

AWS Genomic Workflow Automation Solutions

September 27, 2021

© 2021, Amazon Web Services, Inc. or its Affiliates.



About your speakers



Lee Pang

Principal Bioinformatics Architect, Health AI, Amazon Web Services



Ryan Ulaszek

Worldwide Tech Lead for Genomics, Amazon Web Services







The Genomics challenge

Why AWS for Genomics

AWS Genomics Workflow Automation Solutions

Customer case studies

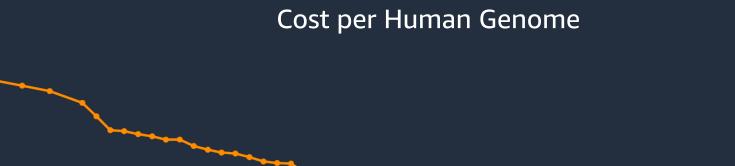
Summary & next steps

© 2021, Amazon Web Services, Inc. or its Affiliates.



The Genomics challenge





2010

2008

20'12

2014

2016

https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data

2004

2006

© 2021, Amazon Web Services, Inc. or its Affiliates.

\$100,000,000

\$10,000,000

\$1,000,000

\$100,000

\$10,000

\$1,000[.]

2002

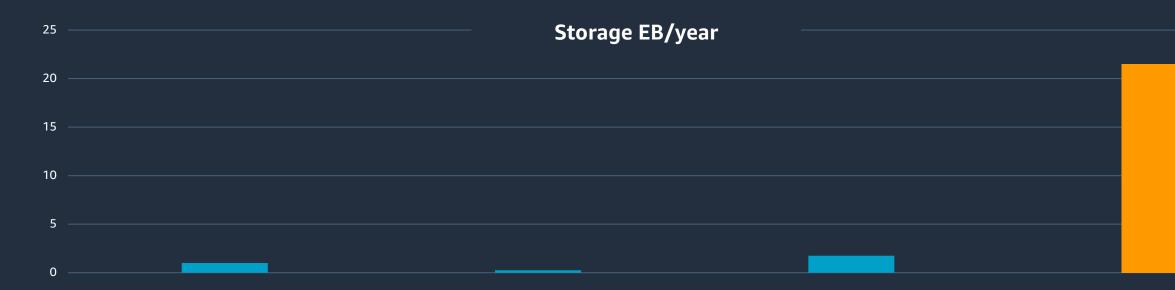


2018

2020

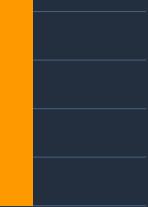


Genomic footprints require scalable storage and compute



Data phase	Astronomy	Twitter	YouTube	Gen
Acquisition	25 zetta/bytes/year	0.5/15 billion tweets/year	500–900 million hours/year	1 zet
Storage	1EB/year	1–17 PB/year	1–2 EB/year	2–40
Analysis	In situ data reduction	Topic and sentiment mining	Limited requirements	Hete
	Real-time processing	Metadata analysis		Varia proce
	Massive volumes			All-pa ~10,0
Distribution	Dedicated lines from antennae to server (600 TB/s)	Small units of distribution	Major component of modern user's bandwidth (10 MB/s)	Many mass
		Source: Stephens, et al., Big Data: Astronomical or Genomical? (2015)		

© 2021, Amazon Web Services, Inc. or its Affiliates.



enomics

- etta-bases/year
- 40 EB/year
- terogeneous data and analysis
- riant calling, ~2 trillion central ocessing unit (CPU) hours
- -pairs genome alignments, 0,000 trillion CPU hours
- ny small (10 MB/s) and fewer ssive (10 TB/s) data movement



Key considerations for Genomics workloads



Data size, gravity, and diversity

- Multiple Genomic data types (WGS, WES, Targeted)
- Genomics data range from 10s of GBs to 100s of GBs per sample
- Growing rapidly due to continuously improving sequencing technologies
- Cost needs to be front and center •



