Spatial Innovation Initiative Delivering Harmonized Analysis-Ready Datasets

Marissa Hirst, Sondra Kopyscinski, Nicole Leyland, Andy Hope, Amrita Bhattacharya, Dzmitry Fedarovich, Kenneth Chan, Dan Rozelle

Rancho BioSciences, LLC

SpatialInnovation Initiative

Abstract

In 2024 the spatial innovation initiative was formed to make spatial transcriptomic data accessible to the biomedical research community. As an emerging technology many organizations have yet to leverage spatial transcriptomic data for gaining insight into tissue organization, cellular interaction, and the interpretation of spatial patterns in a clinically relevant manner.

This initiative focuses on providing analysis-ready datasets by ensuring consistency and harmonization of metadata in computational format. This includes alignment of metadata to an expanded data model, built upon our existing transcriptomic data models, adding spatially relevant terms and concepts.

Three charter members kicked off this initiative with a meeting to prioritize technology, workflow, and format requirements. As part of this meeting members have also designated 10 datasets to begin processing.

Spatial Technology

Technology	Type	Size	Features	Tissue Size
10x Visium	Sequencing	50um (10-25 cells)	Transcriptome (16k)	40 mm²
Nanostring GeoMX DSP	Sequencing	Several cells	Transcriptome (11k)	Triplicate ROIs
Nanostring CosMx	Sequencing	Single cell	Up to 1000 genes	20 mm²
Vizgen MERSCOPE/ MERFISH	Imaging	Single cell	300-500 gene panel	100 mm²
10x Xenium	Imaging	Subcellular	200-400 gene panel	1400 mm²

As an emerging technology, many technology solutions exist. The main types are defined as either sequencing or imaging-based techniques. In sequencing types, traditional high-throughput sequencing is performed on RNA located in a defined physical location which typically includes multiple cells. Imaging-based spatial technologies localize and enumerate gene-specific probes to a high degree of resolution but are limited to fewer genes than whole transcriptome.

