Google Cloud Whitepaper
June 2019

Handling genomic data in the cloud

Google Cloud

Table of contents

Disclaimer

This whitepaper applies to Google Cloud products described at cloud.google.com. The content contained herein is correct as of January 2019 and represents the status quo as of the time it was written. Google's security policies and systems may change going forward, as we continually improve protection for our customers.

Introduction

Significant advancements in genomics analysis have been made possible with the emergence of better and more cost-effective tools. For example, cloud-based technologies, such as those offered by Google Cloud Platform (GCP), provide computational resources capable of analyzing massive amounts of genomic information at unprecedented speeds and in many cases, at a lower cost compared with on-premises solutions. Today, the use of cloud-based tools enables analysis across thousands of genomes to identify patterns and markers for disease predisposition, prediction, and causality. This helps improve how healthcare providers understand and treat disease, and creates better-informed treatment plans for patients.

With cloud services offered by GCP, researchers can securely store, process, explore, and share large genomic datasets. For example, the Stanford Center for Genomics and Personalized Medicine is using Google Genomics and Google BigQuery to safely and securely analyze hundreds of whole genomes in hours — an effort that would have been far more costly and time intensive previously. The Broad Institute has also benefited from using GCP, replacing its in-house genome sequence analysis platform with GCP products such as Google Genomics, Google Compute Engine, and Google Cloud Storage, resulting in greater speed, scalability, and data security. To learn more about how cloud-based tools are helping to advance how researchers process genomic information, refer to Genomic Data Is Going Google.

One of the many advantages organizations can benefit from when handling genomic data in cloud platforms, like GCP, is the ability to leverage state-of-the-art security and privacy features to protect their sensitive data and to help them comply with applicable data privacy/protection regulations. Security and privacy are fundamental concerns for those handling genomic data regardless of where it resides — on a local hard drive, in an institutional data center, or in the cloud. This type of information is highly sensitive in nature and can be subject to various regulatory requirements such as the Health Insurance Portability and Accountability Act (HIPAA) in the United States and the General Data Protection Regulation (GDPR) in the European Union.

The Stanford Center for Genomics and Personalized Medicine is using Google Genomics and Google BigQuery to safely and securely analyze hundreds of whole genomes in hours — an effort that would have been far more costly and time intensive previously.

This whitepaper contains an overview of how GCP helps secure customer data and the products and features we provide to help customers manage and secure their data on GCP. We also provide an overview of regulations that govern the safe handling of genomic data in the United States, along with the GCP products and capabilities that customers can leverage to help meet their operational and applicable regulatory/compliance requirements.

Genomic data and the regulatory environment

Innovation in genomic research opens new possibilities for how human health is understood and how diseases are treated. This section discusses current trends in genomic research and provides an overview of noteworthy regulations that govern the protection of sensitive genomic data in a clinical or research environment.

Important considerations when handling genomic data

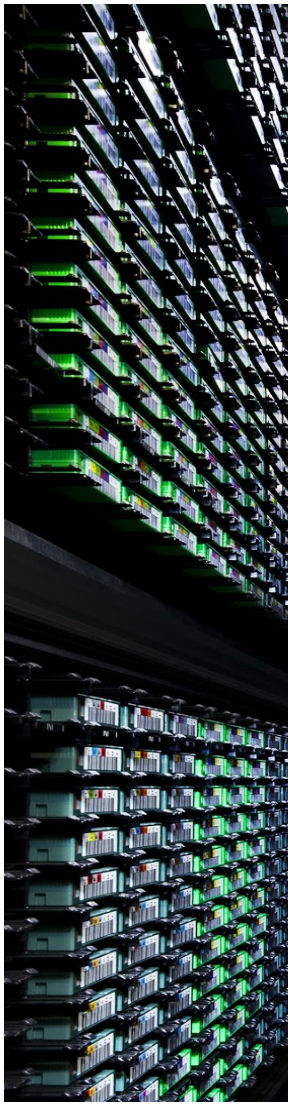
Genomic data can contain personal, and potentially identifying, details about an individual's physical traits and health. Genetic identifiers may signal predispositions to certain diseases or be markers to specific conditions. Organizations must follow security and privacy best practices for protecting genomic data to uphold individuals' privacy.

How are organizations using genomic data today?

Genomic research is leading to profound innovation in the healthcare industry. New insights are making improvements in the risk assessment, diagnosis, prognosis, and treatment of patients. Genomics is helping to personalize healthcare by enabling targeted prevention and treatment plans leading to improved clinical effectiveness and outcomes. Organizations today are leveraging genomic research to develop more advanced pharmaceuticals, provide more accurate diagnostics testing solutions, and transform traditional care delivery models.

How cloud is enabling genomic data usage

Advancements in DNA sequencing and the growing volume of genomic data available have transformed biology and medicine into data-rich fields. The advancements and volume growth have been catalyzed by the reduced costs for data storage offered by cloud service providers, such as GCP. Organizations are adopting cloud-based technologies to collect, process, and store vasts amounts of data in a financially sustainable manner. Cloud offerings like Google Cloud Storage provide large-scale, highly redundant storage solutions for researchers to easily access, share, and store genomic data. GCP offerings like Nearline & Coldline provide long-term storage



solutions at cost-effective prices for data archival. Cloud-based solutions are enabling new entrants and possibilities in genomic research.

