

# 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, Samarth Setty, 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 249 high-quality datasets with metadata harmonized to a 6 entity, 112 attribute data model.


In 2023 the consortium has grown to **seven** 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



- Sparsity of Data**  
Artificial zeros, whether real biological phenomena or artifacts of measurement. Many methods to handle sparsity.
- 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
- Scaling & Resolution**  
High dimensional data with more cells and more data per cell. What level of resolution is needed to answer a particular question?
- Integration**  
Across different types of single-cell measurements. RNA, DNA, protein, methylation, time-points, treatment groups, organisms

## Challenges for Pharma and Biotech

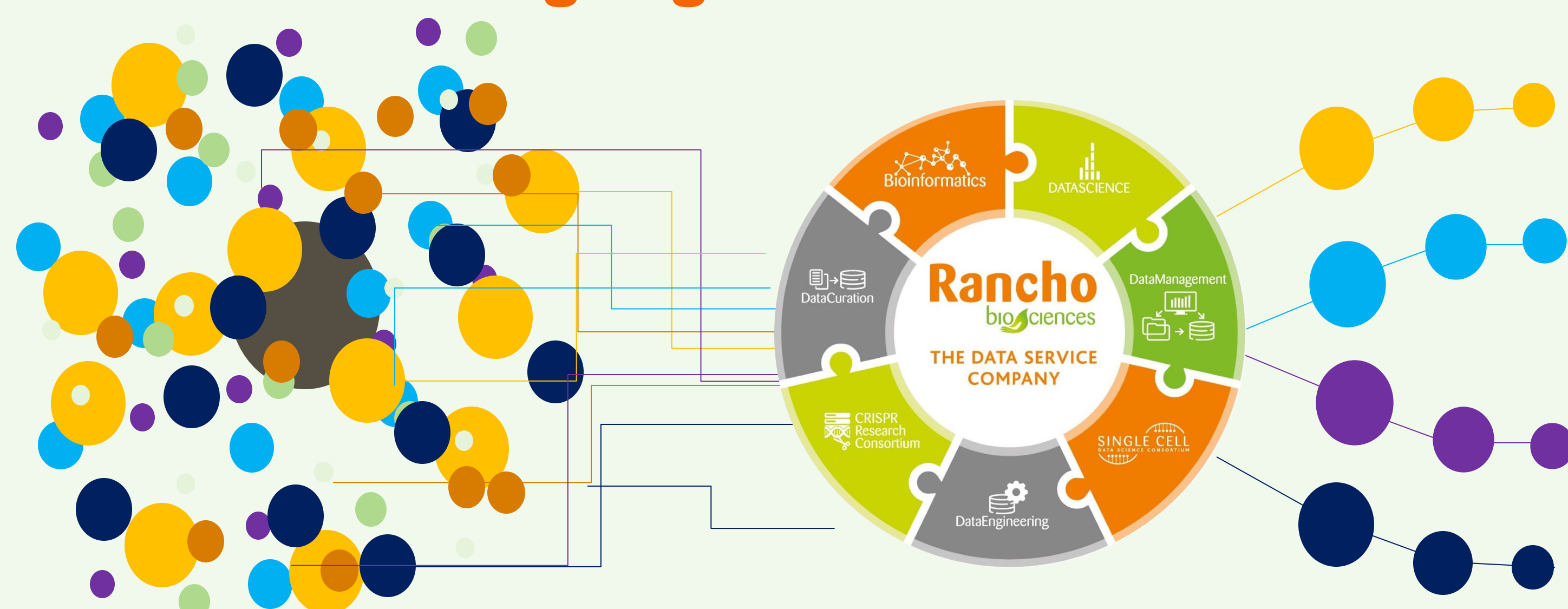


**Lack of Standardization**  
Makes aggregation and meaningful re-use of the data on a larger scale difficult and very time-consuming. Batch correction effects need to be addressed.

