Integrating Bioinformatics and Metadata Harmonization Pipelines to Ensure High-Quality, AI-Ready Datasets in the Single-Cell Data Science Consortium Initiative

SINGLE CELL DATA SCIENCE CONSORTIUM 2

Dimitris Konstantopoulos, Anne Cooley, Panagiotis C. Agioutantis, Andrew Hill, Yang Hu, Mehmet Tekman, Sondra Kopyscinski, Cynthia J. Grondin, Kenneth Chan, Amrita Bhattacharya, Dzmitry Fedarovich, Andy Hope, Nicole Leyland, Dan Rozelle

Rancho BioSciences, LLC

Background

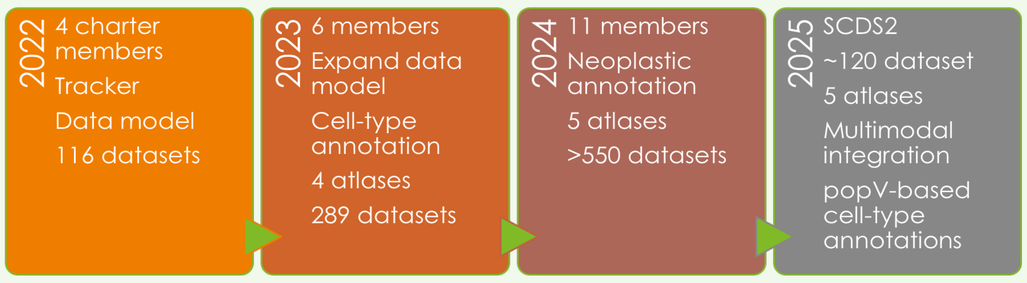
To date the Single-Cell Data Science Consortium (SCDS) has delivered over 800 publications, 1,000 datasets, and **100 million** annotated single cells. In addition to landmark datasets such as the Tabula Sapiens v2, creation of **9 cell-type atlases** has already driven new insights in the biomedical space.

Building on this initial success, consortium work includes refinement of the data model for alignment with Rancho's hyper-model, cutting edge consensus cell-type annotation, and even more automation to deliver exceptional quality AI-ready datasets.

The SCDS data model now includes 83-attributes across 5 concept entities. As an **AI-ready data product**, these datasets accelerate downstream research. A robust harmonization pipeline combines automated mapping with expert subject matter expert curation which includes specifications to public ontologies (DOID, UBERON, CL, MONDO, MeSH, EFO) and custom controlled vocabularies. The OMICs reprocessing pipeline ensures standardized and batch-corrected expression data, and consistently high-quality, well-integrated, and biologically meaningful content. In 2025, SCDS2 is expected to harmonize and process over **120 datasets** and generate **5 atlases.**

What's New in 2025?

- Kick-off of SCDS with 6 members
- Extension of data model to 5 entities and new ontologies
- Unsupervised, data-driven Quality Control pipeline
- New cell-type annotation pipeline using consensus voting (popV)
- Expanding our multi-OMIC analysis to additional modalities (Perturb-seq)
- 5 cell-type atlases from healthy tissue and disease models

