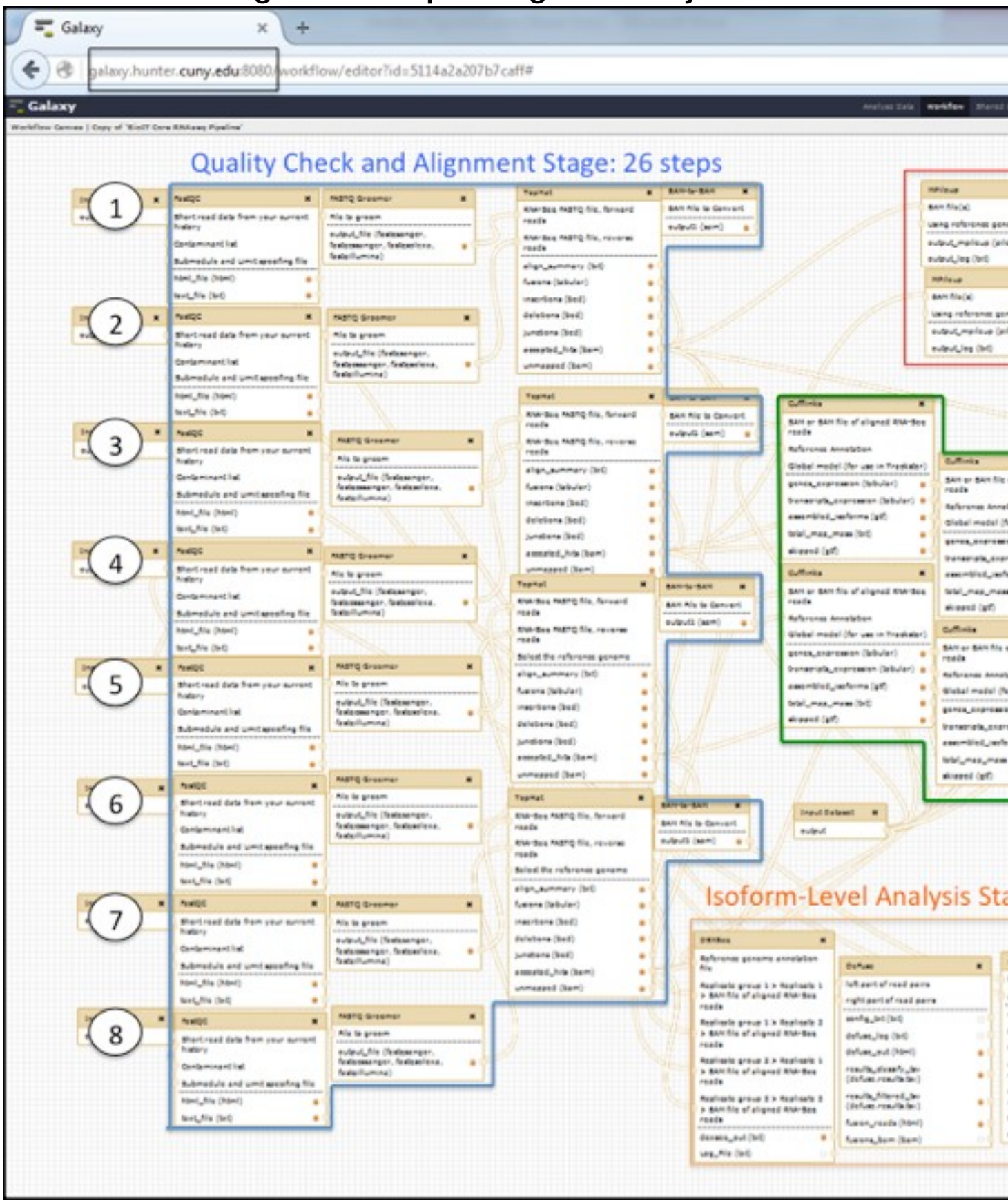
Developing cross-platform, scalable Science as a Service (SciaaS) infrastructures for genomic sequencing data analysis.