In addition to storing genomic data in the cloud, organizations are taking advantage of highly scalable cloud-based data warehouse and analytic solutions. Genomic research requires massive processing power and scalability to analyze enormous datasets. Cloud-based solutions like Google Genomics and Google BigQuery help researchers to analyze the data from thousands of genomes in seconds, far less time than required using other computational infrastructure. As organizations scale their genomic research to larger datasets, cloud solutions can offer advanced distributed processing technologies and big data platforms. Whether organizations require tens or thousands of compute nodes, cloud solutions provide on-demand access so organizations pay only for what they need.

Regulations and guidelines governing the protection of genomic data

When using genomic information for research or clinical purposes, it can be subject to several regulations and guidelines around the world. In some cases, genomic data may be subject to regulations such as the EU's GDPR that mandates personal data protection, or the US HIPAA that establishes clear requirements around the protection and security of protected health information (PHI). In addition, guidelines such as the National Institutes of Health (NIH) Genomic Data Sharing Policy and NIH Security Best Practices for Controlled-Access Data Subject to the NIH Genomic Data Sharing Policy (commonly referred to as "dbGaP Security Best Practices") prescribe the application of specific protections to genomic data being shared with the NIH.

The EU General Data Protection Regulation (GDPR)

The GDPR lays out specific requirements for organizations established in Europe or who serve users in Europe. It regulates how organizations can collect, use, and store personal data. Under the GDPR, genomic data can qualify as personal data. Data controllers are required to only use data processors that provide sufficient guarantees to implement appropriate technical and organisational measures in such a manner that processing will meet the requirements of the GDPR. To learn more about GCP's commitment to meeting GDPR requirements, refer to our GDPR page.

Health Insurance Portability and Accountability Act (HIPAA)

The HIPAA Privacy Rule and Security Rule establish data privacy and security protections for PHI. GCP provides a business associate agreement for organizations subject to HIPAA that want to utilize any Google Cloud products in connection with PHI. For more information about how GCP helps customers meet their responsibilities under HIPAA, and a list of GCP products covered by the BAA, refer to our HIPAA Compliance page and HIPAA whitepaper.

NIH Genomic Data Sharing Policy

The NIH Genomic Data Sharing Policy was devised to foster the responsible sharing of genomic data between NIH entities and NIH-funded research organizations. Organizations are encouraged to submit research-generated human genomic data to the database of Genotypes and Phenotypes (dbGaP). The dbGaP is maintained by the NIH and was developed to store and share data and the results of studies that have investigated the interaction of genotype and phenotype in humans. The Genomic Data Sharing Policy also recommends that organizations implement appropriate “confidentiality, privacy, and data use measures” to protect the privacy of human subjects.[1] These guidelines apply whether the genomic data resides in GCP, a researcher’s computer, or an institution’s network.

dbGaP Security Best Practices

The NIH Security Best Practices for Controlled-Access Data Subject to the NIH Genomic Data Sharing Policy (or “dbGaP Security Best Practices”) provides guidance to organizations granted access to genomic data maintained by the NIH in the dbGaP. Organizations should demonstrate good security and privacy practices that ensure they secure genomic data and limit access to only approved researchers. The dbGaP Security Best Practices provide information for scientific and IT audiences. The guidance spans multiple security domains, including physical security, network security, and cloud security. The GCP and dbGaP Security Best Practices section below highlights GCP’s solutions that help customers meet the NIH guidelines.

We secure and standardize our global infrastructure and services, making it easier for our customers to consume compliant services.

In addition to supporting compliance with the GDPR, HIPAA, and dbGaP Security Best Practices, GCP maintains compliance with numerous other globally recognized data privacy/protection regulations and standards. Furthermore, we secure and standardize our global infrastructure and services, making it easier for our customers to consume compliant services. Refer to our Compliance page and Infrastructure Security Design Overview whitepaper for more information.

GCP security and privacy

This section covers the design and maintenance of GCP’s infrastructure and operations.

GCP’s approach to security

Security and information protection are among our primary design criteria for GCP services. Security is at the core of everything we do; it is embedded in our culture and our architecture, and we focus on improving it every day. Our relentless focus on security and information protection earned us recognition as a Leader for Public Cloud Native Security by Forrester in Q2 of 2018. The following is an overview of the organizational and technical controls we use to protect customer data.

[1] NIH GDS Policy. (2014, August 27). Retrieved from https://osp.od.nih.gov/wp-content/uploads/.NIH_GDS_Policy.pdf

Strong security culture

Security is central to Google culture. It is reinforced in employee security training and company-wide events to raise awareness and drive innovation in security and privacy. To learn more, refer to the Security Culture section of the Google Cloud Security whitepaper.

Security team

Google employs hundreds of security professionals, including some of the world's foremost experts in the domain. This team maintains the company's defense systems, develops security review processes, builds security infrastructure, implements our security policies, and actively scans for security threats.

Our team also takes part in research and outreach activities to protect the wider community of Internet users, beyond just those who choose Google solutions. Our research papers are available to the public. As part of our outreach efforts, our Project Zero team aims to prevent targeted attacks by reporting bugs to software vendors. To learn more, refer to the Security Team, Vulnerability management, and Monitoring sections of the Google Cloud Security whitepaper.

Trusted infrastructure

We conceived, designed, and built GCP to operate securely. Using "defense-in-depth" principles, we have created an IT infrastructure that is more secure and easier to manage than most other deployment options. Our infrastructure provides secure deployment of services, secure storage of data with end-user privacy safeguards, secure communications between services, secure and private communication with customers over the Internet, and safe operation by administrators. The security of the infrastructure is designed in progressive layers starting from the physical security of data centers, continuing on to the security of the hardware and software that underlies the infrastructure, and finally, the technical constraints and processes in place to support operational security. We are an innovator in hardware, software, network, and system management technologies. We custom design our servers, proprietary operating system, and geographically distributed data centers. We design and manufacture purpose-built servers and network hardware without unnecessary components. To learn more, refer to the Google Infrastructure Security Design Overview as well as the GCP Data Processing and Security Terms, Appendix 2: Security Measures.