Availability of Public Data

A. Year of Publication
B. Repository
C. Technology
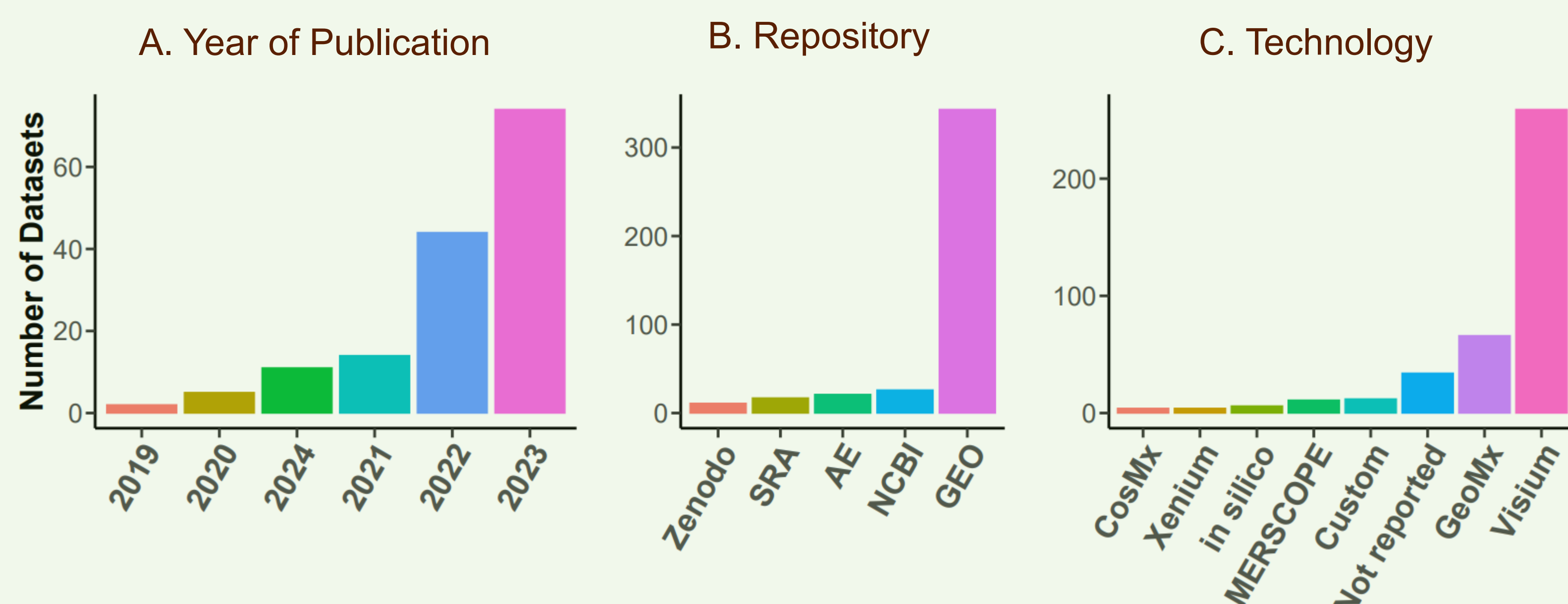
Rancho used their DataCrawler to identify publicly available spatial datasets. A., the number of publications is rapidly increasing each year. B., a majority of public spatial data is found in NCBI's Gene Expression Omnibus (GEO) repository C., 10x Visium is the most popular platform for spatial data.
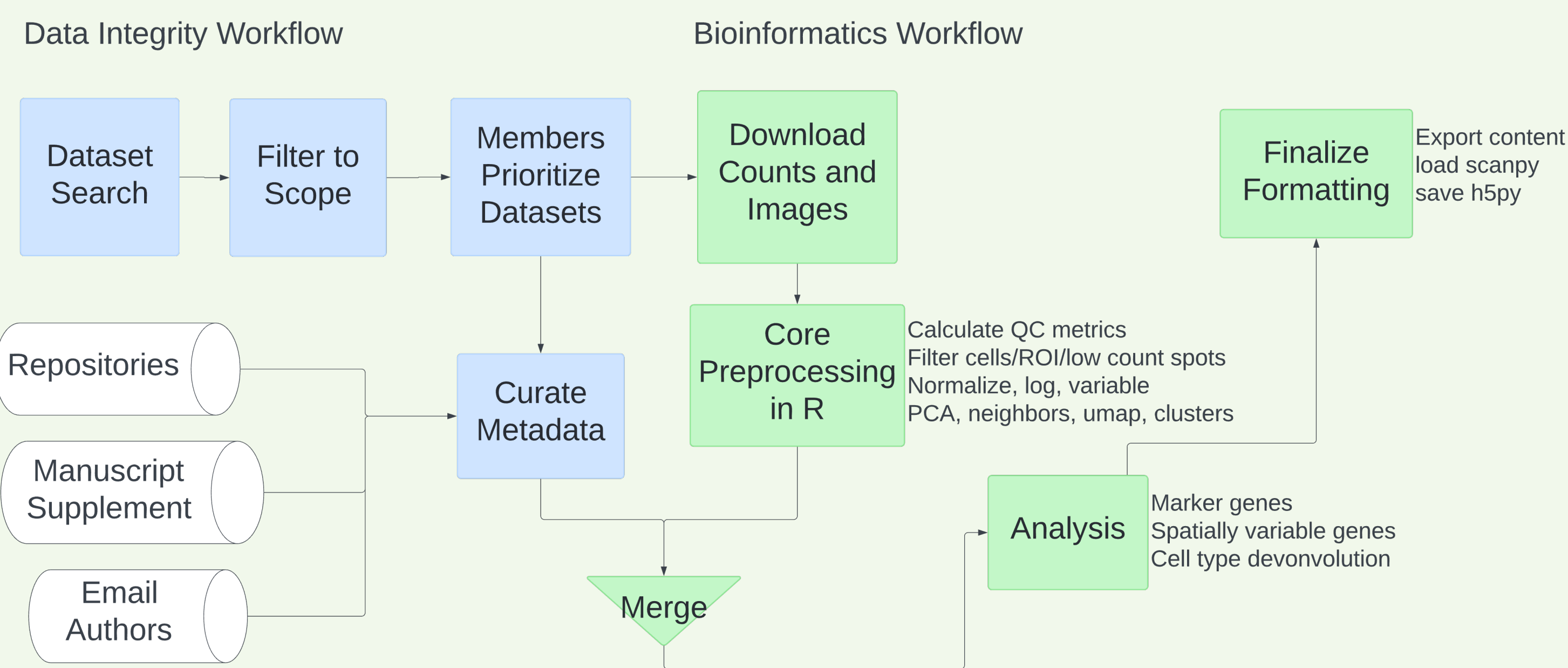
Rancho Initiatives and Consortia

