# *High-Performance Data Management for Genome Sequencing Centers Using Globus Online: A Case Study*

Dinanath Sulakhe
Computation Institute
Argonne National Laboratory and
University of Chicago
Chicago, IL 60637 USA
sulakhe@mcs.anl.gov

Rajkumar Kettimuthu
Computation Institute
Argonne National Laboratory and
University of Chicago
Chicago, IL 60637 USA
kettimut@mcs.anl.gov

Utpal Dave
Computation Institute
Argonne National Laboratory and
University of Chicago
Chicago, IL 60637 USA
dave@ci.uchicago.edu

***Abstract*— In the past few years in the biomedical field, availability of low-cost sequencing methods in the form of next-generation sequencing has revolutionized the approaches life science researchers are undertaking in order to gain a better understanding of the causative factors of diseases. With biomedical researchers getting many of their patients' DNA and RNA sequenced, sequencing centers are working with hundreds of researchers with terabytes to petabytes of data for each researcher. The unprecedented scale at which genomic sequence data is generated today by high-throughput technologies requires sophisticated and high-performance methods of data handling and management. For the most part, however, the state of the art is to use hard disks to ship the data. As data volumes reach tens or even hundreds of terabytes, such approaches become increasingly impractical. Data stored on portable media can be easily lost, and typically is not readily accessible to all members of the collaboration. In this paper, we discuss the application of Globus Online within a sequencing facility to address the data movement and management challenges that arise as a result of exponentially increasing amount of data being generated by a rapidly growing number of research groups. We also present the unique challenges in applying a Globus Online solution in sequencing center environments and how we overcome those challenges.**

***Index Terms*— Globus, Globus Online, GridFTP, sequencing center, data transfer, data management, grid, cloud, next-gen sequencing, translational medicine**

## I. Introduction

Today's research communities in various scientific domains such as physics, astronomy, cosmology, and biology are dealing with an unprecedented data deluge [1]. Technological advances in scientific methodologies and instrumentation are generating massive amounts of data that require sophisticated and high-performance computational capabilities. In the biomedical field, for example, the low-cost [2] availability of next-generation and third-generation [3] sequencing in the past few years has encouraged larger as well as smaller research groups to have many of their patients' DNA and RNA sequenced in order to help improve diagnosis and treatment plans. Dobyns Laboratory [4], a research group at the University of Washington, Seattle, has sequenced hundreds of its patients in the past year, resulting in tens of terabytes of data [5]. The lab uses various sequencing centers (PerkinElmer, Broad Institute, University of Washington) depending on the type of sequencing required. Currently, most of these sequencing centers send the massive raw sequence data back to the research labs on multiple hard disks, using snail mail (Fedex) [1]. It is an extremely inefficient process. Small research labs suffer from a lack of resources and the expertise to use available advanced computational solutions for data handling, and the large sequencing centers require high-performance tools that would allow them to handle data for hundreds of their clients or researchers.

It is an extremely challenging task for sequencing centers to manage hundreds of researchers and their data at the scale of petabytes, as well as to implement user access control mechanisms and security. They all demand a robust yet simple and transparent high-performance data management solution that provides data movement among multiple locations, security and authentication integrated within local settings, and flexible access control.

In this paper, we explore the use of Globus Online [6] to address the needs of a large, multiuser research data facility such as a sequencing center. We highlight the challenges that a typical sequencing center would encounter related to its data management needs, and we discuss how Globus Online can be set up to address these challenges.

The remainder of the paper is organized as follows. Section II provides background on Globus Online. Section III discusses the use cases under consideration. Section IV and Section V describe two approaches to addressing the data movement and access control issues for the use cases described in Section III. We explain the advantage of our approach in Section VI. In Section VII we discuss how Globus Online can be used for further sequence analysis, and in Section VIII we outline future work. We conclude in Section IX with a brief summary.

## II. Globus Online

Globus Online (GO) seeks to provide an easy-to-use and powerful set of services and tools for research data management.

Globus Transfer [9] implements methods for managing the transfer of single files, sets of files, and directories, as well as rsync-like directory synchronization. It can manage security credentials, including cases where transfers cross multiple security domains; select transfer protocol parameters for high performance; monitor and retry transfers when there are faults, and allow users to monitor status. This functionality can be accessed through a variety of methods, including a web-based GUI, a command line interface suitable for scripting and automated workflows, and an API that can be called from applications.

