

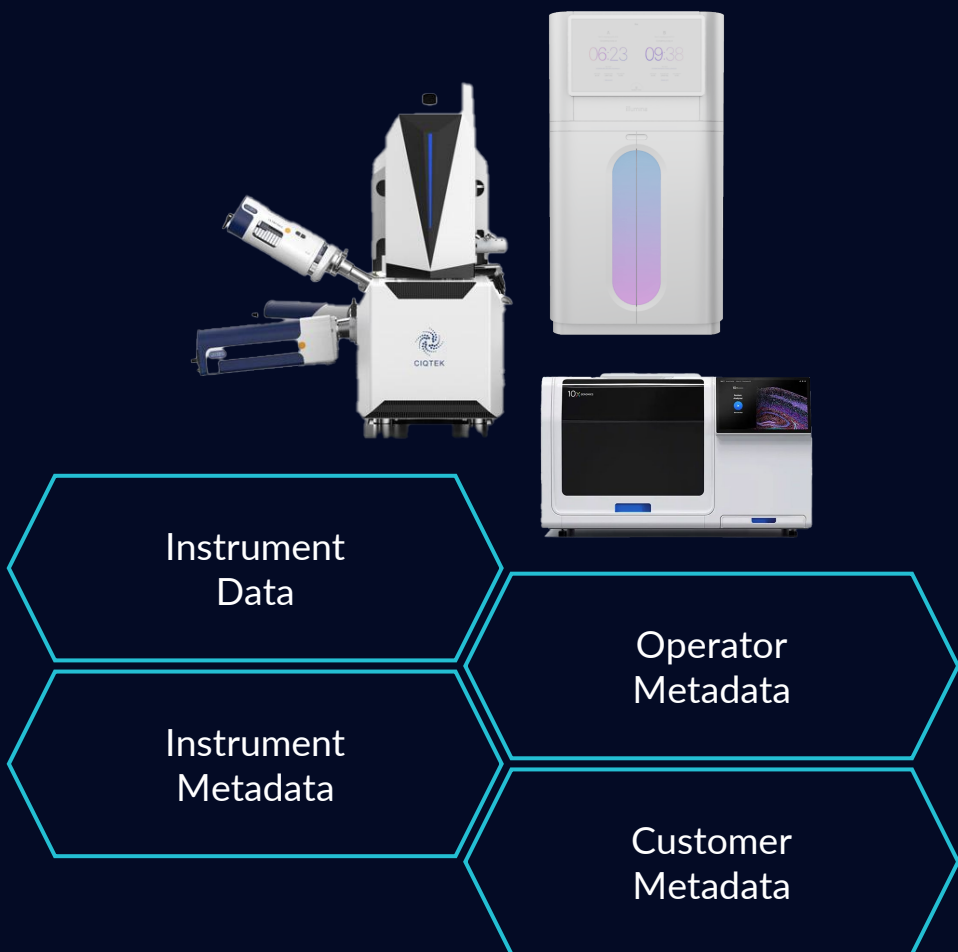
Life Sciences  
Generates One-Third  
of the World's Data

80% Of This Data Is Lost



# Orphaned Data

Custody begins with capture and linkage —  
trust ends without them



# Data Lineage

Automating provenance – preserves trust



The screenshot displays the Cirro web interface, which is used for managing and executing bioinformatics pipelines. The interface includes a sidebar with a "Project" menu and a "Pipeline Catalog" list. The main content area shows the configuration for a pipeline named "Resegment Cells (Proseg)".