**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

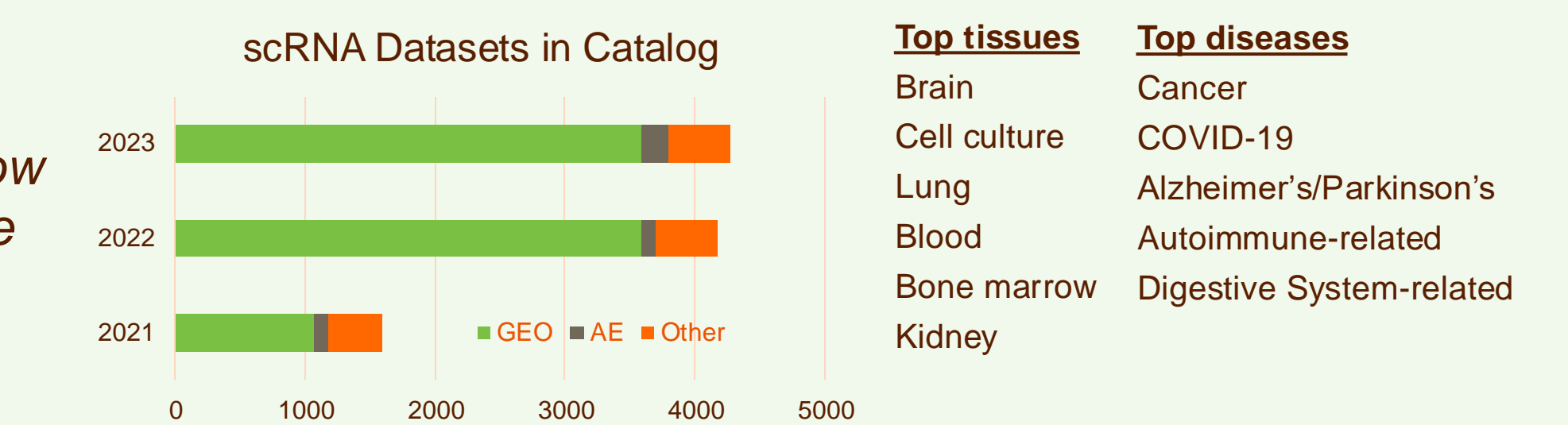
Standardization expertise for transcriptomic metadata

Facilitation and logistics support

## Year 2 Updates

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

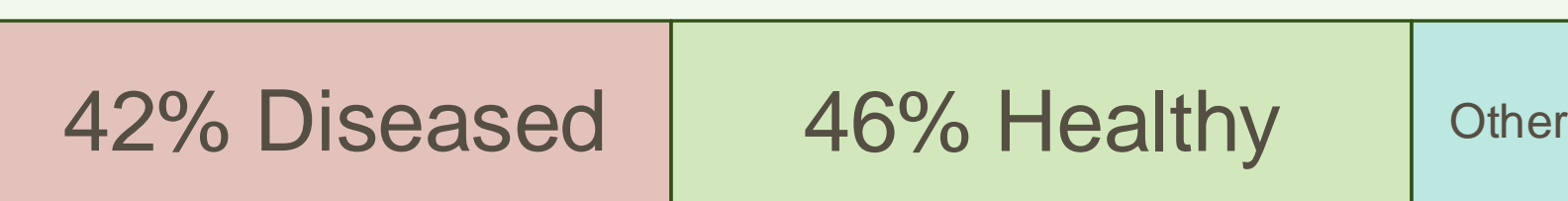
Rancho has developed a simple dataset tracker to allow members to search for single cell datasets and designate their priorities for ingestion.



**Top tissues**  
Brain  
Cancer  
Cell culture  
Lung  
Blood  
Bone marrow  
Kidney

**Top diseases**  
Cancer  
COVID-19  
Alzheimer's/Parkinson's  
Autoimmune-related  
Digestive System-related

- 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, 32 million cells have been delivered. >14M 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)