Globus Storage (currently under development) will enable users to store data wherever it is needed, access it from anywhere via different protocols, update it, version it, take snapshots, and share versions with collaborators.

Globus Collaborate (currently under development) will make it easy for researchers to track their work, share data among their project's members, and publish information via point-and-click web interfaces.

Globus Online adopts a "software as a service" (SaaS) model to simplify research data management. The key idea behind SaaS [10, 11] is that a trusted provider operates the software as a hosted service that can then be used by many clients. Thus, users can instantly use powerful research data management and collaboration capabilities without installing any software. New features are immediately available to users, and GO allows experts to intervene and troubleshoot on the user's behalf in order to deal with more complex faults.

Globus Transfer leverages GridFTP [7, 8] for high-speed data movement. The GridFTP protocol specification [7] extends the File Transfer Protocol (FTP) to provide secure, reliable, and efficient data transfer. GridFTP extensions provide for checkpointing, parallelism (i.e., the use of multiple socket connections between pairs of data movers), and strong security on both control and data channels. Globus GridFTP [8] is an open source GridFTP implementation developed primarily at Argonne National Laboratory and the University of Chicago. GO's support for GridFTP delivers the benefits of grid technology, including support for heterogeneity and local control of access control and resource allocation policies at individual endpoints [12]. Data sources and sinks are termed "endpoints" in Globus Transfer. Globus Transfer acts as a third-party agent and moves data between such endpoints.

Users can sign on to Globus Online using widely adopted federated identity systems such as InCommon [13] as well as from OpenID providers such as Google. Globus Online uses Grid Security Infrastructure (GSI) [14], which is based on X.509 certificates, to authenticate with endpoints. However, users do not have to deal with certificates directly. They can access the endpoints in Globus Online by using username/password, OTP, and the like, which they would normally use to access their endpoints. GO can handle transfers across multiple security domains with multiple user identities. Having authenticated and requested a transfer, a client can disconnect and return later to find out what happened. GO tells the user which transfer(s) succeeded and which, if any, failed. It notifies the user when a transfer completes, or whether a critical fault has occurred such as a deadline not being met, or whether a transfer requires additional credentials to proceed.

Globus Transfer mediates transfers between two GridFTP servers. Traditionally, setting up a GridFTP server has been a complex task. But as part of the Globus Online effort, we have developed tools that make the GridFTP installation relatively trivial. Here we describe briefly the tools that enable easy GridFTP setup for the two most common use cases.

### A. Globus Connect

Globus Connect comprises specially packaged GridFTP server binaries for Windows, Mac OS X, and Linux that turn a personal computer into a Globus Transfer endpoint. A user can easily install Globus Connect with one click and one copy/paste without requiring administrative privileges. Globus Connect makes only outbound connections. Therefore, even if a user's machine is behind a firewall or network address translation device, the user can move data in and out of it using Globus Connect via Globus Transfer as long as outbound connections are allowed from the user's machine.

### B. Globus Connect Multi-User

Globus Connect Multi-User (GCMU) [15] is a multiuser version of Globus Connect designed for multiuser environments such as campus clusters. GCMU combines a GridFTP server and MyProxy Online Certificate Authority (CA) server [16]. GCMU installation is easy: an interactive script prompts the system administrator for ten simple inputs and then sets up a secure Globus Transfer endpoint. When GCMU is installed, MyProxy Online CA ties to a local authentication system such as LDAP [17] via a pluggable authentication module [18] API. Once the endpoint is set up, a user can access it from Globus Transfer using the credential (username/password, OTP, etc.) normally used to access that cluster.

## III. Sequencing Center And Research Lab Use-Case

At the Computation Institute at the University of Chicago, we are working with multiple sequencing centers and research labs in applying our Globus Online solution for their data management requirements. In this paper we use our experience in collaborating with the Dobyns Lab and PerkinElmer in order to highlight typical scenarios at a sequencing center and to show how we address these scenarios by setting up a Globus Connect Multi-User (GCMU) [15] endpoint with different configurations.