**Run Pipeline | Resegment Cells (Proseg)**  
Segment cells spatially using the Proseg algorithm (Setty Lab - Fred Hutch)

**Dataset to use \***  
FFPE Human Ovarian Cancer with 5K Human Pan Tissue and Pathways Panel plus 100 C...

**Dataset Name \***

The results of this analysis will be stored in a new dataset with this name

**Dataset Description**

Details provided will enhance searchability and provide context for the dataset

**Prevents cells from having disconnected voxels**  
☒ Enforce Connectivity

**Run Proseg in 2D mode ignoring z-coordinate of transcripts**

**Status Completed**  
**Type Differential Expression Analysis**  
**Pipeline Identify Differentially Expressed Genes**  
**Created By mzager@cirro.bio**  
**Created On 1/17/2025, 12:57:09 PM**  
**Dataset Size 6.5 MiB**

**Generation**  
**Example Gene Expression An...**  
**Pipeline Documentation**  
**Input Parameters**  
**Execution Logs**  
**Cost Estimation**  
**More Info**

**Run Analysis**  
**Gene Set Enrichment Analysis**

**Proseg: Spatial Transcriptomics**  
**Data Type Single-Cell Seq**

One of the biggest challenges in the analysis of spatial transcriptomics data is how to best identify the location of single cells within the tissue. Proseg (probabilistic segmentation) is a deep learning-based method for the identification of cells from in situ spatial transcriptomics, developed by the [Newell Lab](#) at the Fred Hutch Cancer Center. The Cirro workflow for running Proseg is provided by [the Setty Lab](#). Xenium, CosMx, and MERSCOPE platforms are currently supported.

**a Cellular Potts Model (CPM)**  
Cell simulation methodology optimizing a combined objective function

**b Proseg**  
CPM-adapted segment cells by optimizing likelihood of observed transcripts

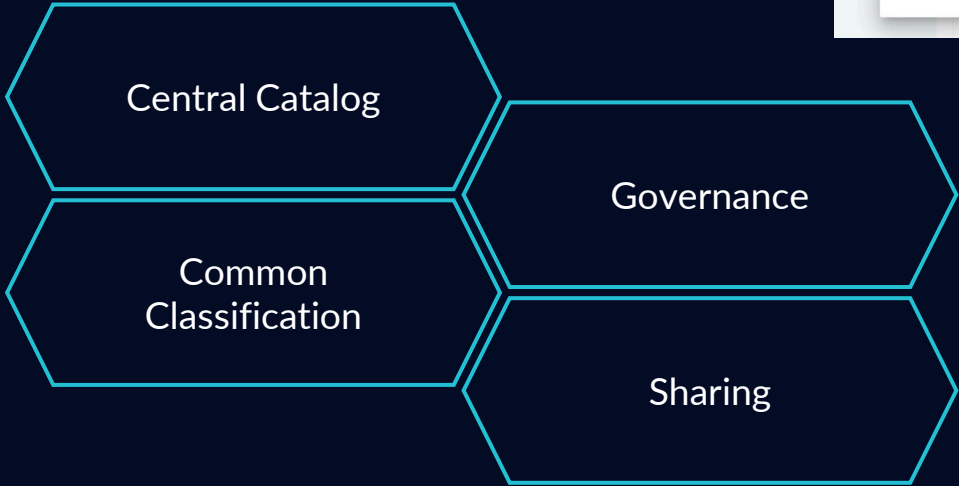
**Execution Log**  
October 11, 2025  
9:41 AM mzager 1000 Genomes - HG00138 - Exome - chr20  
Basecalling looks stable so far — quality scores hovering around Q30 for most reads.  
9:42 AM mzager Cluster density came out higher than expected, but still within optimal range - Michael Zager

**Datasets**  
Discussions  
Pipelines  
Projects  
Samples  
Shares  
Users

**Comment**

# Data Silos

Without a common catalog and classification, federation becomes fragmentation



Share | Create

Project

Overview

Discover

Datasets

Samples

Pipelines

Notes

Works

Shares

Dashboards

Users

Costs

Name \*

Description \*

Search Keywords

Share Details

Share to Project

TCGA

Apply Data Classifications

Public Data

Include Datasets

Specify the conditions upon which a dataset is included within this share

Tags

Created By

PH Data Core

Data Type

DISCOM

☐ Restrict file access to pipeline use only

Share | Subscribe

Subscribe to shares from other projects to access their datasets.