Scalable compute

- Many steps linked for Genomic data analysis and processing
- Varied computational needs depending upon Genomic data type and tool
- Portability and reproducibility





Why AWS for Genomics

aws genomics

AWS core cloud capabilities facilitating Genomics





Managed services



Amazon Elastic Container Service



AWS Step Functions



AWS is global



• Over 1 million active customers across 190 countries

• 2,000+ government agencies

- 5,000+ educational institutions
- 25 regions (+3 Planned)
- 81 availability zones with 3+ data centers per zone

Customer benefits of the AWS Global Infrastructure





Flexibility



Computing as a utility

Focus on applications and not infrastructure

Pay as you go, and only for what you use

On-demand and fit for purpose



Compliance on AWS





HIPAA compliance does not equate to GxP compliance, or non-US data privacy laws

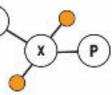




© 2021, Amazon Web Services, Inc. or its Affiliates.

Alignments/Frameworks (industry/function)





HITRUST **CSF** Certified



AWS powers Genomics organizations of all sizes and disciplines

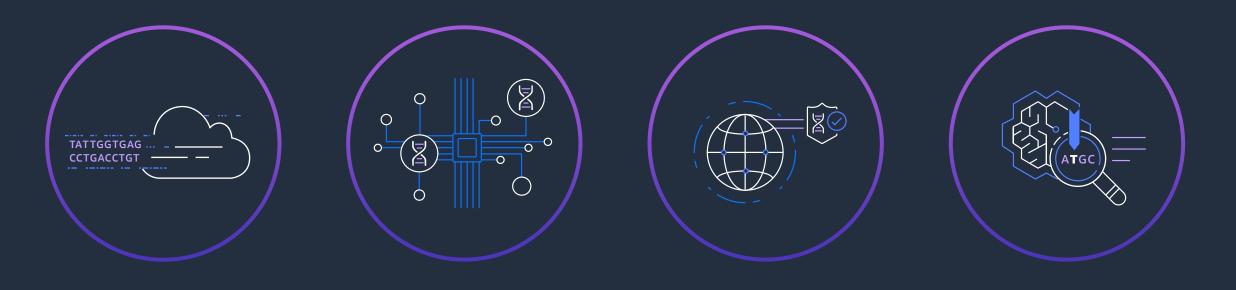






AWS for Genomics solution areas

AWS provides solutions and tools across the Genomics workflow



Data transfer & storage Workflow automation and secondary analysis