Infrastructure redundancy

GCP infrastructure components are designed to be highly redundant. This redundancy applies to server design and deployment, data storage, network and Internet connectivity, and the software services themselves. This “redundancy of everything” creates a robust solution that is not dependent on a single server, data center, or network connection. Our data centers are geographically distributed to minimize the effects of regional disruptions on global products, such as natural disasters and other localized incidents. In the event of hardware, software, or network failure, platform services and control planes are capable of automatically changing configuration so that customers can continue to work without interruption. Our highly redundant infrastructure also helps customers protect themselves from data loss. Customers can create and deploy our cloud-based resources across multiple regions and zones, allowing them to build resilient and highly available systems. To learn more, refer to the Low Latency and Highly Available Solution in the Google Cloud Security whitepaper.

State-of-the-art data center security

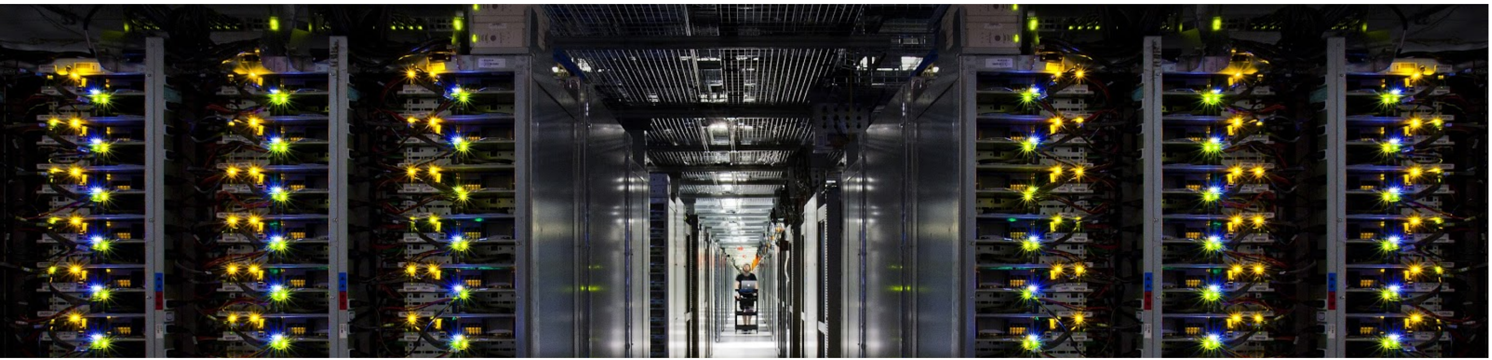
Google data centers feature multiple layers of physical security protections. We limit access to these data centers to only a small fraction of employees and have multiple physical security controls to protect our data center floors, such as biometric identification, metal detection, vehicle barriers, and custom-designed electronic access cards. We monitor our data centers 24/7/365 to detect and track intruders. Data centers are routinely patrolled by experienced security guards who have undergone rigorous background checks and training. To learn more, refer to our Data Centers page.

Cloud-native technology

We continue to invest heavily in security, both in the design of new features and the development of cutting-edge tools for customers to more securely manage their environments. Some examples are the Cloud Security Command Center that brings actionable insights to security teams and Virtual Private Cloud (VPC) Service Controls that help to establish virtual security perimeters for sensitive data.

Data encryption

Google Cloud has a team of world-class security engineers tasked with following, developing, and improving encryption technology, which is a central part of our security strategy. We offer strong encryption on the Google Cloud infrastructure to guard against unauthorized access for data in transit outside of our perimeter, and we are one of the only cloud service providers to offer encryption by default for data at rest. GCP provides a cloud-hosted encryption key management service. In addition, we enable customers to manage their own encryption keys for selected products.



GCP’s approach to privacy

Privacy is fundamental at Google Gloud. We go to great lengths to protect the data customers store on our cloud infrastructure. We incorporate strong security and privacy controls into the design and operation of our products and services. We protect the privacy of our customers’ information by providing them meaningful privacy control options and maintaining and continually evolving our data security features. Moreover, in complying with data protection laws, Google Cloud undergoes regular audits, maintains industry-accepted certifications, provides contractual protections, and shares tools and information to help customers strengthen their enterprises’ compliance abilities.

Google Cloud Trust Principles

Your data is critical to your business, and you take great care to keep it safe and under your control. We want you to feel confident that taking advantage of G Suite and Google Cloud Platform doesn't require you to compromise on security or control of your business's data.

At Google Cloud, we believe that trust is created through transparency, and we want to be transparent about our commitments and what you can expect when it comes to our shared responsibility for protecting and managing your data in the cloud.

Our commitments to you about your data

When you use Google Cloud, you can:

1. **Know that your security comes first in everything we do.**

   We promptly notify you if we detect a breach of security that compromises your data.
2. **Control what happens to your data.**

   We process customer data according to your instructions. You can access it or take it out at any time.
3. **Know that customer data is not used for advertising.**
   You own your data. Google Cloud does not process your data for advertising purposes.
4. **Know where Google stores your data and rely on it being available when you need it.**

   We publish the locations of our Google data centers; they are highly available, resilient, and secure.
5. **Depend on Google’s independently-verified security practices.**

   Our adherence to recognized international security and privacy standards is certified and validated by independent auditors — wherever your data is located in Google Cloud.
6. **Trust that we never give any government entity “backdoor” access to your data or to our servers storing your data.**

   We reject government requests that are invalid, and we publish a transparency report for government requests.

