CASE STUDY

Elucidata

Elucidata Transforms RNAi Drug Discovery: 2X Faster Identification of Potential Target Genes and Accelerated Cell Annotation

Overview

A Cambridge-based RNA interference (RNAi) therapeutics company is developing precision medicine for genetically rare diseases. The company focuses on understanding how genes are naturally regulated within cells. Using RNAi, they 'silence' or turn off specific genes that cause or contribute to disease. To achieve their goal, they use in-house single-cell data and pipelines to process these datasets. To improve the robustness of RNAi drug discovery, they collaborated with Elucidata for high-quality AI-ready public datasets, along with a pipeline for re-annotation of scRNASeq datasets.

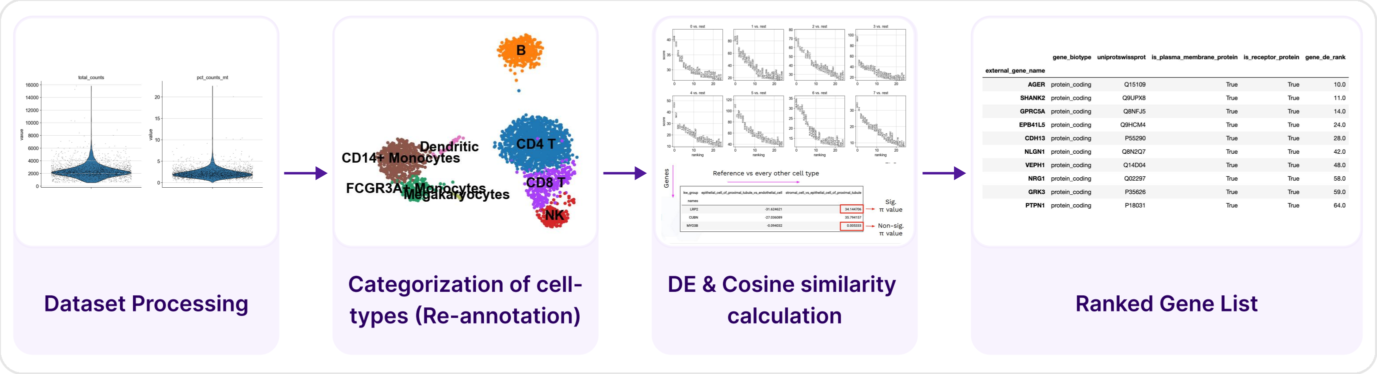
Challenges

To achieve their goals of predicting genes involved in disease for RNAi drug development, the customers required access to high-quality single-cell datasets associated with rare diseases obtained from three organisms (human, mouse, and macaque monkey). This was challenging for the customer's in-house team to accomplish alone due to the following challenges:

1. Finding single-cell datasets in the public domain was time-consuming as it involved navigating multiple repositories and manual reviews in some cases. These datasets were of low quality and not AI-ready for analysis.
2. Even after significant efforts, public datasets suffer from inaccurate identification of cell types based on scRNA-seq data, making it difficult for researchers to generate insight from them.
3. Managing individual similarity files for silenced target genes and conditions (Case vs Control) across various tissues (lung, kidney, skeletal muscle, adrenal, adipose) was challenging due to the lack of a structured storage system, making it difficult for them to efficiently organize and retrieve information.
4. The customer's in-house bioinformaticians were preoccupied with analyzing their prosperity data and had no bandwidth to develop pipelines that could analyze the publicly available dataset.

Elucidata’s Solution

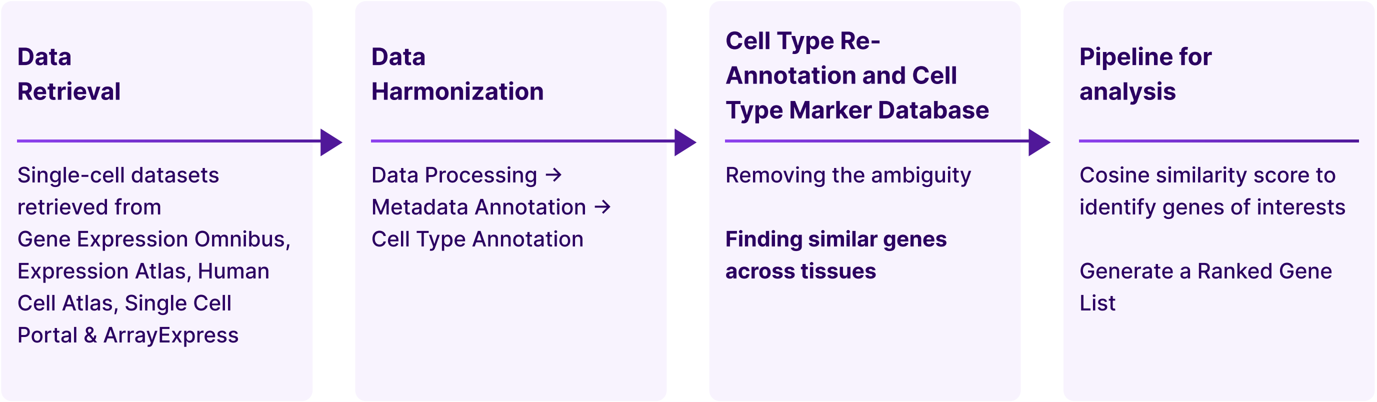
To fulfill the customer's vision of developing RNAi drugs for diseases in 4 therapeutic areas—genetic diseases, cardiometabolic diseases, infectious diseases, and central nervous system (CNS) and ocular diseases. Elucidata assisted them in retrieving high-quality, AI-ready single-cell datasets that are needed to eradicate the affected gene present only in the affected cells from humans, mice, and primates. Given the need for larger-scale validation, the substantial quantity of public data played a key role in complementing the more limited in-house data. Additionally, we developed a pipeline for analyzing these public datasets, aiming to identify and correlate genes with the list of genes associated with disease progression or drug delivery.



