Experiences Building Globus Genomics: A Next-Generation Sequencing Analysis Service using Galaxy, Globus, and Amazon Web Services

Ravi K. Madduri, Dinanath Sulakhe, Lukasz Lacinski, Bo Liu,
Alex Rodriguez, Kyle Chard, Utpal J. Dave, Ian T. Foster
Computation Institute
University of Chicago and Argonne National Laboratory
Chicago, IL

ABSTRACT

We describe Globus Genomics, a system that we have developed for rapid analysis of large quantities of next-generation sequencing (NGS) genomic data. This system achieves a high degree of end-to-end automation that encompasses every stage of data analysis including initial data retrieval from remote sequencing centers or storage (via the Globus file transfer system); specification, configuration, and reuse of multi-step processing pipelines (via the Galaxy workflow system); creation of custom Amazon Machine Images and on-demand resource acquisition via a specialized elastic provisioner (on Amazon EC2); and efficient scheduling of these pipelines over many processors (via the HTCondor scheduler). The system allows biomedical researchers to perform rapid analysis of large NGS datasets in a fully automated manner, without software installation or a need for any local computing infrastructure. We report performance and cost results for some representative workloads.

Categories and Subject Descriptors

H.m [Information Systems]; C.1.4 [Computer Systems Organization]: Parallel Architectures - *Distributed architectures*; J.3 [Computer Applications]: Life and Medical Sciences - *Biology and genetics*

General Terms

Algorithms, Performance, Design, Economics, Reliability, Experimentation, Security, Human Factors, Standardization.

Keywords

Cloud, HPC, HTC, NGS, workflows

1. Introduction

Parallel computing and analysis automation have become vital tools for discovery in biology, as rapidly decreasing sequencing costs have transformed the field from a data-limited to a computationally limited discipline. Increasingly, researchers must process hundreds of sequenced genomes to determine the statistical significance of variants. When genomic datasets were small, they could be analyzed on personal computers in modest amounts of time: a few hours or perhaps overnight. However, this approach does not scale to large next-generation sequencing (NGS) datasets. Instead, researchers require high-performance computers and parallel algorithms if they are to analyze their data in a timely manner.

Yet while high performance computing is essential for progress, few biomedical research labs are equipped to make effective use of parallel computers. Obstacles include the complexities inherent in managing large NGS datasets and assembling and configuring multi-step genome sequencing pipelines; lack of access to parallel computer systems; and difficulties inherent in adapting pipelines to process NGS data on parallel computers.

The Globus Genomics system [1, 2] addresses these challenges, allowing users to specify, with a few mouse clicks, the location of the NGS data that is to be analyzed and the specific analysis pipeline that is to be applied to that data. Globus Genomics then manages the transfer of data to Amazon cloud storage, the deployment of pipeline elements on Amazon cloud computers, and the execution of the analysis pipeline—all in a fully automated manner. Furthermore, due to the use of low-price Amazon spot instances, analysis costs are typically modest.

We provide details on the challenges and opportunities inherent in NGS and the Globus Genomics solution to these challenges. We describe the architecture of the Globus Genomics platform, present several early adopters of the system and demonstrate the advantages to these users when executing common NGS analysis pipelines. We also discuss the relevance of this work to national research infrastructure.

2. Next Generation Sequencing

NGS technologies have revolutionized the process of determining the genetic basis of human ailments [3, 4]. The basic idea is to parallelize the process by which base pairs in a genetic sequence are identified, so that thousands or millions of sequence fragments are produced concurrently—and the overall task of creating a complete sequence is accelerated by a comparable factor. NGS technologies have reduced the cost of sequencing an entire human genome from $2.7B for the first such genome (the cost of the Human Genome Project) to a few $1,000 today (see Figure 1; note the point of inflection around 2007, due to the introduction of NGS). A cost of less than $1,000 is expected in the near future.

The promise of discoveries and the rapid decrease in the cost of sequencing has led to a tremendous increase in the amount of sequencing performed and the need for large scale analysis of the data generated [5]. Indeed, storage and computational costs represent a growing fraction of total sequence analysis costs, a dramatic turnabout relative to just a few years previously.

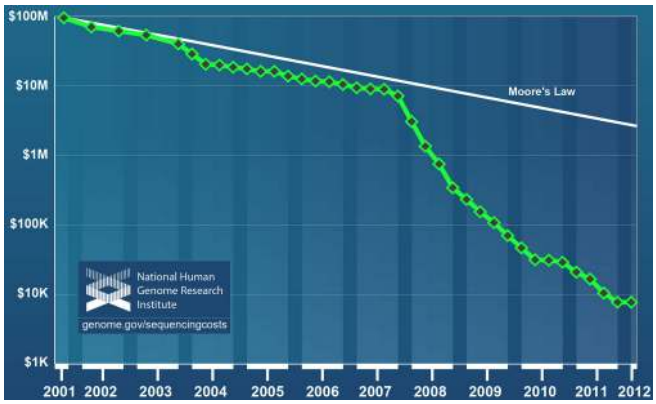

Figure 1: Sequencing costs over time

Stein postulated [5] that cloud computing and associated computing-as-a-service technologies would help researchers deal with the enormous volume of NGS data. However, challenges remain in effectively leveraging cloud capabilities for the typical researcher. These challenges include moving large amounts of data reliably from where it is generated to where it can be analyzed, and running large-scale analyses efficiently and reliably on clouds. Globus Genomics addresses these areas of need.

3. Globus Genomics

In building the Globus Genomics system to address these challenges, we build on powerful capabilities provided by the Galaxy workflow system [6], which we extend to permit convenient and efficient execution in a commodity cloud computing environment, while making it trivial for end users to move large amounts of data into the Galaxy workflow system from various sequencing service providers.

To provide these new capabilities, we integrate the Galaxy framework with several Globus services [7], notably Globus transfer [8] for easy, reliable, and secure movement of large amounts of data and Globus Nexus [9] for identity management and group-based access control. We also integrate computationally efficient data analysis pipelines that leverage the power and flexibility of on-demand cloud computing resources [10]—all without exposing end users to the complexities of managing large scale infrastructure.

Both Galaxy and Globus are Web-based platforms for conducting data-intensive research. Globus makes it easy for researchers to move large amounts of data between endpoints, enabling, for example, transfers from sequencing centers and storage nodes to compute resources on which data can be analyzed. The Galaxy framework allows for the construction of complex biological workflows. It provides for the automatic and transparent tracking of every detail of an analysis, and permits analysis results to be documented, shared, and published with complete provenance, guaranteeing transparency and reproducibility. Importantly, Galaxy is an extensible platform: almost any software tool can be integrated, and an active community of developers ensures that the latest tools are wrapped and made available through the Galaxy Tool Shed, a repository of Galaxy tools that spans every aspect of genomic analysis.

