butler, and providing a new, more dynamic and real-time approach to genomic analysis, in the case of Rheos.