Complete genomics

PRODUCT NOTE

Sample and Data Security

- Highest commitment to confidentiality and security of sample information and data
- Robust processes to protect customer samples and data
- Security throughout the process - from sample receipt, sequencing and assembly to delivery of data

From DNA to Data – Security of Sample Identity and Data

Introduction

This document describes Complete Genomics, Inc. (CGI) best practices on security for handling sample information and data. It includes information and details on how CGI identifies and tracks samples and resulting data.

Security by Design

CGI employs well-established information technology (IT) best practices to ensure confidentiality, security and reliability in all processing of customer samples and derived data (Figure 1). Customer data is stored in robust and secure data repositories, is protected with role-based information security and is segregated from non-production IT infrastructure. IT infrastructure is deployed in a secure and reliable data center, providing uninterruptable power and reliable network connectivity.

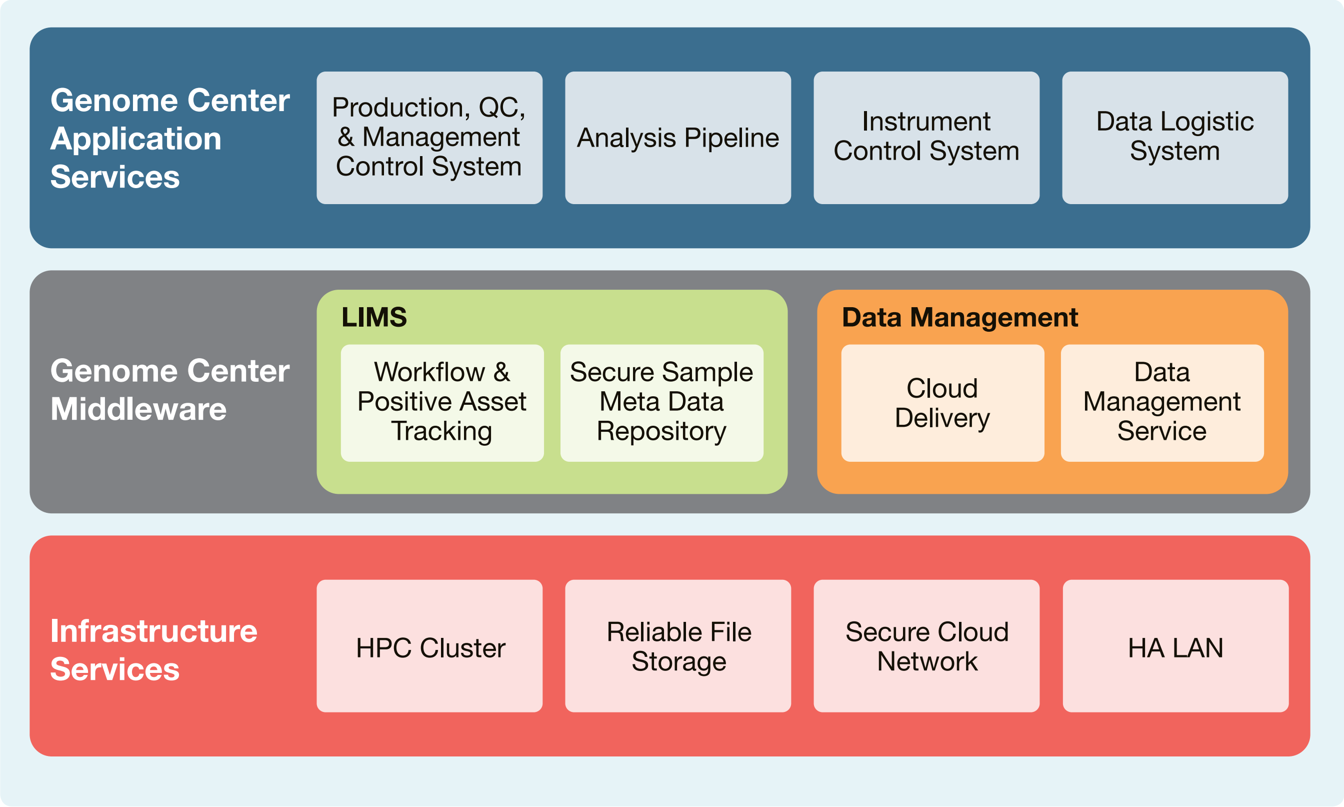

***Figure 1.** CGI's Infrastructure*

Sample Anonymization

All sample information is anonymized with customer-identifiable information accessible only to designated CGI Project Managers and is stored within a secure file system that isolates customer data from the corporate computing environment. Sample data tracking and management is performed by a password-protected Laboratory Information Management System (LIMS), which implements tracking, process controls and workflow automation behind a CGI firewall.

The primary identifier used to track a sample through the sequencing process at CGI is the pre-assigned, unique Complete Genomics Sample ID (e.g., GS00015-DNA_C08). This prevents any identifiable sample information from being obtained by an unauthorized person or system.

In addition, as part of the QC process, Complete Genomics performs a sample identification QC by genotyping 96 markers. The identity of the sample is then confirmed prior to delivering data by comparing these genotypes to the sequencing calls.

Data Processing

Figure 2 illustrates the process from receiving a customer's DNA to delivering data. Data analysis is performed using CGI's secure computing infrastructure, and includes generating reads, mappings, variant files and reports. Once a genome sequence has completed analysis and passes QC, the completed customer data set is automatically transmitted on a secure network connection to CGI's offsite storage location at Amazon Web Services (AWS)[1], a cloud computing solution[2].