The Galaxy team initially operated the public Galaxy server on a large compute cluster that they ran themselves at the University of Pennsylvania. They have recently moved the service to be hosted at the Texas Advanced Computing Center to meet increasing demand. This free service is used by thousands of researchers to perform hundreds of thousands of analyses each month, subject however to some limits on data transfer and computer usage. The Galaxy team created CloudMan [11], which allows researchers to run their own Galaxy server in the cloud. However, CloudMan still requires that end users know and understand the operating complexities of cloud computing infrastructure.

Globus Genomics further improves usability by providing Galaxy as a managed platform capable of elastically scaling compute resources to meet user needs. We leverage on-demand computational infrastructure using Amazon EC2 and the HTCondor scheduler [12] to enable elastic scaling of the compute cluster. We have also developed computational profiles for various analytical tools for optimal and scalable execution on Amazon cloud infrastructure. Globus services provide for the convenient and reliable transfer of big datasets from geographically distributed centers into our cloud computing infrastructure. Figure 2 shows the overall system architecture.

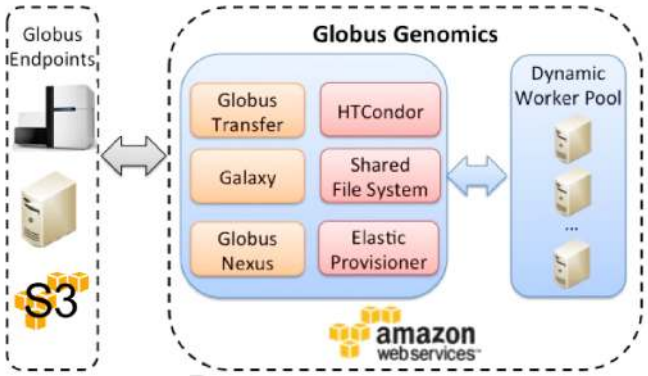

Figure 2: Globus Genomics Architecture

3.1 Architecture

As illustrated in Figure 2, Globus Genomics relies on several services to deliver an integrated, elastic, and scalable cloud-hosted genome analysis platform. A focal point of the system is the Galaxy workflow engine, a Web-based workflow creation and execution platform. Galaxy provides an intuitive Web interface that enables domain scientists to compose arbitrary applications, tools, and scripts. At the completion of an analysis, users may save the analysis steps as a workflow that can be rerun, published, or shared with other users. Galaxy records the steps performed in the analysis as a provenance record for the analysis result.

Galaxy is designed to use local compute resources co-located with the Galaxy server. It can be extended to also use cluster resources through its "runner" interfaces that support PBS, HTCondor, and DRMAA. However, configuring Galaxy to use these interfaces and installing and configuring the cluster, job scheduler, and shared file system is non-trivial.

A limitation of Galaxy is its dependence on a single shared file system. When processing large datasets across distributed compute resources this constraint represents a significant bottleneck. To address this technical limitation, Globus Genomics relies on a large shared file system (NFS in our case) accessible across its infrastructure, accessible to Galaxy, Globus transfer, and the Amazon EC2 nodes that are used for genome analysis.

Globus Genomics also includes a Globus transfer GridFTP Endpoint to transfer data, an elastic provisioner to dynamically scale a pool of worker nodes, and a scheduler to map pipelines to worker nodes. We describe these components in later sections.

A Globus Genomics instance can be deployed in many configurations, depending on whether different components are deployed in the same or different EC2 instances. Our typical deployment bundles all management components (Galaxy, HTCondor master, NFS, GridFTP, elastic provisioner) on a single head node instance (either m1.large or m1.xlarge instance type, to use Amazon Web Services EC2 terminology). This instance is only used to run computationally limited tasks such as data transfers, a Galaxy Web server and a Galaxy database. All other computation is sent to an HTCondor queue for execution on arbitrary worker nodes that are deployed when required (m2.4xlarge to cr1.8xlarge instance types). We use on-demand instances for the Globus Genomics head node and spot instances for all worker nodes. A typical instance of Globus Genomics includes 2-8 TB of total storage accessible across the system. Approximately 10 GB of this storage is used for tools included in Globus Genomics, 55 GB for reference genome indices used by these tools, a small volume for shared Galaxy state such as scripts, and the remainder for user data.

3.2 Globus integration

Globus platform services for identity, group, and data management are used by many large computing facilities such as XSEDE, KBase and others via REST APIs. Globus Genomics leverages these Globus services that support a wide range of security protocols, implement best-practice security approaches, and provide high reliability and availability. We review briefly here the Globus services that are used in Globus Genomics.

3.2.1 Globus Nexus for identities and groups

Globus Nexus provides identity, profile, and group management as a service. Its identity management capabilities allow users to create a Globus identity that can be used for Single Sign-On (SSO) across all Globus services. Users can link external identities to their Globus identity, enabling them to both track and manage their various identities and to authenticate using their preferred identity provider. Nexus's group management capabilities permit users to create and manage their own groups, assign roles to group members, and use groups for authorization.

Globus Genomics leverages these capabilities to outsource all identity and group management to Globus Nexus. Thus, Globus Genomics need not manage users itself, manage user passwords, or implement identity management processes such as password reset. Globus Nexus provides Globus Genomics with SSO across the entire infrastructure and when transferring data to/from other endpoints. It allows users to sign in to Globus Genomics using their preferred identity. Globus Genomics can also leverage groups managed in Globus Nexus for authorization. For example, group membership can be used to control access to a particular project's instance of Globus Genomics or to limit access to data, applications, and pipelines. We are also investigating using group membership as a basis for scheduling policies.

3.2.2 Globus Nexus integration with Galaxy

Globus OAuth is integrated as an external authentication module within the Web Service Gateway Interface (WSGI) middleware used in Galaxy. We extend the WSGI authentication module interface, which is typically used in conjunction with HTTP basic or digest authentication, to support OAuth-based authentication. User authentication then follows the typical three-legged OAuth 2 authentication workflow. When a user attempts to log in, the page redirects the user to authenticate using Globus Nexus. Once the user successfully authenticates (using any of their linked Nexus identities), Globus OAuth redirects the user back to the Globus Genomics instance with a short-lived access token. This access token is then verified and stored in an in-memory database that stores the mapping between the Galaxy session and the Globus username and OAuth token. This information is used subsequently by the Globus transfer tools to perform data transfer on the user's behalf. Importantly, this implementation is not intrusive to the rest of the Galaxy framework, does not disrupt the Galaxy user model, or its management of the session.

