CASE STUDY

Elucidata Delivers 100% Automated AAV-genome Sequencing Pipeline using Polly's Secure Infrastructure

A US-based biopharmaceutical company specializing in RNA-targeted therapeutics and gene therapies for rare diseases partnered with Elucidata **to optimize its AAV-genome sequencing process**. The company sought a custom pipeline to **analyze in-house virion samples and assess cassette quality in virions.** They also **needed a secure infrastructure** to process, store, and analyze in-house generated sequencing data. This solution significantly improved the company's process development, delivering a 4X cost saving in pipeline building and deployment.

Challenges

The team encountered challenges in **building a secure infrastructure** to manage AAV-genome sequencing analysis pipelines due to **limited bioinformatics expertise**, resulting in the following difficulties for their in-house team:

1. Streamlining **pipeline development and deployment** to detect issues like contamination, chimeric sequences, truncations, and point breaks was difficult due to limited in-house resources.
2. **Accurately estimating plasmid purity** within the pipeline became a critical issue, impacting the reliability of their gene therapies.
3. The team, primarily composed of molecular biologists with minimal bioinformatics experience, faced difficulty in **developing, managing, and running the pipeline independently**, hindering the ability to derive actionable insights.
4. Frequent shifts in research priorities further **complicated pipeline management and data analysis.**

Solution

The company leveraged **Polly's** secure infrastructure to process, store, and analyze its in-house AAV-genome sequencing data, seamlessly ingested through an automated ETL pipeline developed by Elucidata. Tailored to address the customer's AAV-genome process development challenges, the pipeline ensured data was high-quality, findable, version-controlled, and secured with role-based access.

Elucidata provided this solution that allowed the customer to independently manage and analyze in-house AAV sequencing data with minimal external support.

Elucidata's biomedical platform, **Polly**, offers comprehensive support for managing the entire data from ingestion to insights generations, offering end-to-end support for enterprises. Polly offers researchers a user-friendly interface that enables them to interact with and manipulate data while benefiting from security and credibility.

Our Approach

The customer requirements were addressed in three steps:

- Building and Managing Contamination Detection Pipeline on Polly
- Deployment of the pipeline on Polly's cloud to achieve processing at scale
- DIY enablement for code and no-code users

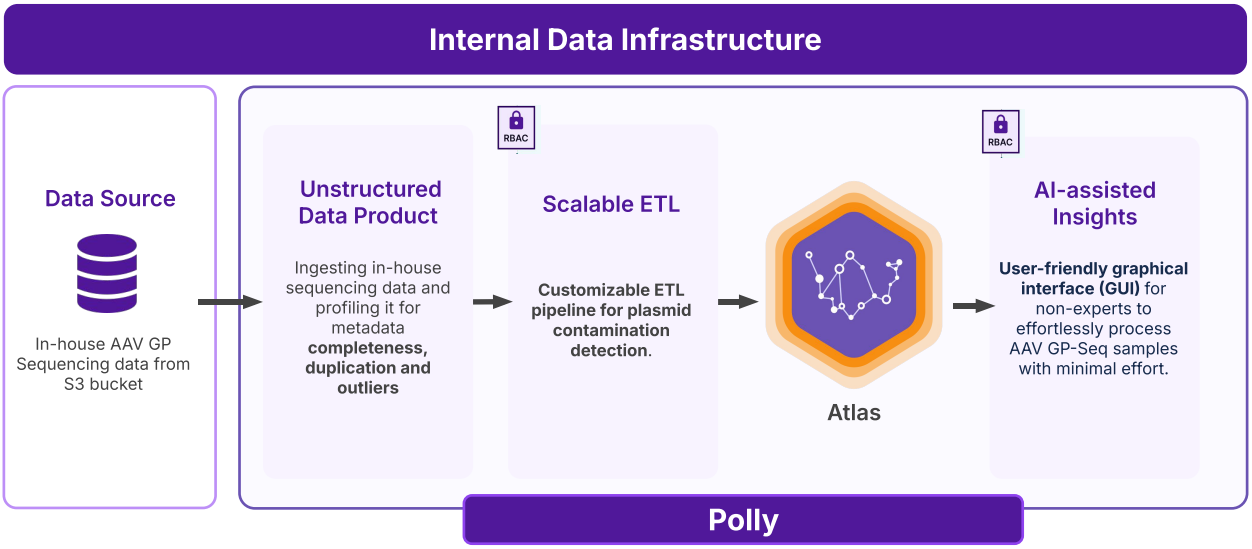
Building and Managing Contamination Detection Pipeline on Polly

Elucidata developed a **Plasmid Purity Pipeline** for **Single-Stranded DNA Virus (SSV-seq)** data, specifically targeting **Adeno-Associated Viruses (AAV)** used in gene therapy. The pipeline begins by ingesting raw in-house sequencing data from the customer's **S3 bucket**, which is then converted into FASTQ format. Our proprietary **harmonization engine** preprocesses the data, curating, standardizing, and enriching it to ensure high-quality input for the next stages of analysis.