As experienced consortium organizers Rancho plays a dynamic and rewarding role that requires leadership, organization, and a passion for bringing people together towards a shared vision.
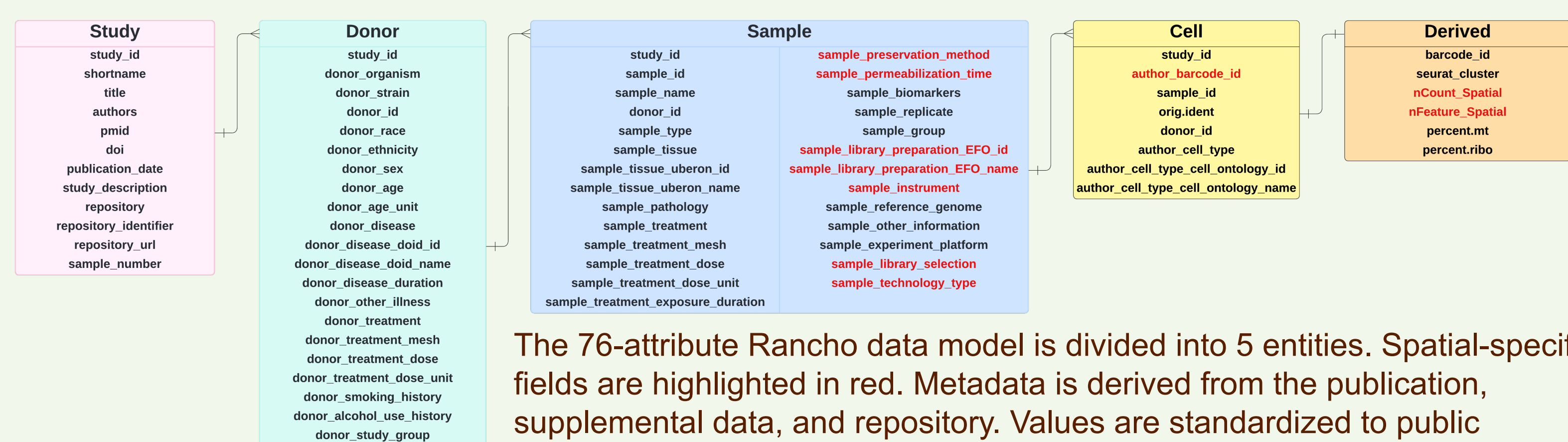
Rancho collaborates with Members to identify their goals and find solutions to ensure we are delivering in line with consortium needs. Through strong planning and communication, we collectively plan, manage timelines, budgets and the logistics for the consortium to keep everything running smoothly.

Ultimately, our goal is to see the collective efforts of the consortium produce impactful results, drive innovation of our Members, and save lives through data.

