L3 Bioinformatics Limited
Life, Logic and Legend

WHITEPAPER

Version: Draft v2, 24 January 2015

Security Overview of the BGI-Online Platform

Cloud computing is a powerful tool to deal with the exponential growth of storage and processing requirements of genomic data. However, cloud computing also raises concern around keeping the data safe. This white paper presents an overview of how BGI-Online approaches security and privacy for genomic data in the AWS cloud in a way that is not only compliant with existing standards, but also meets the specific security requirements of genomic data.

Introduction

Security and privacy are essential when dealing with genomic data and its analysis. Patients are rightfully concerned about the privacy of their genetic information. For researchers in academic and commercial environments, the safekeeping of valuable intellectual property derived from genetic information becomes an additional goal.

The general concerns around genomic data security naturally become amplified when this data is kept and processed in a cloud-computing environment, i.e., an environment that is by design shared between different parties and entities. And even though we would argue that, objectively, "local" server structures are no safer than a well-managed cloud (quite to the contrary in some cases we've observed), recent network security incidents have only contributed to instilling a general sense of distrust toward cloud environments.

To gain the trust of customers to store and process genomic data with BGI-Online on the AWS cloud, we have designed and implemented a comprehensive security framework for keeping this data safe. While our framework certainly needs to ensure compliance with current data protection standards (for example, US HIPAA and the European Data Protection Directive), it also offers concrete implementation suggestions in areas where existing standards do not yet provide certainty with respect to specific security requirements for genomic data and the global nature of cloud environments.

Current Regulatory Environment

On a global scale, the regulatory space for data protection is vast and cannot be sufficiently explored within the scope of this whitepaper. However, several regulatory frameworks are commonly named in discussions around genomic data security and privacy:

1) The US Health Insurance Portability and Accountability Act (HIPAA), which aims to protect all "Protected Health Information" (PHI) and consists of four parts:
    a. Privacy Rule, which protects the privacy of medical records and other personal health information;
    b. Security Rule, which sets national standards for the security of electronic protected health information;
    c. Breach Notification Rule, which requires covered entities and business associates to provide notification following a breach of unsecured protected health information;

   d. Patient Safety Rule, which protect identifiable information being used to analyze patient safety events and improve patient safety.

2) The Clinical Laboratory Improvement Amendments (CLIA), a set of US federal regulatory standards that apply to all clinical laboratory testing performed on humans in the United States, except clinical trials and basic research.
3) The Data Protection Directive, a European Union directive that regulates the processing of personal data within the European Union. (To be superseded by the European General Data Protection Regulation.)

All of these frameworks provide comprehensive guidance for IT and health data security in general even though they do not (yet) provide specific guidelines as to dealing with genomic data or some specific challenges of cloud environments. The current regulatory frameworks and compliance requirements therefore seem to be solid foundation, but not entirely sufficient to deal with security and privacy issues around keeping genomic data safe in a cloud environment.

The BGI-Online Security Framework

We believe that it's our job to provide users with end-to-end security and control over their data and analysis so that they can focus on work rather than having to deal with complex setups, compliance headaches, and security. To achieve this, we have designed a comprehensive security framework for processing genomic data in the cloud that covers three main areas:

1) Data Security: Ensuring that all sensitive data is kept safe during its full lifecycle. This includes data encryption and secure user authentication.
2) Platform and Infrastructure Security: Ensuring that the software platform and its underlying infrastructure (server and network) support the secure architecture.
3) Security Controls: Ensuring security of the system by implementing administrative, technical, and other security controls, while at the same time ensuring compatibility with a broad range of trusted information security frameworks and compliance requirements.

