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, 16955 Via Del Campo #220, San Diego, CA 92127

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 Q1 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 50 high-quality datasets with metadata harmonized to a 4 entity, 75 attribute data model. These datasets are currently focused on oncology, neurobiology and immunology therapeutic areas based on member priorities. The metadata is combined with reprocessed and normalized data into analysis ready computational resources.

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

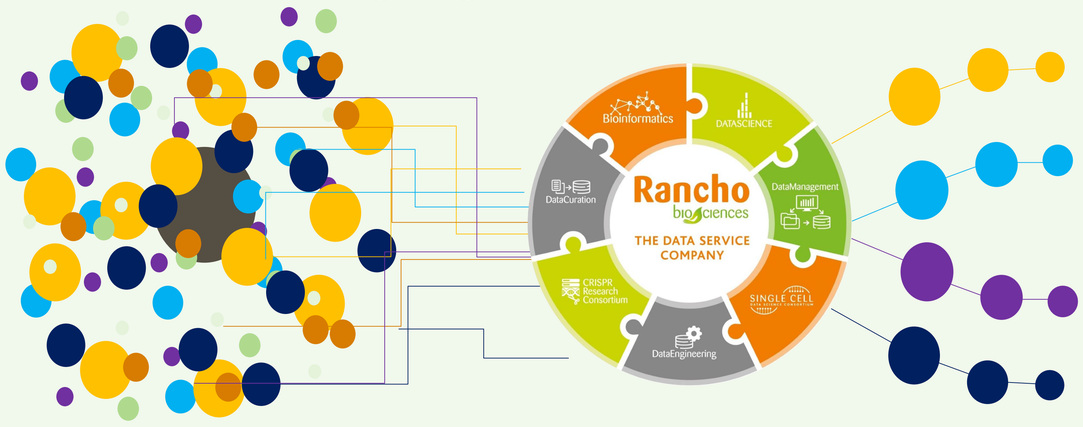
Challenges for Pharma and Biotech

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



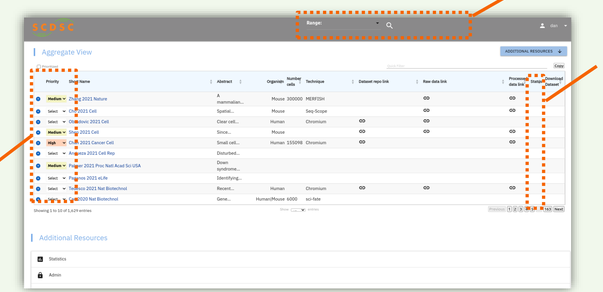
Rancho has created the environment for member collaboration by providing

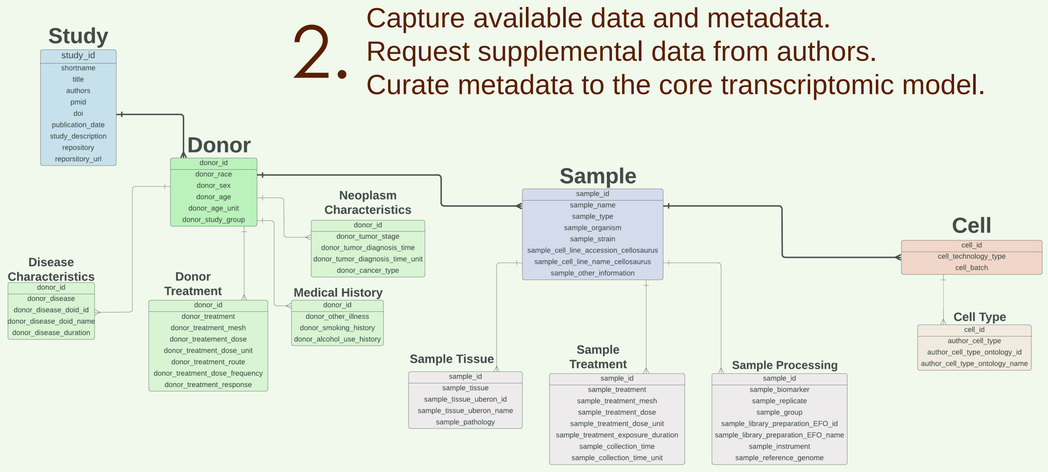
- Coherent single-cell data model
- Leadership in bioinformatics and pipeline support
- Standardization expertise for transcriptomic metadata
- Facilitation and logistics support

