aws genomics

AWS Genomic Workflow Automation Solutions

September 27, 2021

About your speakers

Lee Pang

Principal Bioinformatics Architect, Health AI,
Amazon Web Services

Ryan Ulaszek

Worldwide Tech Lead for Genomics,
Amazon Web Services

aws genomics

Agenda

The Genomics challenge

Why AWS for Genomics

AWS Genomics Workflow Automation Solutions

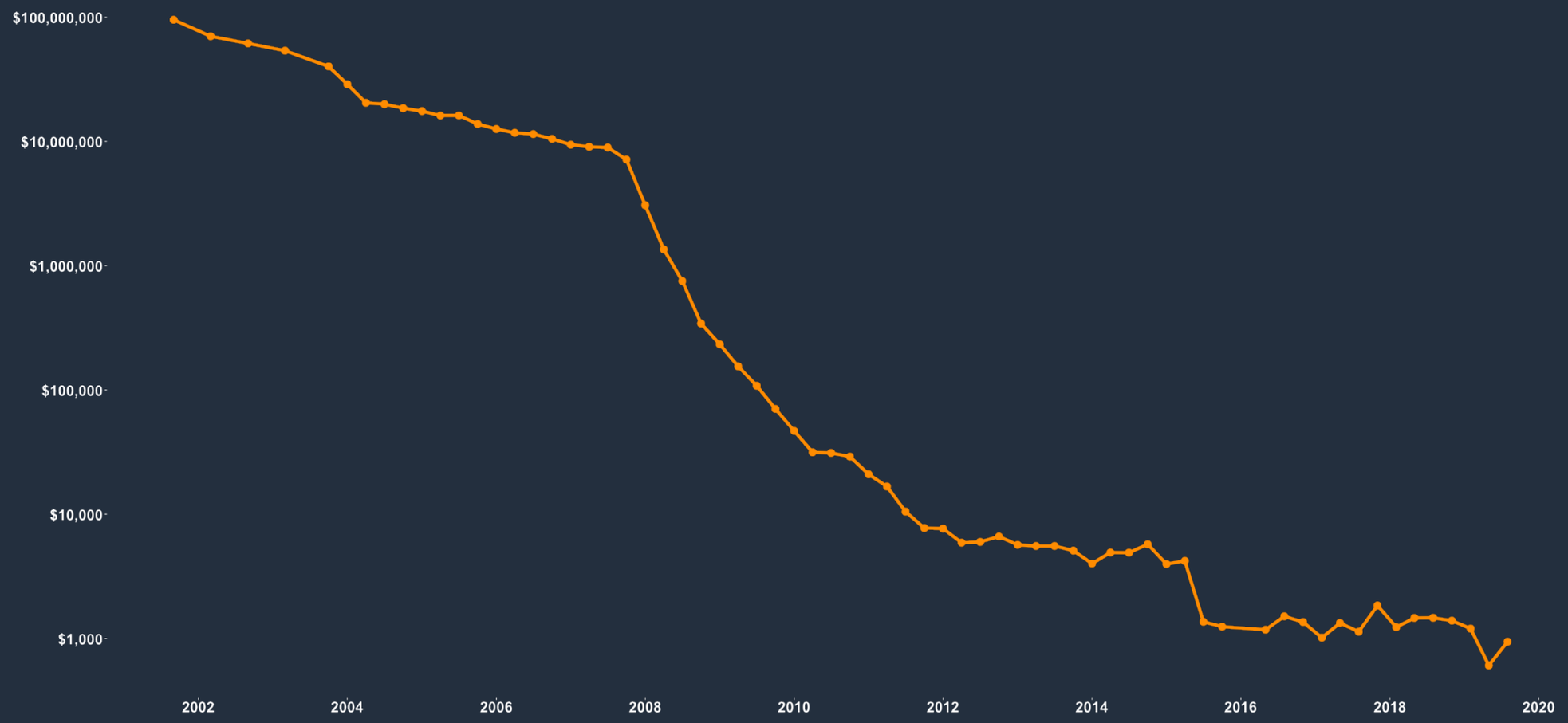
Customer case studies

Summary & next steps

aws genomics

The Genomics challenge

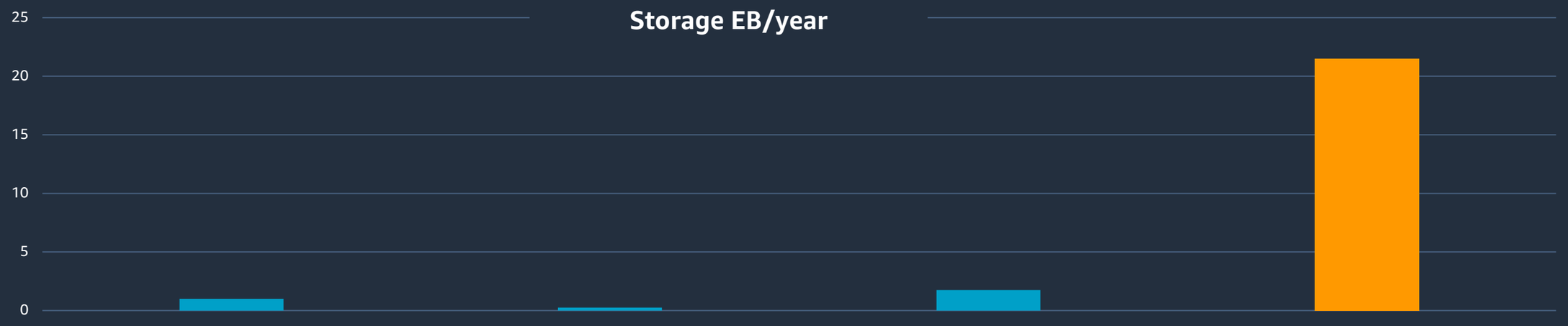
Cost per Human Genome



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

aws genomics

Genomic footprints require scalable storage and compute



Data phase	Astronomy	Twitter	YouTube	Genomics
Acquisition	25 zetta/bytes/year	0.5/15 billion tweets/year	500–900 million hours/year	1 zetta-bases/year
Storage	1EB/year	1–17 PB/year	1–2 EB/year	2–40 EB/year
Analysis	In situ data reduction Real-time processing Massive volumes	Topic and sentiment mining Metadata analysis	Limited requirements	Heterogeneous data and analysis Variant calling, ~2 trillion central processing unit (CPU) hours All-pairs genome alignments, ~10,000 trillion CPU hours
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 small (10 MB/s) and fewer massive (10 TB/s) data movement

Source: Stephens, et al., Big Data: Astronomical or Genomical? (2015)

aws genomics

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

aws genomics

Why AWS for Genomics

aws genomics

AWS core cloud capabilities facilitating Genomics

Compute

Amazon EC2

Storage

Amazon Elastic Block Store (EBS)

Amazon S3

FSX for Lustre

Identity

AWS Identity and Access Management (IAM)

