AWS Architecture Monthly

Genomics
July 2021

Editor's note

The field of genomics has made huge strides in the last 20 years. Genomics organizations and researchers are rising to the many challenges we face today, and seeking improved methods for future needs. Amazon Web Services (AWS) provides an array of services that can help the genomics industry with securely handling and interpreting genomics data, assisting with regulatory compliance, and supporting complex research workloads. In this issue, we have case studies from Lifebit and Fred Hutch, blogs on genomic sequencing and the Registry of Open Data, and some reference architectures and solutions to support your work. We include videos from the Smithsonian, AstraZeneca, Genomic Discoveries, AMP lab, Illumina, and the University of Sydney.

We hope you'll find this edition of Architecture Monthly useful. We'd like to thank Kelli Jonakin, PhD, Global Head of Life Sciences & Genomics Marketing, AWS, as well as our Experts, Ryan Ulaszek, Worldwide Tech Leader - Genomics, and Lisa McFerrin, Worldwide Tech Leader - Bioinformatics, for their contribution.

Please give us your feedback! Include your comments on the Amazon Kindle page. You can view past issues and reach out to aws-architecture-monthly@amazon.com anytime with your questions and comments.

Jane Scolieri, Managing Editor

Table of Contents:

Notices

Ask an Expert:

Ryan Ulaszek, Worldwide Tech Leader - Genomics
Lisa McFerrin, PhD, Worldwide Lead - Bioinformatics

What are some general architecture patterns for genomics in the cloud?

Genomics data generated by research, biopharma, and healthcare organizations is beginning to outpace their ability to cost effectively store, manage, and analyze this data on-premises. Thus, we are increasingly seeing the following four architectural trends that use AWS services for moving genomics workloads to the cloud.

1. **Data transfer and storage.** Organizations are shifting to solutions like AWS DataSync to manage their large-scale data transfers. DataSync provides durable, high throughput data transfer. It can also throttle data to manage network bandwidth. DataSync handles common tasks, minimizing your IT operational burden. Customers use Amazon Simple Storage Service (Amazon S3) because it provides scalable, highly available, and secure storage. It preserves data long term and provides high availability in downstream analysis.

2. **Workflow automation for secondary analysis.** Once in the cloud, raw genomic sequencing data undergoes "secondary analysis." Each sample is processed through orchestrated tasks such as sequence alignment, variant calling, annotation, and quality control. To process these tasks, organizations use AWS Step Functions for serverless orchestration and AWS Batch to provision resources to optimize processing time and cost.

3. **Data aggregation and governance.** To gain insights from data, customers combine sample data and layer inputs such as functional annotations or clinical fields. AWS Glue, Amazon Athena, Amazon Redshift, and AWS Lake Formation are often used to integrate different data sources, enable accessibility and querying, and process data while keeping sensitive data secure. AWS Glue prepares and catalogs data. Athena provides querying for cohort creation. Lake Formation layers data access controls to meet data governance needs and comply with regulatory standards.

4. **Tertiary analysis and machine learning.** Genomics data can be combined with other data modalities to predict disease risk and drug response or inform clinical decision making. Amazon Redshift and Athena are commonly used for variant storage and query. Amazon QuickSight and Jupyter notebooks in Amazon SageMaker are used to query and visualize data. Amazon SageMaker helps build, train, and deploy machine learning models to potentially discover relationships between biomarkers and patient populations to improve treatments and outcomes for patients.

What are some considerations when putting together an AWS architecture to solve business problems specifically for genomics customers?

Realizing the potential of genomics in precision medicine requires data analytics capabilities to increase knowledge of biology and disease, identify new targets for medicines, improve patient selection in clinical trials, and inform treatment strategies to optimize therapeutic benefit for patients.

However, organizations can be challenged by 1) genomics data outgrowing on-premises storage, 2) limited access to local high performance computing (HPC) clusters, and 3) inconsistent processes to manage complex workflows. Key considerations that address these concerns map well to the five pillars of the AWS Well-Architected Framework:

Security is our highest priority when creating an AWS architecture. Genomics data is generally considered the most private of personal data. Privacy, reliability, and security are critical to data creation, collection and processing, and storage and transfer. Amazon S3 and AWS Identify and Access Management (IAM) help maintain a strong security posture. They provide specific controls for authorizing data access, defining data governance, and establishing and maintaining data encryption. Amazon CloudWatch provides logs and events. AWS CloudTrail maintains audit logs. AWS Control Tower applies guardrails for ongoing governance over your AWS workloads.

Operational excellence. The genomics field is still emerging. Many tools are open-source and distributed via code and container repositories like GitHub and Docker Hub. These tools are frequently used by research and development for biomarker discovery, drug development, and association studies. Services like AWS CodeCommit, AWS CodeBuild, and AWS CodePipeline allow you to automate change management through continuous integration/ continuous delivery (CI/CD). Amazon Elastic Container Registry (ECR) and Amazon Elastic Container Service (ECS) help store, manage, share, and deploy container images.

Reliability. Step Functions allow you to handle task failures in bioinformatics workflows, and AWS Batch scales horizontally to meet sample processing demand. Call caching and fault-tolerant pipelines further prevent reprocessing of compute and time intensive work that would be costly to rerun, such as aligning and mapping whole genome sequences during secondary analysis for variant detection.

Performance efficiency. As organizations experience higher volumes and velocity of data generated from genomic sequencers, large-scale data transfer is shifting to solutions like DataSync for durable, high throughput data transfer. Right sizing your tools also allows you to select the Amazon Elastic Compute Cloud (Amazon EC2) instance type for optimal performance. Tertiary analysis, like joint genotyping, single-cell analysis, and genome-wide association studies, can have extensive memory requirements to parse hundreds of thousands to millions of biomarkers. Using efficient data formats such as parquet files and distributed compute via Amazon EMR and distributed query via Athena allows for improved memory management and computation times.

Cost optimization for compute, storage, and data transfer is a key consideration for organizations. Configuring object life-cycling in Amazon S3 optimizes storage costs based on access patterns and storage requirements. Larger files that are rarely accessed can be moved to S3 Glacier Deep Archive for long-term storage and archiving. Amazon EC2 Spot Instances reduce costs by performing genomics data processing and analysis at off-peak hours. This is particularly valuable for secondary analysis workflows. Compute instances optimized for Amazon Redshift are also available. Currently, the AWS Open Data Sponsorship Program covers the cost of storage for publicly available high-value cloud-optimized datasets. This democratizes access to data and encourages development of communities benefiting from shared data access. It also promotes development of cloud-native techniques, formats, and tools that lower the cost of working with the data. There are already over 70 genomics and life science datasets available within the Registry of Open Data on AWS.

Do you see different trends in genomics workflows in cloud versus on-premises?

Organizations running genomic workloads on-premises are typically capacity constrained, which can lead to significant wait times in the available queues and more complicated resource management. When moving to the cloud, most organizations choose to transition to AWS Batch for task and resource management. AWS Batch provides on-demand and spot queues that offer a serverless pay-for-what-you-use approach to scale capacity for increased throughput. Amazon FSx for Lustre is also regularly integrated for data staging and I/O management, which mimics on-premises cluster network file system for data access.

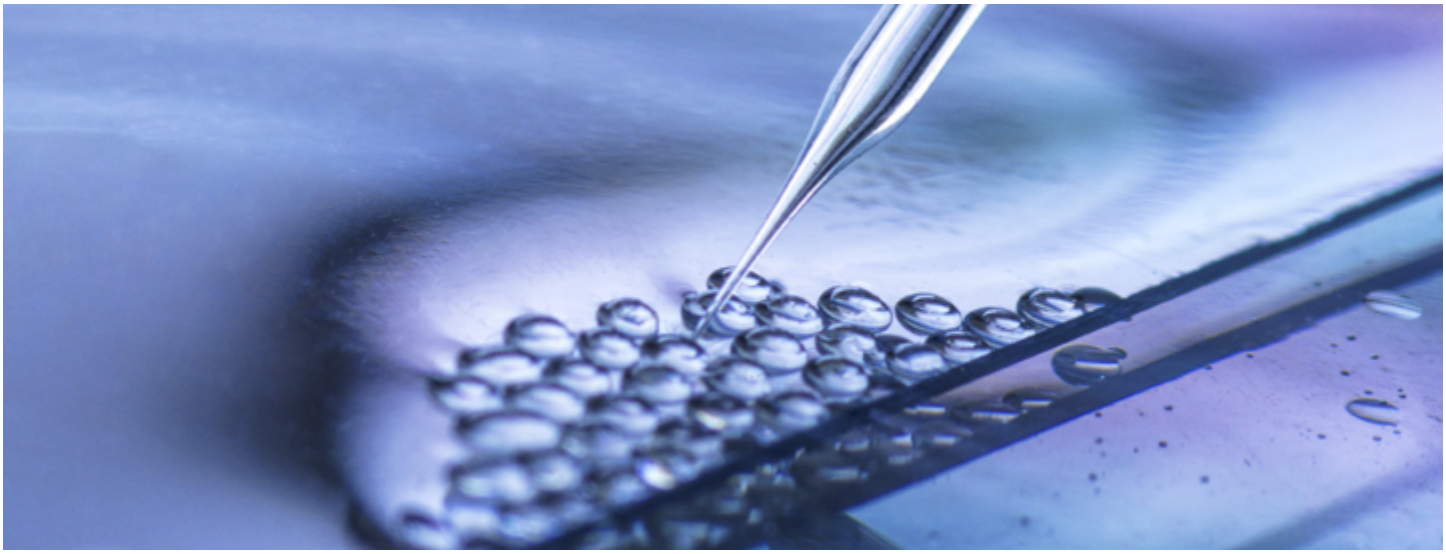
When migrating secondary analysis to the cloud, teams often use AWS compute instance types to optimize their tasks for cost and performance. Teams also move to containers and CI/CD to build and deploy tools independently through their own deployment pipelines. This minimizes the consequence of change and improves operational efficiency.

Downstream genomic analysis frequently uses R and Python languages and community-driven packages. Researchers on-premises usually manage installation and environment dependencies on their own computers, or they'll work with their IT department for version updates. When installed at user, group, or organization levels, application dependencies can be difficult to manage and may break workloads. On AWS, it's fairly easy to migrate to containers that integrate all the tool dependencies. Then you can build, package, and deploy tools independently in Docker images, which minimizes issues for researchers.

What's your outlook for genomics, and what role will cloud play in future development efforts?

Genomics research has led to remarkable advances across industries. This includes accelerating drug discovery, enhancing clinical trials, improving decision support for precision medicine, enabling population sequencing, and powering sustainable and diverse agriculture. Using patients' molecular signatures in research and clinical care will speed up the research and development process, reduce costs, improve patient satisfaction, and ultimately drive the efficacy of clinical trials with up to a 2x higher success rate.

Because sequencing has gotten cheaper over the last decade, there is a rise in population genomics, millions of individuals have been sequenced around the world. The subsequent



increase in data has provided larger and more diverse datasets so we can ask more complex questions about how genes may influence health. This drives progress to improve the prevention, diagnosis, and treatment of a range of illnesses, including cancer and rare genetic diseases. These practices are being similarly applied in agriculture. Genomics is powering sustainable and diverse initiatives that allow farmers to improve plant and crop yield in addition to animal breeding practices.

Genomic analysis and interpretation requires researchers to collaborate with peers, the scientific community, and institutions. The large datasets being analyzed require integrated multi-modal datasets and knowledge bases, intensive computational power, big data analytics, and machine learning at scale.

The cloud enables genomics to innovate by making data and methods more findable, accessible, interoperable, and reusable. The storage and compute services on AWS reduce the time between sequencing and interpretation, with secure and seamless sharing capabilities plus cost-effective infrastructure. Data and tool repositories allow for easier access to existing resources that can be readily deployed in replicate environments, accelerating the modern study of genomics.

