A Data Lineage Platform To Track Clinical Files History In Genomics Datasets

Genivate23 :: Data lineage :: Marquez

ABSTRACT

The volume and complexity of clinical genomics data are growing rapidly each day. We propose a system to address the challenges of tracking genomic files throughout the production workflow. This system clearly designates the origin of files and organizes information about their generation and storage status, ensuring easy access within a cloud-based platform.

Rust Turakulov

www.genivate23-data-lineage.agrf.org.au

Genivate23 Project Report
01/11/2024
Project: A Cloud Based Data Lineage Platform to Track Clinical Files
Team: Rust Turakulov, Application Developer (Bioinformatics); Lesley Gray, Data Strategy Manager; Matthew Tinning, Head of Laboratory Operations; Jacqueline Montgomery, Clinical Project Officer; John Fitzpatrick, Bioinformatics Support Specialist
Alliance Institution: AGRF

CONTENTS

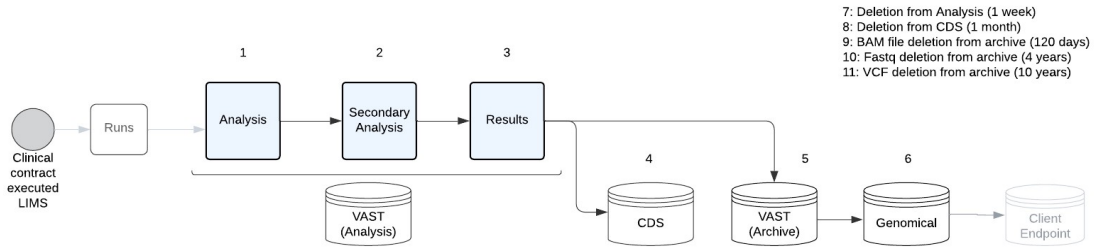
BACKGROUND INFORMATION

A common question when operating large-scale genomics projects or routine clinical genomic services is: 'Where is the data at this particular point in time, and how and when was it generated? What files it derived from?' This query can come from end clients, clinicians, or bioinformaticians, who may seek sequencing data for a specific sample, project, batch, or run (sequencing flow cell). Their interests can range from tracking the progress of a specimen through the pipeline. They may like to obtain results for a batch, or confirming the existence of files in archived folders for review or comparison purposes. The current LIMS or sequencing operation controls system BIOWEB do not have integrated functionality to answer such question.

Data provenance

At AGRF, sequencing data generation begins with the recording of contract and project details in the Laboratory Information Management System (LIMS, represented by the grey circle on the left). After receiving and preparing the DNA for sequencing, it is loaded onto the machine based on a sample sheet prepared by the sequencing operator. Once the sequencing run is complete, the raw FASTQ data is transferred from the machine to a high-performance storage system called VAST. The data is then processed under bioinformatics control, which includes three steps on the VAST drive:

1) Analysis : demultiplexing into individual samples FASTQ files based on sample sheet provided for the run and check for sample yeild and quality control metrics. At this step run data being organized to subfolders according to contract. This step generates FASTQ checksums and QC files sorted for each project.
2) Secondary Analysis. Subfolders/Contracts required additional processing processed with extra bioinformatics pipeline for example variant calling. This step usually generates additional files required for long term storage.
    a. BAM files intermediate alignment files.
    b. VCF files variant calling files (raw, unrefined)
    c. VCF files filtered and annotated (product grade)
3) Results files usually contains reports, graphical and textual outputs for batch QC metrics, and refined files for each sample. After reviewing results those files are normally copied to two places by bioinformatician or sequencing operator.
4) Files from Results folder and FASTQ data copied from VAST for Client Delivery Service (CDS) the company FTP site or cloud system.
5) Another copy from step 5 goes to specially configured VAST Archive area
6) Genomical receives local copy from the stage 5 (VAST Archive) via internet connection and redistribute files to the Client Endpoint.
7) Deletion from Analysis (1 week).
8) Deletion from CDS (1 month)
9) BAM file deletion from archive (120 days)
10) Fastq deletion from archive (4 years)
11) VCF deletion from archive (10 years)



Proposed MVP system is developed for only tracking stage 1,2 and 3 on described dataflow chart. Further system extension can be developed.

USE CASE SCENARIOS

1. Client lost delivery files.

Client lost his delivery files and looking for restoring data traces which was sequenced last year.

2. Project manager or client wander about missed samples.

Project managers want to know how many samples being sequenced for particular contract and what are those missed samples are they being sequenced and delivered, or they are still in pipeline?

3. Bioinformatician collecting information about sample.

Bioinformatician want to know what time sample variant calling (VCF) file was generated and which version of pipeline was used.