Managed services

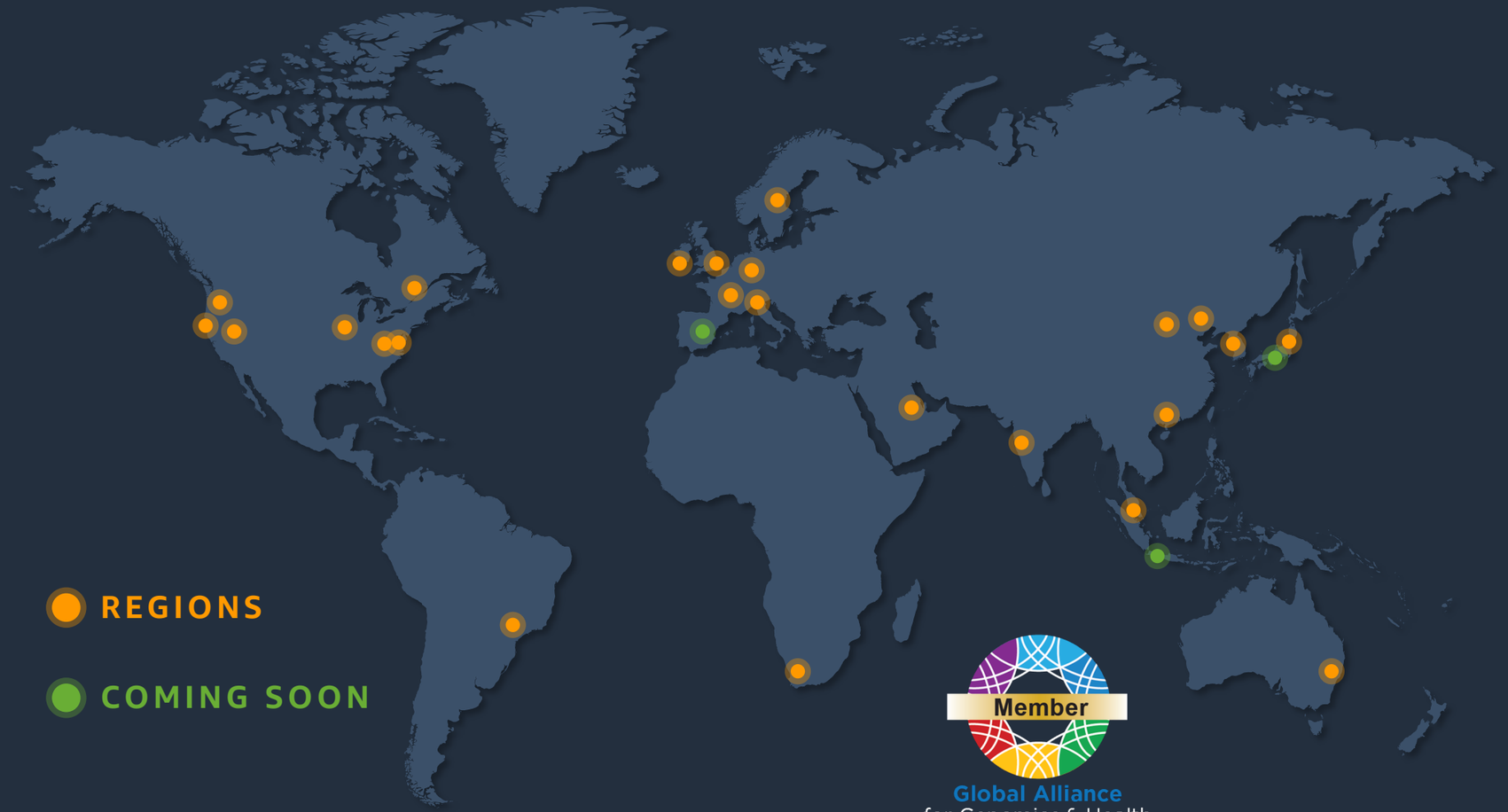
AWS Batch

Amazon Elastic Container Service

AWS Step Functions

aws genomics

AWS is global



Member
Global Alliance
for Genomics & Health

- 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
- 230+ POPs
- 245 countries and territories served

Customer benefits of the AWS Global Infrastructure

aws genomics

Computing as a utility

aws genomics

Compliance on AWS

Certifications/ Attestations

ISO 9001 International Organization for Standardization

ISO 27001 International Organization for Standardization

ISO 27017 International Organization for Standardization

ISO 27018 International Organization for Standardization

AICPA SOC aicpa.org/soc4so SOC for Service Organizations Service Organizations

Laws/Regulations/ Privacy

HIPAA

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

Alignments/Frameworks (industry/function)

NIST

G X P

HITRUST CSF Certified

aws genomics

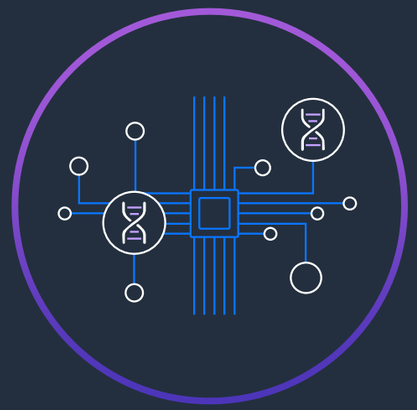
AWS powers Genomics organizations of all sizes and disciplines