Data aggregation & governance Interpretation & ML for tertiary analysis

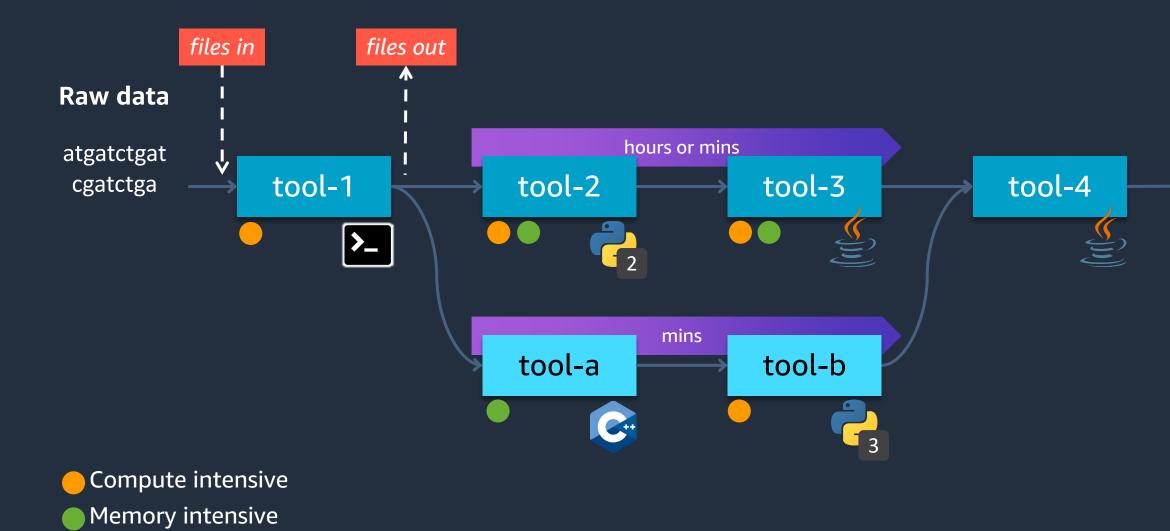


Clinical translation

AWS Genomic Workflow Automation Solutions



Workflow pipelines in a nutshell

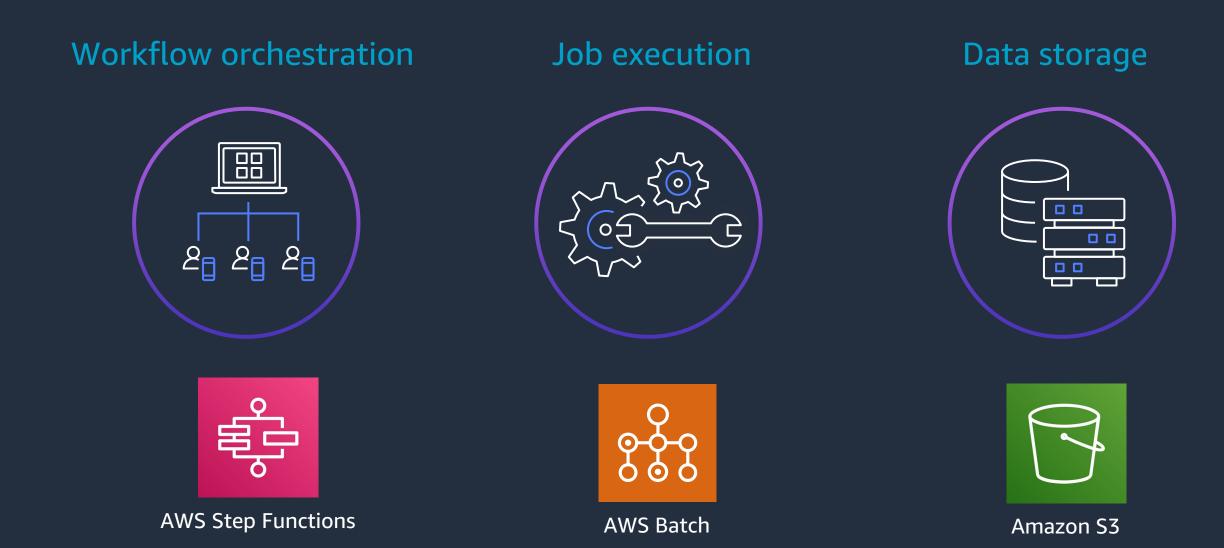


Processed data

0100100101 01001010

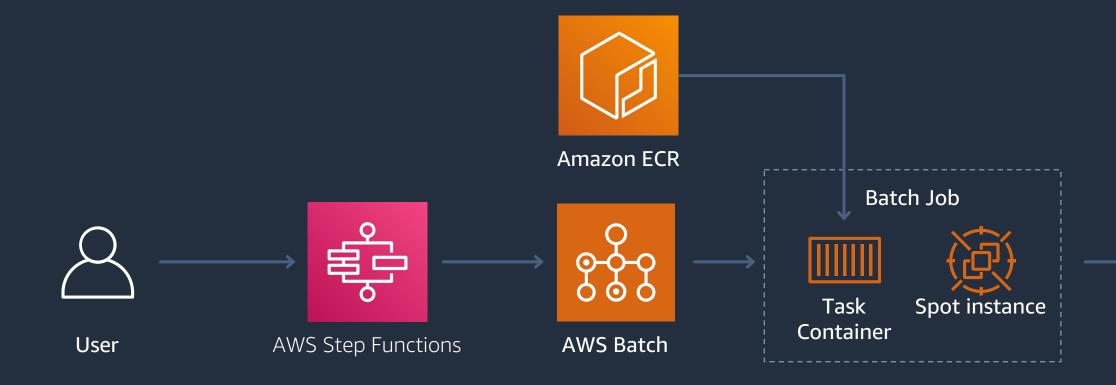


Major infrastructure components





AWS native reference architecture







Putting it all together



Workflow definition

Workflow orchestrator

Develop a workflow using a definition language and containerized tools

Submit your workflow to a workflow engine

Job execution

Batch Job

Spot

Task

Workflow engine submits tasks to cloud compute resources (e.g., AWS Batch)



Amazon S3





Amazon EFS

Amazon FSx for Lustre

Data storage

Tasks retrieve and store data in cloud object storage (e.g., Amazon S3)



Amazon Genomics CLI



Amazon Genomics CLI is an open source command line interface (CLI) that helps customers new to AWS run Genomics workflows in the cloud by automating deployment of best practices infrastructure for workflow engines. Amazon Genomics CLI reduces the time for scientists and developers to start running existing Genomics workflows at scale and speed up iteration cycles as they develop new ones.



Setup a new project and run a Genomics secondary analysis workflow in the cloud with a few CLI commands