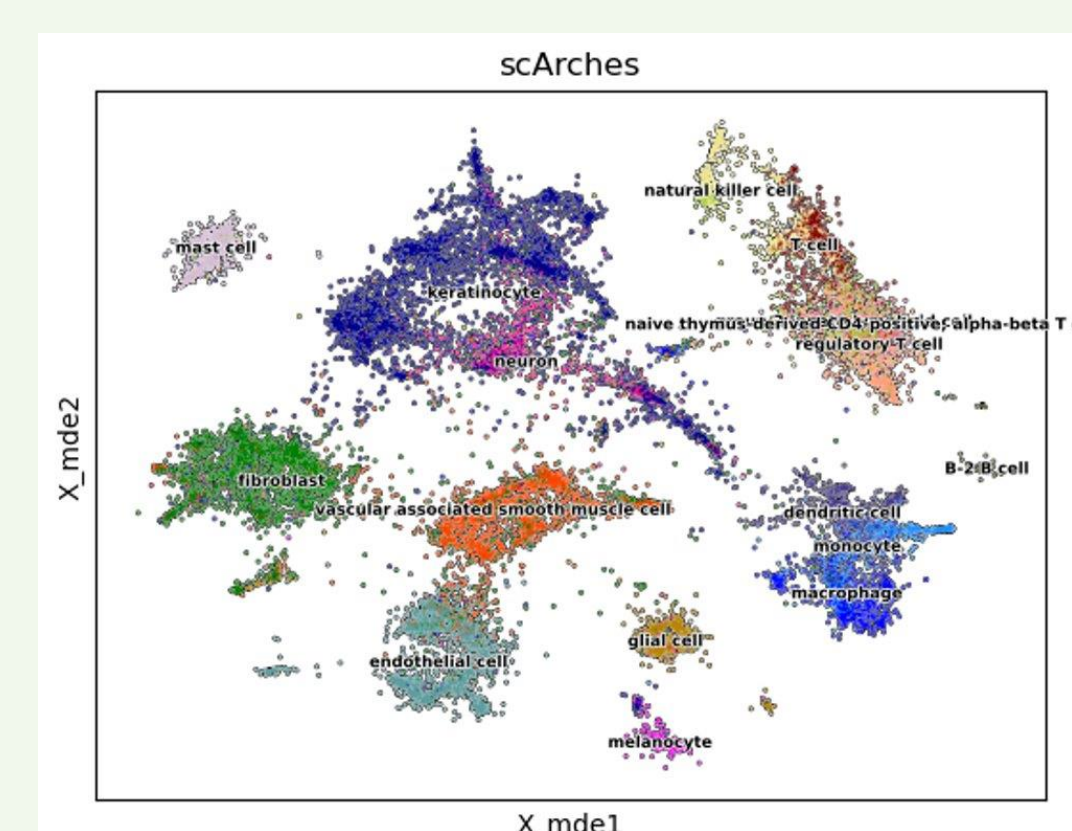
**1.5 million cells are immune related** (1.3M autoimmune) such as psoriatic arthritis (361k), psoriasis (323k), ulcerative colitis (397k) and dermatitis (166k)

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

- SCDS has successfully delivered 249 analysis-ready datasets from 228 studies. Data is provided in a variety of formats: Seurat RDS, scanpy h5 anndata, and as a flat-file csv.

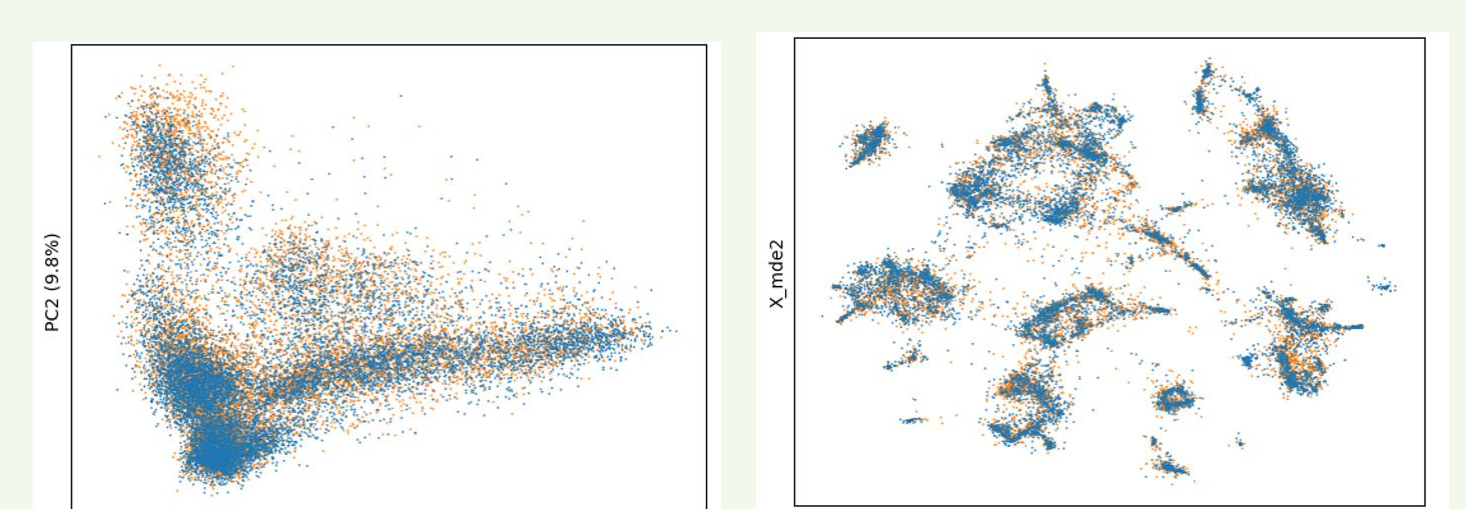
batch	studies	datasets	donors	samples	cells	% author annotation
Batch 1	23	27	326	746	2,680,147	29.9
Batch 2	24	24	251	776	2,981,935	37.6
Batch 3	36	38	426	810	4,625,096	22.8
Batch 4	24	26	566	4,553	7,509,710	76.8
Batch 5	39	42	777	4,470	6,367,311	58.7
Batch 6	35	39	352	1,008	2,642,405	58.9
Batch 7	38	44	698	2296	4,681,098	48.2
Batch 8a	9	9	126	178	463,448	55.1
<b>Total</b>	<b>228</b>	<b>249</b>	<b>3,522</b>	<b>14,837</b>	<b>31,951,150</b>	<b>48.5</b>