aws genomics

AWS for Genomics solution areas

AWS provides solutions and tools across the Genomics workflow

Data transfer
& storage



Workflow automation
and secondary analysis

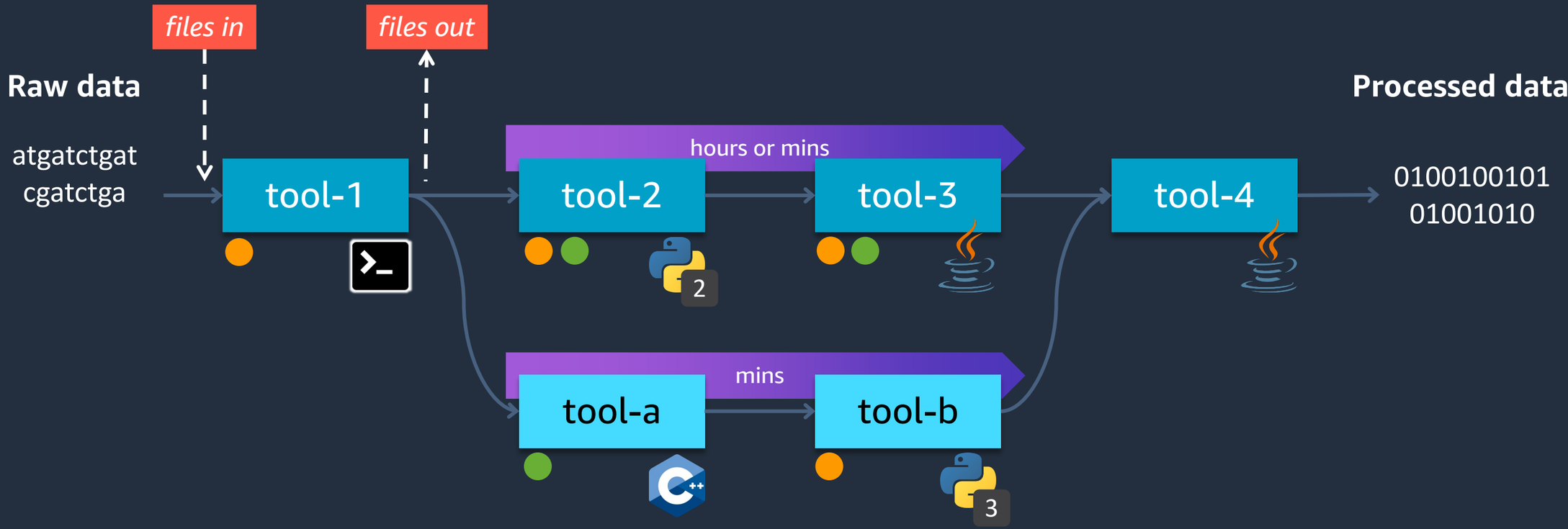
Data aggregation
& governance

Interpretation & ML
for tertiary analysis

Clinical
translation

AWS Genomic Workflow Automation Solutions

Workflow pipelines in a nutshell



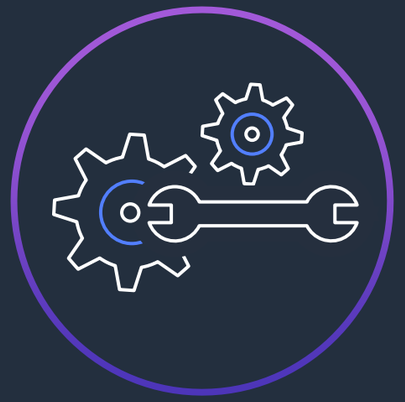
