Automated Discovery and Harmonization of Public Spatial Transcriptomic Datasets Across Multiple Cancer Indications

Amrita Bhattacharya[1], Kenneth Chan[1], Dzmitry Fedarovich[1], Anne Cooley[1], Corey Oravetz[1], Kai Ding[2], Emre Arslan[2], Somnath Bandyopadhyay[2]

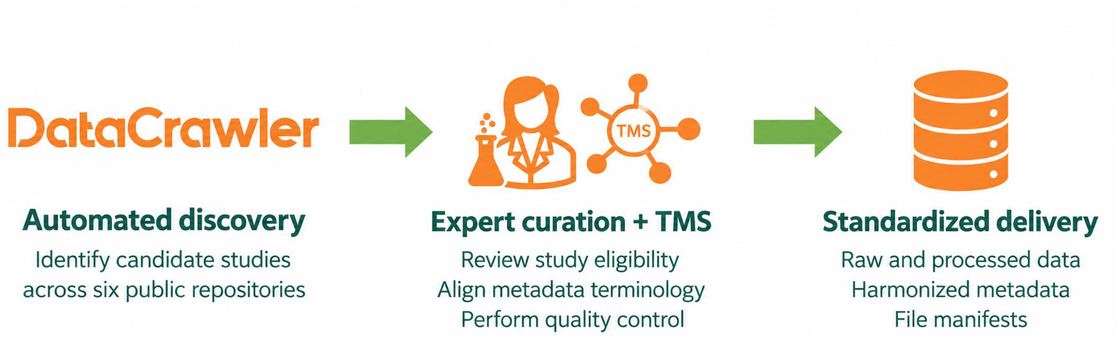
[1]Rancho Biosciences, [2]Takeda

Background and Objective

Spatial transcriptomics reveals how malignant, stromal, and immune compartments are organized within tumors. Public studies can extend this value across indications, but fragmented repositories, platform-specific formats, and inconsistent metadata limit reuse.

Objective: Build a scalable, traceable workflow to discover and harmonize public oncology spatial transcriptomic datasets for downstream translational research

Workflow



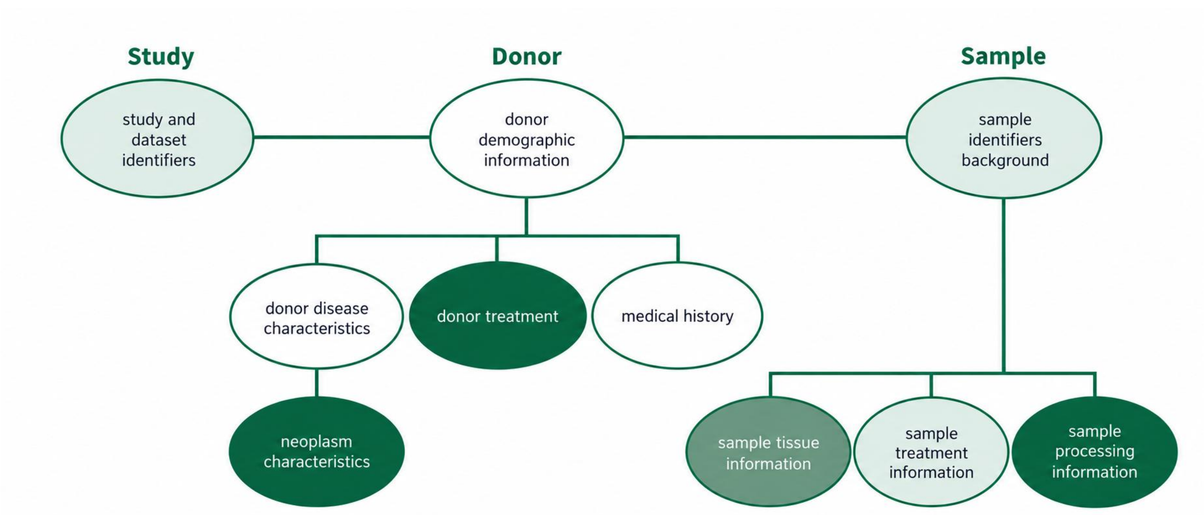
Phase 1: Discovery of candidate datasets

- Disease Indications: 7 Thoracic and Gastrointestinal Cancer Indications
- Repositories: GEO, EGA, dbGAP, Array Express, SRA, Zenodo
- Spatial Platforms: 10x Visium and Xenium, Nanostring GeoMx© and CosMx ©

Data Crawler output: Over 200 potential datasets were identified by the Data Crawler. This candidate inventory was manually reviewed for *eligibility criteria* including *specific tissue type, sample size, and technology* used to produce a list of 46 candidate studies that met requirements.

Phase 2a: Metadata Curation Shortlisted Datasets

A custom 3 entity, 102 attribute data model was developed in collaboration with end users to capture client needs. Metadata was curated from the source repository files, associated publications and if necessary, by
reaching out to authors of the study to gather key information that might be missing.



Phase 2b: Harmonization of Shortlisted Datasets

Curators standardized study-, donor-, and sample-level metadata, aligned biological terms to controlled vocabularies (eg: MONDO for disease, EFO for Technology), and reviewed completeness and consistency. Rancho TMS supports terminology alignment through ontology mapping, custom mapping rules, and API-enabled workflows.

Human expertise + ontology-aware tools

Scientific context remains central while automation improves consistency and scale.

Phase 3: Data Download

Raw and author-processed data were downloaded and delivered using a standardized cloud-based structure.

Results and Deliverables

Automated crawling enabled large-scale, systematic identification of spatial transcriptomic studies across the targeted cancer indications, from which 25 studies were prioritized for detailed curation.

- The curated dataset was enriched for thoracic and aerodigestive tissues, with lung and head-and-neck samples comprising the largest tissue groups, alongside substantial representation of liver, pancreas, colorectum, and stomach.

- Samples were predominantly derived from malignant tumors, with a notable proportion representing advanced and metastatic disease. Tumor-proximal and immune-relevant tissues, including lymph node and peritoneal samples, were also captured.

- Data delivery comprised 27 datasets totaling approximately 4 TB and more than 8,500 files, organized with manifest files to support traceability and data integrity.

Curated Collection at a Glance

25	27	>8,500	~4 TB
studies prioritized	datasets delivered	files organized	data delivered

Conclusion

SYSTEMATIC DISCOVERY

Automated crawling broadened and standardized identification across six repositories.

SCIENTIFIC HARMONIZATION

Expert curation and terminology alignment made metadata comparable across studies and platforms.

REUSABLE DELIVERY

Structured cloud delivery and manifests preserved provenance across 27 datasets and more than 8,500 files.

The Rancho Data Crawler converts a fragmented public-data landscape into a reviewable study inventory. Coupled with expert curation and ontology-aware harmonization, it supports consistent, traceable, analysis-ready spatial transcriptomic collections.

All trademarks are the property of their respective owners

Acknowledgements: The authors thank Sondra Kopyscinski for helpful feedback on the poster draft.

Confidentiality Level: GENERAL/INTERNAL