Open source and built on community open standards



Amazon Genomics CLI

Start running genomics workflows on AWS with a few easy steps and familiar tooling



compute resources and workflows

and container clusters to execute workflow engines



Process genomic data and derive research insights

•••	Amazon Genomics CLI
🗞 Launch and	manage genomics workloads on AWS.
Commands	
Getting Start	ed 🥂
account	Commands for AWS account setup. Install or remove AGC from your account.
Contexts	
context	Commands for contexts. Contexts specify workflow engines and computational
Logs	
logs	Commands for various logs (currently only CloudWate
Projects	
project	Commands to interact with projects.
Workflows	
workflow	Commands for workflows. Workflows are potentially-dynamic graphs of computa
Settings \otimes	
version	Print the version number.
Flags	
-h,help	help for agc
-v,verbose	display verbose diagnostic information
version Examples	version for agc
	nelp menu for the specified sub-command.
`\$ agc account	
→ ~	

fleets to use when running a workflow.

ch).

ational tasks to execute.

Amazon Genomics CLI

Start running genomics workflows on AWS with a few easy steps and familiar tooling



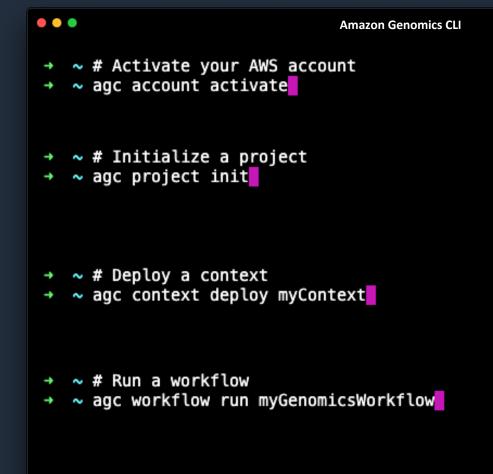
workflows



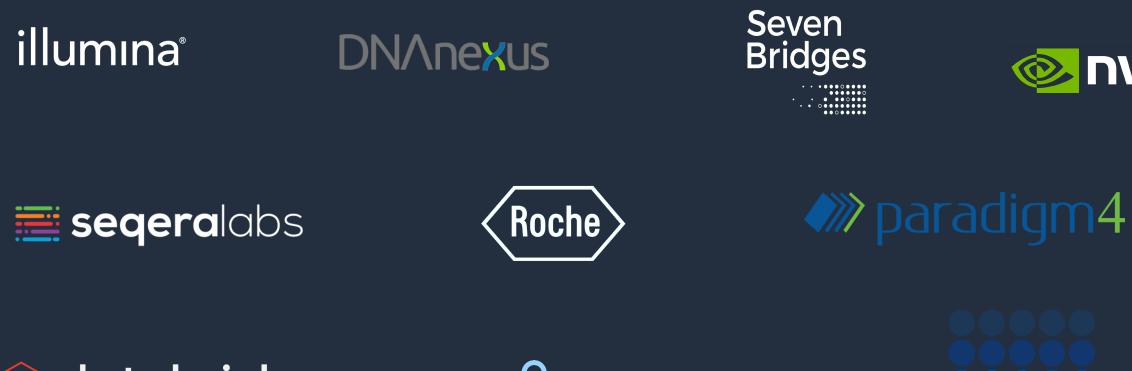
Deploy compute resources and container clusters to execute workflow engines