aws genomics

Major infrastructure components

Workflow orchestration

AWS Step Functions

Job execution



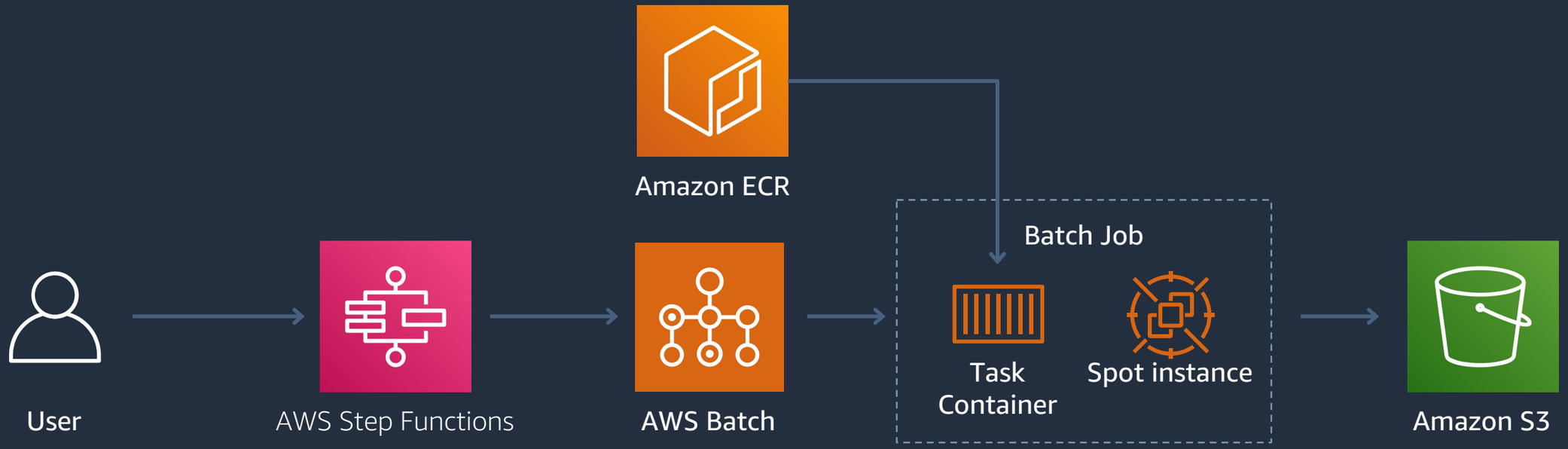
AWS Batch

Data storage

Amazon S3

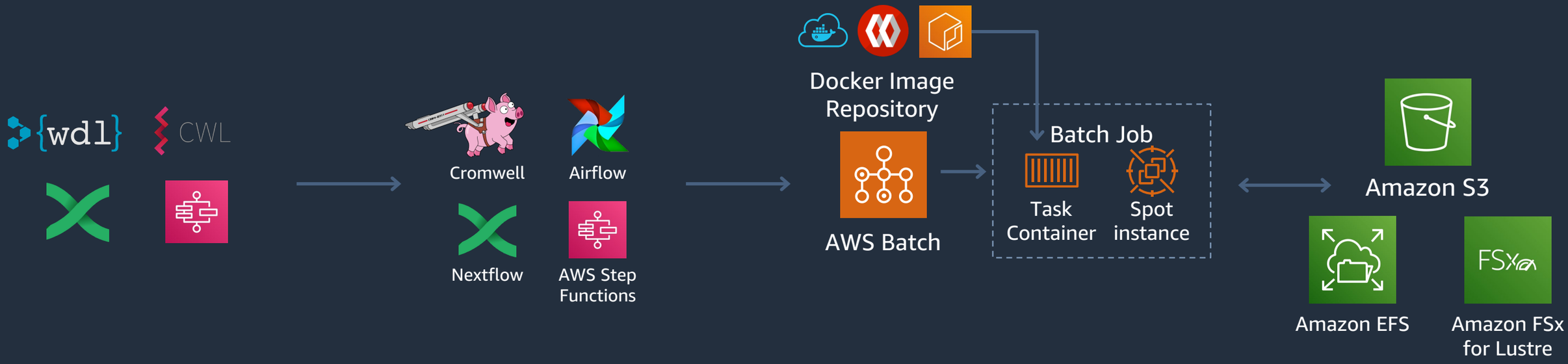
aws genomics

AWS native reference architecture



aws genomics

Putting it all together



Workflow definition

Develop a workflow using a definition language and containerized tools

Workflow orchestrator