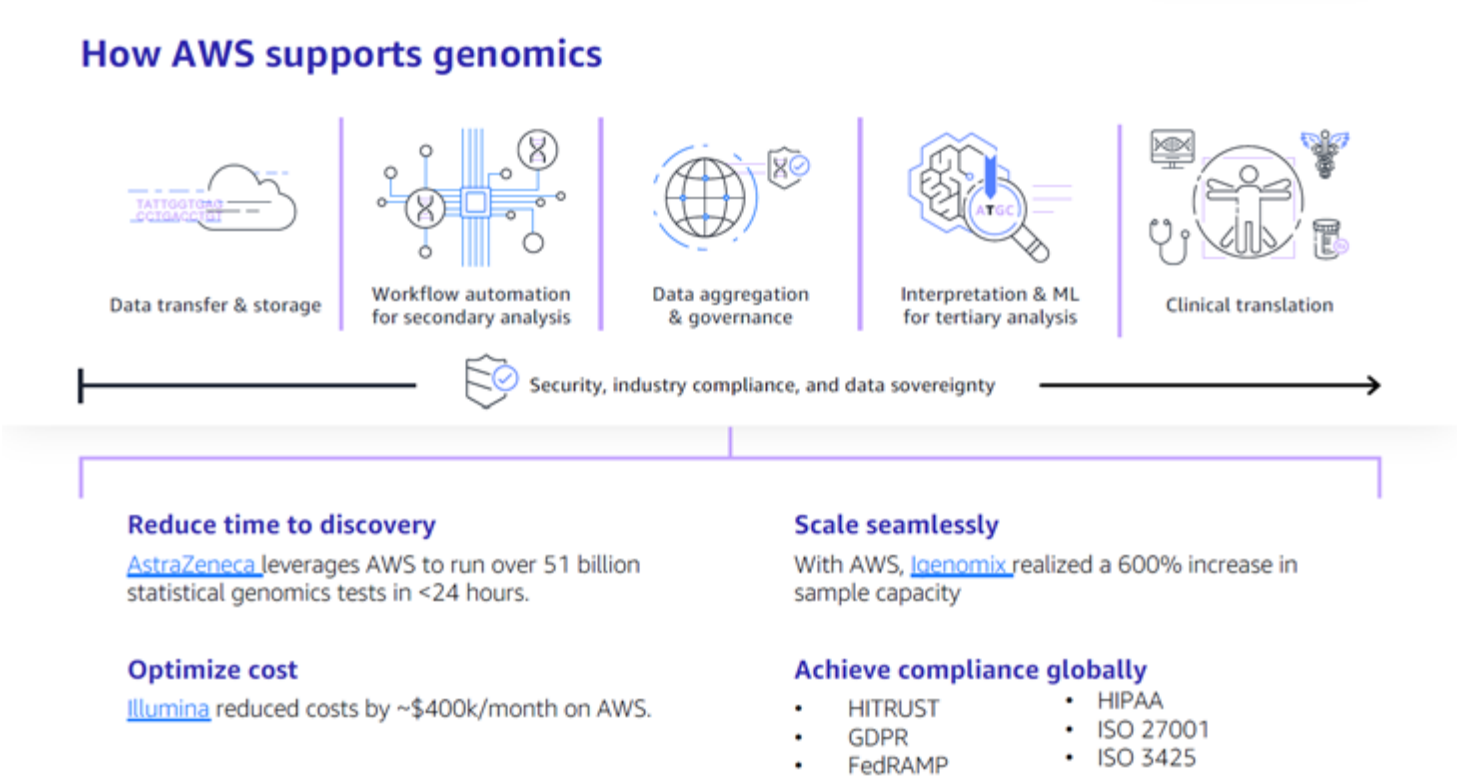
Ryan Ulaszek is Worldwide Tech Lead, Genomics at AWS where he oversees technical initiatives in genomics and acts as liaison between the AWS Service teams and the technical field community, worldwide. Ryan has worked as a genomics specialist in AWS for three years, helping AWS life science customers architect genomics solutions in the Amazon Web Services (AWS) Cloud. Ryan has worked as a software engineer for over twenty years in numerous life science companies, including Human Longevity, founded by J. Craig Venter, where he was the principal architect. He also worked within Amazon as a Senior Software Engineer, leading architecture projects that spanned across teams in Amazon Fresh and Amazon Retail. Ryan holds patents in life science and software engineering and five AWS Solutions Architecture certifications, including the AWS Solutions Architect Professional and AWS DevOps Engineer Professional certifications.

Lisa McFerrin is bioinformatics lead at AWS, where she drives initiatives supporting the genomics, healthcare, and life science industries as part of worldwide business development. Prior to AWS, Lisa worked at Fred Hutchinson Cancer Research Center where she specialized in the development of software and methods that bridge genomic and clinical data to advance the understanding of cancer biology and improve patient care. In these roles, she facilitates collaborative and reproducible research in order to lower the barriers in communication, analysis, and sharing of data, knowledge and methods. Lisa has a background in math and computer science and obtained her PhD in Bioinformatics from North Carolina State University.

Executive Brief:

Genomics on AWS: Accelerating scientific discoveries and powering business agility

Today's genomics landscape is rapidly evolving with the accelerated adoption of genomics by biopharma organizations, infectious disease tracing programs, and healthcare systems. Organizations that are addressing new market opportunities and embracing innovative genomics applications and technologies are looking to the cloud to stay agile, innovative, and economical.For almost a decade, Amazon Web Services (AWS) has helped genomics organizations such as Ancestry, DNAnexus, GRAIL, and Illumina optimize their businesses and build scalable infrastructure to accelerate scientific discoveries. Genomics organizations of all sizes and disciplines choose AWS to meet their unique business needs, which span from using AWS machine learning services to bridge limitations in NGS infectious disease applications, to leveraging AWS global infrastructure to seamlessly deploy globally in a secure and compliant manner.



Access to industry tools and datasets

Technical and business experts from AWS collaborate with genomics customers to address common operational and business challenges, such as providing best practices for deploying analysis workflows at scale and integrating with open-source projects such as Cromwell or Nextflow. AWS works commercially with ISVs, such as Illumina's DRAGEN™Bio-IT Platform and Sentieon, to make their applications available through native and platform services. Finally, the AWS Open Data team works with public data providers such as gnomAD and the NIH to make their datasets available to all AWS customers at the click of a button.

Reduce time to discovery

Genomics is a data-heavy discipline requiring extensive compute resources to analyze samples and extract meaningful insights.

To accelerate discoveries, AWS offers a robust suite of powerful compute and machine learning options that enable scientists to process more samples, run more complex analyses, and query at-scale. The only cloud provider to deliver 100 Gbps of networking throughput, AWS delivers high-throughput data ingestion, analysis, and interpretation services and tools designed to help genomics organizations get more from their data.

Genomics organizations such as Fauna Bio leverage the robust computation power of AWS to analyze multi-omicdatasets, accelerate research, and uncover new discoveries.

Accelerate data analysis

With native integration to workflow tools, including Nextflow and Cromwell, organizations can orchestra AWS Batch processes to accelerate processing time for computational analysis. For example, with AWS Fred Hutch was able to reduce compute time from seven years down to seven days, accelerating the organizations research on developing therapeutics to fight cancer.

Simplify data interpretation

The full value of genomic data is recognized once its put into context. Population genomics programs across the globe leverage the security, flexibility, and scalability of AWS to host population-scale biobanks and provide democratized access to the industry. Using querying and machine learning services from AWS, scientists can rapidly query datasets, including the Cancer Genome Atlas (TCGA) and the Broad Institute's Genome Aggregation Database (gnomAD) hosted on AWS registry of open data, to rapidly extract insights and answers.

Optimize cost

While sequencing costs have fallen, costs associated with compute and storage have grown as organizations continues to increase throughput and explore more data-heavy applications, such as single-cell genomics. In contrast to on-premises setups that require large upfront investments and continuous CapEx, AWS enables genomics organization to use only what they need, when they need it.

AWS offers storage and compute offerings at a variety of price points. Services such as Amazon EC2 Spot Instances offer up to 90 percent discounts in comparison to on-demand compute prices. For long-term storage of infrequently used data, Amazon S3 Glacier provides secure data archiving starting at $1 per terabyte per month.

Genomics organizations also leverage AWS to reduce development and operational costs. By building its AI-based genomics intelligence platform on AWS, Emedgene reduced its costs of applying artificial intelligence to big genomics by 70 percent while accelerating model development and optimization.

Scale seamlessly

Many factors impact spikes and lulls in genomics workloads. From academic grant cycles to flares in infectious disease, genomic organizations must be able to respond quickly to succeed. In today's competitive market, companies that plan their on-premises capacity based on maximum anticipated volume carry an enormous financial burden; whereas those that cannot accommodate demand spikes quickly face costly delays.

The elastic nature of AWS allows organizations to launch as many instances as needed and be ready to work in near real-time. Likewise, they can decrease capacity during down times to reduce costs.

Melbourne Genomics Health Alliance built its GenoVicsoftware, a shared clinical system used by Victoria's decentralized health system, on AWS to enable seamless scalability as more laboratories join the alliance and expand their genomic testing services.

"Data stays in our clients' environments, and all of the AWS safety features keep it secure."

Lifebit reengineered the traditional model for securing data — bringing its compute engine and analytics to the data itself – enabling customers like Genomics England to securely democratize access to its data on the AWS Cloud.

Achieve compliance globally

DNA is the most personal source of data. As genomics becomes an increasingly global practice, organizations must consider both domestic and international regulatory standards.

AWS takes a security first approach to enabling genomics storage and compute, and continues to invest in new availability zones around the globe to enable organization to store and analyze their data locally.

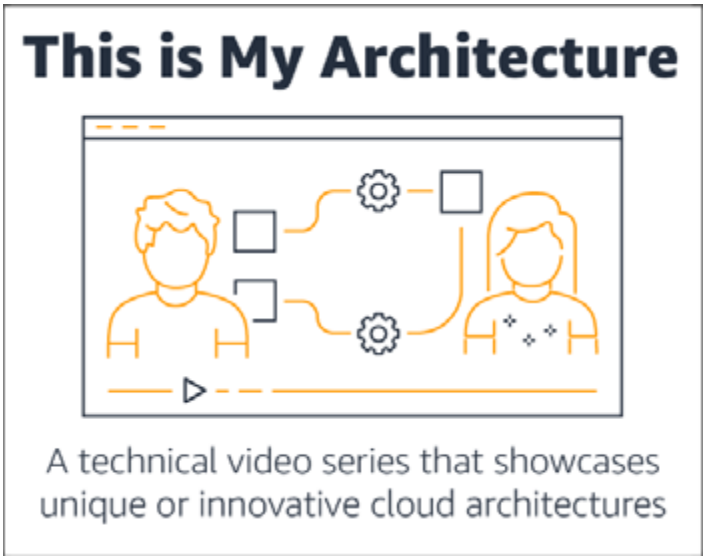
The AWS Global Cloud Infrastructure is the most secure, extensive, and reliable cloud platform, offering over 200 fully featured services from data centers globally.

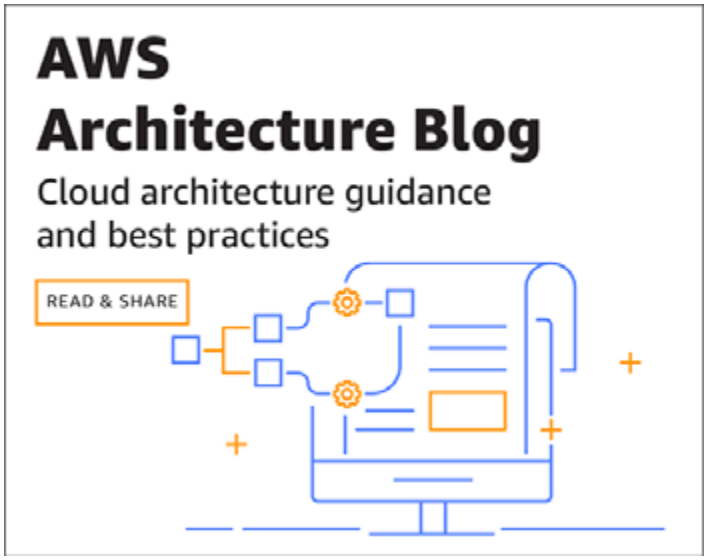
With 77 Availability Zones, 2X more than any other cloud provider, across 245 countries and territories, genomics organizations can take full advantage of cloud and maintain data sovereignty.

To help genomics organization adhere to industry compliance and regulations, AWS maintains the following compliance certifications: HITRUST, GDPR compliance, FedRAMP, HIPAA, ISO 27001, ISO 3425.

View Executive Brief online

View whitepaper online





Case Study:

Fred Hutch Microbiome Researchers Use AWS to Perform Seven Years of Compute Time in Seven Days

2019

At Seattle's Fred Hutch Microbiome Research Initiative (MRI), a team of researchers are engaged in analysis of the microbiome, which is the collection of microbes on and inside the human body. But these researchers aren't just studying the microbiome—they're striving to manipulate microbiomes to make therapeutic cancer drugs more effective.