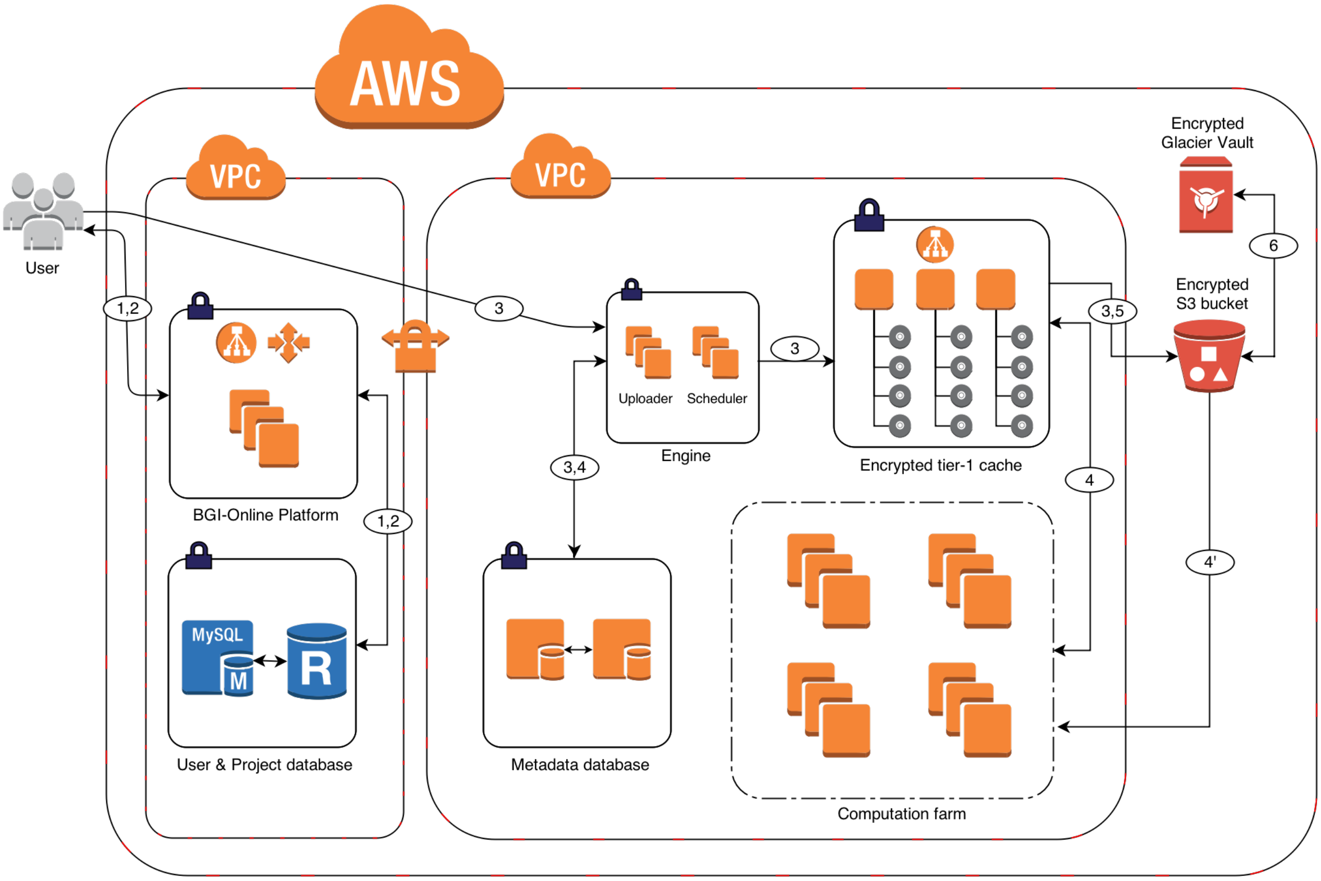
The following three sections present BGI-Online's view on each of these three areas and provide concrete implementation examples from the BGI-Online cloud platform. Since we provide BGI-Online using Amazon Web Services (AWS), we will be using AWS terminology for the remainder of this paper, such as "S3" to denote storage buckets and "EC2" to denote computation instances. Please refer to http://aws.amazon.com/ for details.

1. Data Security

The key security consideration for data of any kind is whether it is safely contained and isolated. Nobody except for an authorized user should be able to access or copy it. This not only means that data needs to be encrypted at all times, but also that tight controls need to be implemented around authentication and encryption key storage. In addition, certain laws and regulations specifically require control over data locality (where data can be stored and transferred).

A phrase that is often heard in regard to encrypting genomic data is that it must be "encrypted during transfer and at rest." While encryption during transfer is fairly straightforward from implementation standpoint (the scheme used by BGI-Online on AWS is outlined below), encryption "at rest" is not uniquely defined. For BGI-Online, there are actually four different types of "at rest" storage:

1) The ephemeral storage used by the EC2 computation instances
2) The Tier-1 cache comprises EC2 instances with multiple ephemeral disks.
3) AWS Simple Storage Service (S3).
4) AWS Glacier.