Typically, after research labs send specimens for exome or complete genome sequencing using next-generation sequencing techniques, the sequencing centers generate raw sequence data ranging from a few gigabytes to hundreds of gigabytes per specimen. With lower costs of next-generation sequencing, research labs are sending hundreds of patients' specimens every few months. The sequencing centers are handling these

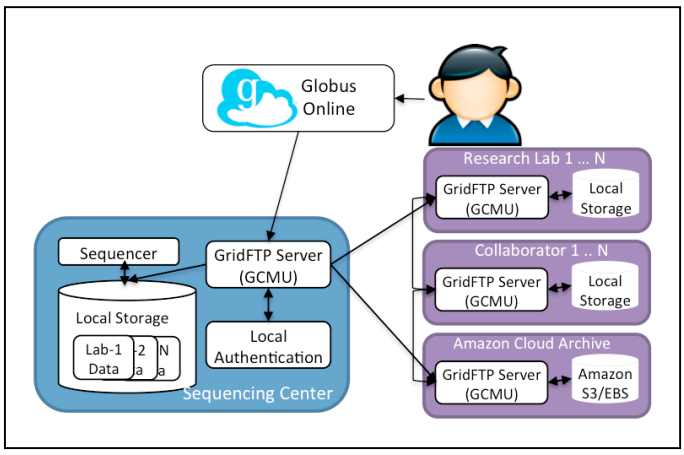


*Figure 1: Sequencing Center and Research Lab Use-Case*

loads for hundreds of their clients/researchers. A sequencing machine writes the raw data on a local file system in different directories created for each researcher. Currently, most of the sequencing centers then copy each researcher's data onto a hard disk and ship the disks using Fedex or UPS (snail mail). The research labs, after receiving the hard disks, mount them onto local compute resources and use the raw sequence data for further analysis [1].

Shipping the data on hard disks to each research lab is very inefficient and adds significant overheads with respect to time and resources involved. The problem is made worse when the research lab needs to share the data with their other collaborators. For example, after copying the data to local storage, the Dobyns Lab ships the hard disks to their collaborators. It usually takes a few weeks to months for each researcher in the collaboration to get a copy of the raw sequence data for further analysis. Any data loss or data corruption in the disks may require restarting the whole process of getting new disks from the sequencing centers.

Globus Online with its SaaS approach and local access control capabilities is an ideal fit for sequencing centers and the research labs involved. We installed a GCMU instance at a sequencing center (PerkinElmer) and installed another GCMU instance at the research lab (Dobyns Lab). By using the GO Web interface, the researchers at Dobyns Lab were able to securely transfer their data from the PerkinElmer GCMU endpoint to their local endpoint within minutes of the data being generated. With an average exome sequence of 10 GB, we were able to transfer sequence data for 10 patients from PerkinElmer to Dobyns Lab in about 3 hours each, achieving transfer rates between 8 Mbps and 10 Mbps, a task that otherwise would have taken weeks via Sneakernet. Once the data is moved to the research lab, the lab can provide its collaborators with access to its local endpoint. The collaborators got access to the data immediately by installing their own GO endpoint (GCMU or GC) and transferring data using Globus Online. The research lab may also set up other GO endpoints for archival storage or at various points of analysis, such as a cluster or cloud resources such as Amazon EC2. Formerly, it could take weeks to months for the research labs to get their data; with GO they are able to get the data within a few hours after it is generated by sequencing centers and use it for further analysis.

While GCMU with its multiuser setup is ideal for sequencing centers that cater to large volumes of data and to a large number of researchers, firewall regulations and local authentication methods at the sequencing centers make the setup nontrivial. For various reasons such as HIPAA compliance and data sensitivity, sequencing centers have strong firewall protection. These firewall and data protection requirements necessitate a special configuration of Globus Connect Multi-User at the sequencing centers. Another issue is the need for creating local user accounts for each user with the typical GCMU setup. In our initial prototype with PerkinElmer and Dobyns Lab, only a few researchers needed access to the data. Thus, it was fairly simple to create user accounts for each of the researchers. But this process can be challenging when a sequencing center wants to provide GCMU access to hundreds of its clients or research labs and with multiple users in each lab. Creating user accounts for all of the hundreds of users can be not only tedious, but unnecessary for a sequencing center when users are not going to log into the system. Users access the endpoint only from Globus Online for getting their data.