Storage on Amazon Web Services

AWS is an infrastructure web services platform that provides remote compute power, storage, and other services. CGI is using Amazon's storage service, called Amazon Simple Storage Service (Amazon S3), for storing genome data sets. Amazon S3's highly scalable and reliable data storage infrastructure allows for secure storage and retrieval of large amounts of data[3]. There are several examples of biological data in AWS:

- ***Ensembl Annotated Human Genome Data*** *- Genome databases for human as well as almost 50 other species.*
- ***Ensembl - FASTA Database Files*** *- Ensembl sequence databases of transcript and translation models*
- ***YRI Trio Dataset*** *- Complete genome sequence data for three Yoruba individuals from Ibadan, Nigeria*
- ***GenBank*** *- An annotated collection of all publicly available DNA sequences including more than 85.7B bases and 82.8M sequence records*

Access to data stored in Amazon S3 is controlled by authentication using an HMAC-SHA1 signature employing a private key. CGI maintains full control over who has access to all data in Amazon S3. All data transfers are performed in a secure manner.

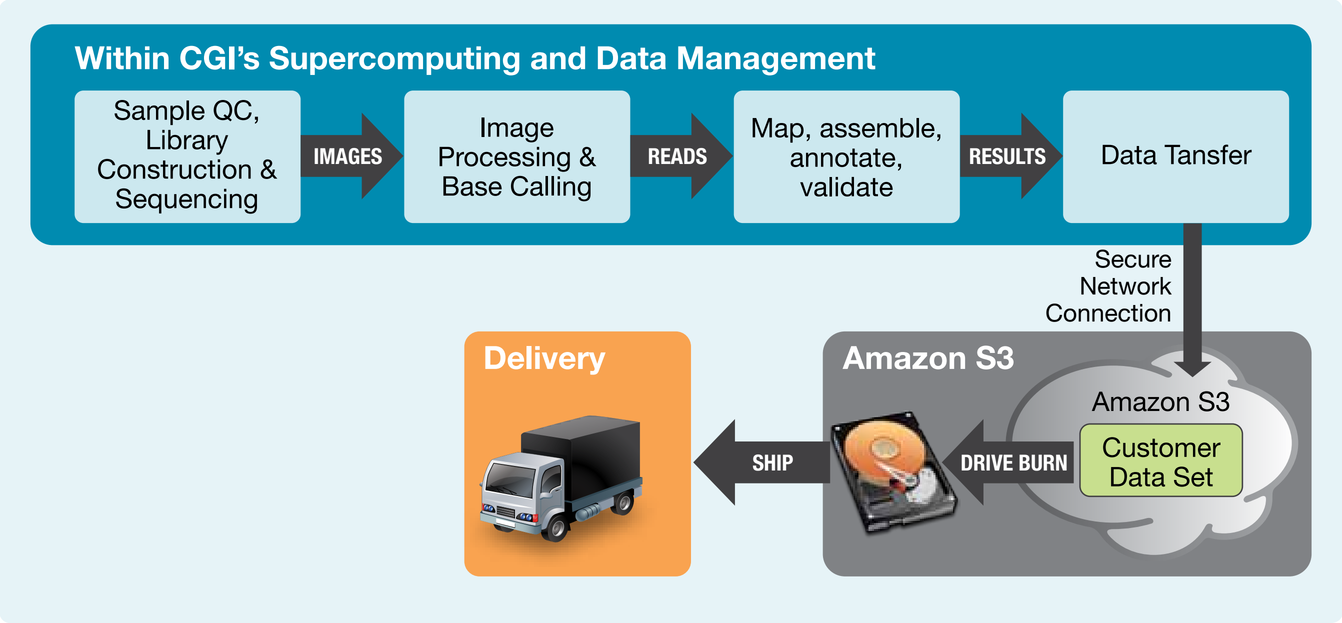
Within CGI's Supercomputing and Data Management
Sample QC, Library Construction & Sequencing
IMAGES
Image Processing & Base Calling
READS
Map, assemble, annotate, validate
RESULTS
Data Tansfer
Secure Network Connection
Delivery
Amazon S3
SHIP
DRIVE BURN
Amazon S3
Customer Data Set

Figure 2. *Process from DNA to Data Delivery*

Secure Delivery and Deletion

After transferring the data set to Amazon S3, CGI will notify AWS to transfer the data set to a hard disk drive and ship the disk drive to the customer. Currently, all completed genomes are delivered to customers on hard disk drives. The customer will receive an email with the tracking number for the package(s). Each hard disk drive will be delivered in a plain box with CGI sticker and a return address label from AWS. The CGI Project Manager will provide an email to the customer that outlines which Amazon Job ID corresponds to which Complete Genomics Sample ID. The data set will also reference the Complete Genomics Sample ID. Inside, the package contains the hard disk drive, power cords, USB 2.0 interface cables, and the disk drive installation guide. Country-specific power adaptors will not be provided by CGI, and so non-US customers need to ensure that they have an adapter for converting their country-specific power for use with US-compatible devices. The data set will be deleted from Amazon S3 30 days after shipment has been sent unless other arrangements have been made with CGI.

References

1) Amazon Web Services - http://aws.amazon.com/what-is-aws/
2) Nature Biotechnology 28, 13 - 15 (2010), doi:10.1038/nbt0110-13
3) AWS: Overview of Security Processes - http://developer.amazonwebservices.com/connect/entry.jspa?externalID=1697&categoryID=175

www.completegenomics.com info@completegenomics.com
2071 Stierlin Court, Mountain View, CA 94043 USA Tel 650.943.2800

Complete Genomics data is for Research Use Only and not for use in the treatment or diagnosis of any human subject.
support@completegenomics.com Toll-free: 1-855-CMPLETE (1-855-267-5383) or 1-650-943-2600
Information, descriptions and specifications in this publication are subject to change without notice.

Published in U.S.A., 01 April 2011