Submit your workflow to a workflow engine

Job execution

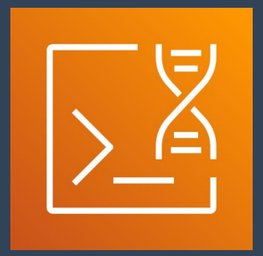
Workflow engine submits tasks to cloud compute resources (e.g., AWS Batch)

Data storage

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

aws genomics

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

Configure

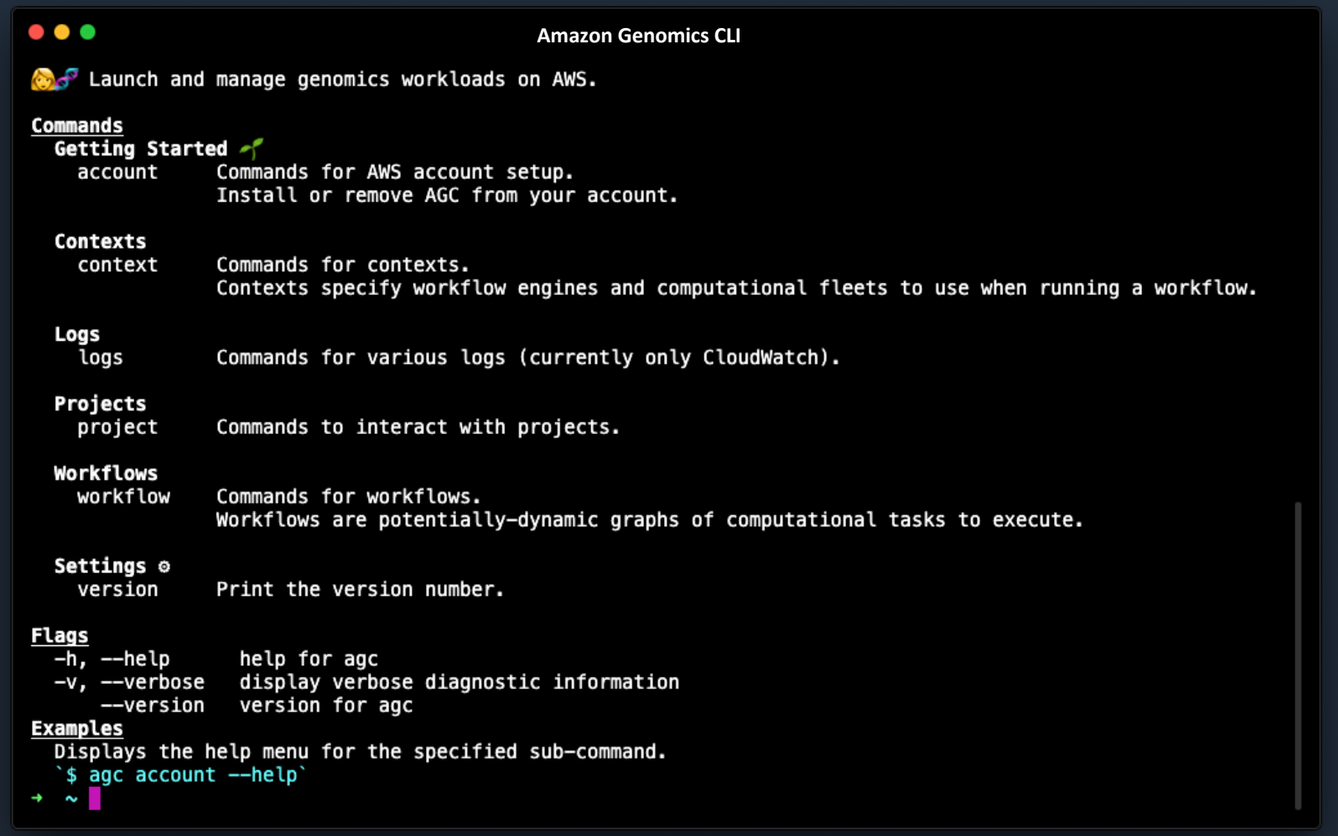
Create a project, define compute resources and workflows