In the next section we describe how we addressed the first issue of strong firewall constraints, and in Section 5 we describe an approach to disseminating data to users without having to create a local account for each of them, while retaining the same level of access control that individual user accounts provide.

## IV. Setting up GCMU With User Accounts

Typically, in a sequencing center all the servers and file systems are behind a firewall and stay within an internal network for security reasons. One of the first requirements in setting up a GCMU endpoint is to get a host machine with a public IP that can be reached by GO and other endpoints (see Figure 1). The sequence data for all the users is stored on an internal file system. Typically, a sequencer writes the sequence data to a user-specific directory on a file system. To allow access to each user's data from outside the firewall via GCMU (for authentication or actual data channel connections), we set up a server with a public IP in a demilitarized zone (DMZ) with appropriate ports opened (Port 2811 for GridFTP server, 7512 for MyProxy authentication requests, Port range 50000-51000 for actual inbound and outbound data channel connections). DMZ is a portion of the network near a facility's local network perimeter with security policies and enforcement mechanisms tailored for remote access and high performance. Individual accounts for each user were created on this DMZ machine. We then mounted the internal file system on the DMZ host with read-only access. A GCMU was then installed on this host and configured in such a way that each user was restricted to access only his or her data.

But this configuration is not straightforward. Since the user accounts are available only on the DMZ machine, the internal file system was mounted as read-only for all users, and the path restriction functionality in GCMU was used to restrict users to access only their data. Specifically, we created a symlink that

points to user data in the mounted internal file system on the user's home directory in the DMZ machine. GCMU is configured to allow each user to access only the symlink available on his or her home directory, while providing the ability to follow symlinks. Thus, the user was able to access only his or her own data and nobody else's data.

In our specific use case, say user Joe's data is available at "/data/joe" and user Mary's data is available at "/data/mary" on the internal file system. "/data" is mounted as read-only on the DMZ machine. User Joe's home directory (/home/joe) on the DMZ machine has a symlink "data" that points to "/data/joe" and user Mary's home directory (/home/mary) has a symlink "data" that points to "data/mary."

Since we created accounts for each user on the PerkinElmer's GCMU host, researchers at Dobyns Lab used the username and password for their account to access the data through the GO Web UI or CLI. Once the PerkinElmer GCMU endpoint was ready, we set up a GCMU endpoint at the Dobyns Lab on top of their local storage and a GCMU endpoint on Amazon EC2 with Amazon S3-based and EBS-based storage for archiving the sequence data. Once all the endpoints were in place, the Dobyns Lab members were able to select the sequencing center endpoint from the GO Web UI, browse their sequence data immediately, and transfer it to their local endpoint as well as transfer the data to a permanent archive created on Amazon S3. The screenshot in Figure 2 shows the PerkinElmer sequencing center's GCMU endpoint on the left-hand side and the Dobyns Lab local GCMU endpoint on the right-hand side.

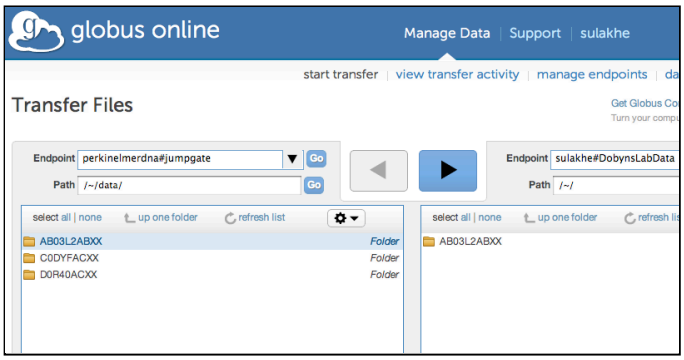


*Figure 2: Screenshot showing the PerkinElmer and Dobyns Lab endpoint in GO Web UI*

## V. Setting up GCMU without Unix accounts

The above approach still requires that an account be created on the system for each user. A sequencing center will have large number of users. And the users access the system only for getting their data. They do not perform any computation on the system. They do not even need to log into the system. Therefore, one would like to avoid creating an account for each user who needs to access the data. In this section, we describe one solution that we have developed to address this issue.