① User logs on BGI-Online.
② BGI-Online creates temporary access token.
③ Using the token, data is uploaded to Engine and being de-identified. Keys to restore the data are stored in Metadata database. De-identified data are stored in Encrypted tier-1 cache and S3 bucket synchronously.
④ Once the user starts a computation, BGI-Online calculates the optimal execution plan. Final results are uploaded to Encrypted tier-1 cache.
⑤ Infrequently accessed data are removed from Encrypted tier-1 cache,
⑥ or being further archived in Encrypted Glacier Vault and removed from S3.

Figure 1 - Dataflow in BGI-Online

The BGI-Online implements all three types of "at rest" encryption. Data is by default uploaded to Tier-1 cache encrypted using an industrial standard AES256 algorithm, and at the same time synchronized to encrypted S3 bucket leveraging S3 server-side encryption. During computation, all data and temporary disk volumes are being encrypted using AES256. Infrequently accessed data would be removed from Tier-1 cache, or further moved to Glacier, which is also encrypted with AES256 on server-side, for archival. Regarding data transfers, all user data is transferred exclusively through encrypted SSL/TLS channels throughout for all data flows shown in Figure 1.

At the end of the data lifecycle, data are wiped with U.S. Department of Energy M205.1-2 Standard to ensure that all data is safely deleted if it is no longer needed on the ephemeral storage or when an authorized user chooses to delete data on the platform. The standard uses three wiping passes:

- Pass 1-2: overwrite the data with a pseudo random values
- Pass 3: overwrite the data with zero-filled pattern

While encryption and the safe wiping of data are key to protect data in the cloud, they are only effective with appropriate authorization and access controls. It's a common argument that proper user authentication is much more important than encryption since, from a risk perspective, an access breach is by comparison much more likely than a physical security breach.

Access controls on the BGI-Online Platform have been implemented in a very fine-grained manner. Rather than establishing principal “file owners,” 6 permissions types including “Admin”, “Upload”, “View”, “Modify”, “Run” and “Share” are set on a per-user-per-project basis, meaning that a user’s access permissions to a given file can depend on the context (project) in which this file is being used. This includes sharing of data, which can only be performed via the platform itself unless a user has the “Share” permission to download a file.

A file could be shared through “link only”, thus prohibiting additional copies. The accessibility of a “link only” shared file could be revoked immediately by unsharing or deleting the file. To account for the two natures of shared file, one is publicly shared such as 1000 genome project data; another is privately shared where the number of recipient should only be one, BGI-Online implements two sharing methods:

1) Public: shared files could be viewed, linked or copied (if allowed) by all projects.
2) Private (hand-shaking): Sharer shares a file to a “Project ID (Recipient)” provided by the recipient. The recipient needs to enter the “Project ID (Sharer)” that owns the shared file to link or copy the shared file.

BGI-Online keeps secure audit logs for all data access for six years to ensure regulatory compliance.
By default, users authenticate on the platform through a username and secure password, in the future BGI-Online will support two-factor authentication.

2. Platform & Infrastructure Security

One clear advantage of running a platform in the cloud is that the cloud provider will usually offer a broad spectrum of built-in compliance and security features for the underlying infrastructure. For example, Amazon Web Services provides a broad spectrum of security features (http://aws.amazon.com/security/) and standards compliance (http://aws.amazon.com/compliance/), ranging from physical datacenter security and network infrastructure security to secure media handling and data encryption.

Naturally, compliance of the cloud provider does not imply compliance of the overall platform and ecosystem, but it is a solid foundation on which to build. Some regulations such as HIPAA even require the compliance of all individual providers within an environment (which, in the case of HIPAA, are linked to each other through so-called Business Associate Agreements).

In addition, BGI-Online secures its infrastructure in two ways:

1) All AWS computation instances run within Virtual Private Clouds (VPC). VPCs are logically isolated networks within the AWS cloud and kept only minimally open for the necessary external and internal access.
2) There is no multi-tenancy of physical resources: all computations are performed on dedicated instances and there is never more than one virtual instance running on the same hypervisor.

Access to the production and development environments in AWS are secured through Virtual Private Networks and IAM.

A second pillar to ensure that the BGI-Online Platform is always secure is to constantly monitor and improve our security by following best-practices of infrastructure stability and security, including

1) Regular software and infrastructure vulnerability assessments to discover vulnerabilities and remediate them;
2) Regular penetration tests in collaboration with third parties to discover vulnerabilities in the system, which may not be noticed in a regular vulnerability assessment;
3) Regular audit log analysis and system-level inspection to look for suspicious behavior, potential attacks, and security breaches;

4) A strict patch-management policy and regular server updates (depending on criticality, the response/fix time is between a few hours and 7 working days), and restriction of access for technical staff to resources on a per-need basis.