Process genomic data and derive research insights



Industry-leading ISVs build on and collaborate with AWS





% lifebit

QIAGEN



NVIDIA



Accelerated Genomics workflows (CPU, GPU, FPGA)

Products integrated with AWS platform and easy to test

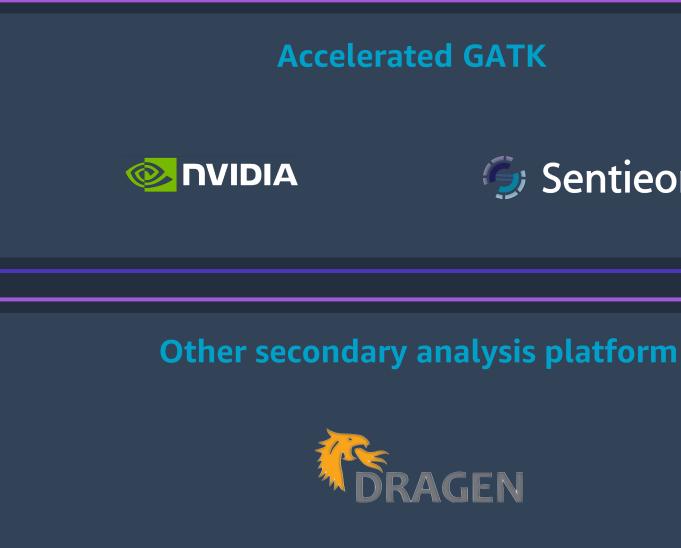


310,000+ AWS active customers are using software from the AWS Marketplace

50+ categories and more than 10,000+ software listings from 1,600+ ISVs

2M+ active SaaS subscriptions

650M hours of usage a month of Amazon EC2 for AWS Marketplace products





Sentieon



Demos



С

Customer case studies



Bayer Crop Sciences Enables scalable Genomics analysis workflow on AWS

Challenge

Bayer crop science wanted to reduce costs and increase modularity for processing and interpreting Genomic analysis data, in order to understand the feasibility of the new "skim" sequencing pipeline.

Solution

The company worked with AWS Professional Services to build a scalable, and modular Genomic analysis pipeline using AWS step functions, with the aim of reducing the costs and turnaround times.

Benefits

- Lowered barriers to entry and expansion for customers
- Reduced costs and turn around times for analysis

Company: Bayor Crop Science Industry: Life Sciences **Country: Germany** Website: www.cropscience.bayer.com/en

What really impressed me was the extraordinary working relationship among all the AWS team members and Bayer scientists involved in the project. We have had many collaborations with external parties, and this one with AWS was truly a partnership.

Tom Osborn, Head of Analytics and Pipeline Design, Crop Science



Bayer CropScience

About Bayer Crop Science

Bayer's Crop Science division is part of Bayer AG, and is the third largest innovative agricultural input company in the world and has businesses in high-value seeds, crop protection and non-agricultural pest control.



Fred Hutch microbiome researchers use AWS to perform seven years of compute time in seven days

Challenge

Fred Hutch is engaged in analysis of the microbiome. 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.

Solution

To accelerate its research, the team uses the Nextflow framework to orchestrate AWS Batch processes and scale the high-performance computing platform to accelerate processing time—reducing 7 years of compute time to 7 days.

