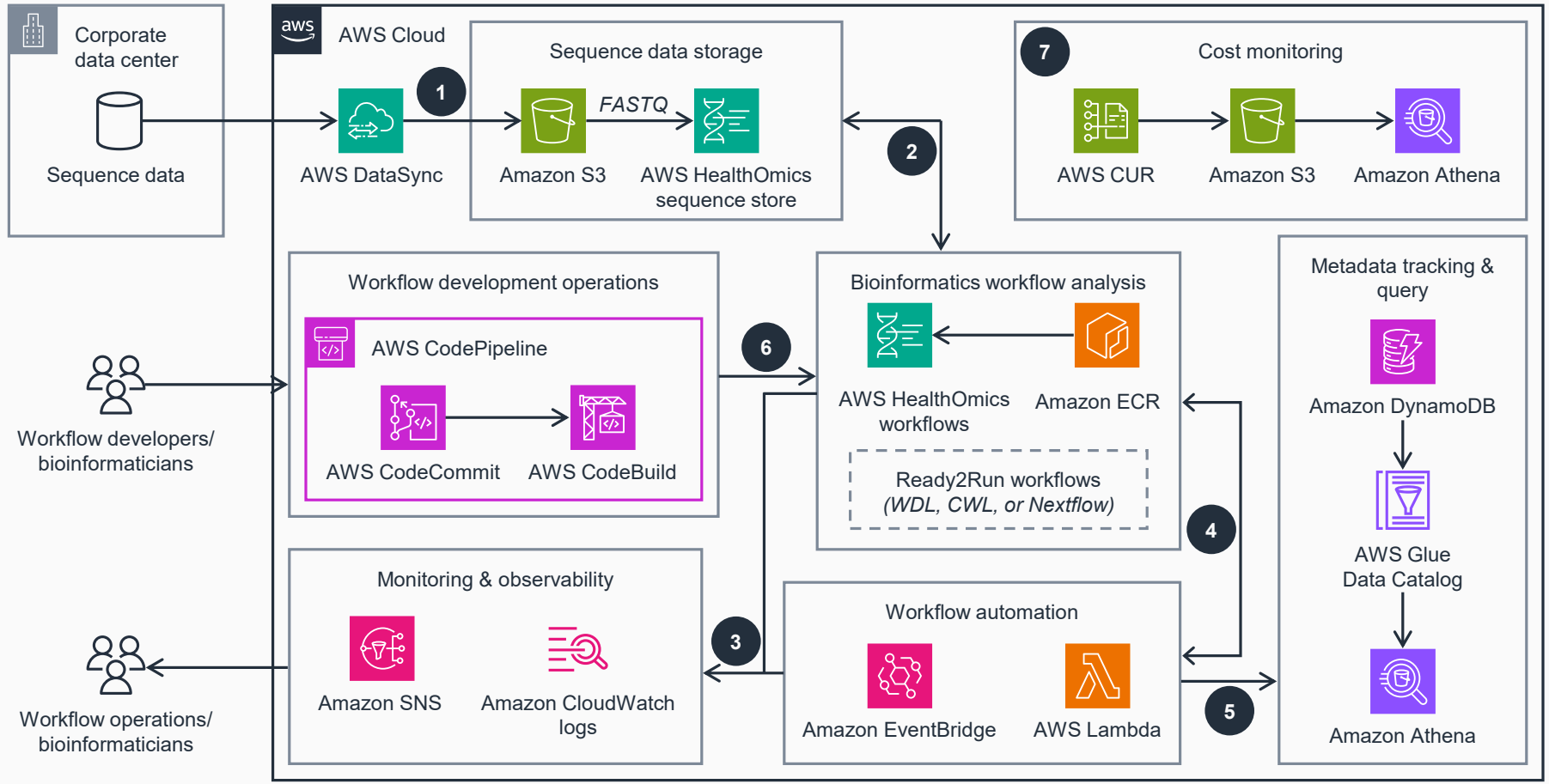
Guidance for Development, Implementation, Automation, and Tracking of Bioinformatics Workflows on AWS

This architecture diagram highlights key considerations and best practices for implementing bioinformatics workflows at scale.



1. Transfer sequence data to **Amazon Simple Storage Service (Amazon S3)** using **AWS DataSync**. If data is in FASTQ format, it can be imported into a sequence store in **AWS HealthOmics** (successor to Amazon Omics) for cost savings.
2. **HealthOmics** runs bioinformatics workflows in languages like Workflow Description Language (WDL), Nextflow, or Common Workflow Language (CWL) to analyze raw data. These workflows can be built as private or Ready2Run (hosted by **HealthOmics**). Tools running within the workflows are stored as Docker images within **Amazon Elastic Container Registry (Amazon ECR)**. Workflow outputs are uploaded to **Amazon S3**.
3. **HealthOmics** publishes workflow engine logs, task logs, and workflow run logs to **Amazon CloudWatch** for troubleshooting and monitoring.
4. **HealthOmics** publishes events using **Amazon EventBridge**, which can automate downstream actions, such as using **AWS Lambda** functions to launch more bioinformatics workflows or notifying users or groups about workflow failures using **Amazon Simple Notification Service (Amazon SNS)**.
5. Useful metadata from **HealthOmics** workflows—such as workflow run ID, tags, sample ID, workflow output file locations—can be tracked in **Amazon DynamoDB** tables. An **AWS Glue** crawler ingests this data into the **AWS Glue Data Catalog**, which can be queried using **Amazon Athena**.
6. Workflow developers and bioinformaticians can iterate on new and existing workflows and maintain version control using continuous integration and continuous delivery with **AWS CodeCommit**. **AWS CodePipeline** can be used to invoke an **AWS CodeBuild** job to automate the creation of new workflows in **HealthOmics**.
7. **AWS Cost and Usage Reports (AWS CUR)** facilitates cost monitoring. This service can be configured to create reports and upload them to an **Amazon S3** bucket. An **AWS Glue** crawler is configured to ingest this data to **AWS Glue Data Catalog**, which can be queried using **Amazon Athena** to derive cost-related insights.

Reviewed for technical accuracy October 24, 2023

AWS Reference Architecture