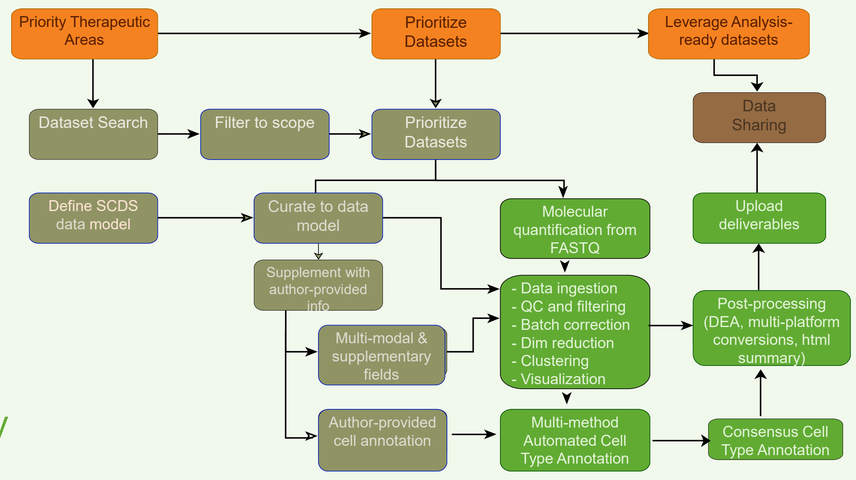


Data Ingestion Workflow

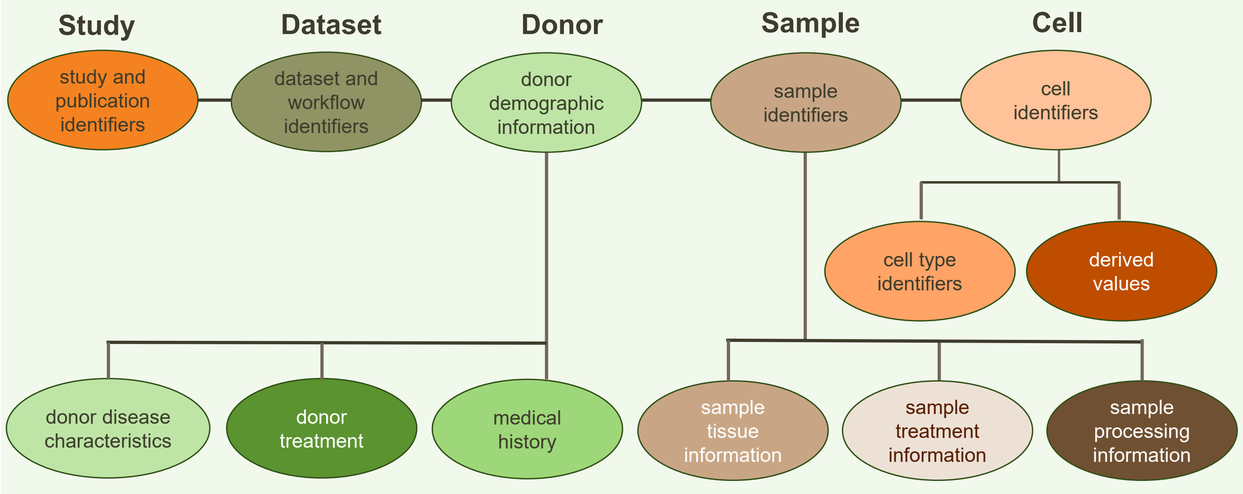
Members
- Define Priorities

Data Integrity
- Maintain catalog
- Prescreen datasets
- Harmonize sample-level metadata

Bioinformatics
- Provide AI/ML-ready datasets



SCDS Data Model



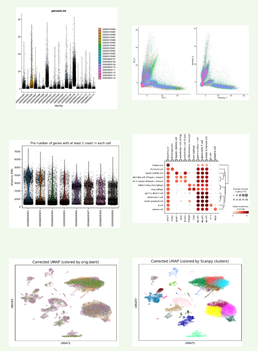
We developed a **custom 5 entity, 90 attribute data** model that is flexible enough to be used for a wide range of experimental purposes

Deliverable Format

Deliverables included

Name	Format
Scanpy analysis object	h5ad
Seurat analysis object	RDS
Metadata table	csv
DGE top table	xlsx
Metadata workbook	xlsx
README	txt
QC directory	misc
Manifest	json

Example summary plots



Datasets are delivered in modular collections and are provided to members in-perpetuity without subscription or limitation.
Analysis-ready datasets include both python and R compatible analysis objects, allowing researchers to work with the data in their language of choice.
Comprehensive QC and primary analysis results are included.

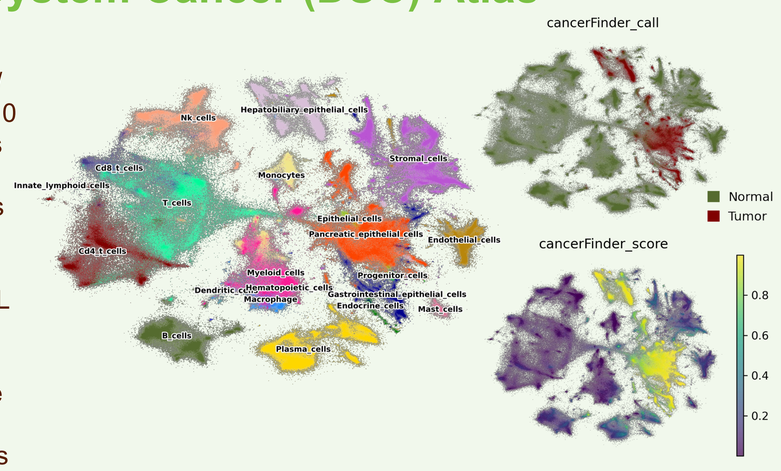
9M cells Autoimmune and Inflammatory Diseases	**24.7M cells** Cancer and Neoplasms	**3.5M cells** Cardiovascular and Blood Disorders	**880k cells** Infectious Diseases
• Crohn's disease • Ulcerative colitis • Psoriatic arthritis • Systemic lupus erythematosus • Atopic dermatitis	• Chronic myeloid leukemia • Pancreatic ductal adenocarcinoma • Lung non-small carcinoma • Glioblastoma • Hepatocellular carcinoma	• Dilated cardiomyopathy • Hypertrophic cardiomyopathy • Cerebrovascular disease • Cardiac arrest • Acute myocardial infarction	• Hepatitis B • Bacterial sepsis • E. Coli infection • HIV disease • COVID-19
1.1M cells Metabolic and Endocrine Diseases	**11M cells** Neurological and Psychiatric Disorders	**2.1M** Respiratory Diseases (non-oncology)	**50.3M cells** Other Conditions and Disorders
• Type 1 diabetes mellitus • Type 2 diabetes mellitus • Obesity • Non-alcoholic fatty liver • Acute pancreatitis	• Alzheimer's disease • Parkinson's disease • Multiple sclerosis • Huntington's disease • Diabetic neuropathy	• Idiopathic pulmonary fibrosis • chronic obstructive pulmonary disease • Allergic asthma • Chronic asthma	• Healthy subject • Unannotated • Endometriosis • Injury • Stroke