Early in the pipeline, rigorous **quality control (QC)** checks are implemented to assess base quality and detect sequencing errors, ensuring data integrity and generating **pre-harmonization QC reports.** This step is followed by an advanced **contamination detection** phase using **ContaVect**, which compares sequencing data to reference genomes (rAAV, plasmid backbones, human genomes). This identifies and quantifies contaminants such as chimeric sequences, truncations, point mutations, and deletions, enabling precise assessment of DNA purity.

Next, sequencing reads are mapped to the **rAAV reference genome** during the **alignment phase**, generating **SAM/BAM files** for downstream analysis. Integration of **mapping statistics** helps assess alignment accuracy, while **read grouping** streamlines multi-sample processing. This **processed data** is then stored in a **structured format** in an **Atlas** for further downstream analysis. Throughout the pipeline, we ensured that unstructured data and metadata were extracted and transformed into standardized products.

With Elucidata's Plasmid Purity Pipeline, the customer enhanced the **sensitivity of impurity detection in AAV sequencing**, improving **purity estimation from 10% to a remarkable 0.01%.**



Plasmid Purity Pipeline on Polly

Deployment of the Pipeline on Polly's Cloud to Achieve Processing at Scale

Elucidata deployed this custom AAV GP-sequencing pipeline on Polly, leveraging its secure and scalable infrastructure to optimize in-house sequencing workflows.

The input and output files are seamlessly managed through an S3 bucket, ensuring secure and efficient data storage. The pipeline was dockerized, and the **Polly Command Line Interface (CLI)** was used to run the pipeline. Polly CLI provides the necessary infrastructure (machines) for data processing using an automated pipeline script. This was integrated directly into their terminal, triggering batch jobs, and enabling the team to easily execute and monitor their processes. For enhanced usability, a **user-friendly graphical interface (GUI)** was also provided, enabling non-experts to effortlessly process AAV GP-Seq samples with minimal effort.

This pipeline is centrally accessible, version-controlled, and secured with role-based access controls. It has a secure SOC2, HIPAA, and GDPR-compliant infrastructure, supporting seamless data integration and analysis. Robust security is maintained throughout the data lifecycle, with role-based access controls applied to unstructured data, ETL processes, structured data products, and insight generation tools.

DIY Enablement for Code and No-code Users

Elucidata enabled their team, including no-code users, to independently process new in-house AAV sequencing data and perform sequencing contaminants quality control checks independently. This capability enabled the generation of actionable insights and contamination purity detection reports that were easily shareable with collaborators. As a result, the process was significantly streamlined, allowing for quicker iterations, more accurate results, and enhanced collaboration.

Outcome

By partnering with Elucidata, the customer achieves the following major outcomes:

Category	Time Saved (hrs)
Pipeline Building and Customization	Without Elucidata, **customers would have to build and validate complex pipelines in-house**, requiring significant resources, full-time bioinformaticians, cloud engineers, further driving up operational expenses.
Plasmid Purity Improvement	Elucidata's plasmid purity pipeline enhanced purity estimation accuracy from **10% to 0.01% impurity**, a critical improvement for downstream applications. Achieving this level of precision would be extremely challenging without specialized tools and methodologies.
In-house Sample Quality Control	The customer's in-house experts leverage these pipelines deployed on Polly infrastructure for data analysis and to perform rigorous quality control on their in-house samples.
Scalable Analysis Pipeline	Robust, scalable pipeline deployed on AWS, tailored to the customer's infrastructure. Achieving 100% automation and significant cost savings with comprehensive end-to-end support.

Impact

100% automation from data ingestion to generating valuable insights

Custom pipelines delivered at **4X lower costs**

Plasmid purity pipeline improved purity estimation by **1,000-fold** from an existing 10% down to 0.01%.

End-to-end support in **building AAV-genome sequence analysis pipelines** for rare disease discovery

About Elucidata

Elucidata's Polly optimizes data quality for pre-clinical drug discovery by harmonizing multi-omics and assay data into ML-ready formats. Polly's powerful harmonization engine is utilized by trained experts to curate diverse data types, annotate metadata, and ensure consistent processing at affordable costs. These ML-ready data are stored on an Atlas on the platform, robust data stores ideal for analysis and management.

Polly's advanced technology accommodates 25+ R&D data types, catering to teams in pre-clinical drug discovery and diagnostics R&D. Polly is trusted by over 25 research organizations, including 4 of the top 10 pharma companies, to accelerate their discovery programs.

Book a Demo