Once GCMU is installed, the Unix super server daemon xinetd listens on port 2811 and invokes the GridFTP server when a connection to port 2811 is made by the client. To start, the server process runs as root. The first step that must happen is mutual authentication. The server verifies that the client is who it claims to be, and the client verifies that the server is who it claims to be. Once the authentication is successful, the server determines the local user id by which the request should be executed and does a "setuid" to the local user id. File system security is handled through normal operating system mechanisms. Once the process is running as an unprivileged user, it is subject to access control and quotas imposed by the operating system.

The real challenge in the use case under consideration is to control access to the data for each GCMU (GridFTP) user without creating an account on the system. An important criterion is that the solution has to be simple (much simpler than creating accounts for each user). Our solution uses the fact that each user has a unique Grid ID issued by Globus Online CA and makes use of that unique Grid ID to map each user to a specific directory path on the local file system. The sequencing center administrator needs to create just one normal (unprivileged) user account on the data distribution machine in the DMZ and run the GridFTP server in GCMU as that unprivileged user. Figure 3 shows the end-to-end flow of our solution. A user creates an account on Globus Online, which then issues a certificate with a unique Grid identifier. The user sends the id to the administrator of the sequencing center, who creates a mapping of the Grid identifier to a directory path for this user. These four steps (steps 1-4 in Figure 3) are one-time activities. Once the user gets the Grid id mapped with the sequencing center, the user can access the sequencing center GCMU from Globus Online and move data in and out of it. When the user accesses the sequencing center endpoint, Globus Online authenticates with the GridFTP server in GCMU on the user's behalf by providing the user's certificate. Once the authentication is done, the GridFTP server checks whether this user has directory path mapping. If so, the user is allowed to access files in that directory path alone. If not, the user is denied access to this endpoint.

Our solution allows the administrator to restrict access for each user to a unique directory path without creating accounts for each user. All transfer requests are executed as an unprivileged user, but each request is restricted to a specific directory based on the authenticated user. This strategy is

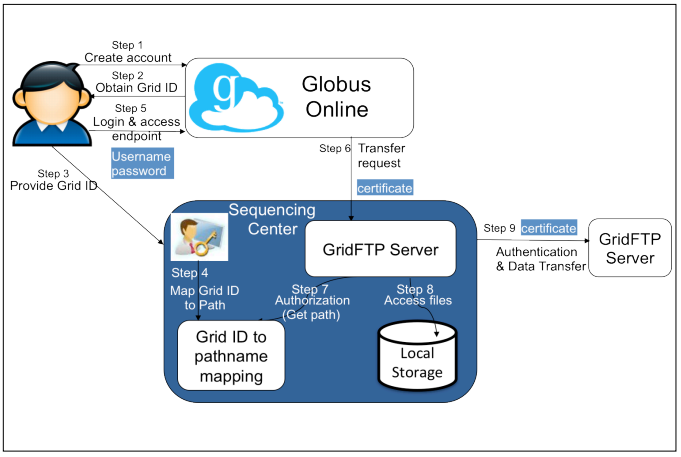


*Figure 3: GCMU setup without Unix accounts*

achieved by using a combination of GridFTP server options and gridmap file. For example, if there are N users, user1, user2, … userN, they can be restricted to access only /home/globus/user1, /home/globus/user2, … /home/globus/userN, respectively. If the data for the users is present elsewhere, the administrator can just create a symlink in each of these directories that points to the actual data and can configure the GridFTP server to allow following symlinks (as described in the previous section).

## VI. Advantages of Using GO

Globus Online eliminates the overheads involved in shipping the sequence data via Sneakernet to various research labs and their collaborators. It allows researchers to immediately start transferring data as soon as a sequencing machine generates it. Typically, most research groups send their sequencing samples in batches, and the sequencing centers send the data back on disks for the whole batch. In our use case with Dobyns Lab, the samples were sent in batches ranging from 40 to 100 exomes to various different sequencing centers including PerkinElmer. Globus Online allows researchers to start transferring each exome as soon as it is sequenced. As a result, the researchers need not wait for all the exomes to be sequenced, copied to disks, and shipped. GO also allows all the involved collaborators to immediately access the data and not wait for the disks to be shipped. Even if the sequencing center were to copy a subset of exomes on a separate disk and ship it on a per day basis, transferring data over the network is still significantly better for the given exome size (25 GB) even with the moderate transfer rate (8-10 Mbps) we are getting: it would take only 6 hours to transfer an exome data. Even if the exome size gets bigger, we can the transfer over the network as the data is being produced.