Rapid SCDS Atlas Development

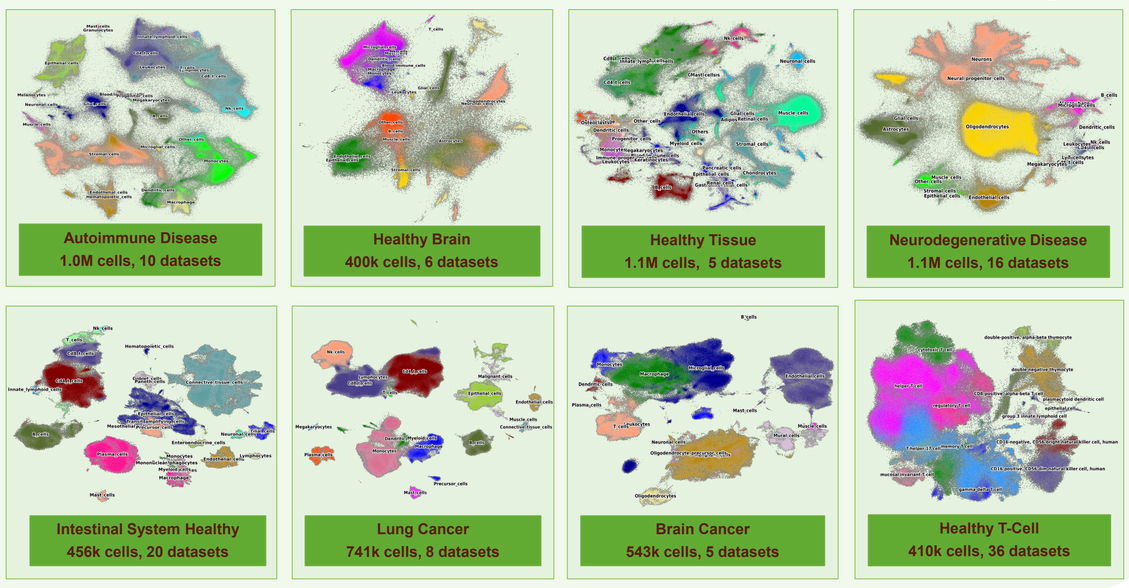
The Rancho scientific staff generates atlases by integrating standardized, analysis-ready datasets across specific tissues to provide insights into cellular heterogeneity across various conditions such as cancer, autoimmune and infectious diseases, thus supporting identification of new therapeutic targets.

1.6M Cell Digestive System Cancer (DSC) Atlas

Our DSC Atlas demonstrates SCDS flexibility. By integrating 18 datasets / 303 samples / 1.6 M cells spanning 10 cancers and 6 tissues, the DSC atlas enables comparative analyses of tumor and microenvironmental states in a unified framework.
We further annotated cell-level malignancy using a marker-based ML classifier (CancerFinder), separating malignant from non-malignant compartments and revealing discrete cancer-cell clusters to support biomarker exploration and hypothesis generation.



SCDS scRNA-seq Atlas catalogue



Conclusions

Curated datasets and atlases delivered as part of this consortium are valuable resources that increase the accuracy and efficiency of downstream analyses, accelerate reproducible science and facilitate joint analysis of public data.

Rancho BioSciences