CASE STUDY

Pharma Company Achieves 4x Faster Target Identification for Inflammatory Disease

A **Boston-based** pharmaceutical company partnered with Elucidata to accelerate target discovery and validation pipeline for inflammatory disease. For the same, they wanted to utilize Elucidata's high-quality curated single-cell RNA sequencing (scRNA-seq) **datasets, data concierge services,** and **expertise** in computational biology.

Problem Statement

The pharmaceutical company sought to **leverage publicly available scRNA-seq data** for identifying/validating gene targets for inflammatory disease in a specific cell type. They wanted to **fast-track the target identification and validation** process which would help them **accelerate their drug discovery** process.

Challenges

The pharmaceutical company encountered multiple challenges which slowed down its discovery process, such as:

- **Finding relevant datasets** for a specific disease type (inflammatory disease).
- **Harmonizing publicly available datasets** to a consistent, analysis-ready format.
- **Lack of expertise** to perform suitable analyses on scRNA-seq data.

Elucidata's Solution

At Elucidata, we approached the problem with a two-fold solution:

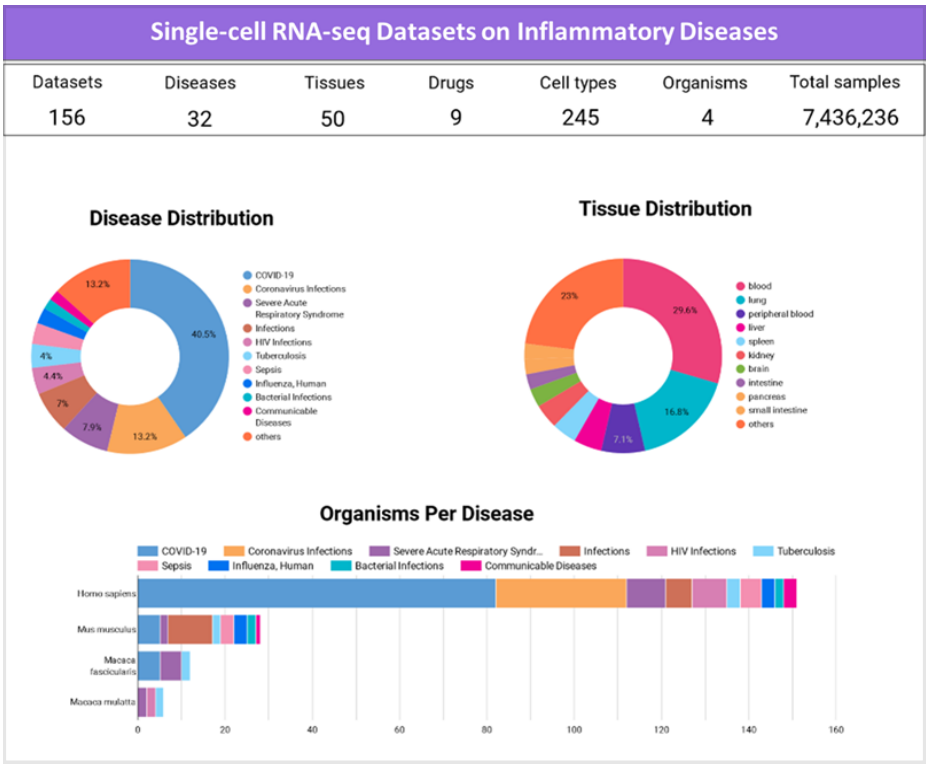
- **Curated data** on our biomedical data platform - Polly, and
- **Elucidata's expertise** in scRNA-seq data management and analysis

Elucidata's data platform, Polly, curates valuable public biomedical data and/or legacy proprietary data into a FAIR (Findable, Accessible, Interoperable, Reusable) resource. It is built on proprietary Bio-NLP technology that cleans and links data at high throughput, with human levels of accuracy, and scales to 26+ R&D data types.