Simple file transfer mechanisms such as secure copy (scp), rsync, and FTP suffer from reliability and performance problems, while the high-performance GridFTP protocol has traditionally been difficult to use. Systems such as Dropbox and YouSendIt provide simplicity and a degree of delivery guarantee, but they are not suitable for the large datasets that pertain here.

PhEDEx [19], CERN's File Transfer Service (FTS), and LIGO's Data Replicator (LDR) [20] have fault tolerance and improved performance capabilities; but installing and operating complex software of this type can present a barrier to easy use.

Globus Online offers a hosted solution to large data transfer challenges over GridFTP, by providing a robust, reliable, secure, and highly monitored environment for file transfers that has powerful yet easy-to-use interfaces. GO provides a simple Web UI designed to serve the needs of less technical users, a command line interface that enables scripting for use in automated workflows, and a REST interface for system builders to integrate file transfer solutions for their end users.

Given these advantages of GO and given the demonstrated success of Globus Online on a wide spectrum of resources—ranging from a user's laptop to small department clusters to regional supercomputer centers to leadership computing facilities and national cyber infrastructure—GO is a logical step for sequencing centers to adopt in order to move and manage the enormous volumes of data generated at these centers. The approach proposed in this paper will allow sequencing centers to exploit the advantages of Globus Online with minimal effort for both the system administrators at the centers and their end users.

## VII. Using GO in Further Analysis of Sequence Data

While in this paper we address the use of Globus Online within the context of sequencing centers, GO usage can be easily expanded to help address the needs of researchers performing the analysis of sequence data. The raw sequence data is analyzed by using various next-generation sequence analysis tools such as Burrows-Wheeler Aligner (BWA) [21], Picard [22], and GATK [23], as well as many others. These analyses may be performed manually or through workflow tools such as Galaxy [24] by using predefined analytical pipelines. Globus Online will be extremely helpful in managing the datasets used and generated during these analyses by transferring the input sequence data to the points of computation and transferring the results of analyses or any intermediate results to permanent storage or archives or sharing them with other researchers and collaborators. In addition to the new capabilities discussed in Section VIII, we are currently integrating Globus Transfer capabilities in analytical platforms such as Galaxy in order to allow researchers the flexibility of managing their data transfers from within these platforms.

## VIII. Future Work

Since we have successfully demonstrated the application of SaaS-based capabilities offered by Globus Online to move sequence data, we plan to apply other research data management capabilities offered by Globus Online for sequencing use cases. We will the make the configurations that we described in Sections IV and V readily available as alternative configurations of GCMU. We plan to leverage the group management capabilities in Globus Online as appropriate in order to enhance the required access control mechanisms. As the sequencing data keeps growing, data storage is becoming a complex task. We plan to use Globus Storage to enable users to store data wherever it is needed, access it from anywhere via varying protocols, update it, version it, take snapshots, and share versions with collaborators. In addition to providing easy access to data, we also need to provide a way for researchers to store metadata related to the sequence data. Metadata management is vital for effective research. Researchers often tend to do ad hoc metadata management and create custom metadata catalogs with narrow functionality. Here again, we plan to leverage the SaaS approach to provide a simple, flexible model that life science researchers can put to use quickly.

## IX. Conclusion

We have outlined the data management challenges that sequencing centers and researchers face especially with the enormous growth in the volume of the data being generated with more and more affordable next generation sequencing techniques. We described how Globus Online, a hosted service,

addresses the data movement challenges by providing high-performance and fire-and-forget data movement capabilities. We then presented some unique challenges that sequencing centers have in adopting the Globus Online solution. We proposed two approaches to address those challenges while still providing system administrators the ability to do fine-grained access control for their users and provide users with all of the simplicity and ease of use that Globus Online provides.

## ACKNOWLEDGMENT

We thank Dr. William Dobyns from the University of Washington and members of his lab and Edward Szekeres from PerkinElmer for their help and support in setting up endpoints.