3.2.3 Globus transfer for reliable data movement

Globus transfer provides high performance, secure, third-party data movement and synchronization between endpoints. It automates challenging aspects of the data transfer process, providing "fire and forget" capabilities, tuning parameters to maximize bandwidth, managing security configurations, providing automatic fault recovery, and notifying users of errors and completion. It also provides in-place data sharing, allowing users to share large datasets from the storage repositories on which data resides. In effect, this sharing capability provides Dropbox-like functionality without the overhead of moving data to the cloud.

Globus Genomics uses the Globus transfer REST API to integrate transfer capabilities into Galaxy. This integration provides for the automation of large-scale data movement between remote endpoints and the Amazon cloud. It allows users to select a data source from which to download data (e.g., an acquisition machine) and/or to select destination endpoints to which data is deposited. Globus Nexus integration handles required authentication operations. As with other Globus Genomics tools, these transfer features can be integrated in a Galaxy workflow, allowing automated download and upload of data to be incorporated within a repeatedly executed process.

3.3 Elastic scalability

NGS analysis is a big data problem: significant compute resources are required to derive significantly useful results from large quantities of data. Analyzing just a single genome can take several days on a single processor, and many researchers wish to analyze hundreds of genomes. Yet few institutions provide biomedical researchers with access to large-scale compute resources, and indeed purchasing such resources is often not cost effective, especially when demand is sporadic. Thus, genome analysis is often stated as a major use case for cloud computing [5].

Globus Genomics addresses this challenge of limited access to computing by implementing a turnkey cloud computing solution that permits genome analysis pipelines to execute efficiently and cost-effectively on Amazon's elastic cloud infrastructure. This solution addresses a range of problems that otherwise hinder use of cloud computing for NGS analysis. In particular, it supports the configuration of compute resources with application software that matches the versions required by analysis pipelines; access from cloud resources to pipeline code and input and reference data; and in order to minimize cost and provide efficient execution, on-demand provisioning of compute resources when required, and the release of those resources when they are not in use.

We build on the CloudBioLinux [13] Amazon Machine Image (AMI) to create a Globus Genomics AMI, a point-in-time snapshot of a running virtual machine that we then clone as often as is required to create Amazon instances for analysis. The base CloudBioLinux AMI includes a collection of leading NGS analysis tools; we expand this set with additional tools identified via periodic user surveys. We also integrate client software such as an HTCondor worker and NFS client that are preconfigured to enable connection to the Globus Genomics head node.

Globus Genomics uses a custom elastic provisioner that runs continuously on a head node. This provisioner uses predefined policies to determine when to instantiate new AMI instances as worker nodes on the Amazon cloud. Specifically, it monitors the HTCondor queue to determine real-time job wait times. If the queue length or wait time exceeds a specified limit, the scheduler requests a new worker node. All worker nodes are hosted on spot instances, a term that Amazon uses to refer to instances that are charged at a rate set by a market place. The provisioner bids a maximum price defined by policy; Amazon delivers an instance at the current market price if that price is less than the maximum that our provisioner bid. If at some later time the spot price increases to the point that it exceeds the original bid, then Amazon terminates the instance. In that latter case, HTCondor will return any jobs that were running on the terminated instance to the queue until the provisioner is able to provision new resources. Any previous execution state is lost.

When allocating a new worker node, the provisioner uses the cloud-init utility to contextualize the Globus Genomics AMI so that it is configured to interact with the appropriate head node (e.g., HTCondor master and NFS server). Cloud-init is a Python based system for installing and configuring software on a newly created instance. We have also used this utility to contextualize a standard AMI from scratch, loading all required software and pipeline components on the fly. In situations with few software dependencies, this approach provides flexibility, but with many dependencies, contextualization time can be considerable.

The resource acquisition and release strategies implemented in dynamic provisioner could, in principle, be applied in other contexts. However, it is currently integrated with Globus Genomics and we have not explored other applications.

4. Adoption within the genomics community

Globus Genomics is in use by many academic biomedical research groups and several commercial companies as shown in Table 1. We have worked with these groups to create, deploy, and apply various computationally efficient, scalable analytical pipelines. Several of these research groups have achieved on the order of 10x speed-ups for various types of analyses, as compared with current approaches in which data is analyzed on desktops or shared departmental compute clusters. For example, researchers at the University of Washington previously processed a single exome on their lab cluster in ~24 hours. Using Globus Genomics, they now regularly process 10 exomes in parallel in ~20 hours due to parallel execution of analysis jobs and the increased power of EC2 instances. While there is room for even greater speed-up, these performance improvements are already transformative for these groups. Additionally, we have run multiple parallel variant calling workflows to evaluate limits on our current architecture. In work to date, we have run up to 48 analysis workflows concurrently with no significant slowdown. Beyond that scale, I/O appears to become a bottleneck as the multiple concurrent applications all access the same shared file system. We are working to improve I/O performance by provisioning high I/O rates using Amazon Web Services Provisioned I/O on the Elastic Block Store (EBS) we use to build the shared file system. We have also consulted with engineers from Amazon Web Services on how to configure EBS blocks for better performance.

To demonstrate the impact that this use of cloud computing can have on the work of a typical research lab, we present an exome analysis use case. The Dobyns lab at the University of Washington researches the nature and causes of a wide range of human developmental brain disorders including malformations of the forebrain, mid-hindbrain (brainstem and cerebellum) and cerebral cortex, as well as a wide spectrum of developmental disabilities such as autism, intellectual disability, early childhood developmental forms of epilepsy, and complex developmental disorders combining several of the above. The group acquires data from samples sequenced at the University of Washington and the commercial sequencing company Perkin Elmer.

Before adopting Globus Genomics, raw sequencing data of 10s of TBs were shipped on hard disks using services like FedEx from the sequencing facilities to the Dobyns lab[14]. Dr. Dobyns and his two postdocs would then analyze the data on a single multi-core machine within their laboratory. Globus Genomics significantly increases the scale at which the group is able to operate and reduces the considerable usage barriers that existed previously. Specifically, the group leverages the Globus Genomics service on Amazon EC2 with its pay-as-you-go billing model for resource usage, to provide a scalable and elastic execution environment for their analysis. They also rely on automated data transfer using Globus transfer from sequencing facilities to Globus Genomics.