Data Ingestion Workflow

Data Integrity Workflow: Dataset Search → Filter to Scope → Members Prioritize Datasets

Bioinformatics Workflow: Download Counts and Images → Finalize Formatting (Export content, load scanpy, save h5py)

Repositories / Manuscript Supplement / Email Authors → Curate Metadata

Core Preprocessing in R (Calculate QC metrics, Filter cells/ROI/low count spots, Normalize, log, variable, PCA, neighbors, umap, clusters)

Merge → Analysis (Marker genes, Spatially variable genes, Cell type devonvolution)
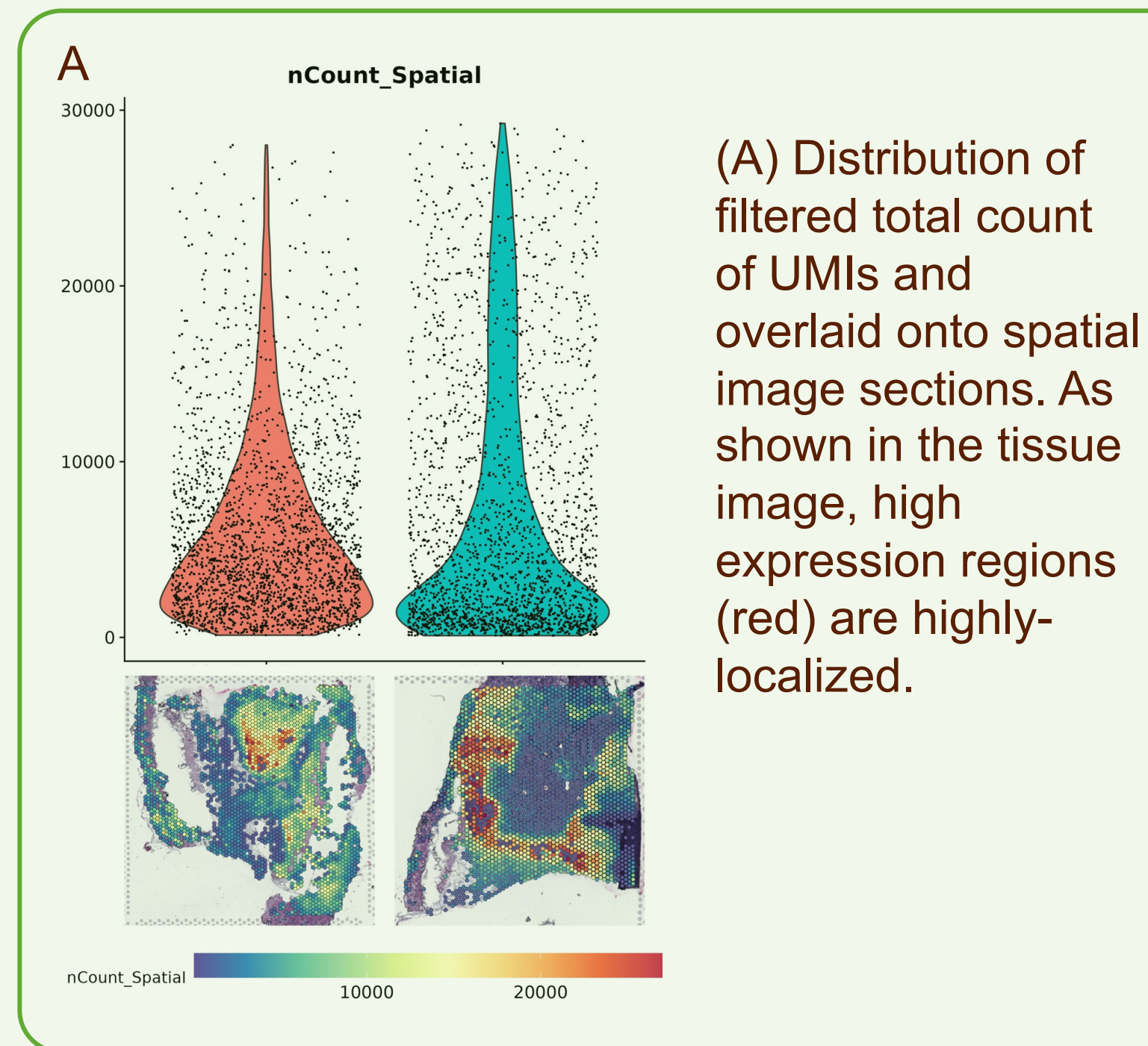
Spatial Transcriptomic Data Model

The 76-attribute Rancho data model is divided into 5 entities. Spatial-specific fields are highlighted in red. Metadata is derived from the publication, supplemental data, and repository. Values are standardized to public ontologies or a controlled vocabulary.

