Security, Compliance and Sharing of Genomic Data on the Cloud

Angel Pizarro
Scientific and Research Computing
Amazon Web Services
angel@amazon.com
The Cancer Genome Atlas

Institute for Systems Biology, Seattle
Oregon H&S U
UCSC, Buck Institute

The Cancer Genomics Hub

MSK Cancer Ctr
The Broad Institute
Harvard

20 Adult cancers
10,000 tumor pairs
Cancer Genomics Hub

- 50K Genomes at ~ $100/year/genome
- Houses genomes for all major National Cancer Institute projects (TCGA, TARGET, etc)
- 1.5PB (growing to 2.5PB in the next year)
- Serving over 1PB of data a month
CGHub Growth: Storage vs. Transfer

- >1PB / month
- Requester must have at least that much storage to analyze
- Requester must have a lot of compute
- Unsustainable
Data held in silos, unshared

No one institute has enough on its own to make progress
Global Alliance
for Genomics & Health

• **Founded on June 5, 2013**
  – Founding partners: 70+ leading healthcare, research, and disease advocacy organizations

• **Now:** more than 170 members from 40 countries

• **Mission:** to enable rapid progress in biomedicine

• **Plan:**
  – Create and maintain the interoperability of technology platform standards for managing and sharing genomic data in clinical samples
  – Develop guidelines and harmonizing procedures for privacy and ethics in the international regulatory context
  – Engage stakeholders across sectors to encourage the responsible and voluntary sharing of data and of methods
GA4GH API promotes sharing
Advantages of Storing and Sharing in a Federated, Cloud-based Commons

- No single owner:
  - Each Institute keeps data on either private or commercial cloud
  - No one entity controls the world’s genome data

- Globalization:
  - Internet data exchange and containerized computation provide global transaction capability and ubiquitous availability

- Reduced cost:
  - Hardware and storage disk costs reduced by bulk purchases
  - Operational costs reduced by building datacenters in optimal locations and deploying automated systems

- Security:
  - Many clouds are already proven for use with sensitive data

- Elasticity:
  - In a large cloud, the cost of 1 computer for 1,000 hours is the same as the cost of 1000 computers for 1 hour (about $1.50/machine/hour). No machines sit idle.
Under the Hood at GA4GH

• We work together in an open source software development environment on the web: https://github.com/ga4gh

• All groups are welcome to participate

• Decision making is done by protocols developed by Apache Open Source Software Foundation

• Leadership is determined by amount of contribution

• Simple Mantra: collaborate on interface, compete on implementation
Interoperability: One API, Many Apps
Do you have any genomes with an “A” at position 100,735 on chromosome 3? 
Yes

No

I can neither confirm nor deny that request.
rs6152
Located in the first exon of the androgen receptor AR gene located on the X chromosome, is *highly indicative* of the ability to develop *male pattern baldness*
Digging Deeper into the Data

Repository Discovery (Public)

- **Does the data exist?**
- **Anonymous user**
- Query: *Do you have any genomes with an “A” at position 100,735 on chromosome 3?*
- Reply: ‘Yes’ or ‘No’

Data Context Queries (Registered)

- **Does the data have the properties I require?**
- **Registered user**
- Shows details of studies with data of interest to user and provides link for requesting full access

Full Data Access (Controlled)

- **Give me the full dataset**
- **Approved user (signed contract)**
- Permits full access to genotype, phenotype and raw DNA sequence
| name: which object do I want? |
| protocol: how do I get it? |
| content: what does it mean? |
## GA4GH Big Data Commons Model

The Web is based on three interlocking standards.

| name: which object do I want? | URLs are universal identifiers for web objects. |
| protocol: how do I get it? | HTTP is the standard API that any web browser can use to request information from any web server. |
| content: what does it mean? | HTML is the single most important content type on the web, connecting data like text, images, videos, and PDFs. |
## GA4GH Big Data Commons Model

The Global Alliance APIs define equivalents

<table>
<thead>
<tr>
<th>name: which object do I want?</th>
<th>Web</th>
<th>Global Alliance APIs</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>URLs</strong> are universal identifiers for web objects.</td>
<td>The GA4GH APIs define <strong>content digests</strong> that uniquely identify genomic and other objects.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>protocol: how do I get it?</th>
<th>Web</th>
<th>Global Alliance APIs</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>HTTP</strong> is the standard API that any web browser can use to request information from any web server.</td>
<td>The GA4GH APIs define <strong>methods</strong> for requesting data. For example, a genome browser can use the <strong>searchReads</strong> method to request genomic reads, and get back a <strong>SearchReadsResponse</strong>, whether it's talking to a genome server at NCBI, EBI, or Google.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>content: what does it mean?</th>
<th>Web</th>
<th>Global Alliance APIs</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>HTML</strong> is the single most important content type on the web, connecting data like text, images, videos, and PDFs.</td>
<td>The GA4GH APIs define logical <strong>schemas</strong> for every object, such as <strong>ReadGroup</strong> and <strong>Variant</strong>.</td>
<td></td>
</tr>
</tbody>
</table>
Data Content Digests as Identifiers