Benefits

- Processes data from more than 15,000 biological samples
- Reduced 7 years of compute time to 7 days
- Increases resolution on microbiome samples to find links to improve health outcomes

Company: Fred Hutch Industry: Life Sciences **Country: United States** Employees: 3,500 Website: https://research.fredhutch.org/

AWS Batch integrates well with Nextflow, so it was easy for us to get Nextflow up and running without having to reinvent the wheel.

Sam Minot, PhD and staff scientist at Fred Hutch MRI



About Fred Hutch

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.



Lifebit: Powering Genomics England's Research **Environment & the UK's COVID-19 Research**

Challenge

COVID-19 has brought the clinical application of genomics to the forefront. Through a partnership with Lifebit and AWS, Genomics England (GEL) launched a large-scale genomics research project that aims to leverage data from 35,000 COVID-19 patients and 100,000 participants from the organization's historical cohort.

Solution

Lifebit's end-to-end data platform will allow global biopharma and academic researchers to query, analyze, augment, and collaborate over these large datasets in seconds so they can accelerate drug and vaccine discoveries.

Benefits

- Organizations bring their private data for joint analyses, and the platform scales to accommodate data from 1+ million individuals, 3+ billion genetic variants, and 1+ million phenotypic and clinical annotations.
- 10X increase in relevant scientific findings, massively improving COVID-19 diagnosis and prevention



We've essentially taken Genomics England, the pioneers of population genomics, and have turned them into the world's most cutting-edge research environment.

Thorben Seeger, VP Commercial

~ lifebit

Company: Lifebit Industry: Life Sciences **Country: United Kingdom** Website: https://lifebit.ai/

About Lifebit

Lifebit is democratising the analysis & understanding of genetic big data to leapforward cures, disease prevention, and our quality and understanding of life.



Summary



How AWS enables scalable genomics workloads



Scalable and secure

Reduce costs and improve turnaround time for genomic analysis



Best fit flexibility Start building with AWS reference architectures, Amazon Genomics CLI, AWS Partner offerings



Infrastructure as code

Maximize results by minimizing operational overhead associated with infrastructure



Accelerate experimentation

Bioinformaticists and Data Scientists modernize and accelerate Genomic research and analysis





Next steps & resources



Resources

AWS for Health aws.amazon.com/health

Genomics in the Cloud aws.amazon.com/health/genomics

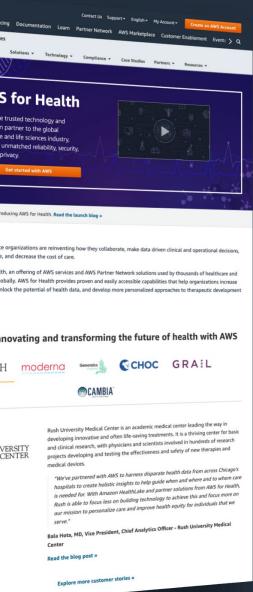
Genomic Solutions aws.amazon.com/health/genomics/solutions