Consortium Data

NIH Funded

Subscribe

Public Data

Subscribe

TCGA

The Cancer Genome Atlas (TCGA) is a pioneering project initiated by the National Cancer Institute and the National Human Genome Research Institute in 2005. It systematically characterized the genomic, epigenomic, transcriptomic, and proteomic changes in 33 different types of human cancers. By generating an unprecedented wealth of molecular data and making it publicly available, TCGA has revolutionized the field of cancer research. Its findings continue to inform the development of targeted therapies and precision medicine approaches.

Governance

+ Add Requirement

Institutional Data Sharing Agreement

04/10/2026

Research Data Retention Policy

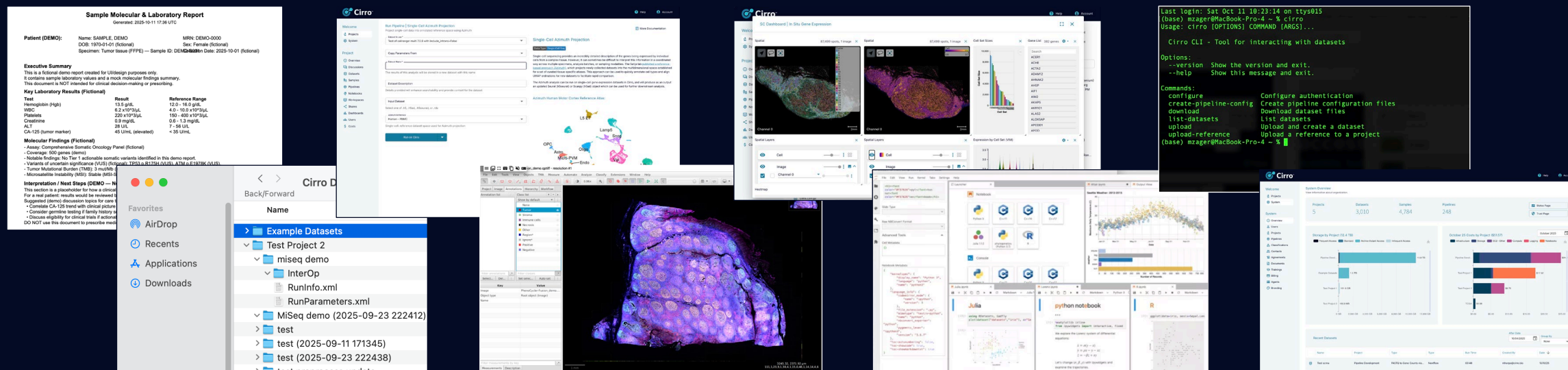
06/01/2026

HIPAA Training

01/01/2027

# Data Accessibility

Bridging diverse audiences and purposes



Clinicians

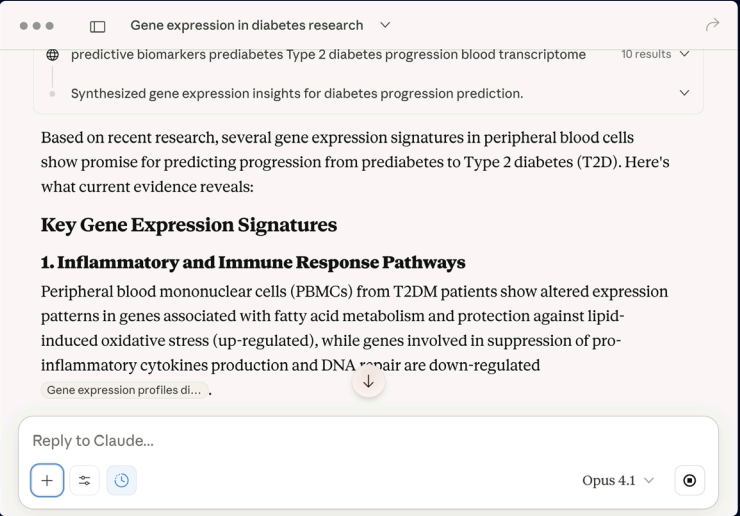
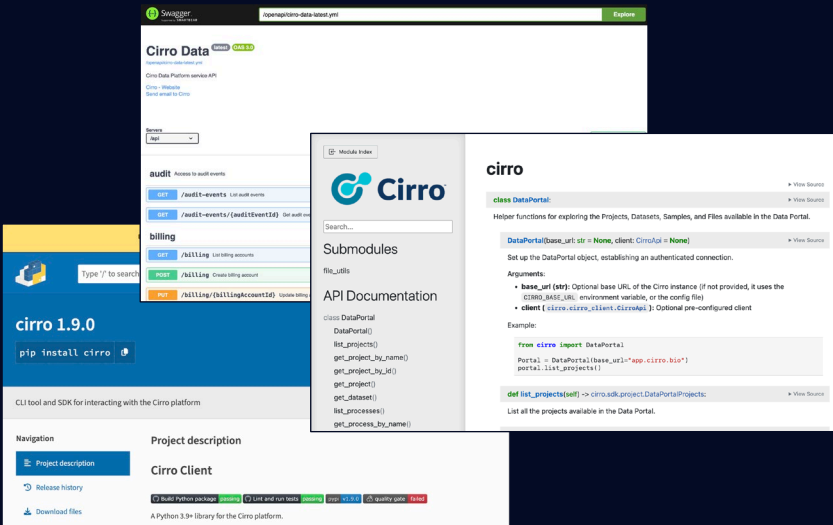
Researchers