To support their efforts, researchers must analyze and process an immense number of whole genome datasets. "There are hundreds of thousands of microbes in each person's microbiome, and they're different for each individual," says Sam Minot, PhD and staff scientist at Fred Hutch MRI. "Translating gigabytes of raw microbiome genomic data into insights about which specific microbes are present in a person is a computationally intensive task requiring highly scalable technology."

Running Scientific Research Workloads on AWS

Dr. Minot chose Amazon Web Services (AWS) to power the high-performance computing (HPC) platform that runs microbiome analysis. He says, "Additional research groups within the organization had been using AWS, and we saw how they benefited from the scalability of the cloud."

"Our goal is to accelerate our research processes on AWS so we can get closer to developing therapeutics to fight cancer."

Sam Minot,
PhD and Staff Scientist, Fred Hutch Microbiome Research Initiative

These researchers execute their computational analysis using the Nextflow framework to orchestrate AWS Batch processes and scale the HPC platform to accelerate processing time. "AWS Batch integrates well with Nextflow, so it was easy for us to get Nextflow up and running without having to reinvent the wheel," says Dr. Minot. The organization runs its HPC workloads on Amazon Elastic Compute Cloud (Amazon EC2) instances, powered by Intel Xeon Platinum 8000 Series processors, and it stores research data in Amazon Simple Storage Service (Amazon S3) buckets.

Recently, Dr. Minot's group began using Amazon EC2 Spot Instances to access compute resources for its HPC environment. Amazon EC2 Spot Instances are unused Amazon EC2 capacity that is available at up to a 90 percent discount compared to On-Demand Instance prices. He says, "We can use the same budget to access more compute resources using Amazon EC2 Spot Instances."

Seven Years of Compute Time in Seven Days

Using AWS, Dr. Minot's group has the scalability to analyze publicly available datasets that contain data on more than 15,000 biological samples, each representing a gigabyte of storage. As a result, they have performed seven years of aggregate compute time in seven days, giving researchers the ability to get results faster and ultimately speed research that will find therapeutics for cancer treatments. "Running our microbiome research on Amazon EC2 Spot Instances, we spend less money and less time to get scientific answers from the analysis," says Dr. Minot. "Our goal is to accelerate our research processes on AWS so we can get closer to developing therapeutics to fight cancer."

"Zooming In" on Genes

Dr. Minot is using the AWS-based research computing environment to increase the resolution of analysis for large collections of microbiome samples. "We can use the scale of AWS to zoom in from the species level to the genes present inside those species, which requires an extremely high level of computational detail," he says. By increasing analytical resolution, researchers more easily find links to health outcomes. "As an example, we can perform global analysis on AWS across thousands of published datasets and study groups of people with inflammatory bowel disease. At the same resolution, we can also analyze microbiomes in people with colorectal cancer, and then identify how inflammation may relate to the development of cancer. Using the scale and resolution we get on AWS, we can make better comparisons across different disease states, many of which interact with the microbiome."

Additionally, Dr. Minot can share his methods with other scientific researchers, who can conduct their own research using these same methods to further extend scientific discovery. "The simplicity of integrating Amazon EC2 Spot Instances with AWS Batch and Nextflow to scale gives us an element of reproducibility," says Vijay Sureshkumar, director of technology partnerships at Fred Hutch MRI. "We can extend this model to other research labs within Fred Hutch or at other institutions. This can lead to more research collaboration, so we can work together to find potential cures for cancer."

To learn more, visit aws.amazon.com/hpc.

View case study online

About Fred Hutch Microbiome Research Initiative

The Fred Hutch Microbiome Research Initiative, funded by Seattle's Fred Hutchinson Cancer Research Center, includes microbiome investigators with expertise in study design, laboratory methods, animal models, human intervention studies, data analysis, and visualization. These researchers are working to predict health outcomes, understand the pathogenesis of disease, and manipulate the microbiota to promote health.

Benefits of AWS

- Processes data from more than 15,000 biological samples
- Reduced 7 years of compute time to 7 days, so researchers can get results faster
- Increases resolution on microbiome samples to find links to improve health outcomes
- Enables collaboration with other scientific researchers

Quick Start:

For rapid deployment

Step 0: Amazon VPC

While you can use an existing "default" VPC to implement deployment of your genomics environment, we strongly recommend utilizing a VPC with private subnets for processing sensitive data with AWS Batch. Doing so will restrict access to the instances from the internet, and help meet security and compliance requirements, such as dbGaP. *NOTE*, these private subnets **must** have a route to the secure route to the internet. A typical method would be to use a NAT Gateway although other options are possible.

Tip

You may also want to review the HIPAA on AWS Enterprise Accelerator and the AWS Biotech Blueprint for additional security best practices such as:

- Basic AWS Identity and Access Management (IAM) configuration with custom (IAM) policies, with associated groups, roles, and instance profiles
- Standard, external-facing Amazon Virtual Private Cloud (Amazon VPC) Multi-AZ architecture with separate subnets for different application tiers and private (back-end) subnets for application and database
- Amazon Simple Storage Service (Amazon S3) buckets for encrypted web content, logging, and backup data
- Standard Amazon VPC security groups for Amazon Elastic Compute Cloud (Amazon EC2) instances and load balancers used in the sample application stack
- A secured bastion login host to facilitate command-line Secure Shell (SSH) access to Amazon EC2 instances for troubleshooting and systems administration activities
- Logging, monitoring, and alerts using AWS CloudTrail, Amazon CloudWatch, and AWS Config rules

The template uses the AWS Quickstart reference for a Modular and Scalable VPC Architecture and provides a networking foundation for AWS Cloud infrastructures, deploying an Amazon Virtual Private Cloud (Amazon VPC) according to AWS best-practices and guidelines.

For architectural details, best practices, step-by-step instructions, and customization options, see the deployment guide.

The VPC quick start template will deploy a VPC with 2 private subnets (the minimum recommended for the core environment) in two Availability Zones (AZs). For production environments we recommend using as many AZs as are available in your region. This allows the AWS Batch compute environments to source workers from more AZs potentially resulting in better pricing and fewer interruptions when using Spot Instances. A simple way to create a CloudFormation template for a complete VPC stack with multiple AZs is to use the AWS CDK. This example Java app will synthesize a CloudFormation template that can be used to generate a VPC with subnets in upto 6 AZs (or the maximum for the region if there are less than 6):

```
public class CdkVpcApp {
    public static void main(final String[] args) {
        App app = new App();

        Environment env = Environment
                .builder()
                .account("my-account-number")
                .region("us-east-1")
                .build();

        new CdkVpcStack(app, "CdkVpcStack", StackProps.builder().
env(cromwell).build());

        app.synth();
    }
}
```

```
public class CdkVpcStack extends Stack {
    public CdkVpcStack(final Construct scope, final String id) {
        this(scope, id, null);
    }

    public CdkVpcStack(final Construct scope, final String id, final
StackProps props) {
        super(scope, id, props);

        // The code that defines your stack goes here
        Vpc vpc = Vpc.Builder.create(this, "vpc")
                .maxAzs(6)
                .build();
    }
}
```

View full quickstart online

Blog

NIH's Sequence Read Archive, the world's largest genome sequence repository: Openly accessible on AWS

by Erin Chu, DVM, Ph.D., Ankit Malhotra, and Lee Pang

AWS and the National Library of Medicine's (NLM) National Center for Biotechnology Information (NCBI) are happy to announce that the Sequence Read Archive (SRA) – one of the world's largest repositories of raw next generation sequencing data, will be freely accessible from Amazon S3 via the Open Data Sponsorship Program (ODP). The SRA is currently hosted by NLM at the National Institutes of Health (NIH). As we publish this blog, the transition to the ODP is under way.

What is the SRA?

Established in 2009 as part of the International Nucleotide Sequencing Database Collaboration (INSDC), the SRA is the NIH's primary repository for raw next generation sequencing data. Currently, the SRA hosts over 36 petabytes of sequence data representing controlled- and public-access data dating back to 2007, and representing sequencing from over 9 million experiments. It is commonly the first stop for scientists looking to validate a research discovery, expand their effective sample population, or test out a new pipeline. In fact, the SRA website saw 1.2 million visitors in 2019 alone. These visitors reflect the changing landscape of genomics; the SRA working group reported that 20% of IP addresses come from cloud-based virtual machines such as those available with Amazon EC2.

With the power of cloud compute, bioinformaticists now have the capacity to analyze the SRA at a comprehensive scale. For example, Serratus, an open science project for rapid discovery of novel and existing coronaviruses, used AWS Batch to call and align over 4 million SRA accessions in parallel for coronavirus sequence.

At the rate it is growing now, the SRA is expected to double every 12-18 months, presenting new challenges for efficient storage and accessibility. To that end, NIH released a Request for Information (RFI) from the biomedical research community to provide input on next steps for the future of the SRA. A major theme of this RFI is to reduce the data footprint of the SRA by eliminating base quality scores (BQS). Results from this RFI are expected to be published in early 2021. In the meantime, steps have been taken to make the SRA even easier to access and use by researchers everywhere.

The new normal

Moving the SRA to the Open Data Sponsorship Program (ODP) provides an avenue to retain BQS, while reducing the complexity by which researchers can locate and retrieve SRA data. AWS users will be able to simply use Amazon Athena to query the publicly accessible SRA metadata bucket `s3://sra-pub-sars-cov2-metadata-us-east-1` for accessions of interest, or directly interrogate the SRA bucket for a specific SRA submission or set of submissions, then call it directly into a cloud-based genomics workflow.

Direct access to SRA data as S3 objects will enable more scalable and cloud native tooling for processing and analyzing genomics datasets. Rather than copying terabytes or even petabytes of data into their own environments, researchers can use SRA in S3 as a single source of truth for their analysis, subsequently making workflows more reproducible and amenable to global research collaborations.

> "Having a single location of sequencing data with complete Base Quality Scores (BQS) is essential for continued development of new and novel methods for genomic analysis," says Benedict

> Paten, Associate Professor and Associate Director UC Santa Cruz Genomics Institute. "We are very interested in the continued support of SRA, and I am glad that the data with full BQS would be available for use by the research community through the AWS Open Data Program."

The NLM also maintains two additional S3 buckets hosted by the Open Data Program (allocated under the NIH STRIDES agreement) that will be used specifically for raw data for newer sequencing technologies such as Pacific Biosciences, Oxford Nanopore, and 10X Genomics, and newer submissions as space allows.

The SRA will join a growing list of key biomedical and genomics datasets such as TCGA, ICGC, Gabriella Miller Kids First, ENCODE, gnomAD, Human Microbiome Project and Human PanGenomics Project that have been provided to the research community free of charge on the AWS platform. Currently, 61 biomedical and genomics datasets are listed in the Registry of Open Data, composing over 9 PB of data.

Get ready for SRA data on AWS

While the SRA is just beginning its transition to the ODP, NLM has already made 250 TB of coronavirus genome sequence data available on AWS ODP; the overall structure of this bucket will echo that of the larger SRA dataset. Get started with this data on AWS—find more on our blog post.

The NLM has also recently held a webinar that demonstrates how to query SRA metadata with Amazon Athena using the same coronavirus genome sequence dataset. See NCBI Minute: SRA in AWS Athena for SARS-CoV-2 Research and More.

Finally, in collaboration with AWS Educate Research Seminar Series, we dug in deeper into the open data ecosystem on AWS, showcasing the SRA. This and all other research seminars are available on demand here.

We look forward to continuing our collaboration with NIH and NLM to increase access and utility of this invaluable resource. Stay tuned for additional resources and releases in coming months.

Learn more about how AWS is supporting research

Ready to start incorporating SRA data into your cloud workflows? Explore AWS genomics solutions.

For more information on how AWS helps solve complex research workloads and enables scientific research, see the AWS Research and Technical Computing webpage.

Read more information about the NIH STRIDES Initiative.