4. Audit or accreditation purposes.

Auditor from accreditation organization or Regulatory body want to extract information from control data from arbitrary date or batch.

5. Cohorting and data reuse for different contracts.

Bioinformatician working on client project need to top up cohort with control samples processed earlier this year with another contract id.

6. Returned client want to combine or top up samples project with previous data.

After running exome sequencing on contracted dataset. Client received extra funds to spend on additional RNAseq assay on subset of samples from the exome batch. To deliver results he asked to combine result for exome and RNAseq data together. This task will require to trace particular samples ID from exome batch and extract available files for those from archive.

7. Duplicated sample or project control.

Project control (duplicated sample) sequenced across different batches and flowcells in different time. One of the control replicates showed borderline quality metrics. Samples process within the same batch need to be flagged for the extra attention of clinician for the result interpretation.

USER BENEFIT

The proposed data lineage system offers significant advantages to a wide range of users, including clients, clinicians, bioinformaticians, and project managers. By enabling clear tracking of genomic files through the production workflow, the system helps users easily identify the current location, generation details, and derivation of files without the need for manual queries. This is especially useful in cases of lost files, project discrepancies, or audit requirements, allowing seamless restoration, retrieval, or validation of data. The system also supports data reuse and cohorting across different projects, enhancing the efficiency of managing large-scale genomics workflows. With detailed tracking in place, users gain transparency and control, ensuring they can access accurate, up-to-date information about their data at any point in the pipeline.

PLATFORM DESCRIPTION

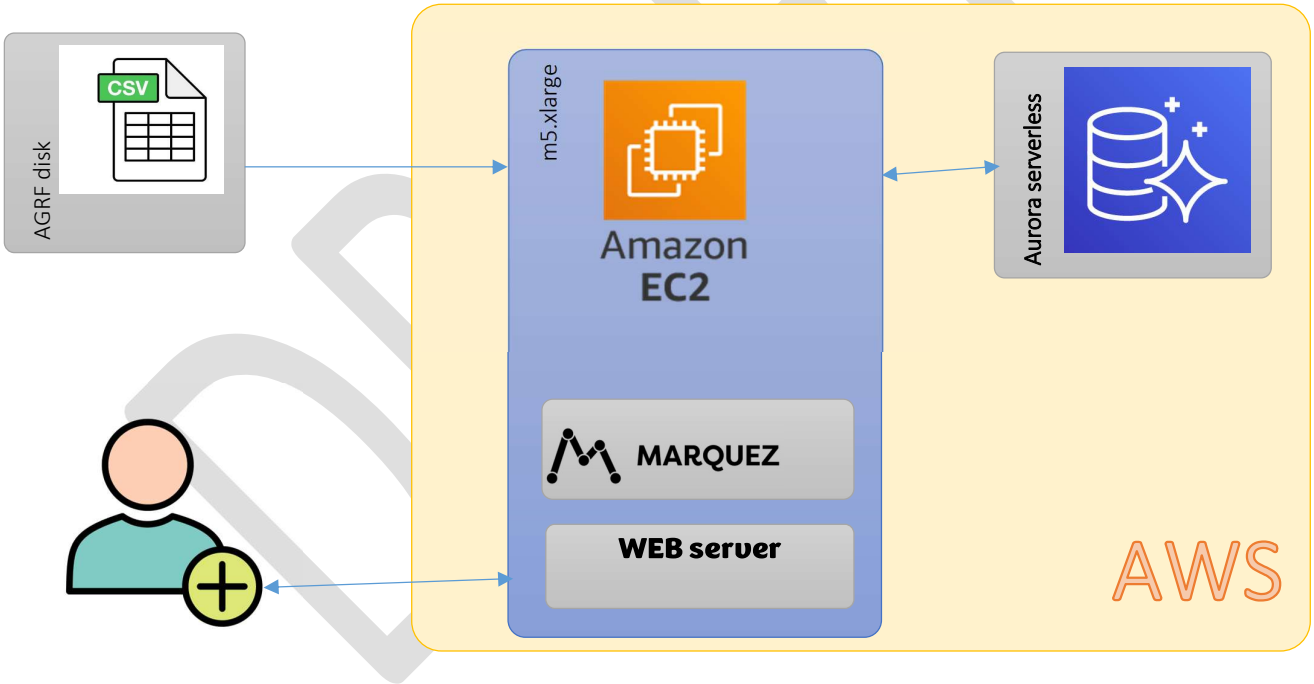
The platform is hosted on AWS, where system connectivity is key to ensuring smooth interactions between various components. At its core, the platform utilizes an EC2 instance for processing and handling application logic, while the database layer, likely hosted on Amazon RDS or another managed database service, stores critical data. This database is tightly integrated with the EC2 instance, facilitating seamless data access and transaction handling. Additionally, the platform leverages AWS S3 for scalable storage solutions, enabling data to be easily transferred between the EC2 instance and long-term storage as needed.

