

# Strand's Metadata Curation Portal

## Leveraging Public Data to Drive Precision Research



Scan to log in to scRNA data curation portal.

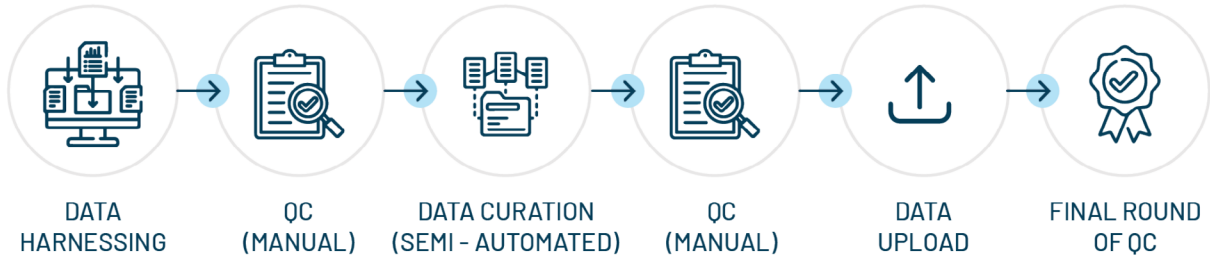
### Challenges Associated with Publicly Available Datasets:

- **Inconsistent Metadata:** Publicly available scRNA-seq datasets often lack standardized metadata or have inconsistent annotations, which makes it difficult to compare or integrate datasets from different sources.
- **Lack of Harmonization:** Datasets often use different terminologies, classifications, and identifiers for cell types, genes, or conditions, making it challenging to harmonize the data for large-scale or integrative analyses.
- **Delays in Research Onset:** Researchers working on a specific disease need to first source relevant datasets, fill in missing pieces, and interpret raw data to extract useful information - all of which can delay a project at the outset.

### Our Solution:

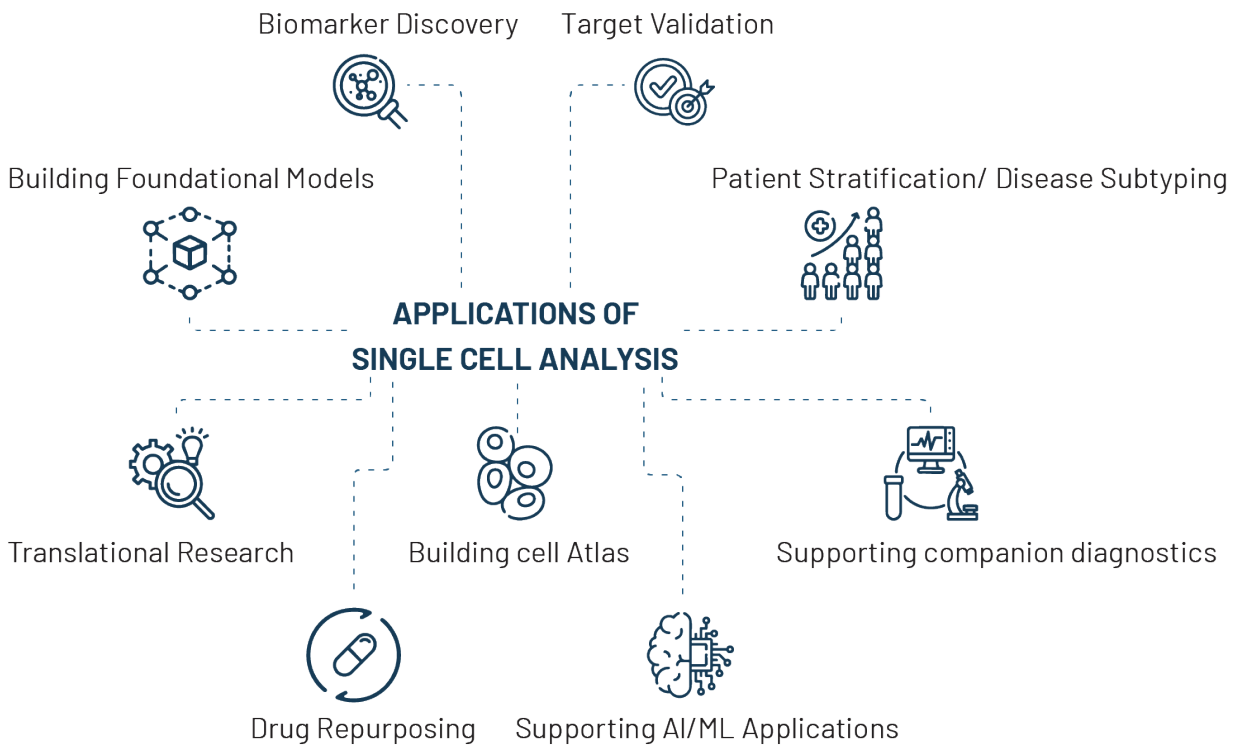
#### Data harnessing, standardization and harmonization:

- Strand has developed a data showcase for scRNAseq datasets with extensive metadata schema and standardized ontologies that can significantly cut down the time, effort, and error associated with each step.
- The showcase focuses on diseases with significant unmet clinical needs like Ulcerative Colitis, Alzheimer's Disease, Frontotemporal Dementia, etc.
- Public datasets, such as those from Gene Expression Omnibus (NCBI), are first harnessed and then meticulously curated and harmonized, with quality control processes applied separately to both stages, using a mix of automation and manual review.
- Turnaround time of 1-2 days.



## Key Features of the Portal:

- **Our datasets have 54 unique metadata fields** organized at 3 levels of curation - study (11), experiment (35), and sample (54) levels.
- 28 metadata fields use controlled vocabulary and **16 metadata fields use defined ontologies**.
- Our portal is equipped with 27 key filters that enable users to set attributes, categories, and values which can pick out relevant datasets within seconds.
- Well organized documentation section with clearly defined SOPs.
- Key metrics of the metadata can be visualized using bar graphs, pie charts and line graphs, and visualizations can be exported by users for easy and effective communication.
- Our showcase, updated monthly, currently contains datasets for Ulcerative Colitis, Alzheimer's disease and Frontotemporal dementia. Other disease types such as Parkinson's disease and Crohn's disease are in the pipeline.





## Example Showcase 1:







### Ulcerative Colitis (UC)

- Ulcerative Colitis is a chronic inflammatory disease with no available treatments for cure at this time. But RNAseq based studies are uncovering the cellular pathways that may be involved in disease pathogenesis.
- Our Ulcerative Colitis showcase currently hosts **44 datasets, 1144 samples, 1,775,138 cells, 65 distinct cell types, 23 different tissue types and 113 metadata fields**, sourced from NCBI's GEO platform.
- About **588 non-inflamed samples and 468 inflamed samples** are included, allowing meaningful comparisons between the two states.
- The samples are primarily from gut tissues, with a few from blood, altogether accounting for 65 unique cell types.

## Example Showcase 2:

### Alzheimer's disease (AD)

- Alzheimer's disease is a progressive, neurodegenerative disorder that gradually impairs memory, cognition, and daily functioning as neurons are damaged and die. Over **416 million** people globally are believed to be on the AD continuum.
- There is currently no cure, and disease pathogenesis is not clearly understood. Its genetic and molecular pathways are being increasingly investigated using scRNA seq studies.
- Our portal showcases data from **48 studies, 607 samples, over 8 million cells, 115 different cell types, 19 different tissue types, 10 model organisms, and over 100 metadata fields**.

	UC Showcase	AD Showcase
 DATASETS	<b>60+</b> sourced from GEO	<b>73+</b> sourced from GEO
 SAMPLES	<b>&gt;1200</b> from 2 organisms*	<b>&gt;800</b> across 10 organisms
 CELLS CURATED	<b>&gt;8.2</b> million	<b>&gt;9.5</b> million
 UNIQUE CELL TYPES	<b>&gt;90</b>	<b>&gt;100</b>
 TISSUE TYPES	<b>30+</b>	<b>20+</b>
 METADATA TYPES	<b>&gt;100</b>	<b>&gt;100</b>

*\*including 588 non-inflamed samples and 468 inflamed samples*

## Applications:

- Enhances data discovery and accessibility.
- Facilitates data analysis and research for downstream applications including biomarker discovery, patient stratification, drug repurposing, etc.
- Supports machine learning, large learning models and AI applications.

**24<sup>+</sup>**

YEARS OF  
EXPERIENCE

**80,000+**  
Genetic Tests  
Reported


**500+** Projects  
Executed for  
Genomics  
Majors Globally

Presence in  
**20+** Countries

**strand** 

 7th Floor, MSR North Tower, #144, Outer Ring Road, Nagavara, Bengaluru - 560045

 +91 9980448044

 [hello@strandls.com](mailto:hello@strandls.com)

 [strandls.com](http://strandls.com)