Learn more about the AWS Open Data Sponsorship Program (ODP)

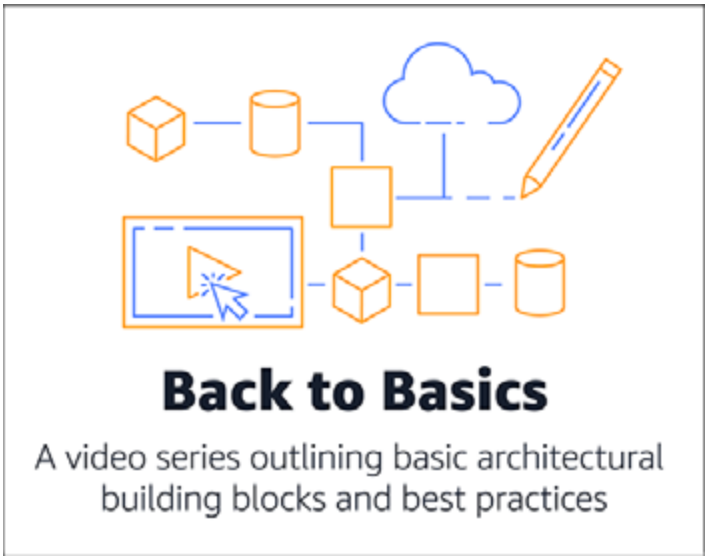
The AWS Open Data Sponsorship Program covers the cost of storage for publicly available high-value cloud-optimized datasets. We work with data providers who seek to:

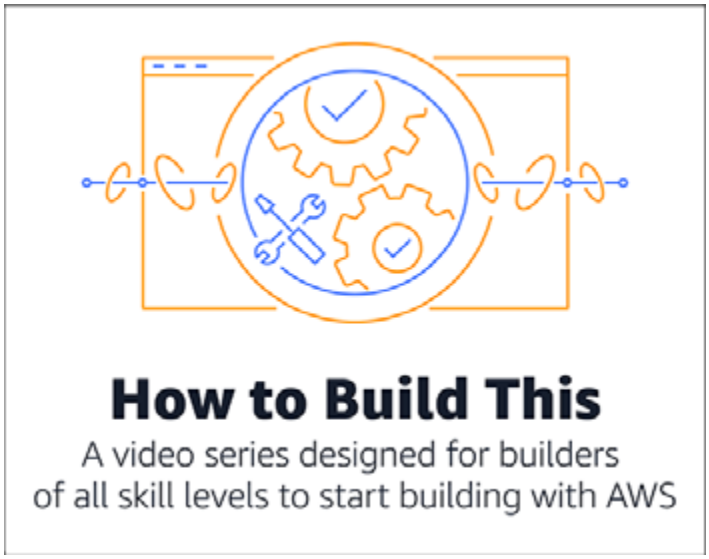
- Democratize access to data by making it available for analysis on AWS
- Develop new cloud-native techniques, formats, and tools that lower the cost of working with data

Encourage the development of communities that benefit from access to shared datasets.

Read blog online

More Life Science blogs





Solutions

Genomics Secondary Analysis Using AWS Step Functions and AWS Batch

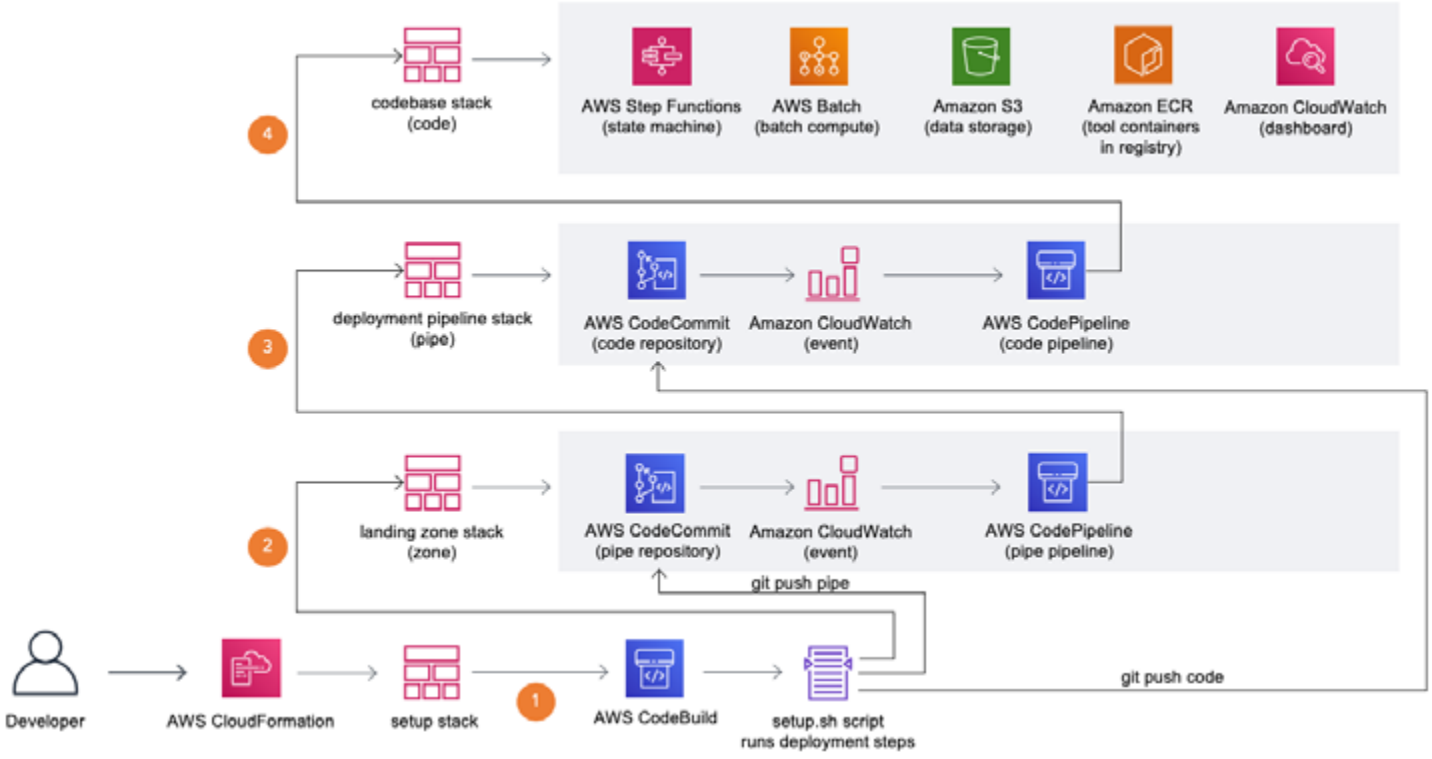
What does this AWS Solutions Implementation do?

The Genomics Secondary Analysis Using AWS Step Functions and AWS Batch solution creates a scalable environment in AWS to develop, build, deploy, and run genomics secondary analysis pipelines, for example, processing raw whole genome sequences into variant calls. This solution includes continuous integration and continuous delivery (CI/CD) using AWS CodeCommit source code repositories and AWS CodePipeline for building and deploying updates to both the genomics workflows and the infrastructure that supports their execution. This solution fully leverages infrastructure as code principles and best practices that help you to rapidly evolve the solution.

Amazon CloudWatch operational dashboards are deployed to monitor status and performance for pipelines and tools. Customers can deploy this solution for their genomics analysis and research projects.

AWS Solutions Implementation overview

The diagram below presents the architecture you can automatically deploy using the solution's implementation guide and accompanying AWS CloudFormation template.



Genomics Secondary Analysis Using AWS Step Functions and AWS Batch solution architecture

The AWS CloudFormation template creates four CloudFormation stacks in your AWS account including a *setup* stack to install the solution. The other stacks include a landing zone (*zone*) stack containing the common solution resources and artifacts, a deployment pipeline (*pipe*) stack defining the solution's CI/CD pipeline, and a codebase (*code*) stack providing the tooling, workflow definitions, and job execution environment source code.

The solution's *setup* stack creates an AWS CodeBuild project containing the *setup.sh* script. This script creates the remaining CloudFormation stacks and provides the source code for both the AWS CodeCommit *pipe* repository and the *code* repository, once they have been created.

The landing zone (*zone*) stack stack creates the CodeCommit *pipe* repository, an Amazon CloudWatch event, and the AWS CodePipeline *pipe* pipeline which defines the continuous integration/continuous delivery (CI/CD) pipeline for the genomics workflow. The deployment pipeline (*pipe*) stack stack creates the CodeCommit *code* repository, an Amazon CloudWatch event, and the CodePipeline *code* pipeline.

The CodePipeline *code* pipeline deploys the codebase (*code*) CloudFormation stack. The resources deployed in your account include Amazon Simple Storage Service (Amazon S3) buckets, CodeCommit repositories for source code, AWS CodeBuild projects, AWS CodePipeline pipelines, Amazon Elastic Container Registry (Amazon ECR) image repositories, an example AWS Step Functions state machine, and AWS Batch compute environments, job queues, and job definitions. An example Amazon CloudWatch dashboard provides operational workload monitoring. In total, this solution enables building and deploying updates to both the genomics workflows, and the infrastructure that supports their execution.

Read online

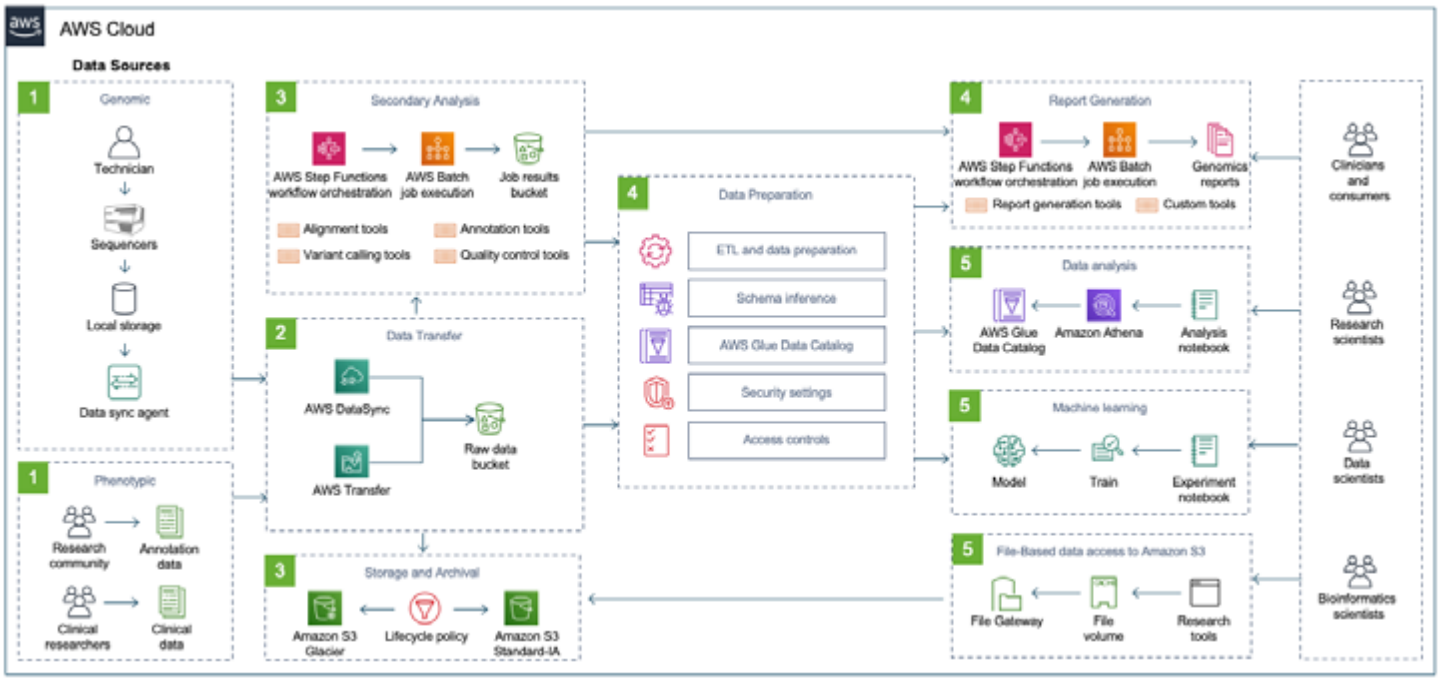
View Implementation Guide here

View all genomics solutions here

Reference Architecture

Genomics data transfer, analytics, and machine learning reference architecture

The following genomics reference architecture describes the AWS services used in this paper to ingest, store, archive, analyze, prepare, and interpret genomics data to gain insights and make predictions.



Genomics data transfer, analytics, and machine learning reference architecture

View reference architecture online

Reference architecture included in this whitepaper

Case Study

Lifebit Powers Collaborative Research Environment for Genomics England on AWS

2021

Less than 1 year after the COVID-19 pandemic outbreak, professionals went from diagnosing the first case to administering a vaccine. Genomic advancement, among other breakthroughs, is widely credited with the rapid understanding of the disease and the expedited vaccine deployment.

Since the first whole human genome was sequenced in 2003, genomics has become commonplace in the healthcare and life sciences industries, resulting in an exponential growth in genomic data. Each human genome contains enough data to fill 200 phone books. Within this data lie life-altering discoveries, including knowledge of the causes of diseases, which can lead to treatments. But disease causes—which are often "typos," or mutations in genetic sequences—can be challenging to find; and genomic data is highly regulated and stored in siloed data lakes, further impeding research.