The use of Globus transfer has allowed the Dobyns group to move and share tens of terabytes of data quickly and reliably. Data is now available for analysis within hours rather than weeks as in the past. The elastic computing environment has allowed them to analyze many exome samples in parallel, which accelerates time-to-results and ultimately, time-to-science. Another advantage of the Globus Genomics approach is that the group's analysis pipelines are now encoded in the system, which means the group can continue to use pipelines that were established by former members. The graphical environment allows newer group members to easily learn how pipelines work and to modify them to use newer methods or tools. The environment makes it easy to adjust or extend the pipelines for comparative analysis and further

Table 1: Some representative Globus Genomics user groups.

Research Domain	Institution	Analysis Type	Scale
Neurodegenerative disorders	University of Washington	Exome	100s of exomes
Tourette's syndrome, diabetes, autism	University of Chicago	Exome	1000s of samples
Breast cancer	University of Chicago	Exome, Whole Genome	100s of samples
Proteomics	University of Chicago	Proteomics	100s of samples
Molecular psychiatry and Autism	UC Irvine	Whole Genome	1000s of samples
Genomics Core	Washington University, St. Louis	Exome, RNA, ChipSeq	Varies
Genomics Core	Georgetown University	Exome, RNA, ChipSeq	Varies
Genomics Core	Boston University	RNA	Varies
Cardiovascular	Cardiovascular Research Grid Johns Hopkins University	ECG Analysis	1000s of time series samples

downstream interpretation.

This example illustrates how Globus Genomics reduces time-to-discovery by automating manually intensive steps in the sequence analysis process. As noted above, we have also designed it to optimize costs via the use of optimized elastic resources such as low-cost spot instances. Cost comparisons with other approaches employed by several Globus Genomics users indicate savings of up to 10x over traditional manual efforts and comparable commercial offerings. (Prices in this field are changing rapidly, so it is difficult to be more precise.)

5. Evaluation

We use two representative NGS analysis workflows: RNA-Seq and ChIP-Seq to examine the time and cost of execution on different cloud configurations. We also present usage data for the production Globus Genomics system.

5.1 RNA-Seq

RNA-Sequencing (RNA-Seq) [15] is a deep-sequencing technique used to explore and profile the entire transcriptome (the complete set of transcripts and their quantities, in a cell) of any organism. Analyzing an organism's transcriptome is important to understand the functional elements of the genome. We created a RNA-Seq analysis pipeline based on [25] as shown in Figure 3. This pipeline is used by researchers from the Progenitor Cell Biology Consortium community to highlight core stem cell signatures, observe changes that occur when stem cells lose or do not have pleuripotentiality, and define transcriptome signatures and pathways that are reflective of stem cells undergoing commitment to heart, lung, and blood lineages.

The pipeline uses Globus transfer to move data from an acquisition machine to Globus Genomics and then to archive resulting output data on a storage endpoint at the conclusion of the pipeline. It includes five analysis tools: TopHat, Cufflinks, Cuffmerge, Cuffdiff, and CummeRbund. TopHat is a fast splice junction mapper that is used to align RNA-Seq reads to mammalian-sized genomes (using the ultra-high-throughput short read aligner Bowtie) and analyze the mapping results to identify splice junctions between exons. Cufflinks is used to assemble these alignments into a parsimonious set of transcripts and then estimate the relative abundances of these transcripts. Cuffmerge merges several assemblies and automatically filters a number of transfrags that are considered artifacts using Cuffcompare. Cuffdiff is then used to find significant changes in transcript expression, splicing, and promoter use. Finally, CummeRbund creates a SQLite database that contains descriptions of relationships between genes, transcripts, transcription start sites, and Coding DNA Sequence (CDS) regions.

To evaluate the execution of this RNA-Seq pipeline in Globus Genomics we use a sample input file of 500,000 paired-end Illumina HiSeq reads. To simulate real usage, we define a workload comprising three independent executions of the resulting pipeline. In the following results we report the cost of executing this workload using three different system configurations: a single large EC2 instance, a single extra-large EC2 instance, and multiple extra-large EC2 instances with elastic scaling via our provisioner. In each case, we allow up to two instances of the pipeline to be executed on a node at a time. When using the provisioner, we define a policy that causes it to create new instances once queue delay reaches five minutes.

Table 2 gives the time and cost required to execute our workload in our three configurations. In each case, we present the cost of running the workload both when using on-demand pricing and when using spot prices. These costs are based on prices for instances in the US East 1A region in October 2013. The base prices for each instance type are shown in Table 3; the total cost is calculated from the total instance time, the number of instances used, and the per-instance price. For baseline comparison, execution of a single pipeline takes 20.5 hours on an m1.large instance and 14.3 hours on an m2.4xlarge instance. (Note that in both cases, two RNA-Seq pipelines can execute simultaneously on a single instance, which reduces total cost relative to sequential execution but also increases elapsed time per pipeline.)