The connection to the client disk (VAST) is an integral part of this system. Data exchanges between the AWS-hosted platform and VAST storage are facilitated through secure file transfer protocols, ensuring that datasets remain accessible

for processing on the EC2 instance while staying compliant with any data residency or security requirements. This system design allows for efficient data processing and storage, ensuring that large datasets can be managed across different AWS services while maintaining the necessary links to client-side storage.
The current data connection process is performed in a semi-automated manner. Data files are cataloged and condensed into a single CSV file, which is then transferred to an EC2 instance on AWS. This file is subsequently ingested into an Amazon Aurora serverless database. This method streamlines the process of managing and uploading large datasets, ensuring that data from multiple sources is consolidated efficiently before being uploaded into the system.
Once the data is ingested into Aurora, it becomes accessible for further querying, analysis, and integration with other components of the platform. This semi-automated approach balances efficiency and control, allowing for periodic updates without full manual intervention.

AWS AND SYSTEM ARCHITECTURE

File belongs to clinical project on AGRF disk listed and searched based on file structure the current disk state captured to the comma separated file contained one line per captured file.
The CSV file copied to AWS EC2 instance in predesigned folder where automatic injection to Aurora relational dataset happened if record existed in database only column with last seen date is updated if new file observed new row is generated in database. In case file is removed from dataset the row corresponding to the record of missed file on user interface table will be greyed.



The marquez flow diagram will be generated and show to user after clicking on file name in
The data stored in Amazon Aurora serverless database which will allow scalable workflow for large number of samples.

KEY FUNCTIONALY OF THE PLATFORM

Two type user access Admin and User level.
User can be registered only by Admin.
User can browse and see entire dataset
Search facility to look for:

- Project / contract id
- Flowcell /run name

Sample id / sample name

All three fields can be combined with “AND” logic to narrow the search results.
Results table can be sorted by any column to easy navigation and data sorting.
File lineage can be navigated with Marquez interactive dashboard.
Information about each node (file or dataset) can be expanded in Marquez.
Picking each data node allow you to expand the dataset and see other files processed or generated within the dataset node (the Marquez feature).
System written in Jango language and potentially can be integrated in AGRF inhouse data tracking laboratory platform “Bioweb” which is also Jungo based.

LIMITATIONS

Limited Scope of Data Tracking (Stages 1-4)

The proposed MVP (minimum viable product) only tracks data through stages 1, 2, and 3 of the described dataflow (up to results generation). If a user needs to track data beyond this point (e.g., archival, deletion cycles, or final storage), the system will not cover those stages, potentially limiting its utility for long-term data tracking or comprehensive audit purposes. Extending the system to include further stages will require additional development.

Semi-Automated Data Handling

The system employs a semi-automated process for ingesting data from client disks (VAST) into AWS. This semi-automated approach could lead to delays or inefficiencies, especially if large datasets are involved. For example, data cataloging into CSV files may be subject to manual errors or require frequent updates, introducing the possibility of missing or outdated information during ingestion.

Basic Search Capabilities

The platform's search capabilities are relatively simple, relying on filtering by project/contract ID, flowcell/run name, and sample ID/name with "AND" logic. While this is useful for narrowing down search results, more advanced search features (such as wildcard searches, fuzzy matching, or complex logical queries) might be needed for users handling large datasets or trying to locate files with incomplete information.

Marquez Dashboard Usability

The interactive Marquez dashboard is a core feature for visualizing data lineage, but it may introduce usability challenges for non-technical users, such as clinicians or project managers unfamiliar with complex data lineage visualization tools.

CONCLUSION

The proposed data lineage platform addresses a critical need in large-scale genomics projects and clinical services by providing a transparent, efficient, and scalable solution for tracking data throughout its lifecycle. By leveraging AWS architecture, including EC2 instances, Amazon Aurora, and secure VAST connectivity, the system enables seamless management of genomic files from initial sequencing to final storage, ensuring data integrity, accessibility, and traceability. This platform enhances operational efficiency, mitigates risks of data loss or mismanagement, and supports a wide range of user needs, from client requests to audit compliance, ultimately delivering a comprehensive and robust solution for managing complex genomic data workflows.

REFERENCES

MVP on AGRF site:
https://genivate23-data-lineage.agrf.org.au/agrf/login

GitHUB documentation page and code:
https://github.com/datalineagerust/clinical_data_lineage/
https://github.com/AGRF/clinical_data_registry