Facing this challenge is Lifebit Biotech Ltd. (Lifebit), an Amazon Web Services (AWS) Select Consulting Partner. Working with biobanks, research institutions, and pharmaceutical companies, Lifebit provides solutions that analyze clinicogenomic datasets to accelerate drug discovery, diagnostics, disease surveillance, drug-response predictions, and wellness models.

Unlocking Access to Siloed Genomic Data

Lifebit CloudOS, a fully federated cloud operating system, uses AWS to unlock clinicogenomic data for drug and biomarker discovery. This facilitates greater research collaboration, enabling a rapid increase in drug development and disease prevention. At the onset of the COVID-19 pandemic, Genomics England (GEL) turned to Lifebit CloudOS. A pioneer of population genomics, GEL oversees the 100,000 Genomes Project, a cohort of cancer and rare-disease whole genomes.

Earlier genomics research relied on fewer, smaller datasets, and the industry could rely on centralized technologies to analyze this data. As a result, data protection regulation was more lenient, and collaboration was more manageable. But because genomic data has since become the largest source of data in history, that system cannot support today's research. "Data centralization is no longer feasible or affordable," says Thorben Seeger, vice president of commercial for Lifebit. "The

We use the whole roster of AWS computations—from general-purpose computation to graphically accelerated units—to run large production pipelines faster and more efficiently."

Thorben Seeger
Vice President of Commercial, Lifebit Biotech Ltd.

data is too big to move efficiently, and many regulations forbid data to leave an organization, state, or nation." As a result, 80–90 percent of these datasets are unavailable to research. "GEL is widely known as the 'Fort Knox' of genomics," Seeger says. "But when you lock data up, it's nearly impossible to access or combine with other data."

Lifebit reengineered the traditional model for securing data—bringing its compute engine and analytics to the data itself. This new model is powered by Amazon Elastic Compute Cloud (Amazon EC2), a web service that provides secure, resizable compute capacity in the cloud. "We are deploying our cutting-edge research in our clients' own environments on AWS," says Seeger. "Each user receives a clean-room environment to access and analyze data separately. The fully managed service provides maximum research utility without sacrificing security or control."

Lifebit uses the highly scalable cloud capabilities of AWS to gain the compute capacity it needs to accommodate the exponential relationship between the size of a dataset and the outcomes. The company works on projects with more than 100 PB of stored data, requiring billions of virtual CPU hours. "We use the whole roster of AWS computations to run production pipelines faster and more efficiently," says Seeger. "That was critical because GEL needed rapid data processing for faster insights."

Standing Up a Secure, Robust Collaboration Service

During the COVID-19 pandemic, GEL launched an initiative with the UK government to deliver a cohort to eight leading pharmaceutical companies—as well as research organizations—to fuel vaccine, treatment, and early-detection research. The cohort included sequenced genomes from 20,000 COVID-19 patients with severe cases and 15,000 patients with mild cases, plus data from the 100,000 Genomes Project. Yet GEL needed a federated data analytics system to make that cohort available to multiple parties. "We were setting up a new research environment, and we needed a company that could go live within 7–8 weeks," says Parker Moss, chief commercial officer at GEL.

Lifebit built upon GEL's existing AWS architecture to deliver the fully live system in under 3 months. Today, pharmaceutical companies and researchers can access the cohort and connect their own private datasets. "The user's external data doesn't move into the GEL environment," says Moss. "However, through federated links, you can research as if that data is in one place. It's a very powerful value proposition." This system saves time and offers extra protection. "Data stays in our clients' environments, and all of the AWS safety features keep it secure," says Seeger.

On the system, researchers use automated tools to securely query, analyze, and collaborate on large datasets in seconds. "We are bridging the dichotomy between security and usability," says Seeger. "This fosters global collaboration between public institutions like GEL, other leading cohorts, research organizations, and private institutions."

Scaling at the Speed of Genomics on AWS

Lifebit CloudOS makes genomic research more accessible. "The cloud, combined with our data environment, is the great democratizer," explains Seeger. "Millions of researchers can access and perform big data analysis on demand—something only a few trained specialists with high-performance computing could do previously."

Critically, Lifebit customers and their users gain virtually infinite storage using Amazon Simple Storage Service (Amazon S3), which offers industry-leading scalability, data availability, security, and performance. One whole human genome equates to 120–300 GB of data, and Lifebit is performing simulations on running

databases on more than 10 million patients with thousands of clinical and phenotypical variables. “Connecting global datasets is driving ethnic genomic diversity,” says Seeger. “This helps us understand diseases in general but also enables us to cater to previously underserved populations.”

On AWS, Lifebit delivered a system that led to one of the most significant cloud computing deals in the history of life sciences. “The prevalence of AWS in the healthcare and life sciences markets is extremely helpful,” Seeger says. “We have seen incredible flexibility from AWS, which, in the London region, is helping us set up the security GEL is famous for. The scale and global presence of AWS is of huge strategic importance for us as we pursue large government initiatives.”

Accelerating Global Collaboration in Drug Research and Disease Prevention

By using AWS, Lifebit enabled GEL to rapidly deliver a research environment for COVID-19 data and analytics. Now, Lifebit is speaking to nations about combining datasets to facilitate research outcomes and speed up drug development for cancer and rare diseases. “Our federated analysis system doesn’t only exist for the singular purpose of serving one country or one disease cohort,” says Seeger. “It works with other cohorts worldwide, making this scientific field the most collaborative it has ever been.”

View case study online

About Lifebit Biotech Ltd.

Lifebit Biotech is a global leader in population genomics software and AI-powered drug discovery. Operating in North America, Europe, the Middle East, Africa, and the Asia-Pacific region, it powers population genomics initiatives, biobanks, research, and pharma companies.

Benefits of AWS

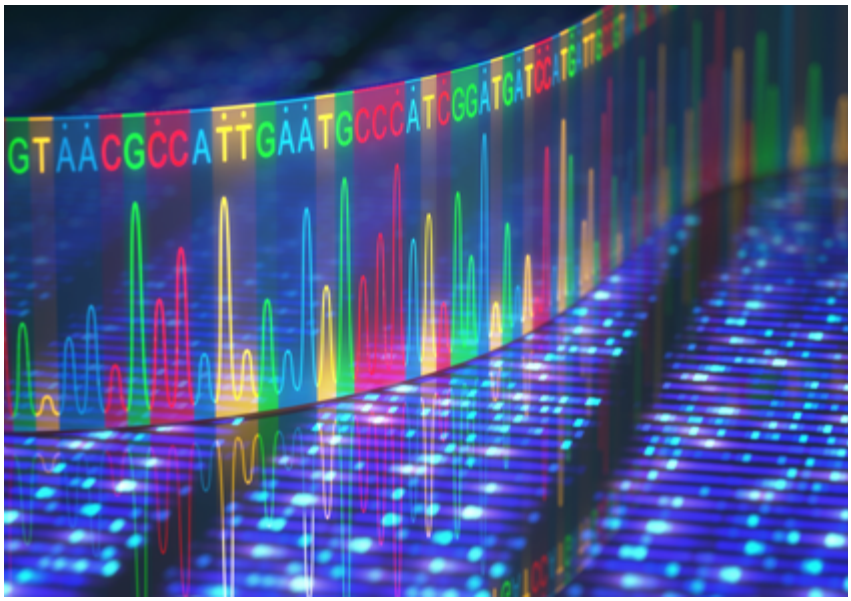
- Launched a federated data analytics system in under 3 months
- Processes more than 100 PB of project data
- Enables collaborative research on disparate datasets worldwide
- Maintains compliance with data privacy regulations
- Performs analysis in clients’ own environments
- Efficiently orchestrates billions of CPU hours
- Democratizes access to bioinformatic analysis
- Enables sustainable self-funding business models

Quick Start

Illumina DRAGEN on AWS

Ultra-rapid analysis of next-generation sequencing (NGS) data with DRAGEN and F1 instances

This Quick Start deploys the Illumina DRAGEN (Dynamic Read Analysis for GENomics) Bio-IT Platform on the AWS Cloud.

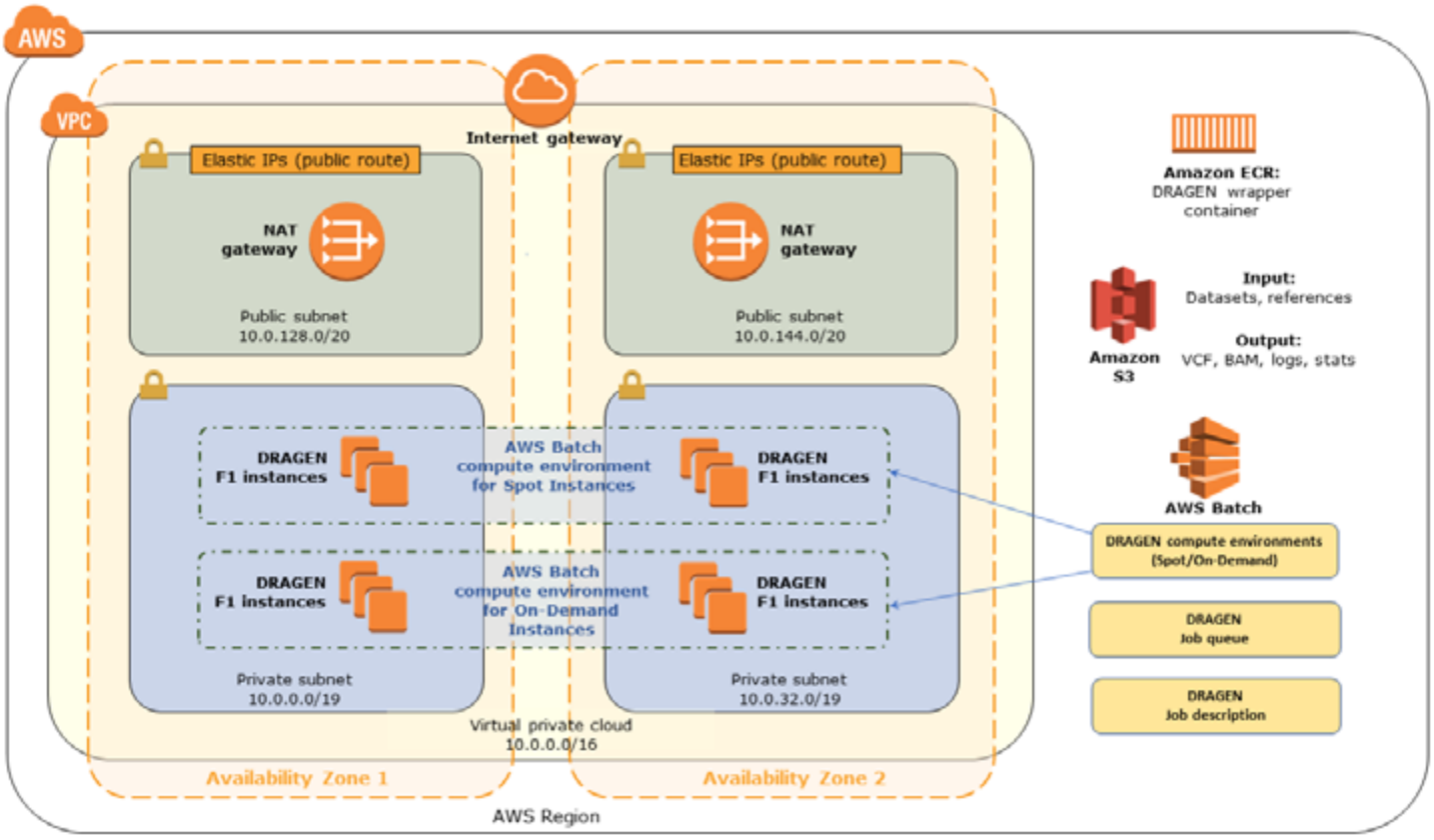


The DRAGEN Bio-IT Platform enables ultra-rapid analysis of next-generation sequencing (NGS) data, significantly reduces the time required to analyze genomic data, and improves accuracy. It includes bioinformatics pipelines that provide optimized algorithms for mapping, aligning, sorting, duplicate marking, and haplotype variant calling. These pipelines include Germline, Somatic (tumor and tumor/normal), RNA, Single Cell RNA, Methylation, Joint Genotyping, and DRAGEN-GATK.

The Quick Start builds an AWS environment that spans two Availability Zones for high availability, and provisions two AWS batch compute environments for spot instances and on-demand instances. These environments include DRAGEN F1 instances that are connected to field programmable gate arrays (FPGAs) for hardware acceleration.