   See the data processing terms for G Suite and Google Cloud Platform for further details.

Dedicated privacy team

GCP's privacy team operates separately from the product development and security organizations, but participates in every product launch by reviewing design documentation and performing code reviews to ensure that privacy requirements are followed. They help release products that reflect strong privacy standards. This includes the transparent collection of user data and providing users and administrators with meaningful privacy configuration options, while continuing to be good stewards of any information stored on our platform. To learn more, refer to the Privacy Team section of the Google Cloud Security whitepaper.

Data access and customer control

Customers own their data stored on GCP, not Google. We process their data in accordance with contractual obligations. We also provide customers with solutions that allow granular control of resource permissions. For example, using Cloud Identity and Access Management, customers can map job functions to groups and roles so users access only the data they need to get the job done.

Restricted access to customer data

To keep data private and secure, GCP logically isolates each customer's data from that of other customers and users, even when the data is stored on the same physical server. For more information on how we logically isolate customer data and workloads on GCP, refer to the Service Identity, Integrity, and Isolation section of the Google Infrastructure Security Design Overview whitepaper. To learn more about how we provide visibility into Google's access to customer data, refer to Access Transparency.

The Shared Responsibility Model

In the pre-cloud IT model, organizations maintained full responsibility of their environment. They managed everything from the networking and infrastructure to the security controls and applications. In the cloud IT model, management of the IT environment, including responsibilities for security and compliance, is shared between the customer and its cloud service provider. This is often referred to as the Shared Responsibility Model.

GCP's part in the Shared Responsibility Model includes providing services on a highly secure and controlled platform and offering a wide array of security features customers can use. Shared responsibility enables our customers to allocate resources more effectively to their core competencies and concentrate on what they do best. Although the Shared Responsibility Model does not remove the accountability and risk from customers using GCP services, we help by operating and controlling system components and physical control of facilities. Moreover, using our cloud services is a more cost-effective approach for customers because we manage a substantial portion of the security and compliance efforts.

GCP and dbGaP security best practices

This section tracks the areas outlined in the NIH dbGaP security best practices. We describe how various GCP products and capabilities can help organizations implement these security best practices. The list of products and capabilities provided under each best practice is by no means exhaustive; it represents the common features that our customers can take advantage of to manage and protect their genomic data. The dbGaP security best practices do not explicitly require that organizations use each of the types of products and services listed below; customers can apply the best practices in various ways.

For most of the products and capabilities identified below, customers can implement them in various configurations to protect their genomic data. This document is intended only to provide context and suggestions, not prescriptive or comprehensive guidance.

Customer responsibility

You are responsible for complying with laws and regulations applicable to you and determining whether your use or intent to use GCP complies with applicable laws. We encourage you to ensure you understand your legal obligations before using GCP and, where appropriate, seek legal advice.

Terminology explained

In the products and capabilities descriptions in this section, we use:

- ***"by default"*** to mean a feature that is turned on without any manual configuration required by our customers
- ***"available by default"*** to mean a feature that is native to a product and can be manually activated by a customer for use

General information security guidelines

Limit Internet-facing exposure of data

Security Best Practice	Customer Responsibility
Genomic data should not be publicly accessible; measures should be implemented to enable secure access, transfer, and downloading of genomic data.	*Customers are responsible for managing the access to sensitive information they place on GCP.*

GCP Products and Capabilities

- **Cloud Identity & Access Management (Cloud IAM)**
  GCP provides Cloud IAM that is ***available by default*** and, when configured, provides customers the ability to centrally manage resource permissions and effectively control access to GCP resources.

- **VPC Service Controls**
  GCP provides VPC Service Controls that are ***available by default*** and, when activated, can be used to define a security perimeter around GCP resources, such as Cloud Storage buckets, Bigtable instances, and BigQuery datasets, to constrain data within a VPC and help mitigate genomic data exfiltration risks. With VPC Service Controls, organizations can keep their sensitive data private as they take advantage of the fully managed storage and data processing capabilities of GCP.

  Private Google Access allows GCP instances with only private IP addresses to access the public IP addresses for Google APIs and services, subject to firewall rules in the customer's network.

Manage authentication and authorization of users to controlled-access data

Security Best Practice	Customer Responsibility
Use strong authentication technology like two-factor authentication solutions for controlling access to genomic data; when granting access, adhere to the principle of least privilege in which users are given only those permissions needed to perform their tasks.	*Customers are responsible for implementing the controls to properly safeguard their genomic data and other sensitive information.*

GCP Products and Capabilities

- **Cloud Resource Manager**
  By default, all GCP customers have access to Resource Manager, which enables organizations to group and hierarchically organize other Cloud Platform resources. Using Resource Manager, customers can centrally configure and control administrative access and privileges for GCP.

- **Cloud IAM**
  GCP provides Cloud IAM that is ***available by default*** and, when configured, lets administrators authorize who can take action on specific resources, giving them full control and visibility to centrally manage access to cloud resources. For organizations with complex structures, hundreds of workgroups, and potentially many more projects, Cloud IAM provides a unified view into access control policies across the customer's entire organization, with built-in auditing to ease compliance processes.

- **Cloud Identity**
  Cloud Identity helps GCP customers manage users, devices, apps, and access from a central console; it enables intuitive access to cloud and on-premises apps for users, helps secure their accounts, and protects company data on devices. Cloud Identity also provides context-aware access capabilities that help our customers further control access to their genomic data.