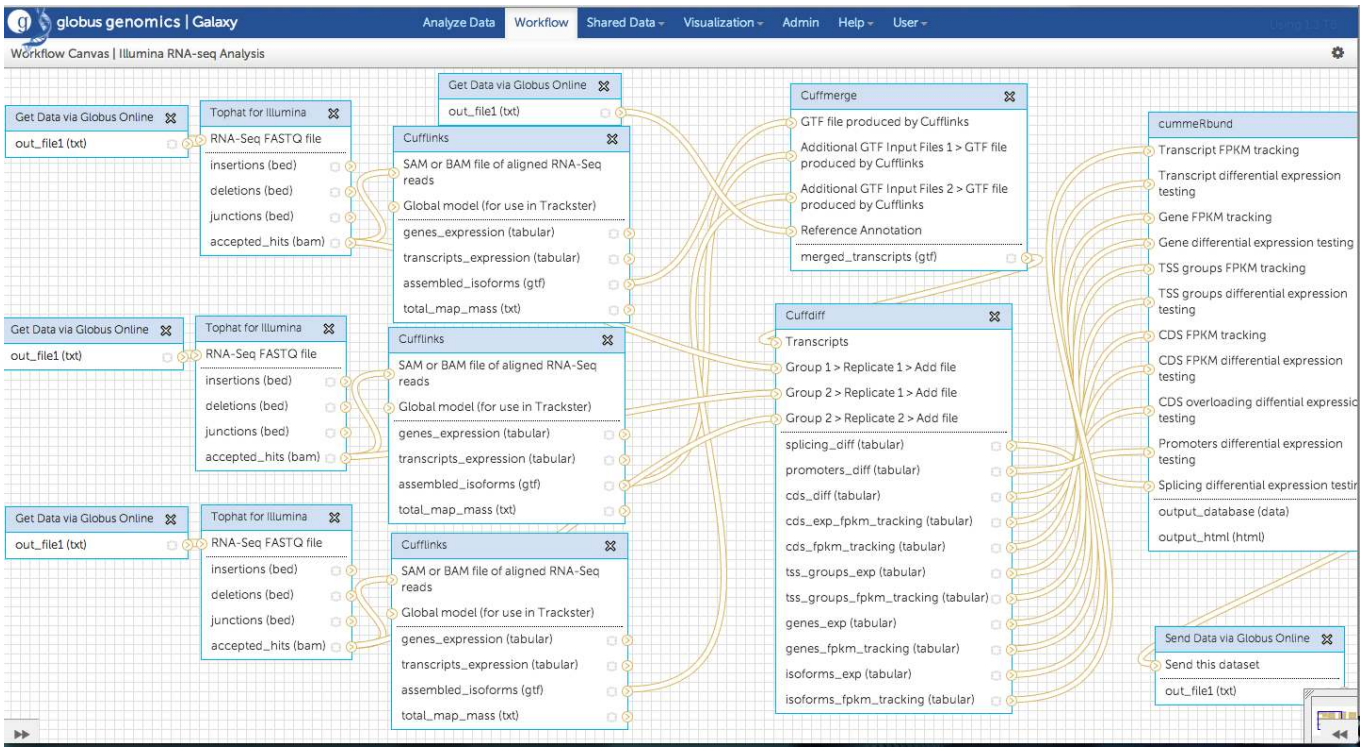

Figure 3 RNASeq Pipeline in Globus Genomics

The execution of three parallel pipelines using the cloud provisioner takes a total of 14.8 hours. First, two pipelines are scheduled on a single instance; then, after five minutes, a second instance is created to execute the third pipeline, as at that point the third job has been delayed for more than five minutes. A small additional cost is thus incurred due to the time spent creating and configuring the instance, but total execution time is reduced as multiple instances are used concurrently.

Table 2: Time and cost of running three RNA-Seq pipelines concurrently using different AWS infrastructure.

	Single M1.large	Single M2.4xlarge	Elastic scaling M2.4xlarge
Total hours	54.8	29.4	14.8 (2 nodes)
On-demand cost	$13.15	$48.23	$48.54
Low spot cost	$1.43	$4.12	$4.14
High spot cost	$109.60	$147.00	$148.00

Table 3: AWS instance prices for October 2013, in $US/hour.

	On-Demand	Low Spot	High Spot
M1.large Cost	0.24	0.026	0.5
M2.4xlarge Cost	1.64	0.14	5

We see that the use of spot instances results in a roughly order-of-magnitude reduction in cost when prices are low. At the other extreme, we find that when the spot price approaches the maximum, the total cost is three times the cost of using on-demand instances. We have found spot instance prices infrequently approach the stated high price; see also Figure 4, which shows the history of spot prices for m2.4xlarge instances in US East 1A during October, 2013. Our elastic provisioner is designed to switch to on-demand instances when spot prices exceed the on-demand cost. However, if spot prices increase beyond the on-demand price during execution, it can be more cost-efficient to continue execution at the increased price rather than restart the pipeline on an on-demand instance.

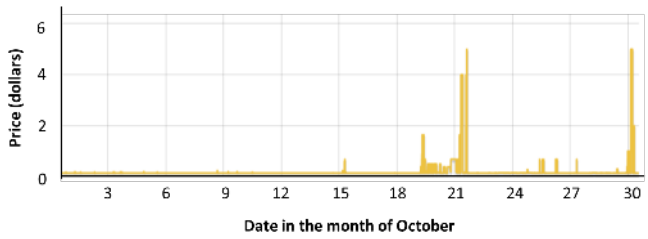

Figure 4: AWS spot instance price for m2.4xlarge instances in US-East-1a during October 2013.

5.2 ChIP-Seq

ChIP-sequencing (ChIP-seq) [16] is a method designed to analyze protein interactions with DNA. ChIP-seq combines chromatin immunoprecipitation (ChIP) with massively parallel DNA sequencing to study the binding sites and interactions between DNA and proteins. It is used to determine how transcription factors and other chromatin-associated proteins influence phenotype-affecting mechanisms, and how proteins interact with DNA to regulate gene expression. It is an important tool for understanding many biological processes and disease states.

We created a standard ChIP-Seq analysis pipeline shown in Figure 5, based on [26]. The pipeline includes Globus transfer to move data and Demultiplexer, Novoalign and SAMTools Sort tools to analyze ChIP-Seq data.

This pipeline's two input branches use Globus transfer to retrieve sequence files and tag files for a control (wild type) and sample (mutant), respectively. Both sequences are demultiplexed, and then aligned using Novoalign. The resulting Sequence Alignment/Map (SAM) file is converted to an indexed binary BAM file using SAMTools [17]. The BAM files are sorted using SAMTools Sort and then the two branches (control and sample) are merged using Model-based Analysis of ChIP-Seq (MACS)[18]. The resulting BED file indicates genomic regions of potential protein binding through differential analysis of the control and sample signals.

To evaluate the execution of this pipeline in Globus Genomics we execute 20 concurrent pipelines using 10 different input sequence files (~30GB each) and one tag file (with nine barcodes). A reference human genome "hg19.fa" is locally cached in Globus