Deploy

Deploy compute resources and container clusters to execute workflow engines

Run

Process genomic data and derive research insights



Amazon Genomics CLI

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

Configure

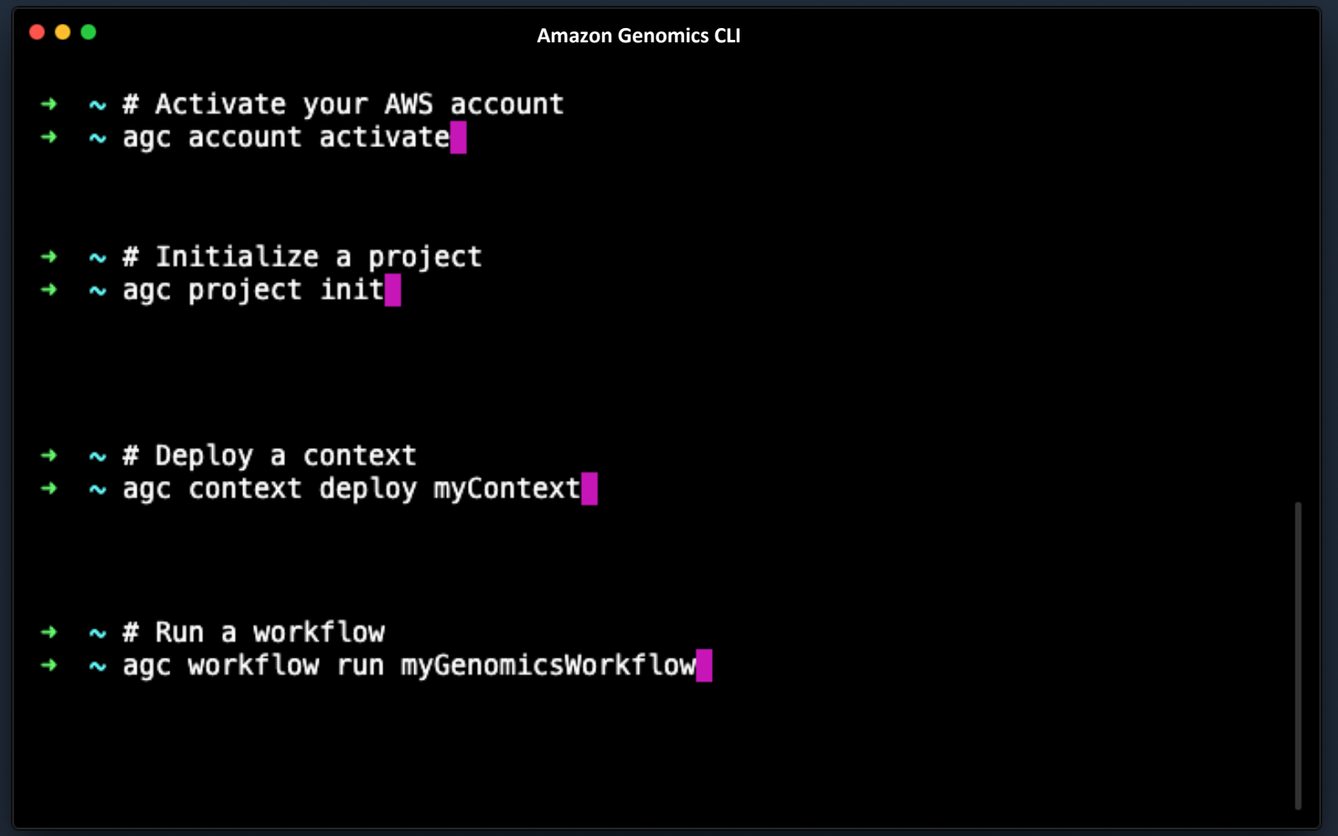
Create a project, define compute resources and workflows