- Cryptographic 1-way hash function used
- Data dependent, not format dependent
- Unique to a data set
- Unforgeable and verifiable
- Privacy preserving
- Decentralized

A file and database with same data encoded within have the same digest.
Methodologies for Data Content Digests

- Hashing allows
  - decentralized generation, privacy-preserving digests, unforgeable, unique
- A digest is computed on data itself, not storage format
  - serves as a distributed Rosetta Stone between repositories, formats, and accessions
  - unique to each specific version of data, used to establish provenance
- At a base level digests are computed on coarse-grained, immutable data objects (if data are changed, a new digest is created)
- Digests are composeable to create hierarchies for larger and larger sets of objects without needing to create additional copies of base level objects
Example Digest: SAM/BAM

<table>
<thead>
<tr>
<th>Col</th>
<th>Field</th>
<th>Type</th>
<th>Brief description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>QNAME</td>
<td>String</td>
<td>Query template NAME</td>
</tr>
<tr>
<td>2</td>
<td>FLAG</td>
<td>Int</td>
<td>bitwise FLAG</td>
</tr>
<tr>
<td>3</td>
<td>RNAME</td>
<td>String</td>
<td>Reference sequence NAME</td>
</tr>
<tr>
<td>4</td>
<td>POS</td>
<td>Int</td>
<td>1-based leftmost mapping POSition</td>
</tr>
<tr>
<td>5</td>
<td>MAPQ</td>
<td>Int</td>
<td>MAPping Quality</td>
</tr>
<tr>
<td>6</td>
<td>CIGAR</td>
<td>String</td>
<td>CIGAR string</td>
</tr>
<tr>
<td>7</td>
<td>RNEXT</td>
<td>String</td>
<td>Ref. name of the mate/next read</td>
</tr>
<tr>
<td>8</td>
<td>PNEXT</td>
<td>Int</td>
<td>Position of the mate/next read</td>
</tr>
<tr>
<td>9</td>
<td>TLEN</td>
<td>Int</td>
<td>observed Template LENgth</td>
</tr>
<tr>
<td>10</td>
<td>SEQ</td>
<td>String</td>
<td>segment SEQUENCE</td>
</tr>
<tr>
<td>11</td>
<td>QUAL</td>
<td>String</td>
<td>ASCII of Phred-scaled base QUALity+33</td>
</tr>
</tbody>
</table>

At a dataset level, an order independent digest of each read & alignment would identify the dataset.
Putting it All Together

Compute Environment
- The compute environments support running arbitrary analysis or processing pipelines.
- The only requirement for pipeline developers is that they use the GA4GH API to access data, so code running in any compute environment can work with data hosted in any data environment.

GA4GH Data Environment
- Data repositories are responsible for:
  - Storing data, using whatever internal representation they deem appropriate
  - Managing globally unique data content digests, so that anyone anywhere can unambiguously refer to a particular data object (e.g. "all the reads from a given sample")
  - Keeping track of authorization: single access control list for who is allowed to see what data
- All data is exposed to the outside world via the GA4GH API.
  - Users of the data don't care about the internals; they just have to know how to call the API.
Security and Compliance

- Shared responsibility
- Emphasis on encryption
- Authentication and authorization at the data object level

<table>
<thead>
<tr>
<th>Service Consumers</th>
<th>Data, Application, and Infrastructure Service Providers*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Access control</td>
<td>Integrity protection</td>
</tr>
<tr>
<td>Cryptographic controls</td>
<td>Non-repudiation</td>
</tr>
<tr>
<td></td>
<td>Access control</td>
</tr>
<tr>
<td></td>
<td>Audit recording and review</td>
</tr>
<tr>
<td></td>
<td>Cryptographic controls</td>
</tr>
<tr>
<td></td>
<td>Business continuity</td>
</tr>
</tbody>
</table>

*For infrastructure service providers, the applicability of each security measure is dependent upon the specific service(s) provided.

<table>
<thead>
<tr>
<th>Data Stewards</th>
</tr>
</thead>
<tbody>
<tr>
<td>Consent management</td>
</tr>
<tr>
<td>Metadata management</td>
</tr>
<tr>
<td>Integrity protection</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Global Alliance Leadership</th>
</tr>
</thead>
<tbody>
<tr>
<td>Global risk assessment</td>
</tr>
<tr>
<td>Security and privacy policy</td>
</tr>
<tr>
<td>Security infrastructure recommendations, guidance, and standards</td>
</tr>
<tr>
<td>Security coordination</td>
</tr>
</tbody>
</table>
GA4GH Working Groups

- Genome Data Working Group (co-led by David Haussler and Richard Durbin)
- Clinical Data Working Group (co-led by Charles Sawyers and Kathryn North)
- Security Working Group (co-led by Paul Flicek and Dixie Baker)
- Regulatory and Ethics Working Group (co-led by Bartha Knoppers, Kazuto Kato and Partha Majumder)
GA4GH Driving Projects

Currently active:

1. **Beacon Project**: (Steve Sherry, Marc Fiume) public querying of simple top level information about genetic variants

1. **Genomic Matchmaker**: (Heidi Rehm, Anthony Philippakis) expert sharing of rare genetic variants

1. **BRCA Challenge**: (Gunnar Rätsch) aggregation of research and clinical data to assess pathogenicity and penetrance of all variants in BRCA1 and 2
Ultimate Goal

Assess clinical outcomes

Select patient treatment

Compare treatment effectiveness

Implement new evidence for treatment prioritization

Learning System
Global Footprint

Everyday, AWS adds enough new server capacity to support Amazon.com when it was a $7 billion global enterprise.

