Project Phase I

Pancreatic Cancer Organoid Profiling – Complete Project Documentation

By: Wahed Shaik – 60302087

Title: Pancreatic Cancer Organoid Profiling for Chemotherapy Response Prediction

---

Problem Description

Pancreatic cancer remains one of the deadliest cancers, with limited treatment success and a 5-year survival rate below 12%. Chemotherapy response varies widely between patients, and clinicians currently lack strong predictive markers that determine which treatment would be most effective for a particular individual.

To help solve this, researchers developed *patient-derived organoids* (lab-grown mini-tumors). These organoids mimic the genetics and drug responses of the original tumors and provide an ideal system for studying chemotherapy response.

The dataset we analyzed contains:

- Whole Genome Sequencing
- Whole Exome Sequencing
- RNA-Seq gene expression
- Sample attributes (tumor vs normal, body site, etc.)
- Clinical metadata

Our goal was to transform the raw RNA-Seq gene count data from organoids and create clean, ready-for-analysis data products usable in Power BI, machine learning, or future bioinformatics pipelines.

---

3. Hypothesis

Certain gene expression patterns and genetic features in pancreatic cancer organoids correlate with chemotherapy response.
If we profile gene activity accurately, we can identify markers that predict whether a patient will respond to treatment.

This project focuses on preparing the data pipeline that makes such analysis possible.

---

4. Data Sources

Primary Dataset

Pancreatic Cancer Organoid Profiling (dbGaP accession: phs001611.v1.p1)

- Dataset description: https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001611.v1.p1

Open Access RNA-Seq Data (No login needed)

- AWS Open Data Registry: https://registry.opendata.aws/organoid-pancreatic
- S3 bucket (public): s3://gdc-organoid-pancreatic-phs001611-2-open/

Gene Reference Annotation (GRCh38 Release 109)

Downloaded from Ensembl:

- https://ftp.ensembl.org/pub/release-109/gtf/homo_sapiens/

We used:
Homo_sapiens.GRCh38.109.gtf.gz
(Not the ab-initio version.)

---

5. Tools Used

Core Platform

- Databricks for the entire ETL pipeline and validation.

Storage

- Azure Data Lake Storage Gen2 (ADLS) using:
    - SAS Token (Shared Access Signature) authentication

Transformation Framework

- Delta Lake Bronze / Silver / Gold architecture

Visualization

- Power BI Desktop & Power BI Service

---

6. The Journey: Step-by-Step What We Did

Step 1 — Configuring ADLS access in Databricks

We configured the connection to Azure Blob Storage using:

- Storage account

- SAS token
- Container
- Mounting / direct ABFSS access

This allowed Databricks to read and write files to:
abfss://project@<storage_account>.dfs.core.windows.net/

Step 2 — Exploring the Raw JSON Matrix

The RNA-Seq gene counts matrix was a wide JSON file, containing:

- 1 column for gene identifiers
- 110+ columns for sample UUIDs
- Many nulls and repeated gene names (non-standard format)

We inspected the schema and discovered:

- Some columns were quality-control artifacts (e.g., N_multimapping)
- Gene column was incorrectly named "Unnamed: 0"
- Sample names included hyphens and needed normalisation

This was ideal for a Bronze table.

---

7. Bronze Layer – Raw but Structured Data

Objective:
Store raw data in Delta format, with minimal cleaning.

Actions:

- Renamed "Unnamed: 0" → gene_id
- Standardised sample column names (replace - with _)
- Stored the cleaned-but-not-transformed matrix in Delta Lake

Output:
gold/goodreadsreviews60302087/bronze/combined_counts_matrix_raw

---

8. Silver Layer – Analytical Data (Normalized Shape)

This is where the real transformation happened.

Key Actions:

✓ Unpivoted Wide → Long format

Converted:

gene_id | sample1 | sample2 | sample3 | ...

→

gene_id | sample_uuid | count

✓ Converted nulls to 0 counts

Because most RNA-Seq matrices encode “no reads” as null.

✓ Verified value ranges

We computed:

- minimum count
- maximum count
  Result: max expression ~896k reads — correct range for raw RNA-Seq.

Stored at:
silver/counts_long

---

9. Adding Gene Annotations (Ensembl GTF)

We imported the human reference gene annotation file:

- Parsed all rows where feature = "gene"
- Extracted:
  - gene_id
  - gene_name
  - chromosome
  - start/end
  - strand orientation

We then joined this table with the Silver counts table.

Result:
ENSG… identifiers became linked to gene symbols like:

- DDX11L1
- WASH7P
- FAM138A

- PHKA1P1
- TP73-AS2
  etc.

Stored at:
gold/counts_with_genes

---

10. Gold Layer – Final Curated Datasets (For PowerBI)

We generated three gold products:

Gene-Level Features

(Gene expression summarised across all samples)
Includes:

- mean log-expression
- standard deviation
- number of samples where gene is expressed
- percent detection

Stored at:
gold/gene_features

Sample-Level Features

(Characteristics of each organoid sample)
Includes:

- library size (sequencing depth)
- number of detected genes
- % of zero-expressed genes
- mean expression intensity

Stored at:
gold/sample_features

Normalized Expression Table

(CPM & logCPM values for each gene in each sample)
Used for:

- heatmaps
- clustering
- PCA
- predictive models

Stored at:
gold/normalized_counts

---

11. Power BI Integration

What we connected to

In Power BI →
Get Data → Azure → Azure Data Lake Storage Gen2

We selected the Gold folder that contains:

- gene_features
- sample_features
- normalized_counts
- counts_with_genes

What we observed

At first, some tables incorrectly showed blank values (backend preview issue), but after loading, the correct numeric columns appeared.

We then created:

- gene-level summary visuals
- sample quality plots
- distribution charts
- zero-inflation heatmaps

---

12. Challenges Faced

- SAS authentication failures
- Resolved by switching from access keys to SAS token in Spark configs.
- JSON matrix schema irregularities

- Many nulls + weird column names required careful normalization.
- GTF file mismatch
- The wrong file (abinitio) caused incorrect annotations.
  Correct version (Homo_sapiens.GRCh38.109.gtf.gz) fixed it.
- Large unpivot requiring type casting
- Because Spark required uniform numeric types.
- Power BI showing blank preview
- Resolved when loading the dataset; visualization worked.

---

13. Conclusion

This project successfully transforms messy, raw RNA-Seq organoid data into clean, structured, fully annotated datasets ready for biomarker discovery.

The resulting Gold tables enable:

- visualization of gene expression patterns
- sample quality control
- identification of genes consistently active
- future machine learning models for predicting chemotherapy response

This is a complete functional foundation for precision oncology analytics.

---

14. References

- AWS Open Data Registry – Organoid Pancreatic Dataset
  https://registry.opendata.aws/organoid-pancreatic
- dbGaP Study phs001611.v1.p1
  https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001611.v1.p1
- Ensembl GRCh38 Release 109 GTF
  https://ftp.ensembl.org/pub/release-109/gtf/homo_sapiens/