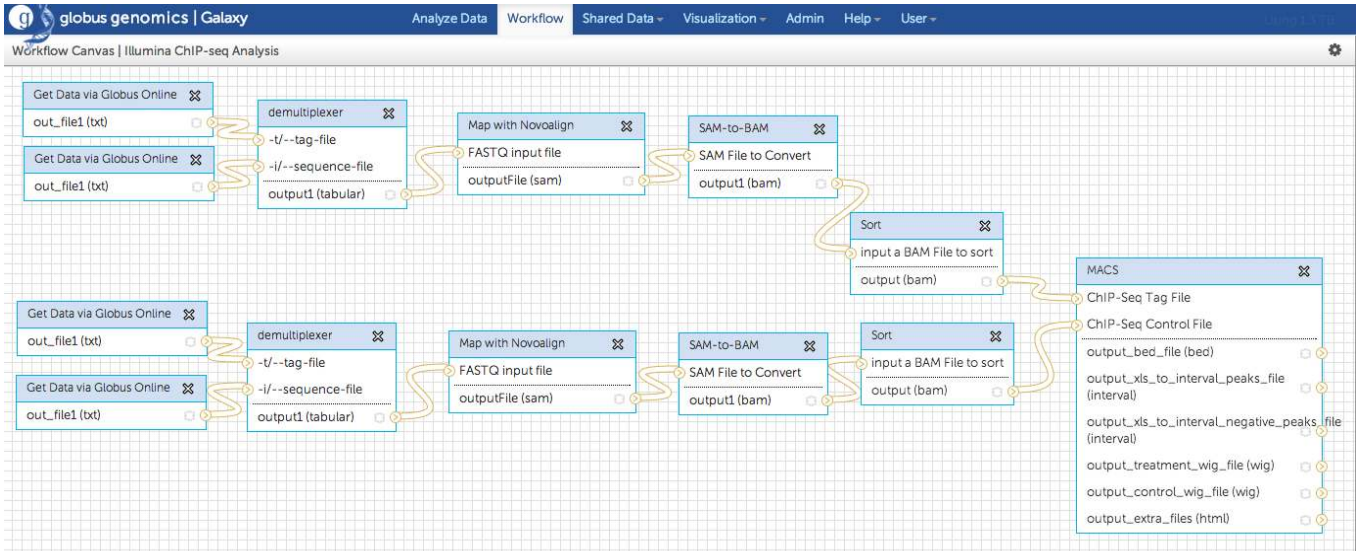

Figure 5: ChipSeq Workflow in Globus Genomics

Genomics. Table 4 shows the time and cost of executing these pipelines using Globus Genomics on the same three configurations as used above. Executing a single pipeline takes 8.2 hours on an m1.large instance and 3.6 hours on an m2.4xlarge instance. Both time and cost are significantly better when using elastic scaling. When using spot instances, 20 ChIP-Seq analysis pipelines can be executed in 4.5 hours and cost only $12.60.

Table 4: Time in hours and cost in $US of running 20 ChIP-Seq pipelines concurrently in different AWS configurations.

	Single M1.large	Single M2.4xlarge	Elastic scaling M2.4xlarge
Total hours	138.5	102.3	4.5 (20 nodes)
On-demand cost	33.24	167.77	147.60
Low spot cost	3.601	14.32	12.60
High spot cost	69.25	511.50	450

5.3 Production Usage

Figure 6 presents usage data for a shared Globus Genomics service for the month of October 2013. This particular Globus Genomics instance is shared by six research groups across the US. The figure shows only the spot instances used for analyses. It reports, for each day of the month, both the total number of instance hours used and the total spent on those AWS instances. The instances in this case are 32-core compute-intensive (cr1.8xlarge) instances. The considerable variation in usage emphasizes the value of elastic provisioning. At its peak, on the 28th of October, 521 instance hours are used for a total daily cost of $277. The average cost per instance-hour over the month is $0.56; the unusually high cost on October 23rd was due to an Amazon competition that briefly increased spot market prices.

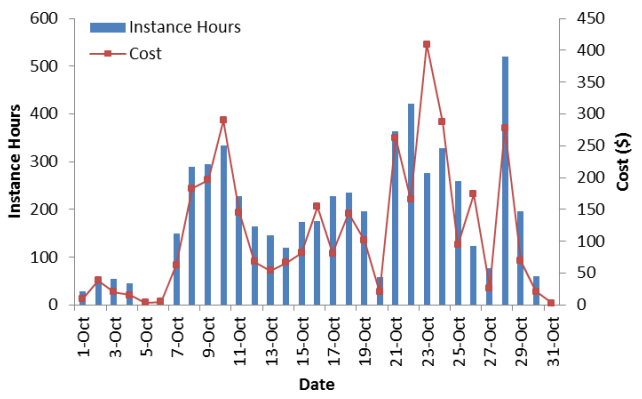

Figure 6: Globus Genomics spot instance usage in hours/day and cost in $/day for October 2013.

6. Relevance to XSEDE

As an early version of this paper was prepared for the XSEDE13 conference, we discuss briefly the relevance of this work to the XSEDE project. XSEDE defines its mission as "accelerat[ing] open scientific discovery by enhancing the productivity of researchers, engineers, and scholars and making advanced digital resources easier to use." Globus Genomics addresses this goal directly, leveraging XSEDE services (Globus data movement and identity management) to integrate a range of advanced digital resources (sequencing centers, a commercial cloud provider, and NGS analysis pipelines) in a manner that greatly reduces the cost and complexity of scientific discovery for a new community (NGS researchers) who have not historically made much use of advanced eScience infrastructure. In so doing, Globus Genomics addresses two key XSEDE goals:

- "Deepen and extend the impact of eScience infrastructure on research and education; in particular, to reach communities that have not previously made use of it; and
- Expand the environment through the integration of new capabilities and resources such as instruments and data repositories based on the identified needs of the community."

Globus Genomics currently achieves these goals without making use of XSEDE supercomputers. Our choice of Amazon cloud services rather than XSEDE systems for Globus Genomics computations is deliberate. At the scales at which our target users operate today, the costs associated with the use of Amazon cloud computers are modest, and Amazon's on-demand, pay-as-you-go storage and computing capabilities match sporadic user needs well. In addition, the genome analysis applications that we have deployed to date run efficiently on cloud resources. Genome analysis is a largely data reduction activity: significant input data is analyzed and only small amounts of output data are produced. Such tasks are well suited to commercial cloud providers such as Amazon that provide free data upload. Common genome analyses also lend themselves well to cloud computation as the analysis tools run efficiently on commodity hardware, and comparative analyses over many genomes can be trivially paralyzed at the pipeline level. The ability to deploy fully configured machine images (building upon the work of CloudBioLinux) alleviates the need for users to install and configure software tools on potentially heterogeneous computing platforms.

As NGS datasets become yet larger and our users progress to larger sample sizes, it may make sense to extend Globus Genomics to execute analysis pipelines on XSEDE computers. Several science gateways, such as the CIPRES Science Gateway [19], provide a successful model for the use of XSEDE resources via internet-accessible interfaces for the bioinformatics community. Similar approaches could be leveraged by Globus Genomics to enable progression from cloud resources to XSEDE resources as datasets become larger without significantly altering the interfaces through which users interact with their analyses. But for now, at least, we demonstrate that XSEDE services can expand access to advanced digital services without increasing demand on XSEDE supercomputers—an approach that, given constrained NSF supercomputer acquisition budgets, can contribute to expanding the number of researchers, engineers, and educators who benefit from advanced digital services.

Several features of the Globus Genomics implementation may both facilitate future integration with XSEDE and suggest strategies that may be helpful for other XSEDE applications.