Curated Data on Polly

With Polly, teams have access to a curated collection of RNA-seq datasets, known as OmixAtlas, saving up to 80% of the time spent preparing data for downstream analysis. Datasets on Polly are processed consistently using pipelines, have accurate and complete metadata associated with them, and follow a standard ontology.

With this solution, we empowered the company to find datasets highly relevant to inflammatory disease and preferred cell type of interest.



The single-cell inflammatory disease datasets on Polly.

Expertise in scRNA-seq Data Management and Analysis

At Elucidata, our experts spanning bioinformatics, computational biology, data science, and machine learning domains are familiar with conventional as well as state-of-the-art techniques to analyze RNA-seq data. The team understands that standard RNA-seq approaches cannot be directly applied to scRNA-seq data and are dedicated to providing valid solutions with accurate results without bias or inconsistency.

Conventionally, differential expression analysis is used to analyze RNA sequencing data. It relies on calculating fold change in gene expression between diseased vs. control and p-value to determine the significance of the fold change. However, this analysis is not directly applicable to scRNA-seq data. A few reasons that could account for this are:

- The fold change may represent an **average gene expression profile** rather than that of the desired cell population.
- Filtering genes based on p-value can be **unreliable** as large sample sizes (individual cells in the case of scRNA-seq) can yield significant p-values even with a small fold change.

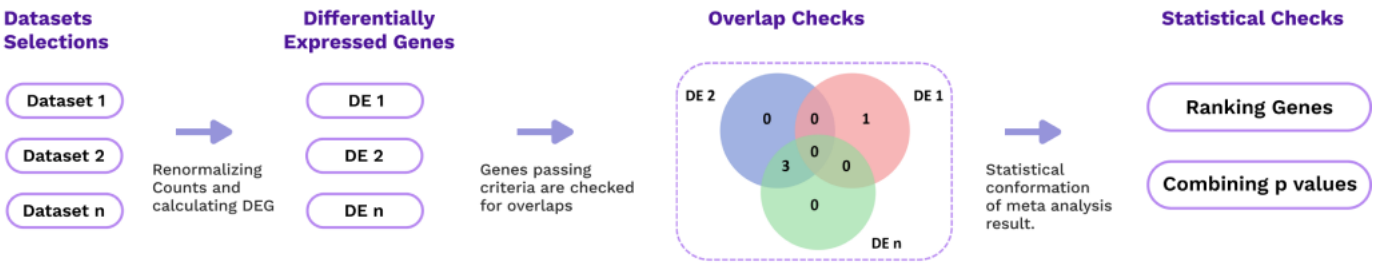
Both of the aforementioned factors could potentially skew the results of differential expression analysis since they fail to consider the inherent variability at the individual cell level in single-cell data.

Since our experts were familiar with the drawbacks of conventionally used differential expression techniques, they suggested that meta-analysis would be the best way to move ahead to ensure the target specificity to cell type and disease of interest. Meta-analysis of single-cell transcriptomics data refers to a systematic approach that combines and integrates multiple independent studies to gain a comprehensive understanding of gene expression patterns at the single-cell level. It involves collecting, harmonizing, and integrating transcriptomic data from different studies to derive new insights or validate existing findings.

We took two approaches to meta-analysis - biased and unbiased. Biased analysis was done to validate the targets pre-identified by the company. Unbiased analysis was done to identify new targets. In the unbiased analysis, we used classifier models and correlation scores to accurately identify targets in the desired cell type for inflammatory disease.

Biased Meta-Analysis

We started by validating the targets the company identified by doing a biased analysis. In this analysis, differentially expressed genes were shortlisted from relevant datasets, then the genes were checked for overlaps and statistical significance. Using this strategy, we validated the 5 targets shortlisted by the company.