Deploy

Deploy compute resources and container clusters to execute workflow engines

Run

Process genomic data and derive research insights



Industry-leading ISVs build on and collaborate with AWS

aws genomics

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

Accelerated GATK

Other secondary analysis platform

aws genomics

Demos

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 turn-around 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

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.

> “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

aws genomics

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

> "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

FRED HUTCH
CURES START HERE™

Company: Fred Hutch

Industry: Life Sciences

Country: United States

Employees: 3,500

Website: https://research.fredhutch.org/

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.

aws genomics

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 leap-forward cures, disease prevention, and our quality and understanding of life.

aws genomics

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

aws genomics

Next steps & resources

Resources

AWS for Health
aws.amazon.com/health

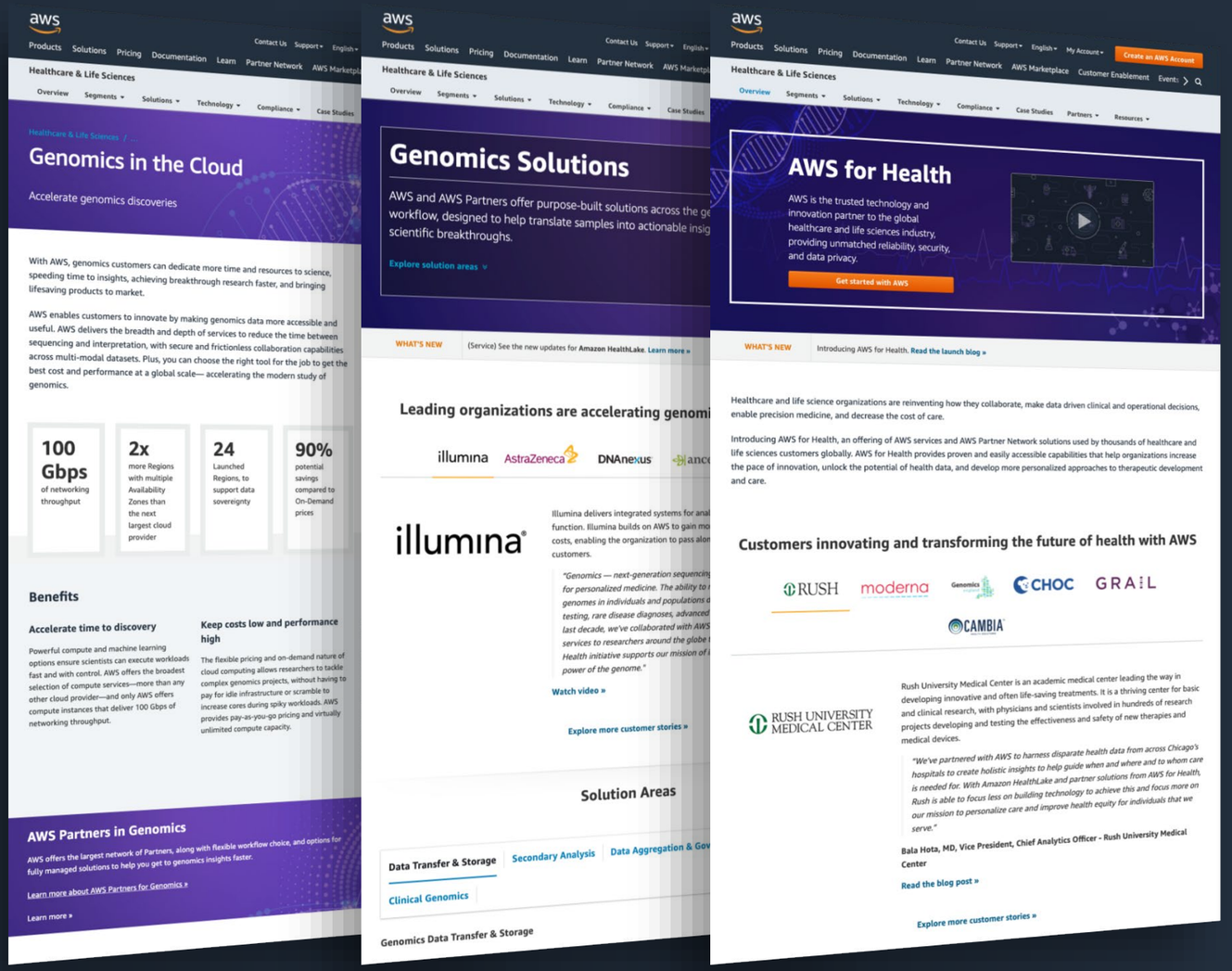
Genomics in the Cloud
aws.amazon.com/health/genomics