Globus Genomics both exposes and builds upon a strong service model. Specifically, Globus Genomics leverages AWS APIs to access compute and storage resources. These powerful APIs enable sophisticated cloud scheduling, dynamic usage of unused computational resources, and use of elastic storage capacity. Globus Genomics also leverages the Globus Nexus and transfer REST APIs to outsource identity, group, and data management. We have found that by using published APIs and outsourcing important functionality to dedicated and reliable providers that we were able to more rapidly develop and operate production quality systems. The availability on XSEDE of APIs similar to those

supported by Amazon would make the deployment of Globus Genomics and other similar applications on XSEDE computers far easier.

Globus Genomics leverages mature and reusable frameworks, such as Chef, that simplify deployment, maintenance and extension. These frameworks enable us to quickly deploy entire infrastructures with minimal user interaction. In addition, these frameworks fit within a broader ecosystem that includes tools for managing and monitoring deployments to rapidly address problems as they occur. Similar approaches are used by enterprises such as Netflix to manage huge global cloud infrastructures. Similar approaches in XSEDE would provide comparable advantages to developers.

7. Sustainability

Scientific software suffers from under-investment, lack of effective economic models, and poor usability [20]. While the cost of developing and supporting scientific software is often high, current funding models for such software have been found to be inadequate. For example, while open source software has several success stories, just as many efforts lack contributions. Meanwhile, research funding is spread over many different projects and typically focuses on developing new approaches rather than supporting existing software.

Software- and platform-as-a-service approaches have been proposed as a means of reducing costs. As only one copy of the software stack need be supported, and users need not deploy or operate any software, total costs can be lower. On the other hand, the service provider incurs not only development costs but also infrastructure and support costs, and as a service becomes more useful and therefore more popular, the provider incurs increased costs associated with operating at larger scales. We believe that new utility models are needed to support such services—models that will involve a philosophical paradigm shift for users and the funding agencies that support them.

Table 5: Globus Genomics cost structure for exome, whole genome and RNA analyses.

Pipeline	Pipeline characteristics	Estimated price per sample
Exome	Quality control, alignment, variant calling, and annotation using the GATK best-practices pipeline for samples of paired-end FASTQ files (~5 GBases)	$5-$30
Whole Genome	Quality control, alignment, variant calling, and annotation for samples of paired-end FASTQ files (~80 GBases)	$20-$100
RNA	Quality control, alignment, exon count using cufflinks, and HT-Seq count for samples of paired-end FASTQ files (~5 GBases)	$5-$10

Globus Genomics represents a hybrid model, in that it leverages open source software, such as Galaxy and HTCondor, to create a utility-based service for the community. To support such projects, we have contributed extensions made for Globus Genomics back to their open source projects. For example, we have contributed various tool wrappers to the Tool Shed maintained by the Galaxy project so that users of Galaxy can also leverage these tools in their own deployments.

Globus Genomics is also an example of a utility model. We, the Globus Genomic operators, resell compute and storage capacity (in this case from Amazon), but also incur additional costs involved with usage that must be recouped in order to provide a sustainable resource for the community. To address these challenges we are exploring a pre-paid blended subscription model, in which users are charged for various components of costs that are incurred such as compute and storage resources that are directly consumed for their analysis, as well as a fractional contribution towards operations and technical support, new feature development and other enhancements to the service. Table 5 describes how the subscription model roughly translates to expected per sample pricing for various types of analysis. The range in pricing is driven by several factors including size and complexity of input data, depth of coverage, complexity of pipelines, and duration of storage using AWS. What we find interesting is the low costs that can be achieved via this hybrid open source/public cloud/utility approach.

8. Related work

Researchers are increasingly turning to the cloud to conduct a wide range analyses at scale without requiring the acquisition and operation of in-house resources. We can only cite a few representative papers [21-23].

In the genomics area, developments such as CloudBioLinux make it easier for researchers to run individual analyses on the cloud. Several groups have developed tools that leverage cloud infrastructure to run at scale, for example Crossbow [24], CloudBurst [25], Myrna [26], and CloudMan [11]. Yet others have investigated the use of technologies such as Hadoop to scale-out bioinformatics analyses on cloud resources [27]. Several companies offer commercial genome analysis services that operated on cloud resources: for example, DNANexus, Appistry, Illumina Basespace, and Seven Bridges Genomics.

Projects such as Rainbow [28] and Atlas2 Cloud [29], like Globus Genomics, focus more broadly on the end-to-end genome analysis lifecycle from data acquisition to cluster management, through to results retrieval. However, unlike Globus Genomics these projects do not support elastic provisioning of instances nor do they employ utility software-as-a-service models of usage.

9. Summary

The rapid decrease in sequencing costs has resulted in a huge increase in NGS data acquisition and transformed the field from a data-limited to a computationally limited discipline. Researchers now want to process hundreds and even thousands of sequenced genomes in a single study, and thus require thousands of compute hours to derive results. These researchers require efficient ways to manage large NGS datasets and to execute multi-step genome analysis pipelines across high performance computer systems.

Globus Genomics implements a turnkey, cloud-hosted solution to these challenges. Hosted on commercial Amazon cloud resources enables reliable and highly scalable execution of NGS analysis workflows. Integrated data management capabilities address the challenges associated with managing the movement of big data from acquisition through analysis and storage. Globus Genomics incorporates a wide range of leading analysis tools, preconfigured for immediate use by researchers.

The Globus Genomics architecture extends the Galaxy system with both superior data management capabilities and a novel scheduling architecture that can scale analyses elastically across a dynamic pool of cloud nodes. This approach allows researchers to

execute pipelines transparently on Amazon cloud computers, with parallel analyses executed in a fraction of the time that would be required on a local workstation. Researchers pay only for resources used. We have found costs to be modest, particularly when using spot instances.

Initial uptake has been promising. Several research groups now rely on Globus Genomics as their primary analysis platform. We have seen use across a range of NGS analysis domains including neurodegenerative disorders, molecular psychiatry, oncology, and cardiovascular research. The lessons learned developing and operating Globus Genomics as a platform for the research community are widely applicable to XSEDE. Globus Genomics leverages XSEDE services for data management, and like XSEDE it focuses on improving the usability of advanced digital resources and reducing cost and complexity for researchers.

Acknowledgements

This work was supported in part by the NIH through the NHLBI grant: The Cardiovascular Research Grid (R24HL085343) and by the U.S. Department of Energy, Office of Science, under contract DE-AC02-06CH11357. We are grateful to Amazon, Inc., for an award of Amazon Web Services time that facilitated early experiments. We thank Globus Genomics users for their invaluable contributions.

References

[1] L. Bo, B. Sotomayor, R. Madduri, K. Chard, and I. Foster, "Deploying Bioinformatics Workflows on Clouds with Galaxy and Globus Provision," in *High Performance Computing, Networking, Storage and Analysis (SCC), 2012 SC Companion:*, 2012, pp. 1087-1095.