Pilot Datasets Analysis

(A) Distribution of filtered total count of UMIs and overlaid onto spatial image sections. As shown in the tissue image, high expression regions (red) are highly-localized.

After kicking off the Spatial Innovation Initiative, members selected an initial dataset to perform a pilot analysis. The aim of this pilot was to finalize the processing workflow and provide deliverable files to confirm suitability of object features and format.

This pilot analysis included:
- 10x Visium and Nanostring CosMx datasets.
- Multiple healthy and diseased tissue samples.
- Delivery in both Seurat and scanpy formats.
- Additional files with spatially variable gene list and QC plots.
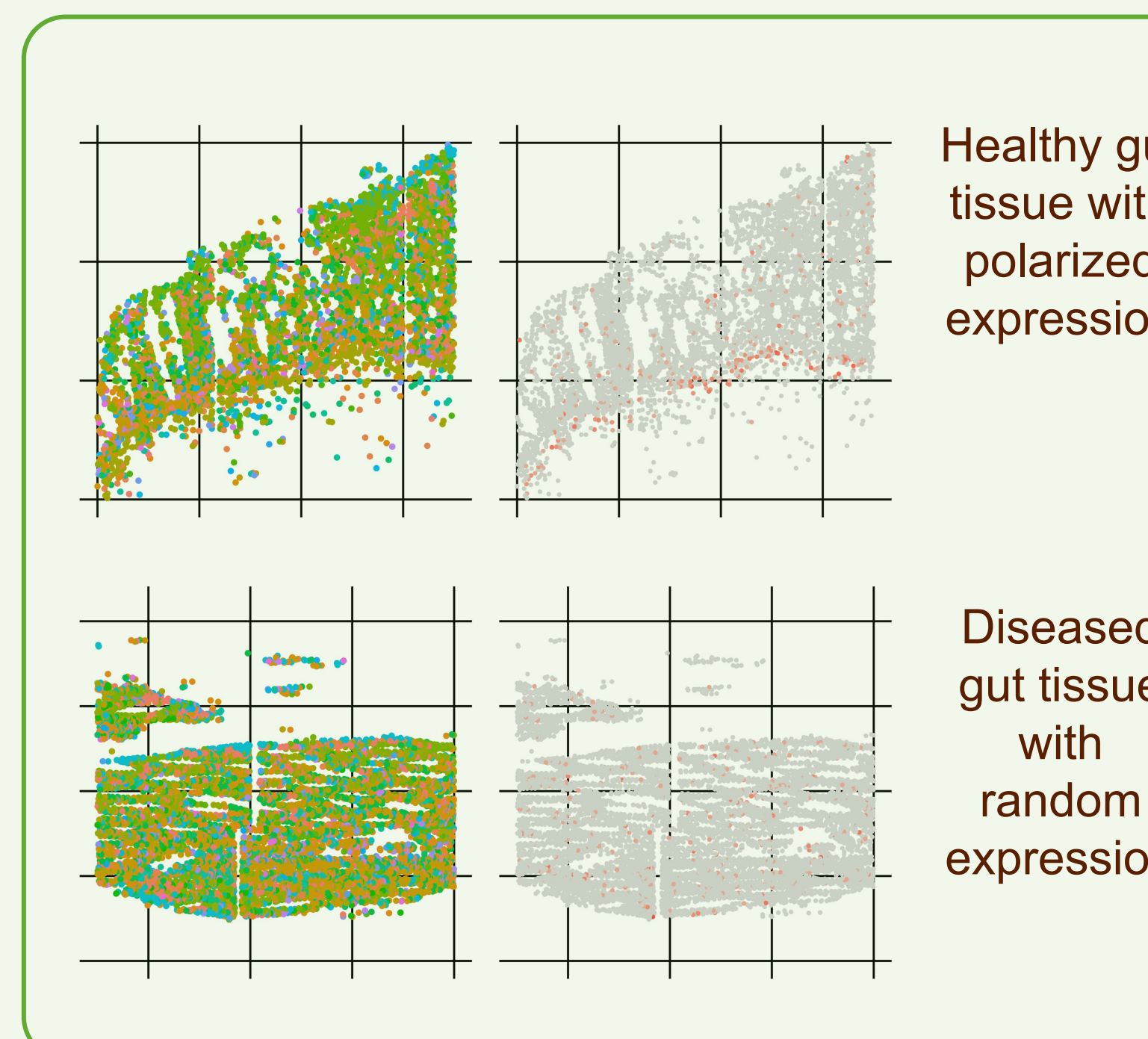
Healthy gut tissue with polarized expression