3. Security Controls

As in any IT security framework, security of the system must at the same time be ensured by implementing administrative, technical, and other security controls. The BGI-Online security framework aims to establish these in such a way that compatibility with a broad range of trusted information security frameworks and compliance requirements (such as HIPAA and ISO27001) is ensured at the same time.

A general aim should be to make all software and infrastructure fully compliant with the NIST 800-53A moderate profile, which BGI-Online considers to be the most complete set of controls which can be easily mapped to the majority of accepted information security and compliance frameworks. These controls cover all areas of information security, namely: Access control, Security awareness and training, Auditing and accountability, Security authorizations, Configuration management, Contingency planning, Authentication, Incident response, Dealing with equipment maintenance, Secure media handling, Physical and environmental security, Risk management and security planning, Personnel security, Systems and network security, Dealing with supply chain security and System and information integrity.

Proper implementation of these controls requires a broad range of policies that need to be effectively implemented as shown in Table 1.

Purpose	Policy/Control	Main content
Ensure proper controls	Risk Assessment and Management	How to treat risk and what controls to implement in order to support the required level of information security and minimize risk to business operations and customer data.
Keep data safe	Information Security	Detailing the information security stance, regarding infrastructure and customer data.
	Encryption	How to use encryption to secure data, both in transit and at rest. How to properly manage secret encryption keys.
	Workplace Security	How to secure the workstation, clean desk policy, expected employee behavior.
Ensure compliance	Security Awareness and Training	How to train staff and make them aware of security issues and procedures.
	Sanctions	How to deal with employees and subcontractors who violate the policies and procedures.
Maintain ongoing platform security	Change Management	How to manage changes in software and infrastructure to minimize information security and operations risk.
	Asset Management	How to manage information assets, maintain a good inventory and information system border.
	Vulnerability Management	How to manage vulnerabilities in software and infrastructure, namely discover and remediate them without disrupting business operations.
	Logs and Auditing	What to audit, how and why; how to analyze logs and provide reporting.
Keep infrastructure safe	Network Security	How to secure the network, how to connect remote systems and users, which protocols to use and how in order to maintain proper level of security.
	Antimalware	How to defend from malware.
	Acceptable Use	Details employee and subcontractor obligations.
	Patch Management	How to manage patches to software and underlying infrastructure.
Plan ahead	Security Incident Response	How to treat security incidents and potential breaches; how to notify customers of potential breaches.
	Disaster Recovery (including Recovery Plan)	How to manage backups and how to recover infrastructure and customer data in a case disaster strikes.

Table 1 - Policies required for implementing comprehensive security controls

Conclusion and Outlook

The explosion of genomic data and rise of cloud computing are recent developments. Currently BGI-Online has implemented all infrastructure, policies and controls to ensure that customers can process Protected Health Information (PHI) on the BGI-Online in full compliance with HIPAA.

Over the coming months and years, we expect regulatory standards on a national and international level to change and evolve to reflect the specific security and privacy challenges that are now emerging. Likewise, we expect our security framework to iterate and constantly evolve as well.

External References

1. Amazon Web Services Compliance Center
   http://aws.amazon.com/compliance/
2. Amazon Web Services Risk and Compliance Whitepaper
   http://media.amazonwebservices.com/AWS_Ri sk_and_Compliance_Whitepaper.pdf
3. Amazon Web Services Security Center
   https://aws.amazon.com/security/
4. New Guidance on De-identification Methods under the HIPAA Privacy Rule
   http://www.tricare.mil/tma/privacy/downloads/New%20Guidance%20on%20De-Identification_13March13.pdf
5. Creating HIPAA-Compliant Medical Data Applications With AWS
   http://media.amazonwebservices.com/AWS_HI PAA_Whitepaper_Final.pdf
6. NIST SP 800-30 Rev 1 "Guide for Conducting Risk Assessments"
   http://csrc.nist.gov/publications/nistpubs/800- 30-rev1/sp800_30_r1.pdf
7. NIST SP 800-53 Rev 4 "Security and Privacy Controls for Federal Information Systems and Organizations"
   http://nvlpubs.nist.gov/nistpubs/SpecialPublica tions/NIST.SP.800-53r4.pdf
8. NIST SP 800-66 rev1 "An Introductory Resource Guide for Implementing the Health Insurance Portability and Accountability Act (HIPAA) Security Rule"
   http://csrc.nist.gov/publications/nistpubs/800- 66-Rev1/SP-800-66-Revision1.pdf