[2] R. K. Madduri, P. Dave, D. Sulakhe, L. Lacinski, B. Liu, and I. T. Foster, "Experiences in building a next-generation sequencing analysis service using galaxy, globus online and Amazon web service," presented at the Proceedings of the Conference on Extreme Science and Engineering Discovery Environment: Gateway to Discovery, San Diego, California, 2013.

[3] M. J. Bamshad, S. B. Ng, A. W. Bigham, H. K. Tabor, M. J. Emond, D. A. Nickerson*, et al.*, "Exome sequencing as a tool for Mendelian disease gene discovery," *Nat Rev Genet,* vol. 12, pp. 745-755, 2011.

[4] M. Meyerson, S. Gabriel, and G. Getz, "Advances in understanding cancer genomes through second-generation sequencing," *Nat Rev Genet,* vol. 11, pp. 685-96, Oct 2010.

[5] L. D. Stein, "The case for cloud computing in genome informatics," *Genome Biol,* vol. 11, p. 207, 2010.

[6] J. Goecks, A. Nekrutenko, and J. Taylor, "Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences," *Genome Biol,* vol. 11, p. R86, 2010.

[7] I. Foster, "Globus Online: Accelerating and democratizing science through cloud-based services," *IEEE Internet Computing,* pp. 70-73, 2011.

[8] B. Allen, J. Bresnahan, L. Childers, I. Foster, G. Kandaswamy, R. Kettimuthu*, et al.*, "Software as a service for data scientists," *Commun. ACM,* vol. 55, pp. 81-88, 2012.

[9] R. Ananthakrishnan, J. Bryan, K. Chard, I. Foster, T. Howe, M. Lidman*, et al.*, "Globus Nexus: An identity, profile, and group management platform for science gateways and other collaborative science applications," in *Science Gateway Institute Workshop, co-located with IEEE Cluster*, 2013.

[10] Amazon. (2013). *Amazon Web Services*. Available: http://www.aws.amazon.com

[11] E. Afgan, D. Baker, N. Coraor, H. Goto, I. M. Paul, K. D. Makova*, et al.*, "Harnessing cloud computing with Galaxy Cloud," *Nat Biotech,* vol. 29, pp. 972-974, 2011.

[12] D. Thain, T. Tannenbaum, and M. Livny, "Distributed computing in practice: the Condor experience: Research Articles," *Concurr. Comput. : Pract. Exper.,* vol. 17, pp. 323-356, 2005.

[13] K. Krampis, T. Booth, B. Chapman, B. Tiwari, M. Bicak, D. Field*, et al.*, "Cloud BioLinux: pre-configured and on-demand bioinformatics computing for the genomics community," *BMC Bioinformatics,* vol. 13, p. 42, 2012.

[14] D. Sulakhe, Kettimuthu, R., & Dave, U, "High-performance data management for genome sequencing centers using Globus Online: A case study. In E-Science (e-Science), ," in *2012 IEEE 8th International Conference on, pp. 1-6, 2012*, 2012.

[15] Z. Wang, M. Gerstein, and M. Snyder, "RNA-Seq: a revolutionary tool for transcriptomics," *Nat Rev Genet,* vol. 10, pp. 57-63, Jan 2009.

[16] D. S. Johnson, A. Mortazavi, R. M. Myers, and B. Wold, "Genome-Wide Mapping of in Vivo Protein-DNA Interactions," *Science,* vol. 316, pp. 1497-1502, June 8, 2007 2007.

[17] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer*, et al.*, "The Sequence Alignment/Map format and SAMtools," *Bioinformatics,* vol. 25, pp. 2078-9, Aug 15 2009.

[18] Y. Zhang, T. Liu, C. A. Meyer, J. Eeckhoute, D. S. Johnson, B. E. Bernstein*, et al.*, "Model-based analysis of ChIP-Seq (MACS)," *Genome Biol,* vol. 9, p. R137, 2008.

[19] M. A. Miller, W. Pfeiffer, and T. Schwartz, "The CIPRES science gateway: a community resource for phylogenetic analyses," presented at the Proceedings of the 2011 TeraGrid Conference: Extreme Digital Discovery, Salt Lake City, Utah, 2011.

[20] I. Foster, V. Vasiliadis, and S. Tuecke. (2013, Software as a Service as a path to software sustainability. Available: http://dx.doi.org/10.6084/m9.figshare.791604

[21] X. Qiu, J. Ekanayake, S. Beason, T. Gunarathne, G. Fox, R. Barga*, et al.*, "Cloud technologies for bioinformatics applications," in *2nd Workshop on Many-Task Computing on Grids and Supercomputers* 2009, p. 6.

[22] R. L. Grossman, Y. Gu, J. Mambretti, M. Sabala, A. Szalay, and K. White, "An overview of the open science data cloud," in *19th ACM International Symposium on High Performance Distributed Computing*, 2010, pp. 377-384.

[23] K. Chard, M. Russell, Y. A. Lussier, E. A. Mendonça, and J.C. Silverstein, "Scalability and Cost of a Cloud-based Approach to Medical NLP," presented at the 24th international Symposium on Computer-Based Medical Systems, Bristol UK, 2011.

[24] B. Langmead, M. C. Schatz, J. Lin, M. Pop, and S. L. Salzberg, "Searching for SNPs with cloud computing," *Genome Biol,* vol. 10, p. R134, 2009.

[25] M. C. Schatz, "CloudBurst: highly sensitive read mapping with MapReduce," *Bioinformatics,* vol. 25, pp. 1363-9, Jun 1 2009.

[26] B. Langmead, K. D. Hansen, and J. T. Leek, "Cloud-scale RNA-sequencing differential expression analysis with Myrna," *Genome Biol,* vol. 11, p. R83, 2010.

[27] J. Ekanayake, T. Gunarathne, and J. Qiu, "Cloud Technologies for Bioinformatics Applications," *Parallel and Distributed Systems, IEEE Transactions on,* vol. 22, pp. 998-1011, 2011.

[28] S. Zhao, K. Prenger, L. Smith, T. Messina, H. Fan, E. Jaeger, *et al.*, "Rainbow: a tool for large-scale whole-genome sequencing data analysis using cloud computing," *BMC Genomics,* vol. 14, p. 425, 2013.
[29] U. Evani, D. Challis, J. Yu, A. Jackson, S. Paithankar, M. Bainbridge, *et al.*, "Atlas2 Cloud: a framework for personal genome analysis in the cloud," *BMC genomics,* vol. 13, p. S19, 2012.