- **Cloud Identity-Aware Proxy**
  Cloud Identity-Aware Proxy (Cloud IAP) controls access to cloud applications running on GCP. Cloud IAP works by verifying user identity and context of the request to determine if a user should be allowed to access the application.

- **Security Keys**
  Security Keys is a feature that our customers can enable to help protect their Google accounts and help prevent account takeovers. A security key helps prevent phishing, and uses cryptography to enable 2-step verification. This feature (1) makes sure the user is logging into the service with which they originally registered the security key, and (2) verifies that the user is using the correct security key.

  Titan Security Keys are phishing-resistant two-factor authentication (2FA) devices from Google that help protect high-value users such as IT admins. Titan Security Keys work with popular browsers and a growing ecosystem of services that support Fast ID Online (FIDO) standards. They are built with a hardware chip that includes firmware engineered by Google to verify the integrity of the key. Titan Security Keys are available on the Google Store and through the customer's GCP representative.

Physical security

Secure infrastructure where genomic data resides

Security Best Practice	Customer Responsibility
Implement controls to restrict, monitor, and manage access to physical servers and other infrastructure on which genomic data resides.	*As described in the Shared Responsibility Model section, Google is responsible for securing the underlying infrastructure. Customers retain responsibility for securing on-premises or third-party services they use.*

GCP Products and Capabilities

- **Data center security**
  By default, all GCP products and services inherit the multiple layers of controls universally implemented across our data centers to physically secure infrastructure where data may reside. These controls include biometric identification, high-resolution video surveillance, and laser-based intrusion detection systems. Access to these data centers is limited to only a small fraction of Google employees; only approved employees with specific roles are authorized to enter. Find more information in this paper's GCP's Approach to Security section.

  Additionally, for instances where GCP hosts some servers in third-party data centers, we ensure Google-controlled physical security measures are implemented on top of the security layers provided by the data center operator. For example, in such sites we may operate independent biometric identification systems, cameras, and metal detectors. For more information, refer to the Security of Physical Premises section of the Google Infrastructure Security Design whitepaper.

Controls for servers

Provide network security for servers

Security Best Practice	Customer Responsibility
Keep servers that host genomics data from being directly accessible from the Internet; disable all non-essential services.	*Customers are responsible for securely using GCP to manage access to and control over the GCP resources they use.*

GCP Products and Capabilities

- **Secure cloud networking technology**
  By default, all customers benefit from GCP's cloud networking technology stack that represents years of cutting-edge research, leading to highly scalable, optimized, and secure network solutions. When using GCP, our customers have access to Google's expansive global high-speed private network that uses state-of-the-art software-defined networking and distributed systems. For more information, refer to our enterprise networking security best practices page.

- **Data access and restrictions**
  By default, GCP logically isolates each customer's Cloud Platform data from that of other customers and users, even when it's stored on the same physical hardware.

- **VPC Service Controls**
  GCP provides VPC Service Controls that are ***available by default*** and, when activated, enable customers to enforce a security perimeter around their resources on GCP,

reducing the risk of genomic data exfiltration. VPC Service Controls can help enterprises protect against data exposure due to misconfigured access controls, malicious users copying data to unauthorized cloud resources, and attackers attempting to access genomic data in GCP resources from the Internet.

- **Virtual Private Cloud**
  Get private access to GCP services, such as storage, big data, analytics, or machine learning, without having to give the service a public IP address. Configure an application's front end to receive Internet requests and shield back-end services from public endpoints, all while being able to access GCP services.

- **Operating system image management**
  GCP customers using Google Compute Engine can choose public operating system (OS) images that, by default, disable non-essential services and root or admin accounts. Customers can also create custom OS images that do the same. Additionally, we provide customers the ability to manage what applications are run on which images.

Establish IAM controls

Security Best Practice	Customer Responsibility
Implement measures to manage access to genomic data; control what data is accessible by whom and limit user permissions to only those required to perform their assigned tasks.	*Customers are responsible for properly authenticating and authorizing users who are permitted to access genomic data.*

GCP Products and Capabilities

- **Cloud IAM**
  GCP provides Cloud IAM that is ***available by default*** and, when configured, can be used to manage resource permissions by mapping job functions within the organization to groups and roles. Users get access only to what they need to get the job done, and admins can easily grant default permissions to entire groups of users. A full audit trail history of permissions authorization, removal, and delegation helps the customer to easily monitor compliance.

- **Cloud Identity**
  With Cloud Identity, GCP customers can create or import user accounts into a cloud-based directory; they can also quickly provision and deprovision accounts as people join the organization, change roles, and leave. Cloud Identity enables customers to implement screen locks, wipe data, and secure devices from the same integrated console where users and apps are managed. In addition, Cloud Identity integrates with hundreds of applications out of the box so GCP customers can manage account identities across cloud and on-prem directories.

- **Cloud IAP**
  Cloud IAP provides secure web access without the need to implement a VPN. With authentication and authorization evaluated by identity and context, only authorized

users are able to access applications. With Cloud IAP, customers can simply manage remote workforces while ensuring their applications and genomic data remain safe.

Ensure genomic data is securely accessed and protected against exfiltration

Security Best Practice	Customer Responsibility
Manage permissions for what data can be exported and by whom; leverage data loss prevention (DLP) technology to prevent accidental and malicious data leakage.	*Customers are responsible for implementing their own controls to limit the copying or exfiltration of genomic data.*

GCP Products and Capabilities

- Cloud **IAM products**
  GCP customers can use Cloud IAM, which is ***available by default*** to all customers, and Cloud Identity to control access to genomic data. Find more information in the General Information Security Guidelines above.