Over 1 million active customers across 190 countries

800+ government agencies

3,000+ educational institutions

11 regions

28 availability zones

52 edge locations
AWS Regions and Availability Zones

Customer Decides Where Applications and Data Reside

Note: Conceptual drawing only. The number of Availability Zones may vary.
Full stack sequence analysis platform

Storage → Compute → Databases → Analytics

Sequence data → Upstream analysis → Variants Expression Phenotypes → Data mining
1+ Million Cancer Genome Data Warehouse
Security is a Shared Responsibility
Shared responsibility model

- Customers implement their own set of controls
- Multiple customers with FISMA Low and Moderate ATOs

Amazon

Foundation Services

- Compute
- Storage
- Database
- Networking

AWS Global Infrastructure

- Availability Zones
- Regions
- Edge Locations

Customer

- Platform, Applications, Identity & Access Management
- Operating System, Network & Firewall Configuration

- Customer Data

- Client-side Data Encryption & Data Integrity Authentication
- Server-side Encryption (File System and/or Data)
- Network Traffic Protection (Encryption/Integrity/Identity)

SOC 1/SSAE 16/ISAE 3402
SOC 2
ISO 27001/ 2 Certification
Payment Card Industry (PCI)
Data Security Standard (DSS)
NIST Compliant Controls
DoD Compliant Controls
FedRAMP
HIPAA and ITAR Compliant
Shared responsibility model

- Re-focus your security professionals on a subset of the problem
- Take advantage of high levels of uniformity and automation

First global public cloud provider to achieve certification for security & quality management system
NIH dbGaP security best practices

- Physical security
  - Data center access and remote administrator access

- Electronic security
  - User account security (for example, passwords)
  - Use of Access Control Lists (ACLs)
  - Secure networking
  - Encryption of data in transit and at rest
  - OS and software patching

- Data access security
  - Authorization of access to data
  - Tracking copies; cleaning up after use
Amazon Virtual Private Cloud (VPC)

Create **secure network** configurations for working with sensitive data
AWS Identity and Access Management (IAM)

- Create and manage users in the AWS services
- **Identity federation** with Active Directory
- Control passwords, access keys, and **multi-factor authentication** (MFA) devices
  - Hardware or software (OAuth)
- **Fine-grained permissions**
- Very familiar security model
  - Users, groups, permissions
- Integrated into the following:
  - AWS Management Console
  - AWS API access
  - IAM resource-based policies
  - Amazon S3, Amazon SQS, Amazon SNS

Roles:
- Administrators: Jim, Alyson
- Developers: Shandra, Xiao, Susan, Anand
- Applications: Reporting, Console, Tomcat

Multi-factor authentication

AWS system entitlements
Segregate duties between roles with IAM

You get to choose **who** can do **what** in your AWS environment and from **where**

Manage and operate
Data encryption in transit and at rest

Amazon S3
- HTTPS
- AES-256 server-side encryption
- AWS or customer managed keys
- Each object gets its own key

Amazon EBS
- End-to-end secure network traffic
- Whole volume encryption
- AWS or customer managed keys
- Encrypted incremental snapshots
- Minimal performance overhead (utilizes Intel AES-NI)
Use Amazon CloudTrail to track access to APIs and IAM

- **Records** API calls, no matter how those API calls were made (console, SDK, CLI)
- Who did what and when and from what IP address
- Logs saved to Amazon S3
- Includes EC2, Amazon EBS, VPC, Amazon RDS, IAM, AWS STS, and Amazon RedShift
- Be **notified** of log file delivery by using the Amazon Simple Notification Service (SNS)
- **Aggregate log information** across services into a single S3 bucket
- **Out of the box integration** with log analysis tools from AWS partners including Splunk, AlertLogic, and SumoLogic
AWS Config is a fully managed service that provides you with an inventory of your AWS resources, lets you audit the resource configuration history and notifies you of resource configuration changes.
NIH dbGaP security best practices

• Physical security
  – Data center access and remote administrator access

• Electronic security
  – User account security (for example, passwords)
  – Use of Access Control Lists (ACLs)
  – Secure networking
  – Encryption of data in transit and at rest
  – OS and software patching

• Data access security
  – Authorization of access to data
  – Tracking copies; cleaning up after use
HIPAA: Protected Health Information on the Cloud
HIPAA Security and Compliance on AWS

• We sign Business Associates Agreements
  – As do other cloud providers
• All data (including PHI) is under the complete control of customer
• Encrypt data
• Lock down resources and applications
• Monitor anything and everything
Thank you!

Architecting for Genomic Data Security and Compliance in AWS


Creating Healthcare Data Applications to Promote HIPAA and HITECH Compliance