OncoDM and Spatial Innovation Initiatives Provide High-Quality, F.A.I.R. and AI-Ready Data to Fast-Track Drug Discovery Insights

Sondra Kopyscinski[1,2], Hillary Mosso[1], Cassin Williams[1], Amrita Bhattacharya[2], Dzmitry Fedarovich[2], Konstantin Bobkov[1], Marissa Hirst[2], Nicole Leyland[1,2], Dan Rozelle[1,2]

Rancho Biosciences Initiatives: [1]OncoDM and [2]Spatial Innovation Initiative

Abstract

Rancho BioSciences is a trusted authority in data integrity and bioinformatics and is committed to staying at the forefront of biomedical data science and technology, with a focus on innovation and continuous improvement. Providing data services and consulting to accelerate research in areas like oncology, genomics, and other biomedical sciences Rancho is offering several unique pre-competitive initiatives to maximize return on investment and deliver a greater number of standardized, harmonized datasets, processed with a custom bio-informatics pipeline. One of these initiatives, **Oncology Data Mart** (OncoDM), identifies, downloads, integrates and processes **bulk gene expression oncology** data, including data from TCGA, GTEx (as a healthy reference) and additional datasets from Gene Expression Omnibus (GEO) across a variety of cancer types. The second initiative, **Spatial Innovation Initiative** (SII), identifies and processes high-quality **spatial transcriptomic** datasets that use cutting edge technology to map gene expression within the spatial context of tissue architecture across various indications. The first round of SII consisted of 30 datasets including 10x Visium, Nanostring GeoMx and CosMx workflows.

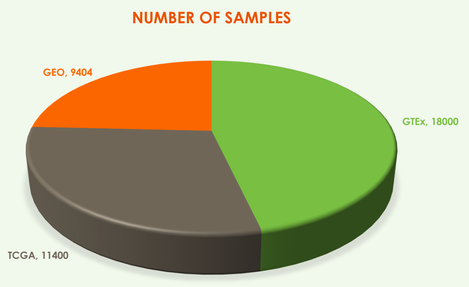
Both initiatives' deliverables feature **high-quality F.A.I.R harmonized metadata** aligned with Rancho-generated data models, and are structured in consistent, machine-readable formats suitable for AI/ML applications. This includes standardized annotations and harmonized metadata, which enable integration into computational pipelines for model training, benchmarking, and large-scale analysis. The reprocessed data is delivered in a ready-to-use format to seamlessly integrate and compare with other datasets and accelerate insights into drug discovery. Similar initiatives are available in neurology, immunology, and gastrointestinal spaces with the same emphasis on F.A.I.R principle and preparations for both human, and computational and AI-based analyses.

Oncology Data Mart

- ✓ **Collaborative Effort:** Generate large-scale, disease-specific content across multiple datasets.
- ✓ **FAIR Data:** Content that is integrated, re-analyzable, harmonized, and portable.
- ✓ **Unlocked Value:** Extract value from vast amounts of disparate public domain oncology data.
- ✓ **Seamless Integration:** Process, harmonize, and integrate oncology data into current systems and workflows.

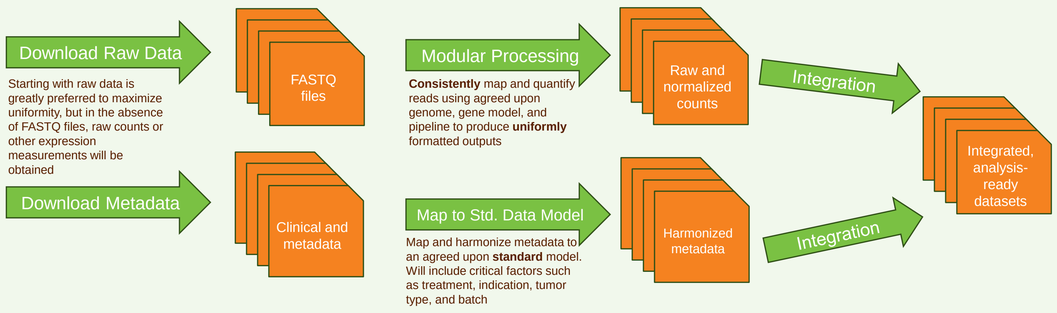
What is Oncology Data Mart?

Oncology Data Mart, also known as OncoDM, is a pre-competitive, community initiative with an initial focus on bulk RNA-sequencing. Eventually, the effort could expand to gene variants, phenotypical analysis, and beyond.



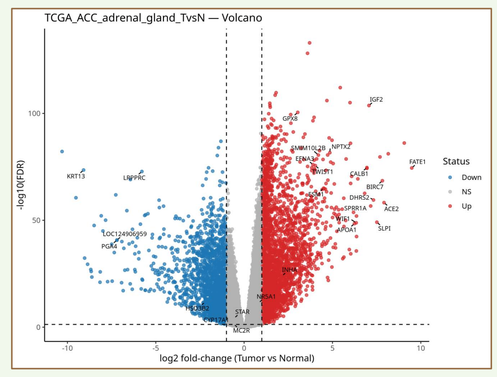
What are the Data Sources?