external_gene_name	gene_biotype	uniprotswissprot	is_plasma_membrane_protein	is_receptor_protein	gene_de_rank
AGER	protein_coding	Q15109	True	True	10.0
SHANK2	protein_coding	Q9UPX8	True	True	11.0
GPRC5A	protein_coding	Q8NFJ5	True	True	14.0
EPB41L5	protein_coding	Q9HCM4	True	True	24.0
CDH13	protein_coding	P55290	True	True	28.0
NLGN1	protein_coding	Q8N2Q7	True	True	42.0
VEPH1	protein_coding	Q14D04	True	True	48.0
NRG1	protein_coding	Q02297	True	True	58.0
GRK3	protein_coding	P35626	True	True	59.0
PTPN1	protein_coding	P18031	True	True	64.0

Workflow

Elucidata’s Approach



Data Retrieval

Single-cell datasets retrieved from Gene Expression Omnibus, Expression Atlas, Human Cell Atlas, Single Cell Portal & ArrayExpress

Data Harmonization

Data Processing → Metadata Annotation → Cell Type Annotation

Cell Type Re-Annotation and Cell Type Marker Database

Removing the ambiguity

Finding similar genes across tissues

Pipeline for analysis

Cosine similarity score to identify genes of interests

Generate a Ranked Gene List

Flowchart to show our work

Retrieving Relevant Datasets

Elucidata’s data harmonization engine can store metadata from numerous public datasets sourced from diverse repositories. This enables us to effortlessly conduct advanced queries across platforms like Gene Expression Omnibus, Expression Atlas, Human Cell Atlas, Single Cell Portal, EMBL Expression Atlas, and Genotype-Tissue Expression (GTEx) —all in significantly less time than usual. Leveraging programmatic search, we identified over ~1 million cells and 5000 samples associated with rare diseases.

Data Harmonization of Single-cell Datasets

1. Data Processing

Raw, unfiltered single-cell RNA-seq datasets were uploaded onto our proprietary data harmonization engine. Subsequently, our single-cell AI model conducted quality control checks to filter out poor-quality cells and genes, normalized the data for meaningful comparisons, selected features, corrected batch effects, and reduced dimensionality. Following these steps, cell types were comprehensively annotated for each dataset. Leveraging known marker genes and reference datasets, our single-cell AI model assigned each cell its respective biological identity, ensuring thorough annotation. There was particular emphasis on annotating kidney cells and facilitating comparisons across heterogeneous cell types in kidney datasets.