- **Cloud Data Loss Prevention (DLP)**
  Cloud DLP helps our customers better understand and manage sensitive data. It provides fast, scalable classification and redaction for sensitive data elements including social security numbers, patient information, and genomic data. Automatically choose the most suitable storage system and the right set of access controls based on the presence of sensitive content. Cloud DLP gives customers the power to scan, discover, and report on data from virtually anywhere.

- **Cloud Virtual Private Network (VPN)**
  Cloud VPN securely connects on-premises networks to GCP Virtual Private Cloud (VPC) networks through an IPsec VPN connection. Traffic traveling between the two networks is encrypted by one VPN gateway, then decrypted by the other VPN gateway. This protects genomic data as it traverses the Internet.

Source data and control of copies of data

Support version control of genomic data

Security Best Practice	Customer Responsibility
Retain the original version of genomic data and track all copies.	*Customers are responsible for tracking all copies of genomic data and securely retaining original data when necessary.*

- **Object Versioning**
  Cloud Storage offers the Object Versioning feature, which is ***available by default***, to support the retrieval of objects that are deleted or overwritten. To control copies of genomic data, we offer Object Lifecycle Management tools to establish file lifespan, change storage classes for cost management, and archive older versions of objects.

- **Periodic backups**
  Backups provide a way to restore a Cloud SQL instance to recover lost data or recover from a problem with the customer's instance. A feature is ***available by default*** that, when activated, enables Cloud SQL to retain up to seven automated backups for each instance. In addition to a full-sized backup, customers can maintain several incremental backups at a cost-effective price to support their genomic research efforts. Backups can be created on demand or be scheduled periodically.

- **Cloud Source Repositories**
  GCP provides Google Cloud Source Repositories, which are fully featured, private Git repositories hosted on GCP. If customers prefer to use their current version control solutions, they can do so by hosting and running them on GCP or by connecting to an externally hosted or managed service such as GitHub or Bitbucket.

- **Cloud archival storage**
  To meet customers' storage needs, GCP's Cloud Storage offers four storage classes that differ by their availability, minimum storage durations, and pricing:

  Multi-Regional Storage is a low-cost, highly durable storage service used for storing data that is frequently accessed around the world. This option is especially useful for genomic data that will be accessed on a regular basis by researchers from multiple geographic regions.

  Regional Storage is a low-cost, highly durable storage service for storing data that is frequently accessed in the same region as a customer's Google Cloud DataProc or Google Compute Engine instances. This is a great choice for genomic data being accessed by researchers within a specific region.

  Nearline is a low-cost, highly durable storage service for storing infrequently accessed data. This is a great option for customers wanting to continuously add genomic data to Cloud Storage, but that plan to access the data only once a month for analysis.

  Coldline is a low-cost, highly durable storage service for data archiving, online backup, and disaster recovery. Unlike other "cold" storage services, our customers' data is available within milliseconds, not hours or days. For organizations looking to retain original copies of other copies of genomic data for long periods of time, Coldline storage is a great low-cost solution.

Provide logging of actions related to the use or access of genomic data

Security Best Practice	Customer Responsibility
Log who accessed genomic data, what data was accessed, where it was accessed from, and when the data was accessed.	*Customers are responsible for implementing and maintaining audit logs to identify access to genomic data.*

GCP Products and Capabilities

- **Audit logs**
  Cloud Audit Logging maintains three audit logs, which are ***available by default***, to provide visibility into the actions of a GCP customer's own administrators. These logs include: Admin Activity, System Events, and Data Access — while Admin Activity and System Event logs are on by default, the Data Access logs are available by default. When activated, GCP services write audit log entries to these logs to help answer the questions "who did what, where, and when?" to genomic data and related applications.

- **Stackdriver Logging**
  Cloud Platform offers a powerful integrated suite of logging-oriented services. In the Cloud Platform stack, Stackdriver Logging serves as the centralized collection and indexing service, aggregating logs from across the customer's GCP resources.

- **Access Transparency logs**
  Access Transparency gives customers near real-time logs when GCP administrators access customer data. Our administrators access customer data only when necessary to fulfill GCP's contractual obligations. In the limited situations in which data access is required by GCP administrators, technical controls are in place to restrict access and in near real-time generate logs to customers without them needing to implement the technology themselves. Only Admin Activity logs include extensive information such as accessor location, access justification, and the action taken on a specific resource. ***By default***, Access Transparency logs are available for Compute Engine, Cloud Storage, Cloud IAM, Cloud KMS, and several other GCP product offerings.

Destruction of data

Dispose of genomic data properly

Security Best Practice	Customer Responsibility
Wipe systems used for hosting genomic data repositories when they are no longer needed; retain only encrypted copies of the minimum data necessary to comply with institutional data retention policies.	*Customers are responsible for properly disposing of genomic data and securing encrypted copies of the minimum amount of genomic data needed to be retained.*

GCP Products and Capabilities

- **Data deletion**
  By default, when deleting data on GCP, we confirm the deletion request and then begin eliminating the data iteratively from application and storage layers, from active and backup storage systems. To learn more, refer to the Data Deletion on GCP whitepaper.

- **Equipment disposal and reuse**
  By default, all GCP products and services inherit the robust controls implemented as part of our IT asset management program, including procedures around equipment disposal and reuse. When a hard drive is retired, individuals verify that the disk is erased by writing zeros to the drive and performing a multiple-step verification process to ensure the drive contains no data. If the drive cannot be erased for any reason, it is stored securely until it can be physically destroyed. Physical destruction is a multi-stage process that includes the use of a crusher and a shredder.