Workflow of Biased Meta-Analysis

Unbiased Meta-Analysis

Next, we also did an unbiased analysis so no potential targets were left out and to ensure target specificity to cell type and disease of interest. The following are the steps of unbiased meta-analysis:

- **Dataset Selection** - Datasets were selected via Polly's single-cell RNA-seq OmixAtlas based on the client's specified criteria: cell type, disease, and sequencing method
- **Merged Datasets** - Datasets were renormalized, batch-corrected, selected, and integrated using single-cell variational interference (scVI). scVI was used to reduce technical noise and bias. It uses deep neural networks to aggregate cell and gene information by approximating expression distributions while addressing batch effects and sensitivity constraints.The integration was validated with single-cell integration benchmarking (scIB). It was used to evaluate and validate the effectiveness of dataset integration using scVI. Combining datasets reduces technical variations, providing richer insights with greater dependability.
- **Metadata Classes** - The merged datasets have 250,000+ cells. The cells were sorted into:
  - Cell type of interest - Normal and Diseased
  - Other cell types - Normal and Diseased

  Categorizing cells ensures the selection of specific markers for diseased cells.
- **Feature Selection** - The top 250 differentially expressed genes were prioritized under each class. The genes were filtered based on their expression on the cell membrane (as the company was looking for cell surface markers). Then we used Random Forest (RF) classifier model to predict gene/class associations.The RF classifier employs multiple decision trees to generate classification results and improve predictive performance. It can effortlessly manage unbalanced data, outliers, and non-linear features.
- **Surface Marker Prioritization** - After the gene/class association prediction, each gene had 4 correlation scores based on the Point-biserial correlation. Point-biserial correlation score was utilized to evaluate the connection between each gene and its 4 classes. This measure is a metric of association strength between an uninterrupted variable and a binary factor, indicating how strongly related genes were with their respective classes.
- **Validation** - The top 10 genes from the previous step were validated by confirming their expression in 7 other datasets specific to the cell type and inflammatory disease. These genes were then explored for their function and expression through an extensive literature survey. After these validation steps, 5 targets were shortlisted, out of which 1 was common with biased analysis, and 4 were novel.

Workflow of Unbiased Meta-Analysis

Outcome

Elucidata's innovative approach to finding suitable targets has provided significant benefits for the company:

- **Enhanced Workflow Efficiency** - Through a systematic analysis approach, Elucidata has efficiently validated five targets and generated four novel leads for further investigation. This meticulous process has significantly improved the company's research workflow.
- **Accelerated Target Identification** - By leveraging harmonized analysis-ready datasets tailored to their specific disease of interest, Elucidata has achieved an impressive four-fold acceleration in target identification. Within a span of 2.5 months, they have successfully identified and validated targets, a process that typically takes 8-10 months.
- **Proficiency in Analyzing scRNA-seq Data** - Elucidata has demonstrated exceptional expertise in handling and analyzing scRNA-seq data using state-of-the-art classifier models and correlation scores. This proficiency has enabled them to accurately identify and validate targets, establishing their reputation as leaders in the field.

The Impact

Shortlisted **4 novel targets** and **validated 5 pre-identified targets** for further exploration in the inflammatory disease.

Polly's curated single-cell datasets and our team's expertise in single-cell analysis resulted in a **4x acceleration in target identification.**

About Elucidata

Elucidata transforms biological discovery by providing high-quality bulk RNA-seq and single-cell data, among other data types. They support discovery programs at top pharma companies and have 35+ research partners from premier biopharma companies and research labs.
Their FAIR biomedical data platform, Polly, makes data easily findable and more reusable. Elucidata has helped R&D teams scale up and has enabled 10x faster identification of therapeutic assets with high odds of success in the clinic. Having aided the detection of multiple validated drug targets across immunology, oncology, and metabolic disorders, Elucidata looks forward to helping more teams reach their R&D goals quicker!

Request a Demo

For more information, visit Elucidata | Home or reach out to us at info@elucidata.io.

Elucidata

San Francisco | 114 Sansome St, 250, San Francisco, CA 94104, USA
Cambridge | 625 Massachusetts Ave 2nd Floor, Cambridge, MA 02139, USA
Delhi | Avanta Business Center, Saket, New Delhi, Delhi 110017, India
Bangalore | UrbanVault HSR Layout 761, Bengaluru, Karnataka 560102, India