Year 1 Dataset Ingestion Workflow

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

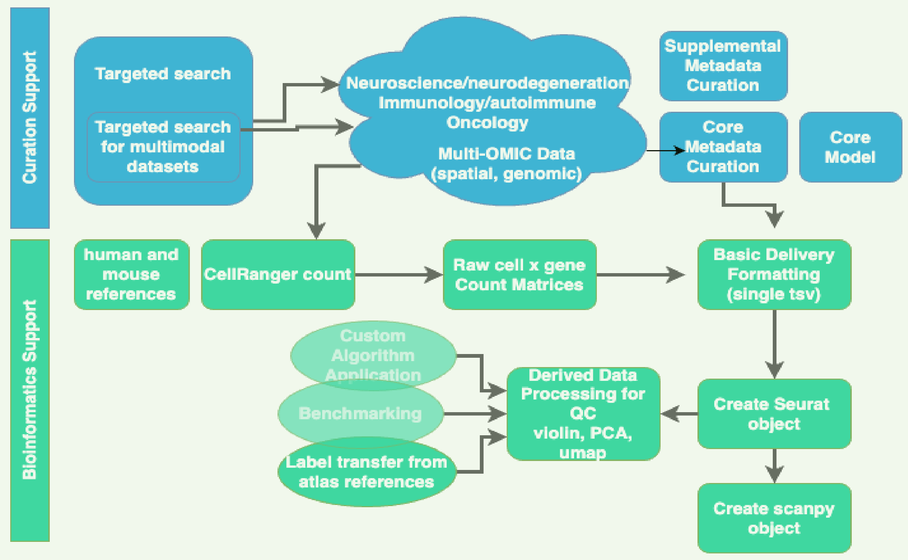
Rancho has developed a simple dataset tracker to allow members to search for single cell datasets and designate their priorities for ingestion.

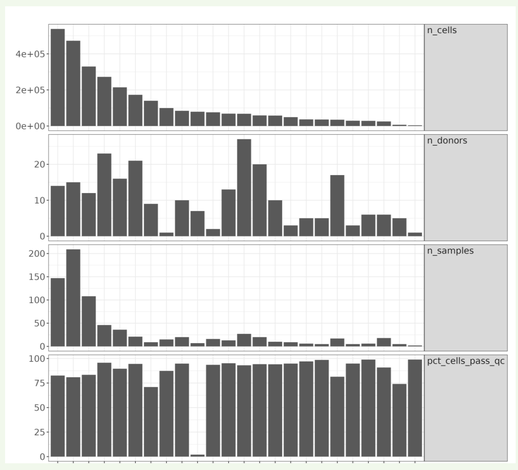




2. Capture available data and metadata.
Request supplemental data from authors.
Curate metadata to the core transcriptomic model.

3. Align FASTQ, quantify, routine preprocessing.
Combine with metadata.
Convert to deliver Seurat, anndata, and flat matrix formats.





4. Deliver analysis-ready datasets.

After harmonized processing and formatting with metadata, each dataset is provided to members.

Members can use these analysis-ready objects with R, Python or ETL into an internal data lake.

Consistent formatting means datasets can be integrated and merged to perform analysis on rare cell types or provide better analysis power.

Years 2+

To date, SCDS has successfully delivered 50 high-quality datasets with metadata harmonized to a 4 entity, 75 attribute data model. Over the next 2.5 years we are on track to deliver more than 300 curated and reprocessed datasets. These will prove valuable resources for consortium members, enabling broad cell type comparisons, rapid target validation assessment, and hypothesis generation capabilities.