Accelerating Biomedical Discovery with Public Data: Single-Cell Data Science Consortium Harnesses Collective Expertise to Deliver Even More High Value Harmonized Datasets in Year 3

Rancho BioSciences, LLC

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

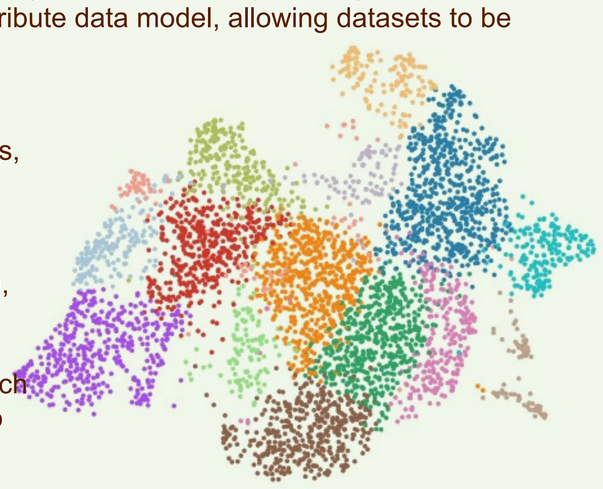
Abstract

Due to their enormous potential for advancing drug discovery, there continues to be an exponential growth in the use of single cell sequencing methods, and a corresponding increase in datasets in publicly available repositories. While these datasets are freely available, they come with hidden costs that hinder the ability of companies to exploit them to their maximum potential. These costs typically result from a lack of metadata standards and significant variation in the processing approach.

The Single Cell Data Science (SCDS) Consortium was formed in 2022 with four charter members as a multi-year effort to harmonize single cell experiments more quickly and cost effectively. This pre-competitive collaboration is being led by Rancho BioSciences, with its deep expertise in single cell data curation, processing, and analysis. To date, SCDS has successfully delivered 417 high-quality datasets with metadata harmonized to a 6 entity, 112 attribute data model, allowing datasets to be used in all kinds of downstream analyses.

In 2024

the consortium has grown to nine members, increasing the return on investment significantly. Due to this collective investment, members are now realizing extensive metadata curation, reprocessing, and cell type annotation for an excellent value per dataset, and more cost effective than they could achieve on their own at such scale. Additional members are expected to join throughout 2024, further reducing the cost per dataset for all Members.



As well as dataset additions, the consortium has already delivered four tissue, disease and organ-specific reference atlases. This includes atlases focused on autoimmune diseased cells, neurodegenerative disease cells, healthy tissues, and a healthy brain cell atlas. Each atlas is composed of a collection of cells from the individual source datasets and exceed 1 million cells each. This exciting work is already accelerating reproducible science, rapid discovery, and joint analysis of valuable public data.

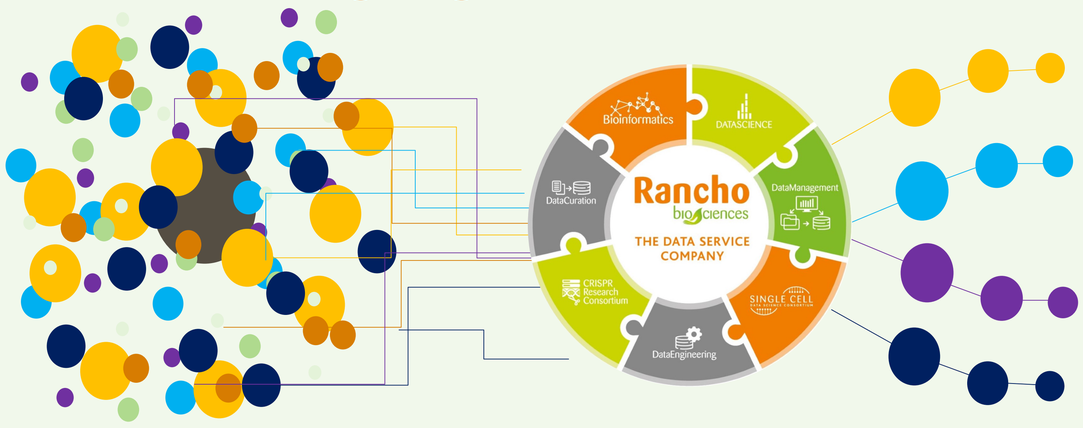
Challenges For Pharma And Biotech

Lack of Standardization
Makes aggregation and meaningful re-use of the data on a larger scale difficult and very time-consuming. Batch correction effects need to be addressed.

Explosion of new analysis algorithms
Monitoring and staying current with the number of new analysis algorithms that continue to be published. Understanding and prioritizing what are valid use cases where new algorithms could be applied to provide meaningful insight

Integration
Combining multiple single cell datasets along with multimodal orthogonal data can provide richer datasets but requires harmonized metadata and processing methods.