The initial deliverables for OncoDM include approximately 11,400 samples across 93 datasets from **The Cancer Genome Atlas** (TCGA and approximately 18,000 samples from the **Genotype-Tissue Expression** (GTEx) project as normal reference samples. An additional 40 **Gene Expression Omnibus** (GEO) datasets were processed which includes 9,404 samples across various cancer types.



What data analysis was performed?

A differential gene expression (DEG) analysis was conducted to compare each cancer/normal pairing from TCGA and GTEx as well as at least one specific comparison for each GEO dataset. A custom outlier detection analysis was also performed on both TCGA and GTEx.



What is the OncoDM Pipeline?

OncoDM leverages the talent of both our Data Integrity Specialists and Bioinformatics team to produce integrated datasets (gene expression data + harmonized metadata) along with individual flat files. All data will have been processed uniformly, better enabling cross comparison between and/or combining datasets.

Want to know more? Ask me or see our Booth at 501!

Rancho is interested in forming DataMart Initiatives for Neurology, Immunology, or Gastro-intestinal areas of focus!

Spatial Innovation Initiative

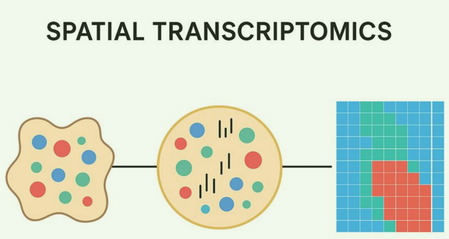
SpatialInnovation Initiative

What is spatial transcriptomics?

Spatial transcriptomics is a collection of methods that measure gene expression while preserving information about the *physical location* of those transcripts within a tissue section. Methods include MERSCOPE/MERFISH, 10x Visium and Xenium platforms, as well as Nanostring GeoMx and CosMX assays.

Why use spatial transcriptomics?

Spatial transcriptomics can be used to understand the gene activity patterns within their native tissue context, revealing cellular organization, interactions, and molecular changes associated with health and disease.



What is the Spatial Innovation Initiative (SII)?

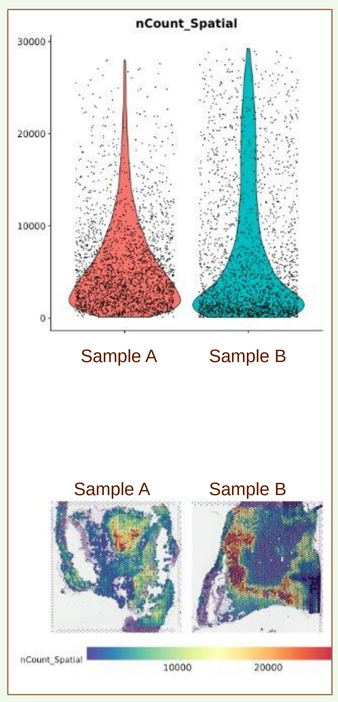
SII is a collective value initiative to develop industry standards and designed to deliver ready-to-use high-quality spatial transcriptomic datasets. Valuable data to enrich existing biomedical research programs using public domain data.

Sharing the cost delivers better value and ROI. Year 1 has been completed and can be deliver immediately.

Year 1
- Completed in 2024
- 33 datasets delivered

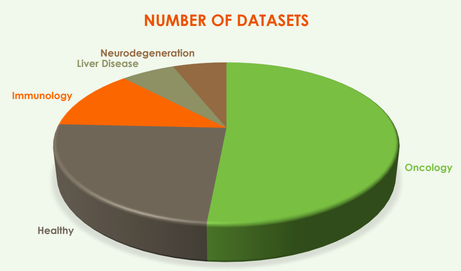
Year 2
- **Join Now**
- Select 10 datasets of interest
- Receive in 2025

Year 3
- Next Year
- Build with new studies
- Select 10 more to expand your data and increase ROI



What are the SII Deliverables?

- ○ Catalog of spatial data
  - Annotated list of available datasets with key metadata categories. This catalog is used to select datasets as well as inform members on the state of the industry.
- ○ Extended spatial data model compatible with other Rancho offerings but includes a special emphasis on spatial-specific data features.
- ○ High-quality F.A.I.R harmonized metadata and analysis-ready molecular data
  - 10 spatial datasets you designate.
  - + 33 (21 human & 12 mouse) spatial datasets selected by previous members from healthy, cancer and other diseased tissues.



Deliverable	Format
Scanpy analysis object	h5ad
Seurat analysis object	RDS
Spatially Variable Genes	csv
README	txt
Metadata workbook	xlsx
QC directory	misc
manifest	json