General cloud computing guidance

Maintain secure protocol and encryption solutions

Security Best Practice	Customer Responsibility
Use end-to-end encryption technology to securely transfer genomic data between systems; encrypt genomic data at rest; ensure cryptographic keys are properly managed and stored.	*Customers are responsible for properly using GCP to protect genomic data at rest and in transit.*

GCP Products and Capabilities

- **Encryption in transit**
  By default, GCP applies encryption to data in transit. The type of encryption varies depending on the connection being made and if the data is being transmitted outside a physical boundary controlled by or on behalf of GCP.

- **Encryption at rest**
  By default, GCP encrypts data stored at rest, without any action required from the customer, using one or more encryption mechanisms.

- **Flexible encryption key management**
  GCP offers several options for encryption key management. A fully managed encryption key service is ***on by default*** that manages server-side encryption keys for customers; no setup or configuration is required. We also provide options for customers to supply their own keys and to fully manage their own encryption keys.

- **Cloud Key Management Service (KMS)**
  Cloud KMS lets customers manage cryptographic keys for their cloud services the same way they do on-premises. Cloud KMS supports advanced features like key hierarchy, broad symmetric and asymmetric key support, and hardware integration support for Cloud HSM.

Protect against common and emerging web application vulnerabilities

Security Best Practice	Customer Responsibility
Ensure the applications used to handle genomic data are protected against Top 10 Web Application Security Risks and emerging vulnerabilities as reported by information sources such as SecurityFocus and the NIST National Vulnerability Database.	*Customers are responsible for properly using GCP services and for protecting their systems and applications against security threats, pursuant to the GCP Shared Responsibility Model.*

GCP Products and Capabilities

- **Security Scanner**
  Cloud Security Scanner is ***available by default*** for customers building applications on App Engine and can be used at no additional cost; it is a web security scanning tool for detecting common vulnerabilities. Security Scanner can automatically scan and detect four common vulnerabilities, including cross-site-scripting (XSS), Flash injection, mixed content (HTTP in HTTPS), and outdated/insecure libraries. This scanning tool makes it easy to set up, run, schedule, and manage security scans.

Protect access to cloud environment and resources

Security Best Practice	Customer Responsibility
Establish access control lists (ACLs) to resources where genomic data resides; assign permissions following the principle of least-privilege approach.	*Customers are responsible for maintaining up-to-date ACLs to prevent unauthorized users from accessing genomic data. Customers are also responsible for enforcing principles of least privilege in their organization.*

GCP Products and Capabilities

- **VPC Service Controls**
  GCP provides VPC Service Controls that are ***available by default*** and, when activated, enable customers to define fine-grained perimeter controls and enforce that security posture across numerous GCP services and projects. Please refer to previous sections of this whitepaper for more information about VPCs and VPC Service Controls. To learn more about our networking solutions, refer to our Networking Products page.

Audit and accountability

Monitor and validate access controls are properly enforced

Security Best Practice	Customer Responsibility
Validate access permissions to genomic data are properly enforced; maintain logging mechanisms to trigger notifications of violations; regularly review and audit access controls.	*Customers are responsible for regularly reviewing access and permission settings to validate safeguards are adequately protecting access to genomic data.*

GCP Products and Capabilities

- **Cloud Security Command Center**
  Cloud Security Command Center gives organizations consolidated visibility into their cloud assets across services such as App Engine, Compute Engine, Cloud Storage, Datastore, and many others. With it, customers can quickly understand the number of projects they have, what resources are deployed, where sensitive data is located, and how firewall rules are configured.

  Cloud Security Command Center can be used to help prevent unintended exposure of genomic data, to ensure the appropriate access control policies are in place across cloud resources, and to get alerts when policies are misconfigured or unexpectedly change.

Image-specific security

Protect systems against vulnerabilities and control administrative access

Security Best Practice	Customer Responsibility
Verify machine images to be used for genomic research do not contain known vulnerabilities, malware, and viruses; ensure root and admin accounts are secured and default passwords are changed; ensure unnecessary user accounts are disabled.	*Customers are responsible for ensuring their systems are protected against security threats and administrator-level access is tightly controlled.*

GCP Products and Capabilities

- **Shielded VMs**
  Shielded VMs are virtual machines (VMs) on GCP hardened by a set of security controls that help defend against rootkits and bootkits. Using Shielded VMs helps protect enterprise workloads from threats like remote attacks, privilege escalation, and malicious insiders. They leverage advanced platform security capabilities such as secure and measured boot, a virtual trusted platform module (vTPM), Unified Extensible Firmware Interface (UEFI), and integrity monitoring. Compute Engine offers a broad range of VM operating systems.

Customers can select VMs shared from the GCP community or bring their own. While we continue to update our product offerings to provide protection against the latest threats, it is the customer's responsibility to provide up-to-date protection for their managed environments.

- **Container security**
Containerization allows development teams to move fast, deploy software efficiently, and operate at an unprecedented scale. As organizations create more containerized workloads, security must be integrated at each stage of the build and deploy lifecycle. Running infrastructure as containers offers our customers a new way to incorporate security into applications. Containerization can simplify the patch management process, reduce the host system attack surface, and further isolate resources run on GCP.

Scale seamlessly with GCP

In addition to giving customers the assurance that the data stored with GCP is safe, our experience in delivering technology at scale provides customers with unique capabilities to expand their genomic research efforts. GCP reduces operational overhead by taking on a large portion of the responsibility in handling the performance, scalability, availability, and security needs of big data solutions so our customers can focus on analysis and less on managing servers. Customers pay only for the resources they use.