Use this Quick Start to set up the following configurable environment on AWS:

- A highly available architecture that spans two Availability Zones.*
- A virtual private cloud (VPC) configured with public and private subnets according to AWS best practices. This provides the network infrastructure for your deployment.*
- An internet gateway to provide access to the internet.*
- In the public subnets, managed NAT gateways to allow outbound internet access for resources in the private subnets.*
- An AWS CodePipeline pipeline that builds a Docker image and uploads it into an Amazon Elastic Container Registry (Amazon ECR) repository.
- Two AWS Batch compute environments: one for Amazon Elastic Compute Cloud (Amazon EC2) Spot Instances and the other for On-Demand Instances.
- An AWS Batch job queue that prioritizes submission to the compute environment for Spot Instances to optimize for cost.
- An AWS Batch job definition to run DRAGEN.
- AWS Identity and Access Management (IAM) roles and policies for the AWS Batch jobs to run.



* The template that deploys the Quick Start into an existing VPC skips the tasks marked by asterisks and prompts you for your existing VPC configuration.

View Quick Start online

View deployment guide

This Quick Start was developed by Illumina in collaboration with AWS. Illumina is an AWS Partner.

Executive Brief

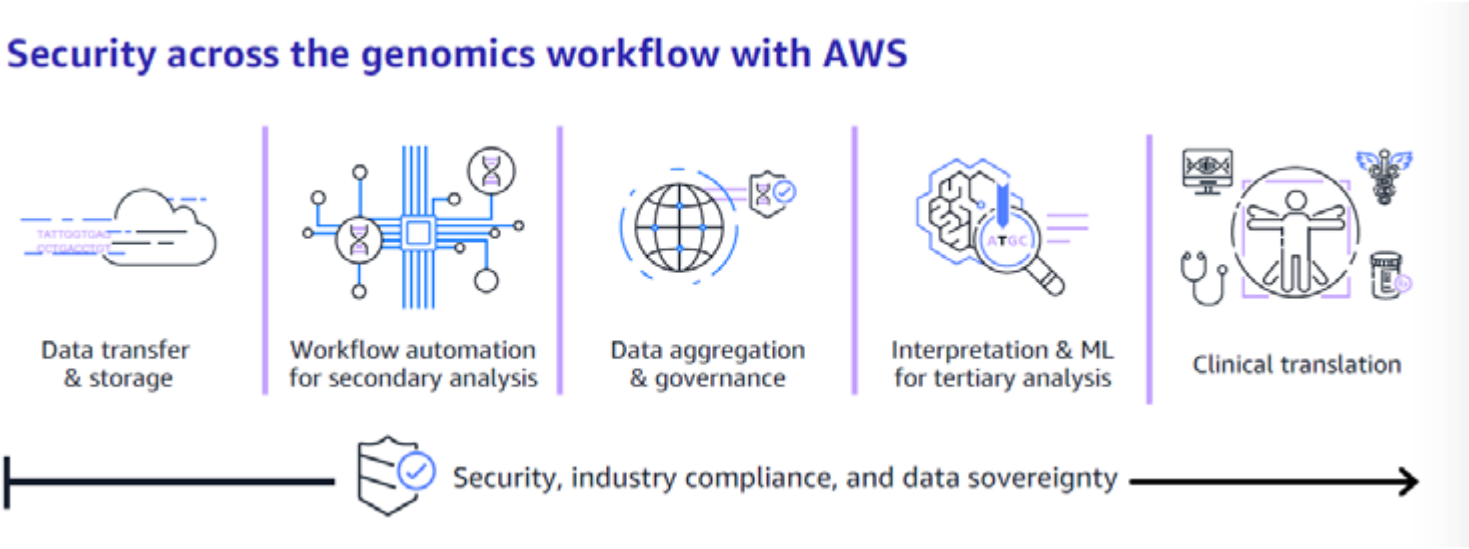
Genomic data security and compliance on the AWS Cloud

Speed and scalability as you want it. Security and compliance as you need it.

DNA is the most personal source of data. The handling, storing, interpreting, and analyzing of genomic data requires a commitment to security and regulatory compliances across the entire genomics workflow.

From clinical genomics to population-scale programs, the genomics industry builds on Amazon Web Services' (AWS) secure cloud environment to create compliant genomics applications. For almost a decade, genomics organizations have leveraged the robust suite of security tools and shared responsibility model of AWS to reduce operational burdens, ranging from hosting operating systems to ensuring physical security.

With AWS, you get the breadth and depth of capabilities needed to create a highly secure cloud environment that simplifies and improves the way you manage, analyze, and share genomics data. An internal team of genomics experts work closely with AWS product engineers to understand the implications of industry-specific regulations on data sovereignty, aggregation, ownership, and provenance—letting you focus on the science.



Data sovereignty
AWS provides a wide offering of regional instances to enable in-country data hosting. With 80 Availability Zones across 245 countries and territories, genomics organizations can take full advantage of the cloud and meet domestic and international regulatory standards, including GDPR.

Data aggregation
Apply FAIR governance policies over stored, pooled data using built-in metadata tags and identity-based permission controls to make data findable and accessible to authorized users only.

Data ownership
Maintain existing security and compliance postures including inter-team data enclaves. As the sole owner of the encryption key for each unique data set, no one can see or access your data on AWS but you.

Data provenance & governance
Improve data auditability, and monitor and control authorization using integrated AWS services on the Cloud that make it easy to automatically receive alerts and granularly track who accesses and changes data, while controlling data versions.

Enhance your security posture on AWS
The AWS shared responsibility model helps you manage data privacy, reliability, and security both within the Cloud and of the Cloud itself. Work with a team of industry specialists to understand and improve your security posture at every stage of the genomics lifecycle—from data creation, to collection and processing, to storage. AWS provides a range of services to help you secure and manage your data, including:

View executive brief online

Solutions

Genomics Tertiary Analysis and Data Lakes Using AWS Glue and Amazon Athena

What does this AWS Solutions Implementation do?

The Genomics Tertiary Analysis and Data Lakes Using AWS Glue and Amazon Athena solution creates a scalable environment in AWS to prepare genomic data for large-scale analysis and perform interactive queries against a genomics data lake. This solution can help IT infrastructure architects, administrators, data scientists, software engineers, and DevOps professionals build, package, and deploy libraries used for genomics data conversion; provision data ingestion pipelines for genomics data preparation and cataloging; and run interactive queries against a genomics data lake.

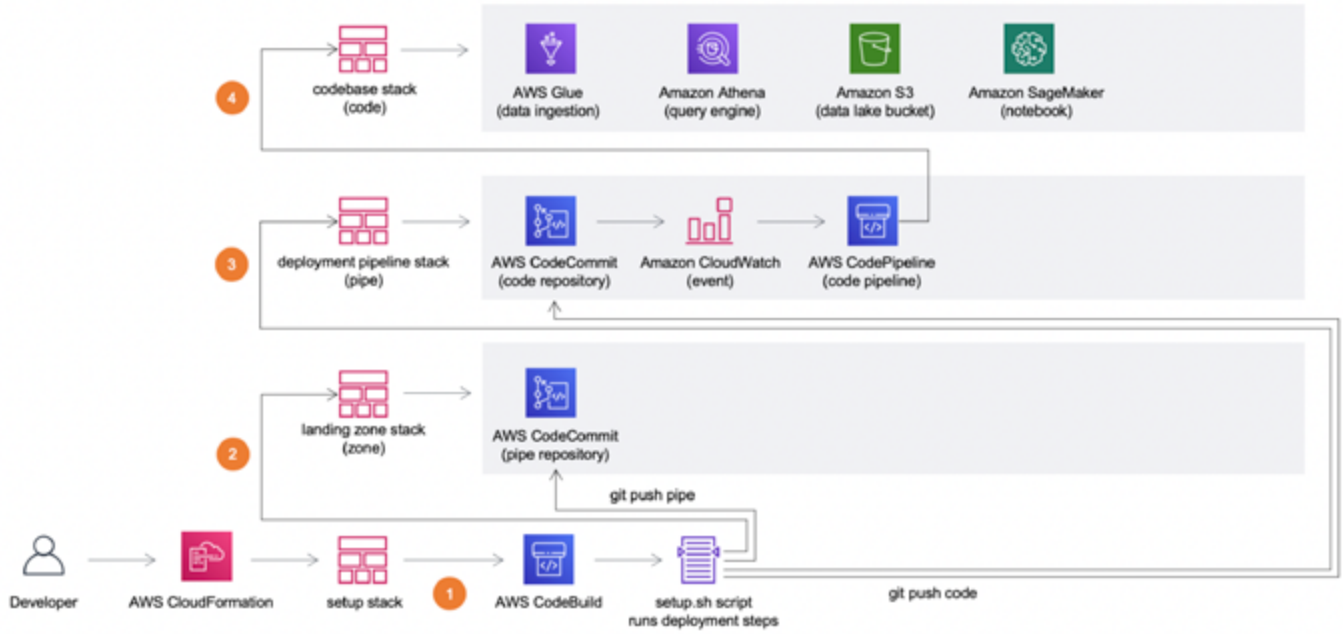
Data outputs from secondary analysis can be large and complex. For example, Variant Call Files (VCFs) must be converted to big data optimized file formats (like Parquet) and incorporated into existing genomics datasets. A data catalog must be updated with the appropriate schema and version so that users can find the data they need and operate within a defined data model that is semantically consistent. Annotation datasets and phenotypic data must be processed, cataloged, and ingested into an existing data lake in order to build a cohort, aggregate the data, and enrich the result set with data from annotation sources. Data governance and fine-grained data access controls are necessary to secure the data while still providing sufficient data access for research and informatics communities. The Genomics Tertiary Analysis and Data Lakes Using AWS Glue and Amazon Athena solution simplifies this process.

This solution provides a genomics data lake and sets up genomics and annotation ingestion pipelines using AWS Glue ETLs and crawlers to populate a genomics data lake in Amazon Simple Storage Service (Amazon S3). The solution demonstrates how to use Amazon Athena for data analysis and interpretation on top of a genomics data lake and creates a drug response report from within a Jupyter notebook.

AWS Solutions Implementation overview

The diagram below presents the architecture you can automatically deploy using the solution's implementation guide and accompanying AWS CloudFormation template.

Genomics Tertiary Analysis and Data Lakes Using AWS Glue and



Amazon Athena solution architecture

The AWS CloudFormation template creates four CloudFormation stacks in your AWS account including a *setup* stack to install the solution. The other stacks include a landing zone (*zone*) stack containing the common solution resources and artifacts, a deployment pipeline (*pipe*) stack defining the solution's CI/CD pipeline, and a codebase (*code*) stack providing the ETL scripts, jobs, crawlers, a data catalog, and notebook resources.

The *setup* stack creates an AWS CodeBuild project containing the *setup.sh* script. This script creates the remaining CloudFormation stacks and provides the source code for both the AWS CodeCommit *pipe* repository and the *code* repository.