Introduction:

With the growth in sequencing throughput and reduction of sequencing cost tens of TeraBytes of data have been released in the recent five years. Acquiring the reads during a whole genome shotgun sequencing project is only the first step, and must be followed by bioinformatic analysis that involves running post-sequencing data analysis software, such as read quality checks, trimming and barcode deconvolution, in addition to assembly, annotation. Furthermore, with the recently available bench-top genome sequencing instruments, such as MiSeq from Illumina, the technology for sequencing small to medium-sized genomes has become affordable for researchers in smaller laboratories. Given the scale of genomic datasets, scientific value cannot be obtained from an investment in a sequencer, unless it is accompanied by an equal investment in bioinformatics infrastructure. Processing large amounts of genomic data requires access to a range of technical capabilities including expertise in information technology, bioinformatics, and software engineering, as well as to large-scale computational and storage capabilities. For smaller laboratories these requirements can become a significant impediments, as in addition to coming up with the funds for building an informatics infrastructure with capacity to handle large-scale sequence data, they also need funds for hiring trained professionals competent to install, configure and run the bioinformatics tools, which after all can present a higher expense than that of acquiring the hardware.

Virtual Machines (VMs) can address one of the main problems for making open-source software with complex dependencies and installation procedures widely accessible to the research community. Furthermore, virtualization technologies have led to the development of cloud computing services, where remote computer server farms can be rented on an hourly basis by researchers and used for scalable, on-demand computation. This has provided researchers with the potential to eliminate many of the upfront capital and effort expenses of building technology infrastructure for genome sequencing informatics, and result in transformation of the analysis and data processing tasks into well-defined operational costs. Furthermore, adoption of cloud computing technologies is providing an opportunity to democratize the sequencing informatics field, meaning that individual researchers and labs can now have access to the resources that were previously only available to institutions with large bioinformatic core facilities.

Research Approach:

By leveraging virtualization technology, we can develop a novel model for bioinformatics infrastructure called Science as a Service (SciaaS). Following this model, bioinformatics developers can build software and data analysis pipelines inside a VM, that can run on computing environments ranging from lab computers, data center clusters and cloud computers. A SciaaS computing model for bioinformatics, can remove the bottleneck faced by smaller, independent laboratories for the use of genomic technologies, since sequencing data analysis requires access to a range of technical capabilities including expertise in information technology, software engineering, as well as to large-scale computational and storage capabilities with the use of cloud computing. We believe that with this computing computing model we can enhance translational research by facilitating access to analysis tools to dispersed groups working on common projects.

From a technical perspective a bioinformatics infrastructure that follows the SciaaS model should: require minimal or zero installation effort; provide a mechanism for software and algorithm version tracking, update and configuration; allow for management of large research

data collections or other required datasets such as genome indexes and assemblies; provide a mechanism for sharing research data along with pre-configured tools, independently of differences in computing setups used among collaborators; provide a rich graphical interface for managing the data, accessing and running the tools, in addition to capability for constructing data analysis workflows and pipelines from existing tools and finally; the complete a sequencing informatics infrastructure ideally should be possible to be installable, replicated, and reused across computing platforms of all scales from institution-wide compute clusters and Clouds to small data centers, or even laboratories that have only personal desktop computers.

Aim 1: Implement a SciaaS bioinformatics platform that is portable, scalable and easy to use on personal computers, computing clusters and clouds.

Our goal is to implement a portable, platform-agnostic tool suite for sequencing bioinformatics following a SciaaS model that provides a plug-and-play, cross-platform genome sequence data analysis platform equipped with a single-click installer, and runs on any available computing system with zero configuration required. With the increasing number of compute core and higher capacity CPUs available on standard personal computers, development of such a self-installable, desktop-based VM will enable analysis of sequencing datasets generated by bench-top, small factor sequencers or low-coverage datasets by larger capacity sequencers locally at each laboratory. Furthermore, the portable format of our VM will allow users to seamlessly port their VM and run the tools on large-scale infrastructures such as private clouds and institutional clusters.

Aim 2: Implement a set of sequencing data analysis pipelines running in Virtual Machines (VMs) under the SciaaS computing model.

Aim 2: Implement a set of sequencing data analysis pipelines running in Virtual Machines (VMs) under the SciaaS computing model. We will make available a set of next-gen sequencing data analysis workflows pre-configured and ready to execute in VMs, including gene expression (RNAseq), exam genome and variation discovery, CHIPseq for transcription factor binding site discovery and bi-sulfite sequencing for epigenetics. During the course of this project, we will curate the literature for published protocols on sequencing data analysis, and implement the data analysis workflows based on these protocols. Our approach will be to select the most frequently cited publications, and following implementation and testing of the workflows in the VMs, make them available for use on local computers or through cloud computing.