We offer a proven, integrated end-to-end Big Data solution, based on years of innovation at Google, that lets our customers capture, process, store, and analyze their data within a single platform. Our Artificial Intelligence platform offers a range of products to bring scale and speed to customers' genomic research. GCP's reliable and robust information security infrastructure can help organizations meet their genomic data security and privacy requirements. We provide the following additional resources to support our customers with successfully building secure applications atop GCP.

Documentation	We share documentation including how-to guides, best practices, blog posts, FAQs, and whitepapers like this one to help customers access the information they need at any time.
Audit logs	GCP services write audit logs that help customers answer the questions of "who did what, where, and when?"
Technical support services	We offer different support options including free support resources and access to online communities of GCP enthusiasts, experts, and Google employees to choose from.
Training and certifications	We offer training and certifications for customers to learn the technical skills and best practices that will help them make the most of GCP product offerings.
Tutorials	We provide tutorials to help customers get started with GCP products and services.
Case studies	We share case studies to highlight the GCP success stories of our Health and Life Sciences customers.

Frequently asked questions

Data control

Who controls access to genomic data on GCP?

GCP customers control access to their own genomic data. To learn more, refer to the GCP's approach to privacy section.

How is access to genomic data managed on GCP?

Cloud Identity & Access Management (Cloud IAM) lets administrators authorize who can take action on specific resources, giving you full control and visibility to manage cloud resources centrally. For example, GCP provides customers with the ability to control who has access to their genomic data stored in Cloud Storage buckets and objects as well as what level of access they have. To learn more, refer to our Access Control Options page for storage products.

What happens if a customer wants to leave GCP and take their genomic data?

GCP provide tools that enable customers to take their data with them if they choose to stop using our services. To learn more, refer to our How-to Guide for Creating and Managing Data Transfers.

Data security

How does GCP protect against hackers and other intruders?

GCP employs a robust defense-in-depth approach to security. Refer to the GCP's approach to security section for more information about how we provide a secure cloud platform to protect customers' genomic data from hackers and other intruders.

What controls are offered to secure genomic data stored on GCP?

GCP offers several products and services that can serve as controls to help our customers secure genomic data; these are discussed throughout the GCP and dbGaP security best practices section. Visit our products page for up-to-date information on our full list of identity and security products. In addition, security controls are imbedded in each of our products and services and include capabilities such as encryption, data loss prevention, and identity and access management.

What controls are offered to secure the transfer of genomic data?

GCP provides controls such as encryption for our customers to securely transfer their genomic data. Refer to the GCP's approach to security and General cloud computing guidelines sections for more information. For customers looking to migrate large volumes of data to GCP, we provide additional information in the Migration section of our Cloud Computing products overview.

Data privacy

Who has access to my genomic data on GCP?

GCP provides customers with the tools to manage who has access to their genomic data. Refer to the GCP's approach to privacy section for more information.

How is genomic data on GCP protected from unauthorized access?

GCP provides highly secure infrastructure and controls to prevent unauthorized access, which are described in Google Infrastructure Security Design Overview and the GCP's approach to security section of this whitepaper.

Data availability

Will my genomic data on GCP always be available?

GCP's application and network architecture is designed for maximum reliability and uptime. Data is distributed across our servers and data centers. We own and operate data centers around the world helping to keep our services running 24/7/365. To learn more, refer to our Reliability page. Customers can take advantage of various reliability and replication features of our products; please refer to our SLAs page.

What happens when a disruption to GCP services is experienced?

GCP plans for our services to always be available. We perform tests to validate that our services are resilient under heavy workloads and unusual circumstances. In the unlikely instance when a GCP server experiences an outage, depending on the customer configuration and where possible, our platform will use robust software failover to withstand the disruption and continue delivering the services being used by our customers. To learn more, refer to our Reliability page.

Data compromise

What happens if an incident involving a customer's genomic data on GCP takes place?

We have a rigorous data incident management process that specifies actions, escalations, mitigation, resolution, and notification of any potential incidents impacting the confidentiality, integrity, or availability of customer data. We will promptly notify customers if we detect a security breach leading to the accidental or unlawful destruction, loss, alteration, unauthorized disclosure of, or access to their data on systems we manage. Moreover, we will assist with investigative efforts via our support team. Refer to our Data Incident Response Process whitepaper and our Data Processing and Security Terms for more information.

Contractual commitments

What does Google do with my genomic data?

Customers own their data. As described in the GCP Terms of Service, GCP does not process, use, or access customer data for any reason other than to fulfill contractual obligations to our customers. To learn more, refer to our Data Processing and Security Terms.

How does GCP guarantee that a customer's genomic data is deleted?

As described in our Data Processing and Security Terms, Google will comply with Customer's instructions to delete Customer Data as soon as reasonably practicable and within a maximum of 180 days. When you delete your Customer Data, Google's deletion pipeline begins by confirming the deletion request and eliminating the data iteratively from application and storage layers, from both active and backup storage systems. This process is described generally in Google's statement on deletion and retention.

Where can I learn more about performance and customer service commitments?

GCP is committed to offering a state-of-the-art cloud services platform, securely built from the ground up. We continue to innovate to bring our customers new products and features so they can solve their biggest challenges. To learn more, refer to our Google Cloud Difference, our SLAs, and our GCP Terms of Service.

Conclusion

Organizations can take advantage of GCP products and services not only to process, analyze, store, and control access to genomic data, but also to help meet their regulatory compliance, security, and privacy requirements. We hope that you learned how GCP products and capabilities align with the NIH dbGaP Security Best Practices and other important regulations governing the processing of genomic data. We encourage customers to use the information herein as they consider designing, building, and deploying applications on GCP that will handle their genomic data in a secure manner, protecting it from unauthorized access, loss, damage, and destruction.