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

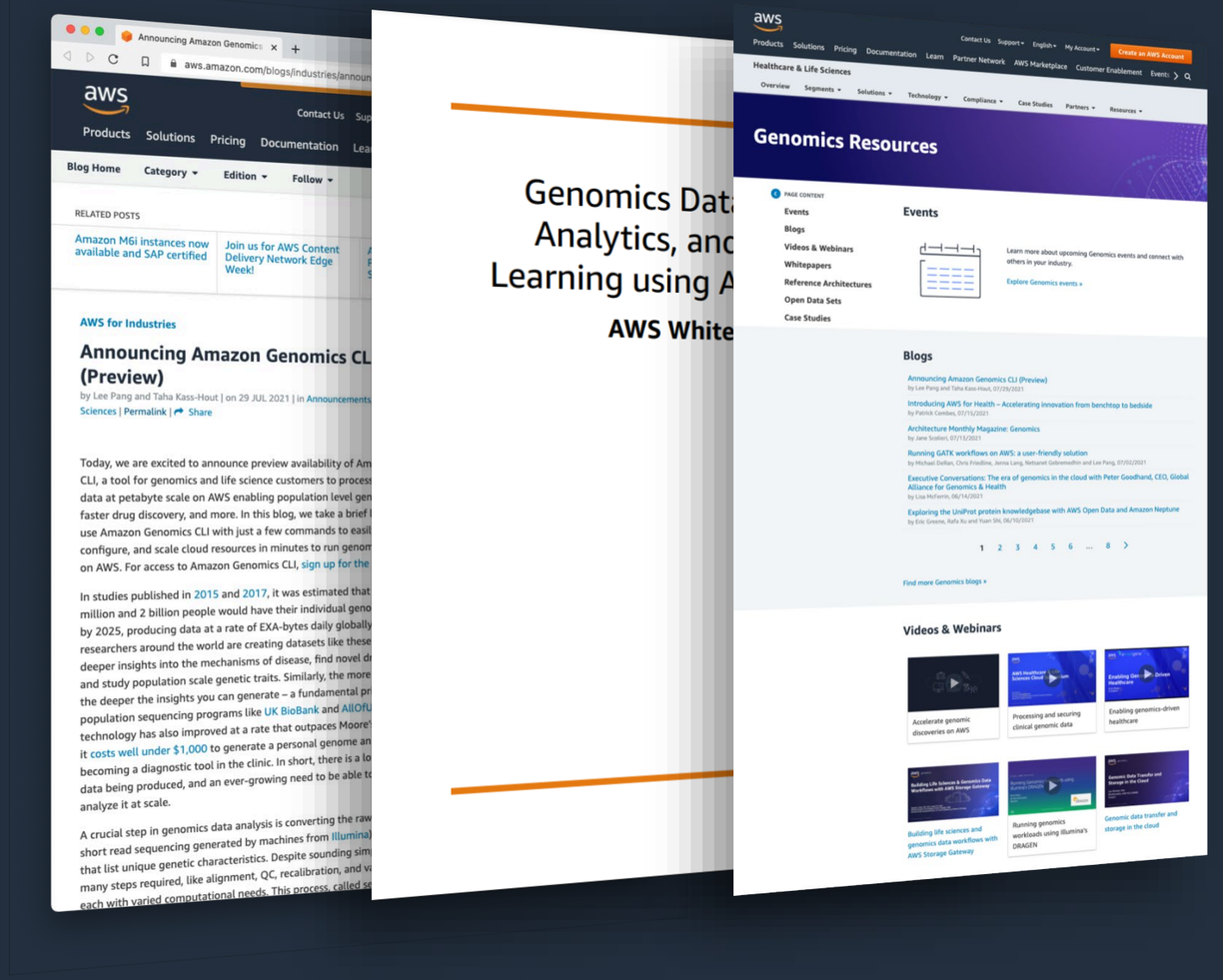


Resources continued

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

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

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



Thank you!

aws genomics

Questions and answers