Diseased gut tissue with random expression
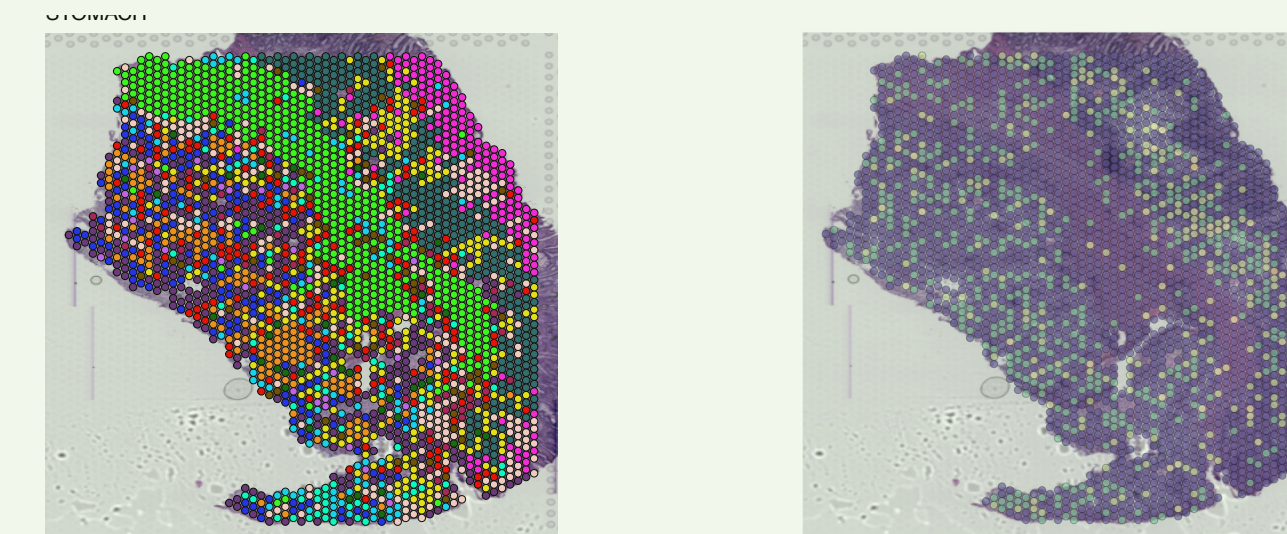
10x Visium data showing image sectioning from the breast tissue overlaid with cluster (left) and marker gene expression (right).

Next steps

After finalizing process requirements and deliverable format we are now working on the main ingestion phase. This aims to deliver 10 priority datasets defined by each member.

When additional members join this initiative, their prioritized datasets will be ingested and added as deliverables for all initiative members.

Deliverables

Deliverables
Seurat object (RDS file)
Python anndata object (h5ad file)
Metadata by entity (xlsx)
Spatially variable gene table (csv)
QC plots (png)

Contact Us

Rancho biosciences