Working Together For A Solution

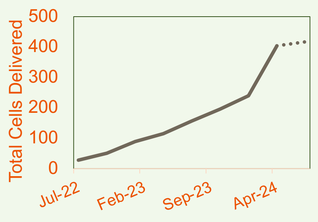


Rancho has created the environment for member collaboration by providing

- Coherent single-cell data model
- Leadership in bioinformatics and pipeline support
- Standardization expertise for transcriptomic metadata
- Facilitation and logistics support

Recent Updates

1. With 3 new members joining since November 2023, we've been able to prioritize delivery of even more datasets.



In February 2024, we delivered the **400th dataset** to consortium members.

Totals delivered to date:

347 Studies
417 Datasets
5,376 Donors
18,754 Samples
47,940,844 Cells

2. Datasets are processed from raw sequencing files, include comprehensive metadata aligned to our extensive transcriptomic data model, and include 3 sources of cell type annotation. These are provided as analysis ready datasets in various formats

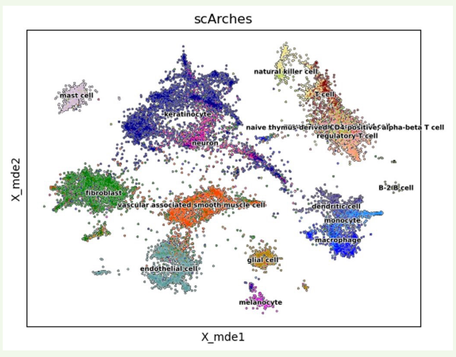
To date SCDS has delivered 48.8M cells

49% Diseased	45% Healthy

This includes a variety of clinical indications

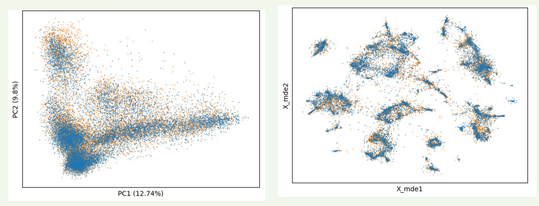
3.1M cells Autoimmune and Inflammatory Diseases • Crohn's disease • ulcerative colitis • psoriatic arthritis • systemic lupus erythematosus • psoriasis	**11.9M** cells Cancer and Neoplasms • chronic myeloid leukemia • pancreatic ductal adenocarcinoma • lung non-small cell carcinoma • glioblastoma • lung adenocarcinoma	**2.2M** cells Cardiovascular and Blood Disorders • dilated cardiomyopathy • hypertrophic cardiomyopathy • cerebrovascular disease • cardiac arrest • acute myocardial infarction	**333k** cells Infectious Diseases • hepatitis B • bacterial sepsis • E. Coli Infection • HIV disease • COVID-19
1.4M cells Metabolic, Endocrine, Nutritional Diseases • diabetic neuropathy • type 1 diabetes mellitus • type 2 diabetes mellitus • Obesity • non-alcoholic fatty liver	**3.6M** cells Neurological and Psychiatric Disorders • Alzheimer's disease • Parkinson's disease • Huntington's disease • frontotemporal dementia	**24k** Respiratory Diseases (non-oncology) • idiopathic pulmonary fibrosis • respiratory failure • asthma • allergic asthma • chronic asthma	**26.1M** cells Other Conditions and Disorders • Healthy Subject • Unannotated • endometriosis • Injury • end stage renal disease

3. During year 2 we developed and delivered four single cell atlases. These have already shown tremendous value as both investigational datasets and reference sources for new datasets annotation.



Integration of Systemic Scleroderma (SS) datasets from Tabib' 21 and Khanna' 22

Atlas	Version	Dataset#	Cell#	Feature#
Autoimmune	v2	10	1,082,275	36,601
Healthy Tissue	v1	5	1,122,197	36,601
Nervous System Disease	v1	7	1,130,027	36,601
Healthy Brain	v1	6	398,952	36,601



Datasets show good overlap, indicating integration was successful

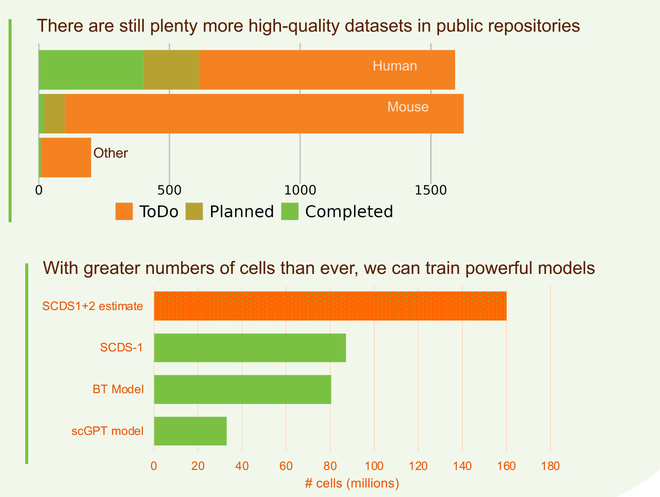
Planning For Year 3 And Beyond (SCDS2)

775 scRNA datasets, an excellent ROI for members

8 integrated atlases

Collaborate with industry leaders

★★★ Highest quality, deeply annotated, fully reprocessed



Rancho bio sciences