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, Yang Hu, Anne Cooley, Amrita Bhattacharya, Kenneth Chan

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 168 high-quality datasets with metadata harmonized to a 4 entity, 99 attribute data model.

In 2023 the consortium has grown to six members 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. As well as dataset additions , we are building 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

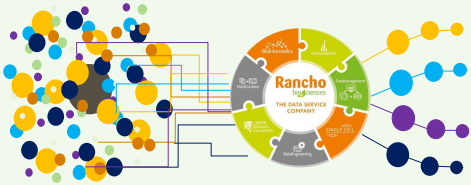
Challenges for Pharma and Biotech

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 2 Updates

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.

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.

To date, 25 million cells have been delivered. 11.3M originating from healthy sample tissues

>500k cells each from
Blood, lung, liver, heart, left ventricle, colon, pleural effusion

>250k cells each from
Skin, bone marrow, lymph node, dermis, skin epidermis, mammary gland, ileum, heart right ventricle, interventricular septum, substantia nigra, pars compacta, pluripotent stem cell, lung parenchyma, apical region of left ventricle, anterior cingulate cortex

4.7 million cells are cancer-related with top types from lung (1.4M), hematological (900k), g.i. (817k), and breast (200k) types.

1.4 million cells are nervous system disease related including HD (134k), PD (229k), AD (271k), MS (84k)

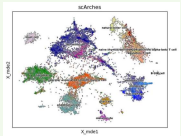
1.5 million cells are immune related (1.3M autoimmune) such as psoriatic arthritis (361k), psoriasis (323k), ulcerative colitis (397k) and dermatitis (166k)

1.1 million cells are derived from g.i. system dysfunction including Crohn's (194k), IBD (44k), cirrhosis (67k) and intestinal cancers (775k)

3. SCDS has successfully delivered 168 analysis-ready datasets from 159 studies. Each is provided in 3 formats: Seurat RDS, scanpy h5 anndata, and as a flat-file csv.

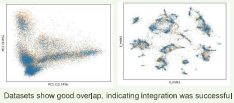
batch	studies	datasets	donors	samples	cells
batch1	23	27	326	746	2680147
batch2	24	24	251	776	2981935
batch3	36	38	426	810	4625096
batch4	24	26	566	4553	7509710
batch5	41	42	777	4470	6367311
batch6	11	11	99	174	807030
Total	**159**	**168**	**2445**	**11529**	**24971229**

4. With a growing list of high-quality harmonized datasets, we have begun work to build a collection of domain and tissue-specific atlas resources for the SCDS Consortium.



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

Our first atlas resource is focused on cells derived from **autoimmune disease** subjects. This work includes optimization of integration methods to combat residual batch effect.



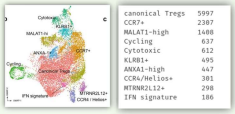
Datasets show good overlap, indicating integration was successful

5. Our new Year 2 pipeline now supports automated cell type annotation with both CellTypist and scArches. This supplements our manually curated author-provided cell type labels with a more systematic level of annotation.

We were able to map author labels to 24.4% of our delivered cells. Most are granular T-cell subsets since sorted for CD3^{+}CD45RA^{-}CD25^{-}CD127^{low} memory Tregs.

To provide systematic annotations we used a pretrained CellTypist model along with a scArches reference dataset.
scArches: Domínguez Conde et al. (2022) Science, Cross-tissue immune cell analysis reveals tissue-specific features in humans

CellTypist: Adult_Human_Blood celltypist.org/organs



Simone_2021_Commun_Biol - Single cell analysis of spondyloarthritis regulatory T cells identifies distinct synovial gene expression patterns and clonal fates (Simone D et al. PMID: 34907325).

Contact Us

Rancho BioSciences