Read online

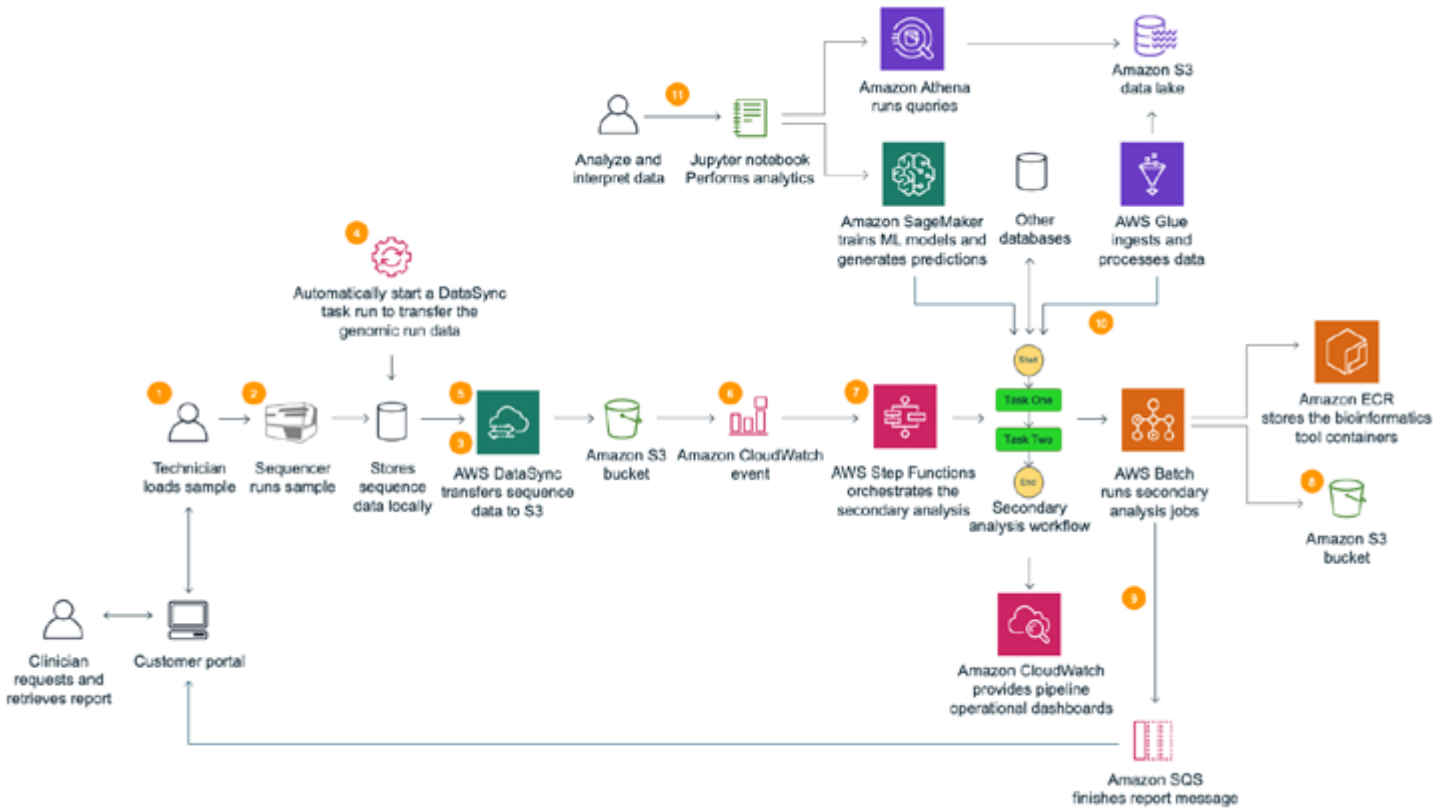
View implementation guide

View all genomics solutions here

Reference Architecture

Genomics report pipeline reference architecture

The following shows an example end-to-end genomics report pipeline architecture using the reference architectures described in this paper.



Genomics report pipeline reference architecture

1. A technician loads a genomic sample on a sequencer.
2. The genomic sample is sequenced and written to a landing folder that is stored in a local on-premises storage system.
3. An AWS DataSync sync task is preconfigured to sync the data from the parent directory of the landing folder on on-premises storage, to a bucket in Amazon S3.
4. A run completion tracker script running as a cron job, starts a DataSync task run to transfer the run data to an Amazon S3 bucket. An inclusion filter can be used when running a DataSync task run, to only include a given run folder. Exclusion filters can be used to exclude files from data transfer. In addition, consider Incorporating a zero-bite file as a flag when uploading the data. Technicians can then indicate when a run has passed a manual QA check by placing an empty file in the data folder. Then, the watcher application will only trigger a sync task if the success file is present.
5. DataSync transfers the data to Amazon S3.

6. An Amazon CloudWatch Events is raised that uses an Amazon CloudWatch rule to launch an AWS Step Functions state machine.
7. The state machine orchestrates secondary analysis and report generation tools which run in Docker containers using AWS Batch.
8. Amazon S3 is used to store intermediate files for the state machine execution jobs.
9. Optionally, the last tool in the state machine execution workflow uploads the report to the Laboratory Information Management System (LIMS).
10. An additional step is added to run an AWS Glue workflow to convert the VCF to Apache Parquet, write the Parquet files to a data lake bucket in Amazon S3 and update the AWS Glue Data Catalog.
11. A bioinformatic scientist works with the data in the Amazon S3 data lake using Amazon Athena via a Jupyter notebook, Amazon Athena console, AWS CLI, or an API. Jupyter notebooks can be launched from either Amazon SageMaker or AWS Glue. You can also use Amazon SageMaker to train machine learning models or do inference using data in your data lake.

View reference architecture online

Reference architecture included in this whitepaper

Blog

Broad Institute gnomAD data now accessible on the Registry of Open Data on AWS

by Erin Chu, DVM, Ph.D.

Co-authored by Grace Tiao, Associate Director of Computational Genomics at the Broad Institute and Erin Chu, DVM, Ph.D., Life Sciences Lead, AWS Open Data Program

Today we announce that data from the Genome Aggregation Consortium (gnomAD) is available for the first time on Amazon Web Services (AWS) as part of the Registry of Open Data on AWS. gnomAD is the world's largest public collection of human genetic variation and a near-ubiquitous resource for basic research and clinical variant interpretation. It is used in virtually all clinical genetic diagnostic pipelines worldwide, with over 20 million page views of the Broad Institute website to date.

The entire trove of gnomAD data, including data stretching back to the earliest release, is now accessible to AWS users at no cost via the AWS Open Data Sponsorship Program. AWS users will no longer need to pay transfer fees or long-term storage costs to access gnomAD data, or to maintain a personal copy of gnomAD data. By democratizing access to gnomAD data through this collaboration, the Broad Institute hopes to accelerate breakthrough genomic discoveries that enhance the scientific community's understanding of human genetics and result in solutions that improve the lives of people all over the world.

The mission of the AWS Open Data Sponsorship Program aligns closely with the Broad's commitment to make genomics tools available to the world. As the industry anticipates further exponential growth of human genomic datasets over the next few years, the Broad and AWS believe that the computational genomics community can benefit from free access to shared datasets. By reducing unnecessary duplication of terabyte- and petabyte-

scale genomic datasets, we as a community save scarce environmental, capital, and human resources that would otherwise be spent maintaining many copies across separate institutions. With this collaboration the Broad Institute hopes to provide an avenue for more individuals and organizations to participate in creative research in human genomics, with potential downstream benefits to us all.

What's included

- All official gnomAD release data, comprising summary statistics and annotations for over 241 million unique short human genetic variants and 335,000 structural variants observed in over 141,000 healthy adult individuals across a diverse range of genetic ancestry groups
- Standard "truth" sets used to assess and validate variant calls
- Interval lists and other resources used in the creation of gnomAD releases
- Data from the Broad's latest collection of papers in Nature

How to access it

To browse the bucket, download the AWS Command Line Interface and type:

```
$ aws s3 ls s3://gnomad-public-us-east-1/
```

If you don't yet have an AWS account set up, you'll need to add a `--no-sign-request` flag before `s3` to browse the bucket.

For a tutorial on using Hail to run computational pipelines on gnomAD data, see the Hail on AWS Quick Start.

Better together

Researchers are applying gnomAD data to diverse ends, such as:

- Annotating their own genomic sequencing data with gnomAD's high-quality allele frequency data;
- Using gnomAD's germline variant calls to separate driver and passenger mutations in cancer samples;
- Building statistical models to help determine the disease-causing potential of genetic variants;
- Comparing mutational tolerance of human genes to their orthologs in other species

View full blog online

Quick Start

Workflow orchestration for genomics analysis on AWS

nextflow

This Quick Start deploys a genomics analysis environment on the Amazon Web Services (AWS) Cloud, using Nextflow to create and orchestrate analysis workflows and AWS Batch to run the workflow processes.

Nextflow is an open-source workflow framework and domain-specific language (DSL) for Linux, developed by the Comparative Bioinformatics group at the Barcelona Centre for Genomic Regulation (CRG). The tool enables you to create complex, data-intensive workflow pipeline scripts, and simplifies the implementation and deployment of genomics analysis workflows in the cloud.

This Quick Start is for teams or individuals who manage informatics infrastructure and genomics analysis for a biotech company.

The Quick Start deploys Nextflow into the infrastructure set up by the Biotech Blueprint core Quick Start. If you want to use an existing virtual private cloud (VPC) or create a new VPC, follow the Genomics Workflows on AWS instructions instead. If you're new to AWS or don't have a strong VPC architecture already, we recommend that you first use the Biotech Blueprint core Quick Start to set up the landing zone for future AWS usage. This environment is automatically configured for identity management, access control, encryption key management, network configuration, logging, alarms, partitioned environments, and built-in compliance auditing to help meet your security and compliance requirements.

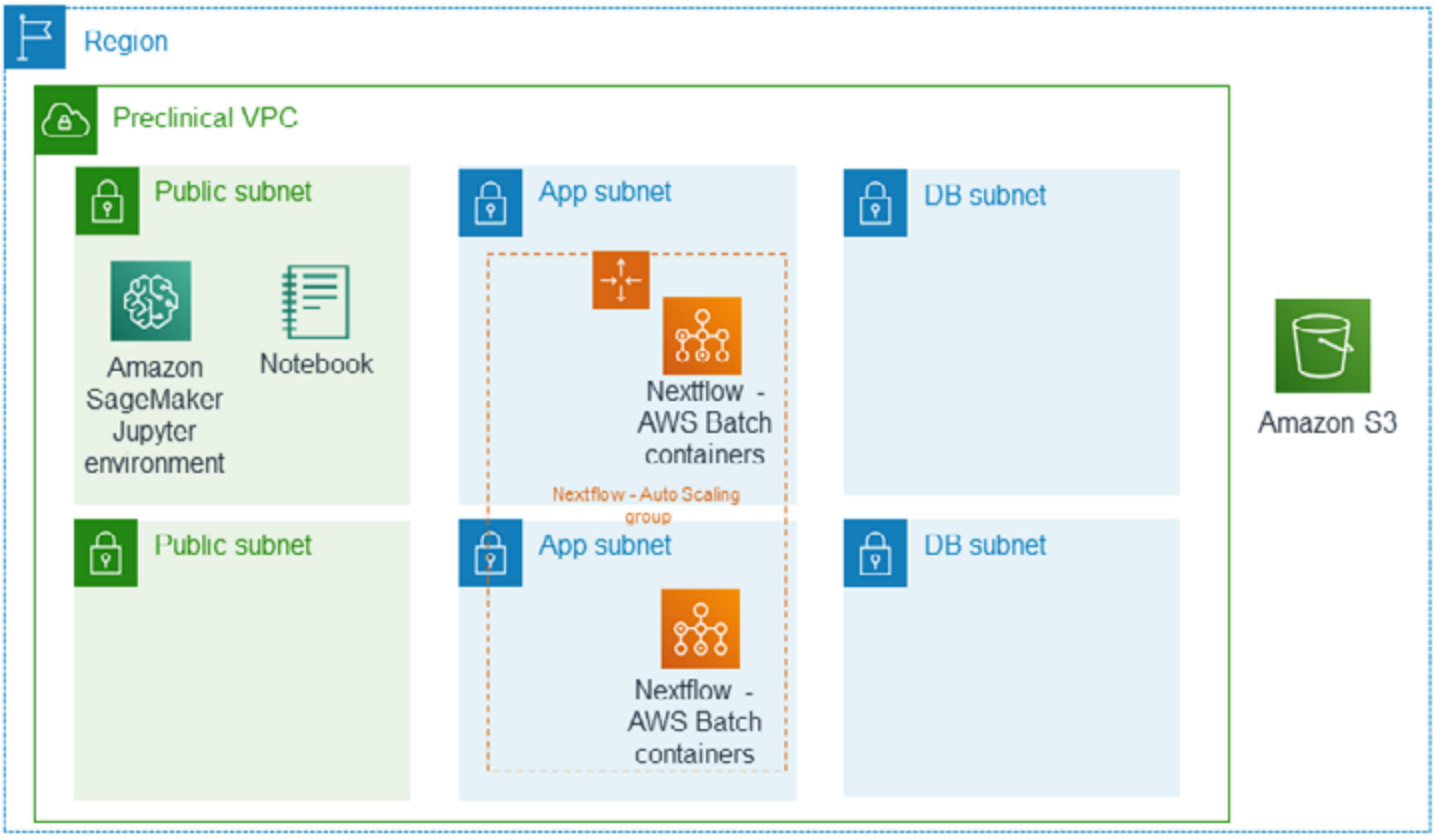
This Quick Start sets up the following environment in a preclinical VPC:

- In the public subnet, an optional Jupyter notebook in Amazon SageMaker that is integrated with an AWS Batch environment.
- In the private application subnets, an AWS Batch compute environment for managing Nextflow job definitions and queues, and for running Nextflow jobs. AWS Batch containers have Nextflow installed and configured, in an Auto Scaling group.
- Because there are no databases required for Nextflow, this Quick Start does not deploy anything into the private database (DB) subnets created by the Biotech Blueprint core Quick Start.
- An Amazon Simple Storage Service (Amazon S3) bucket to store your Nextflow workflow scripts, input and output files, and working directory.

