Single Cell Data Science Consortium Enables Rapid Analysis of High Value Public Datasets

Dan Rozelle, Sondra Kopyscinski, Nicole Leyland, Andy Hope, Andrew Hill, Panagiotis C. Agioutantis, Dzmitry Fedarovich, Cynthia J. Grondin

Rancho BioSciences, LLC

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 (3 large Pharma and 1 Biotech) as a multi-year effort to harmonize single cell experiments more quickly and cost effectively. This **pre-competitive organization is being led by Rancho BioSciences**, with expertise in single cell data curation, processing, and analysis. To date, SCDS has successfully delivered 115 high-quality datasets with metadata harmonized to a 4 entity, 75 attribute data model.

In 2023 the consortium has increased to six members companies and has increased to six members companies and added several defined functions to the scope. Updates to the ingestion pipeline to adapt to these changing needs is currently in progress and seeks to increase both the processing capacity and features provided to analysts. In addition to dataset additions, we plan to build tissue, disease and organ-specific reference atlases. **Curated datasets delivered as part of this consortium are already accelerating reproducible science, rapid discovery, and joint analysis of valuable public data.**

Challenges for Data Science

1 Sparsity of Data
Artificial zeros, whether real biological phenomena or artifacts of measurement. Many methods to handle sparsity.

2 Correction Effects
Measurements in high throughput technologies are affected by biological and non-biological conditions that need to be "corrected" to avoid producing faulty conclusions

3 Scaling & Resolution
High dimensional data with more cells and more data per cell. What level of resolution is needed to answer a particular question?

4 Integration
Across different types of single-cell measurements. RNA, DNA, protein, methylation, time-points, treatment groups, organisms

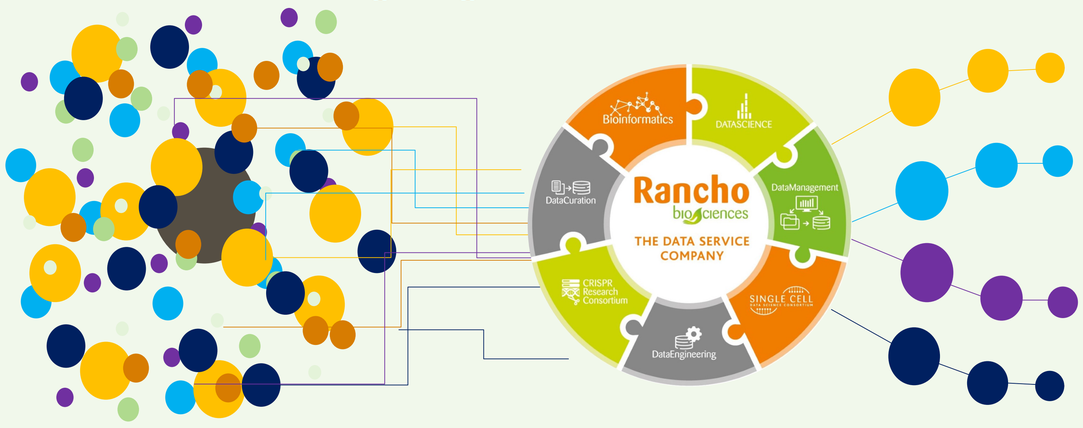
Challenges for Pharma and Biotech

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



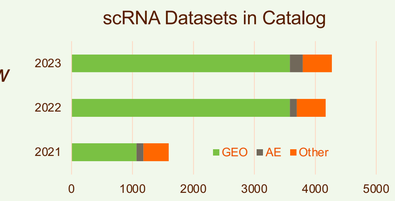
Rancho has created the environment for member collaboration by providing

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

Year 1 Successes

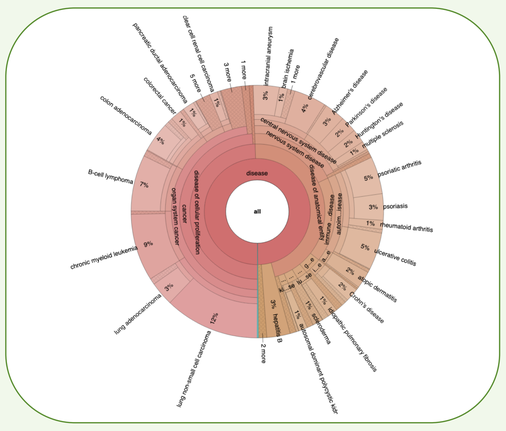
1. Populate tracker application with new single cell datasets. Identify priority datasets for members.

Rancho has developed a simple dataset tracker to allow members to search for single cell datasets and designate their priorities for ingestion.



Top tissues	Top diseases
Brain	Cancer
Cell culture	COVID-19
Lung	Alzheimer's/Parkinson's
Blood	Autoimmune-related
Bone marrow	Digestive System-related
Kidney	

2. High quality metadata is curated to a core transcriptomic data model. Disease, tissue and cell type fields are mapped to official ontologies, supporting both harmonized usage and computational aggregation.



7.5 million diseased cells were ingested

49% are cancer-related with top types from lung, hematological, and g.i. types.

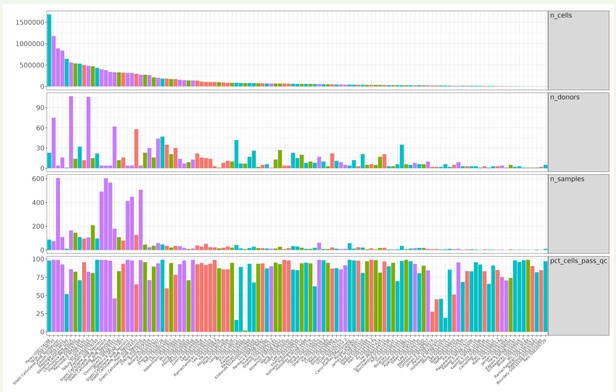
19% are nervous system disease related including HD, PD, AD

18% are immune related (15% autoimmune) such as psoriatic arthritis, ulcerative colitis and dermatitis

45% is derived from g.i. system dysfunction (overlaps with both cancer and immune subsets)

3. SCDS has successfully delivered 115 analysis-ready datasets from 96 studies. Each is provided in 3 formats: Seurat RDS, scanpy anndata, and as a flat-file csv.

batch	studies	datasets	donors	samples	cells	pct_author_annotation
1	23	27	326	746	2,680,147	29.9
2	24	24	251	776	2,981,935	37.6
3	31	38	426	810	4,625,096	22.8
4	18	26	566	4553	7,509,710	76.8
Total	**96**	**115**	**1,569**	**6,885**	**17,796,888**	**49.1%**



Plans for Year 2

Migration to a python-centric ingestion pipeline
For year 1 the bioinformatics team selected R as the ingestion language, mainly for the excellent Seurat interface and extensive analysis modules. The R language was a significant hurdle for many of the larger datasets, adding manual work and process deviation.

Comprehensive automated cell type annotation
While author-provided annotations are becoming more common, most public datasets still lack any cell type annotation, and use inconsistent methods. We will be providing systematic cell type annotation using several reference and ML annotation tools, each with terms mapped to the cell ontology.

Creation of SCDS tissue and disease atlases
With a growing significant corpus of annotated cell types members have expressed an interest in creating tissue and disease specific reference atlas datasets.

Expanded logistics support and versioning of datasets.
Additional features added to each dataset will necessitate new release versions. An improved strategy for delivering, and tracking version information is essential for our members. We will be exploring various options to host and track this information.

Contact Us