Bioinformaticians

Administrators

# Data Interoperability

Bridging diverse audiences and purposes



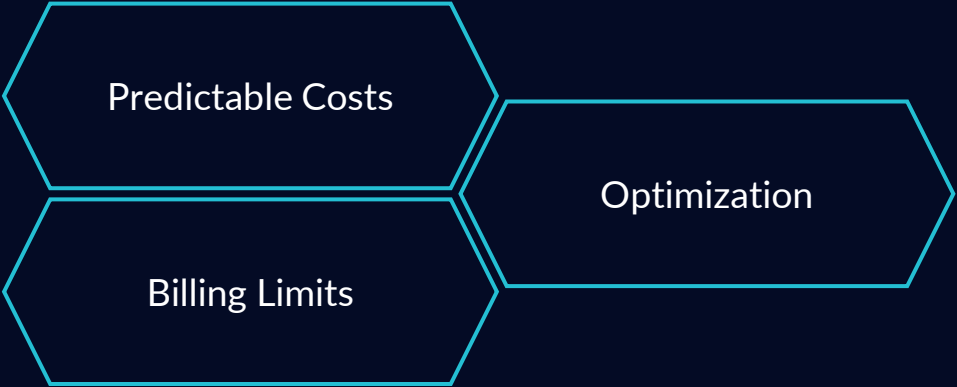
Applications

Agents



# Data Financial Management

## Cost Optimization



Dataset | Cost Estimation

View cost for this pipeline execution

Total Cost

Estimated cost to produce this dataset based on AWS On-Demand instance pricing. This does not include temporary storage, network traffic, regional pricing differences, or GPU costs. Actual costs may be lower if spot instances were used.

\$6.58

Cost by Status

Breakdown by task status. Cached tasks were reused from previous executions and did not incur additional costs for this dataset.

Completed (282 tasks)

\$6.57701

Cirro

HelpAccount

Welcome

ProjectsSystem

Project

OverviewDiscussionsDatasetsSamplesPipelinesNotebooksWorkspacesSharesDashboardsUsersCosts

Costs | Pipeline Development

View cost breakdown and set spending limits.

Edit Budget

Month to Date

\$24.72

Average Month

\$94.46

Total Project Cost

\$1,063.78

The amount the project has cost over its entire lifetime

Total Project Storage

10.8 TB

Rolling 12 month breakdown of charges

StorageInfrastructureComputeNotebooks

\$295.26	\$131.04	\$60.74	\$47.51	\$150.45	\$61.49	\$59.17	\$53.97	\$49.20	\$75.61	\$54.62	\$24.72
----------	----------	---------	---------	----------	---------	---------	---------	---------	---------	---------	---------

# Thank you!

Michael Zager

Linkedin/in/zager

mzager@cirro.bio



UW Medicine