## REFERENCES

[1] *S*chadt, E. E., Linderman, M. D., Sorenson, J., Lee, L., and Nolan, G. P. Computational solutions to large-scale data management and analysis. *Nature Reviews. Genetics*, *11*(9), 647-657, 2010.

[2] Drmanac, R., Sparks, A. B., Callow, M. J., Halpern, A. L., Burns, N. L., Kermani, B. G., et al. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science*, *327*(5961), 78-81, 2010.

[3] Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., et al. Real-time DNA sequencing from single polymerase molecules. *Science*, *323*(5910), 133-138, 2009.

[4] http://depts.washington.edu/dlab/home.php

[5] Rivière, J. B. De novo mutations in the actin genes ACTB and ACTG1 cause baraitser-winter syndrome. *Nature Genetics, 44*(4), 440-444, 2012.

[6] Foster, I. Globus Online: Accelerating and democratizing science through cloud-based services. *IEEE Internet Computing,* May/June, 70-73, 2011.

[7] Allcock, B., Bresnahan, J., Kettimuthu, R., Link, M., Dumitrescu, C., Raicu, I. and Foster, I., The Globus striped GridFTP framework and server. In SC'2005, 2005.

[8] Allcock, W. GridFTP: Protocol extensions to FTP for the Grid. GFD-R-P.020, Global Grid Forum, 2003.

[9] Allen, B., Bresnahan, J., Childers, L., Foster, I., Kandaswamy, G., Kettimuthu, R., Kordas, J., Link, M., Martin, S., Pickett, K. and Tuecke, S. Software as a service for data scientists. *Communications of the ACM*, *55*(2), 81-88, 2012.

[10] Dubey, A., and Wagle, D. Delivering software as a service. *The McKinsey Quarterly,* May 2007.

[11] Waters, B. Software as a service: A look at the customer benefits. *Journal of Digital Asset Management, 1*(1):32-39, 2005.

[12] Foster, I., Kesselman, C., and Tuecke, S. The anatomy of the Grid: Enabling scalable virtual organizations. *International Journal of Supercomputer Applications, 15*(3):200-222, 2001.

[13] Barnett, W., Welch, V., Walsh, A., and Stewart, C. A. A roadmap for using NSF cyberinfrastructure with InCommon, http://hdl.handle.net/2022/13024, 2011.

[14] Foster, I. Kesselman, C. Tsudik, G., and Tuecke, S. A security architecture for computational grids. In 5th ACM Conference on Computer and Communications Security Conference, 1998, pp. 83-92.

[15] R. Kettimuthu, L. Lacinski, M. Link, K. Pickett, S. Tuecke and I. Foster, Instant GridFTP. In 9th Workshop on High Performance Grid and Cloud Computing, May 2012.

[16] Koutsonikola, V., and Vakali, A. LDAP: Framework, practices, and trends. *IEEE Internet Computing, 8*(5):66-72, 2004.

[17] Novotny, J., Tuecke, S. and Welch, V. An online credential repository for the Grid: MyProxy. In 10th IEEE International Symposium on High Performance Distributed Computing, San Francisco, IEEE Computer Society Press, 2001.

[18] Samar, V., and Schemers, R. Unified login with pluggable authentication modules (PAM). OSF RFC 86.0, 1995.

[19] Rehn, J., Barrass, T., Bonacorsi, D., Hernandez, J., Semeniouk, I., Tuura, L., and Wu, Y. PhEDEx high-throughput data transfer management system. In Computing in High Energy Physics (CHEP), Mumbai, India, 2006.

[20] Chervenak, A., Schuler, R., Kesselman, C., Koranda, S., and Moe, B. Wide area data replication for scientific collaborations. In 6th IEEE/ACM Int'l Workshop on Grid Computing, 2005.

[21] Li, H., and Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler Transform. *Bioinformatics, 25*, 1754-1760, 2009.

[22] http://picard.sourceforge.net

[23] McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky A, Garimella K, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res., 20,* 1297-1303, 2010. Epub July 19, 2010

[24] Giardine, B., Riemer, C., Hardison,, R. C., Burhans, R., Elnitski, L., Shah, P., Zhang, Y, Blankenberg, D., et al. Galaxy: A platform for interactive lare-scale genome analysis. *Genome Research*, *15*(10):1451-1455, 2005.