For more information about the preclinical VPC and other infrastructure components, see the Biotech Blueprint core Quick Start.

View full quickstart online

View deployment guide



Solutions

Genomics Tertiary Analysis and Machine Learning Using Amazon SageMaker

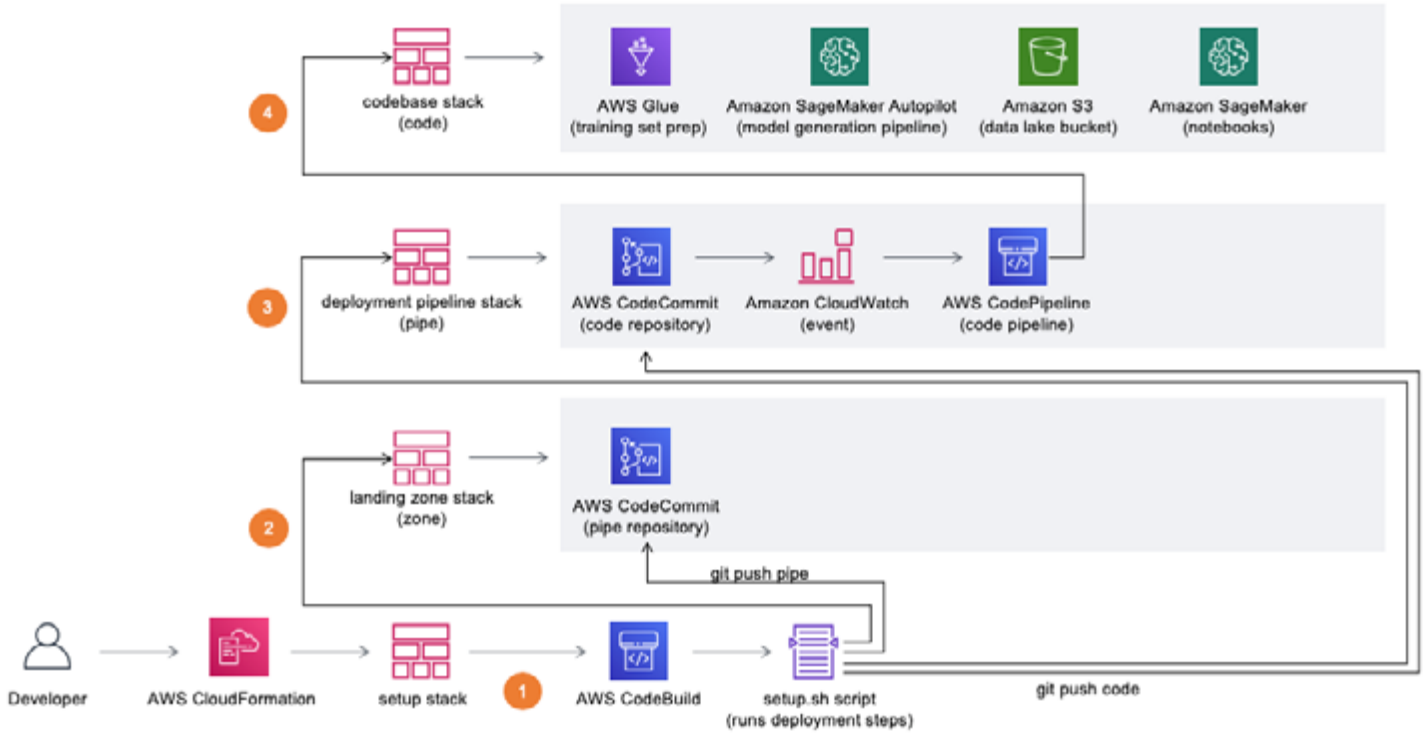
What does this AWS Solutions Implementation do?

The Genomics Tertiary Analysis and Machine Learning Using Amazon SageMaker solution creates a platform in the AWS Cloud that can be used to build machine learning models on genomic datasets using AWS managed services. We define *tertiary analysis* to be the interpretation of genomic variants and assigning meaning to them. This solution provides a broad platform for genomic machine learning in AWS, using variant classification as an example of a scientifically meaningful problem that can be solved using this platform. In the example, we solve the specific challenge of competing clinical definitions when examining genomic variants. Our example is based on the following Kaggle challenge. We create a model to predict if a variant annotated in ClinVar has a conflicting classification or not. A model that can predict the existence of a conflicting classification for a variant can save valuable time that researchers have to spend looking for such conflicts.

This solution demonstrates how to 1) automate the preparation of a genomics machine learning training dataset, 2) develop genomics machine learning model training and deployment pipelines and, 3) generate predictions and evaluate model performance using test data. These steps can be repeated or edited by users for their specific use cases.

AWS Solutions Implementation overview

The diagram below presents the architecture you can automatically deploy using the solution's implementation guide and accompanying AWS CloudFormation template.



Genomics Tertiary Analysis and Machine Learning Using Amazon SageMaker solution architecture

The AWS CloudFormation template creates four CloudFormation stacks in your AWS account including a *setup* stack to install the solution. The other stacks include a landing zone (*zone*) stack containing the common solution resources and artifacts; a deployment pipeline (*pipe*) stack defining the solution's continuous integration and continuous delivery (CI/CD) pipeline; and a code base (*code*) stack providing the ETL scripts, jobs, crawlers, a data catalog, and notebook resources.

View full Quick Start online

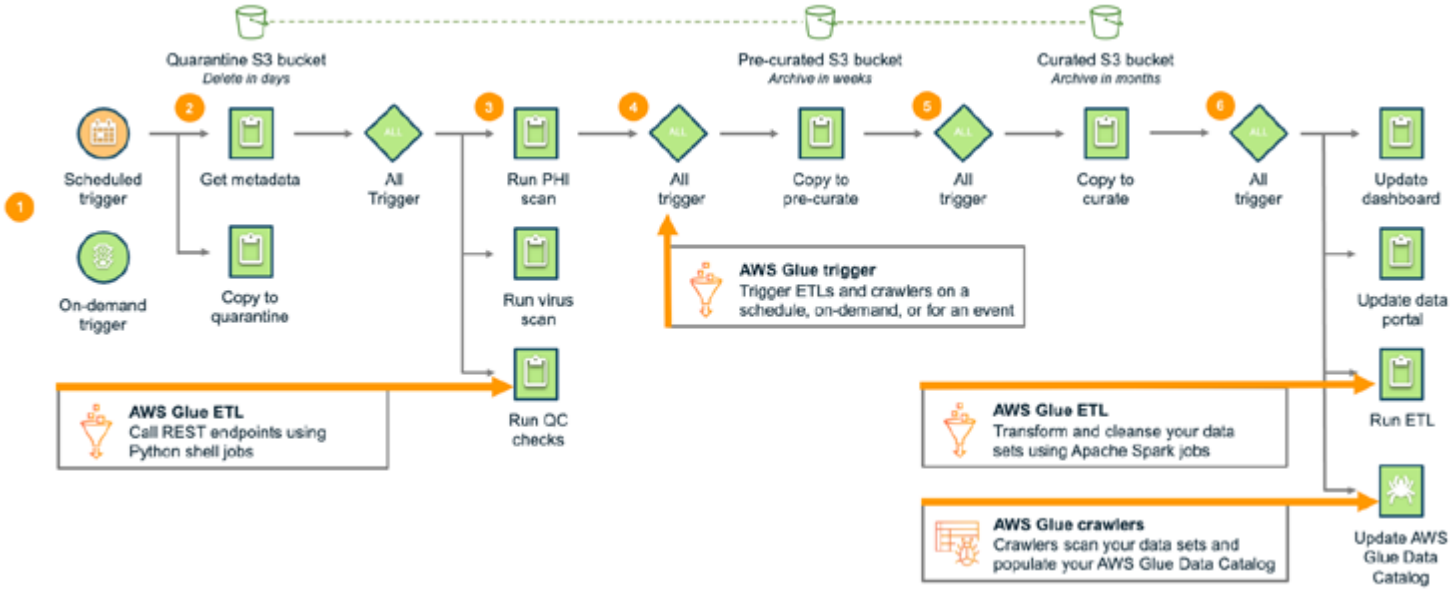
View implementation guide

View all genomics solutions here

Reference Architecture

Research data lake ingestion pipeline reference architecture

The following reference architecture shows an example end-to-end research data lake data ingestion AWS Glue pipeline using the data lake reference architectures described in this paper. The AWS Glue workflows enable you to construct data pipelines using extract, transform, and load (ETL) functions, crawlers, and triggers.



Data pipeline using AWS Glue workflows

1. An AWS Glue trigger is run either on-demand or on a schedule.
2. The dataset is first copied to a quarantine bucket to be scanned for personal health information (PHI) or viruses.
3. A trigger then launches Glue Python shell jobs that make REST calls to Personal Health Information (PHI) and virus scanning services that scan the dataset which resides in the Amazon S3 quarantine bucket. A quality control (QC) process is run to confirm that the data is in the agreed upon format and schema.
4. If the dataset passes the scans and QC validation, the data is copied to a pre-curated bucket where the dataset resides without changes.
5. The dataset is then copied to a curated bucket where it is reorganized and filtered based on the study or research project.
6. A trigger then launches jobs to update a research project dashboard, a research portal database, extract transform, and load (ETL) processes to transform the data and write it to the data lake in Apache Parquet format. AWS Glue crawlers crawl the data, infer the schema, and update the AWS Glue meta data catalog. The data is made available for query using big data query engines such as Amazon Athena. Data governance is managed with Identity and Access Management (IAM) or with AWS Lake Formation.

View reference architecture online

Reference architecture included in this whitepaper

Videos:

The Smithsonian Institution Improves Genome Annotation for Biodiverse Species Using the AWS Cloud

The Smithsonian's Data Science Team's mission is to implement solutions that will accelerate science and lower the bar for entry to genomics research, not only for Smithsonian scientists but for biodiversity researchers in general. They are working to improve a critical part of the genome analysis pipeline – annotation. By using the AWS Cloud for annotation, different parts of a genome assembly can be annotated in parallel, with the results being knitted together in a final step. The ability to scale up to many instances for brief periods will make annotation fast while remaining inexpensive. The cloud has enabled the Smithsonian Institution to share their research and increase knowledge through open data science.

AstraZeneca Genomics on AWS: A Journey from Petabytes to New Medicines

AstraZeneca is on a mission to analyze 2 million genomes/exomes by 2026 for integration with clinical data and use in R&D, clinical trials, and stratified medicine. To achieve this milestone, AstraZeneca built a world-leading genomics pipeline using high performance computing technologies such as AWS Batch and AWS Step Functions that is capable of processing more than 1,600 exomes per hour.

To learn more about how the BioPharma industry builds on AWS, visit: http://amzn.to/3ppOPRz

Accelerate Genomic Discoveries on AWS

With AWS, genomics customers can dedicate more time and resources to science, speeding time to insights, achieving breakthrough research faster, and bringing lifesaving products to market.

Learn more about Genomics in the Cloud at: http://amzn.to/3t2QQpe

UC Berkeley AMP Lab Genomics Project on AWS - Customer Success Story

The AMP Lab at the University of California Berkeley builds scalable machine learning and data analysis technologies to turn data into information. Among the many experiments run by the AMP Lab, one area of concentration is in the field of genomics and cancer research. Due to the vast amount of data that genome sequencing produces, the AMP Lab leverages Amazon Web Services (AWS) to quickly scale the compute resources needed to analyze the algorithms that are used in genomics work. As a result, researchers are able to use many machines in the cloud simultaneously to process genome data faster and more cost effectively. To learn more about how AWS helps universities, researchers and analytics teams solve the big data problem, visit http://aws.amazon.com/publicsector

Helix Uses Illumina's BaseSpace Sequence Hub on AWS to Build Their Personal Genomics Platform

Helix is a personal genomics platform company that is trying to lower the barrier for anyone who wants to develop software that utilizes genomics information. Learn how Helix uses BaseSpace Sequence Hub, an integrated software platform designed for genomic data analysis, from APN Partner Illumina to power its platform.

Visit https://amzn.to/2CPUxVO to learn more about what AWS is doing in Genomics

University of Sydney Accelerate Genomics Research with AWS and Ronin

In partnership with Ronin, the University of Sydney's Wildlife Genomics Group scales on demand to reduce genomics data analysis time from 10 days to six hours with Amazon Web Services (AWS). The group studies non-model species to better understand how they can protect wildlife animal populations in Australia.