Guide to Genomics Workflows on AWS docs.opendata.aws/genomics-workflows

AWS Marketplace aws.amazon.com/marketplace

AWS Partner Network aws.amazon.com/partners/find

aws		aws		aws
Documentation Learn Partner M	Us Support + English +	Products Solutions Pricing Documen	Contact Us Support & English * tation Learn Partner Network AWS Marketpl	Products Solutions Pricing
Duendari	etwork AWS Marketpla	Overview		Healthcare & Life Sciences Overview Segments •
Solutions • Technology • Complia	ance - Case Studies	Stativeew Segments + Solutions +	Technology + Compliance + Case Studies	
Genomics in the Cloud		Genomics Solutions		AWS
Accelerate genomics discoveries		AWS and AWS Partners offer workflow, designed to help tr scientific breakthroughs.	purpose-built solutions across the ge anslate samples into actionable insig	AWS is the tru innovation pa healthcare an providing unr
With AWS, genomics customers can dedicate more time and resour speeding time to insights, achieving breakthrough research faster, a lifesaving products to market.	rces to science, and bringing	Explore solution areas 😼		and data priv
AWS enables customers to innovate by making genomics data mon useful. AWS delivers the breadth and depth of services to reduce th sequencing and interpretation, with secure and frictioniess collabo across multi-modal datasets. Plus, you can choose the right tool for best not near document of the secure and secure the secure of the secur	he time between ration capabilities	WHAT'S NEW (Service) See the new	updates for Amazon HealthLake. Learn more »	WHAT'S NEW Introdue
best cost and performance at a global scale— accelerating the moo genomics.	dern study of	Leading organizations are accelerating genomi		Healthcare and life science or enable precision medicine, an
100 2x 24 Gbps with multiple Regions, to	Launched potential	illumına AstraZı	eneca DNAnexus Aance	Introducing AWS for Health, a life sciences customers global the pace of innovation, unlock and care.
		illumına	Illumina delivers integrated systems for ana function. Illumina builds on AWS to gain mo costs, enabling the organization to pass alon customers.	Customers inno
Benefits			"Genomics — next-generation sequencing for personalized medicine. The ability to I genomes in individuals and populations o testing, rare disease diagnoses, advanced	€RUSH
Accelerate time to discovery Keep costs low and Powerful compute and machine learning			last decade, we've collaborated with AWS services to researchers around the globe t	
options ensure scientists can execute workloads The flexible pricing and o fast and with control. AWS offers the broadest celection of compute services—more than any complex genomics project	researchers to tackle cts, without having to		Health initiative supports our mission of i power of the genome." Watch video »	
other cloud provider—and only AWS offers compute instances that deliver 100 Gbps of networking throughput. provides pay-as-you-go pricing and virtually unlimited compute capacity.			Explore more customer stories »	RUSH UNIVE MEDICAL CE
		Solution Areas		
AWS Partners in Genomics	pice, and options for		1 - 8 COV	
AWS offers the largest network of Partners, along with flexible workflow cho fully managed solutions to help you get to genomics insights faster.		Data Transfer & Storage Secondary Analysis Data Aggregation & Gou		
Learn more about AWS Partners for Genomics a		Clinical Genomics		
		Genomics Data Transfer & Storage		



Resources continued

Speak to a team member: https://pages.awscloud.com/ GenomicsContactSales.html

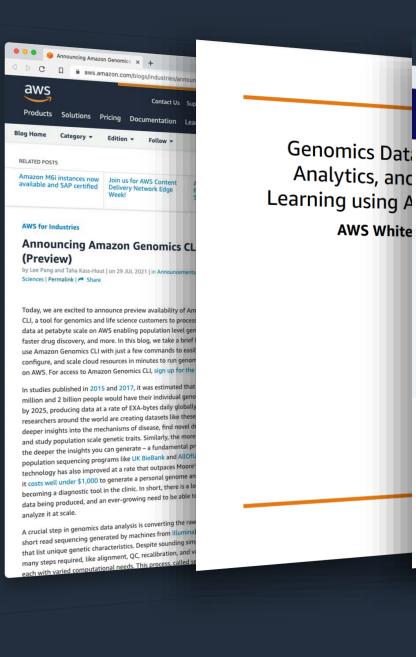


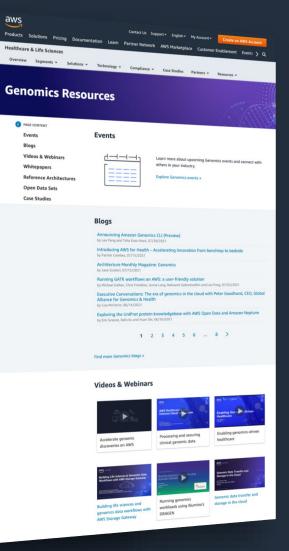
Announcing Amazon Genomics CLI: https://aws.amazon.com/blogs/ industries/announcing-amazongenomics-cli-preview/



View Genomics Resources: https://aws.amazon.com/health/ genomics-resources/







Thank you!



Questions and answers