- 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 Sclerosis (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.



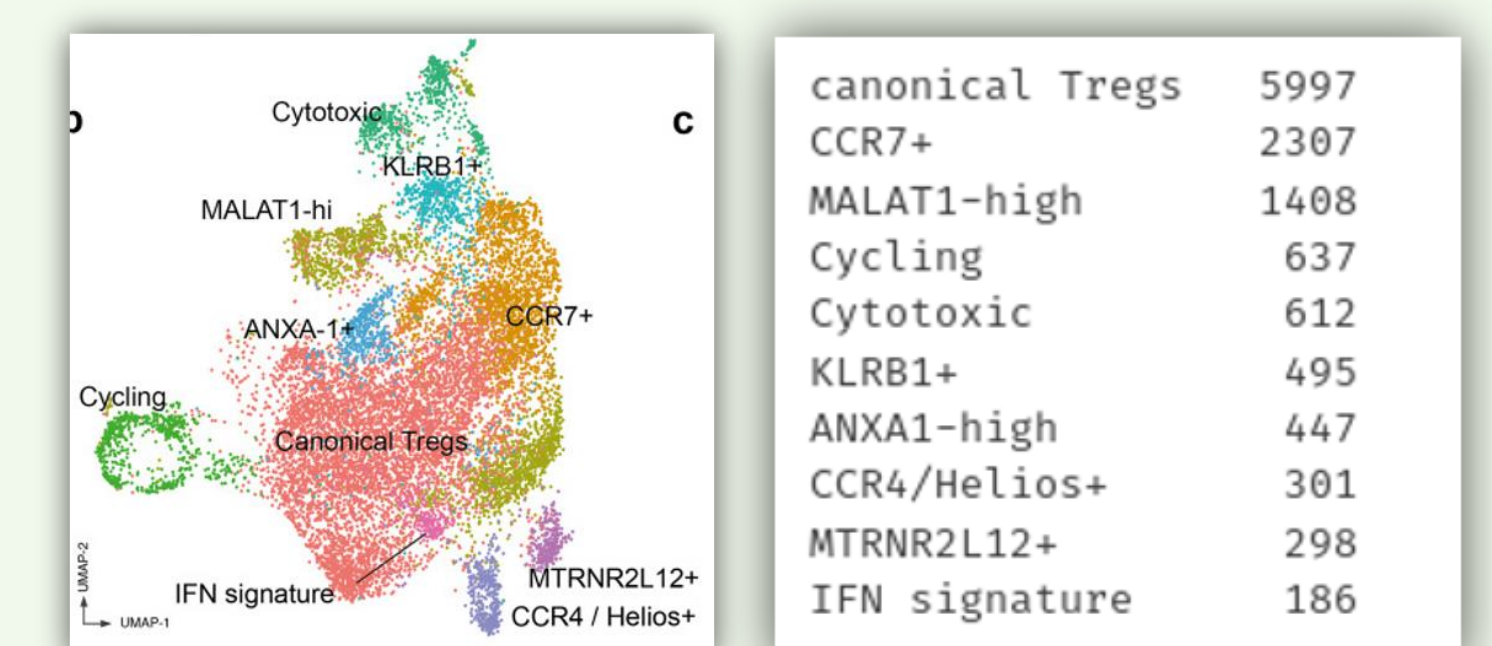
Datasets show good overlap, indicating integration was successful

- 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<sup>+</sup>CD45RA<sup>-</sup>CD25<sup>+</sup>CD127<sup>low</sup> 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](https://pubmed.ncbi.nlm.nih.gov/34907325/)).

Contact Us