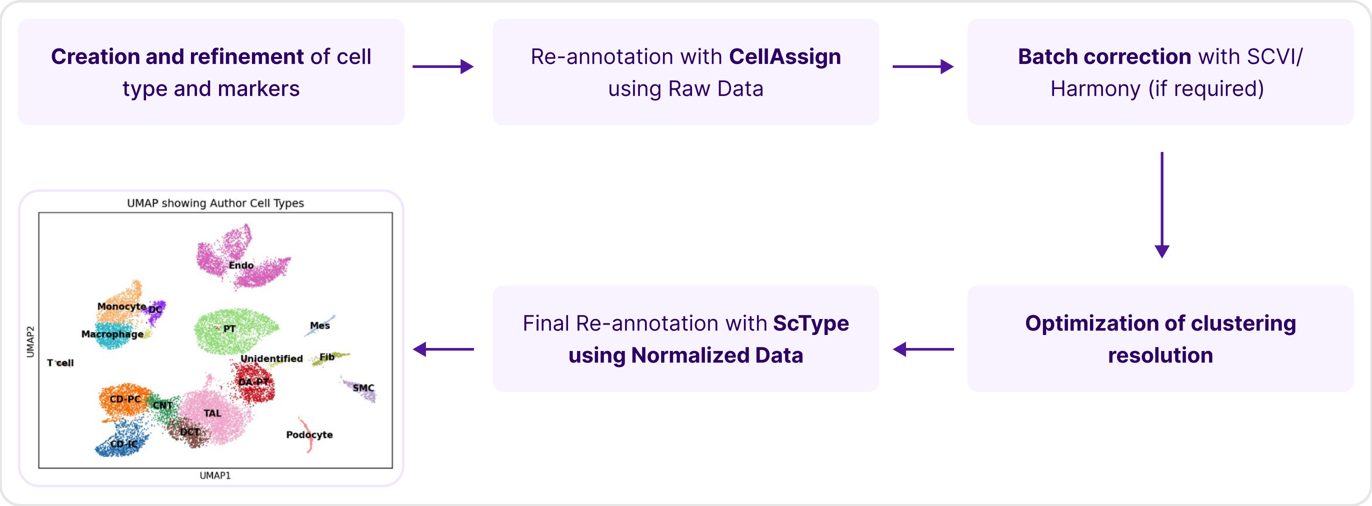
2. Metadata Annotation

Elucidata's harmonization engine annotates datasets with metadata by mapping standard fields like disease, tissue, organism, cell line, cell type, and drug to their respective standard ontologies, MeSH, BRENDA Tissue Ontology, NCBI taxonomy, Cellosaurus, Cell Ontology, and PubChem.

A detailed overview of the dataset, sample, and feature-level metadata fields can be accessed here.

Cell Type Re-Annotation

To eliminate manual parameter tweaking and ensure consistency across multiple single-cell datasets, we performed cell re-annotation to generate better insight. We used the unified cell type and marker dictionary and a cluster-independent cell annotation method (CellAssign) for the re-annotation of the cells.



Creation and refinement of cell type and markers

Re-annotation with **CellAssign** using Raw Data

Batch correction with SCVI/ Harmony (if required)

Optimization of clustering resolution

Final Re-annotation with **ScType using Normalized Data**

Re-Annotation Workflow

This iterative process:

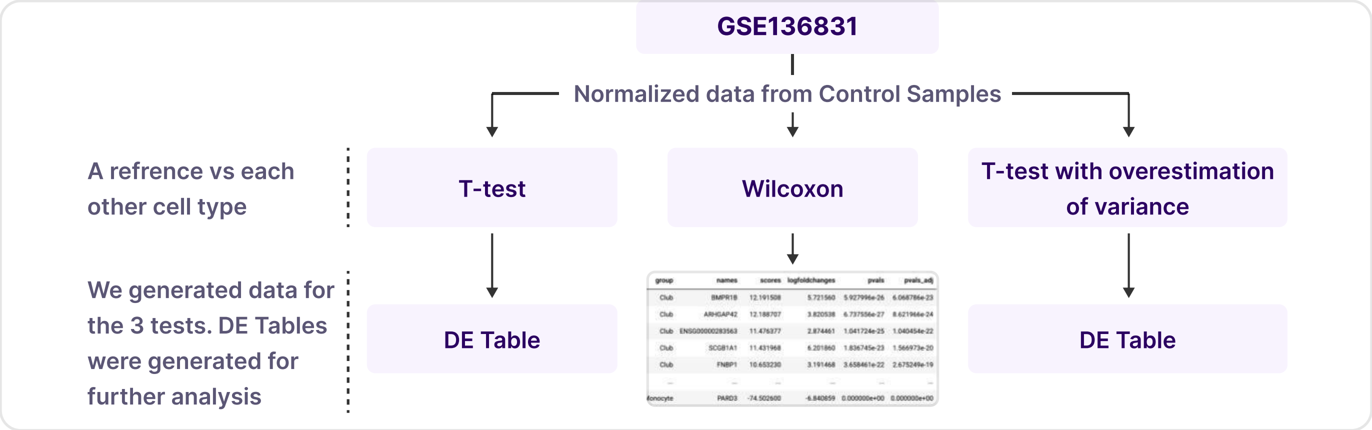
- **Creation and Refinement of Cell Type Markers:** Initially, a database or dictionary of cell types and their associated gene markers is constructed. This database is refined through iterative testing across multiple public datasets.
- **Re-annotation with CellAssign:** Raw data from the scRNA-seq experiments are re-annotated using the CellAssign tool, which leverages the created cell type marker database to identify cell types.
- **Batch Correction:** If necessary, batch effects are corrected using tools like SCVI or Harmony to ensure the data is harmonized across different conditions or experiments.
- **Optimization of Clustering Resolution:** The resolution of clustering is optimized to improve the accuracy of cell type identification.
- **Final Re-annotation with ScType:** The datasets are re-annotated using the cluster-dependent ScType tool on normalized data, ensuring that the final annotations are accurate and consistent.

Target Receptor Identification Pipeline for Analysis

We developed a pipeline to analyze re-annotated single cells, aiming to identify genes with similar expression patterns across disease cohorts. This pipeline relied on cosine similarity scores to correlate genes with those found in literature mining, GWAS, and the GTEx databases, which are associated with disease progression.

This AI-enabled pipeline employed a workflow to generate a differential expression profile for a gene of interest from various cells in the kidney. We extended the target receptor identification process to muscle, kidney, and adrenal tissues, where we generated and delivered ranked lists of similar genes for each corresponding target gene. To enable these operations at scale and improve decision-making on genes of interest, we developed a series of enablers and accelerators.

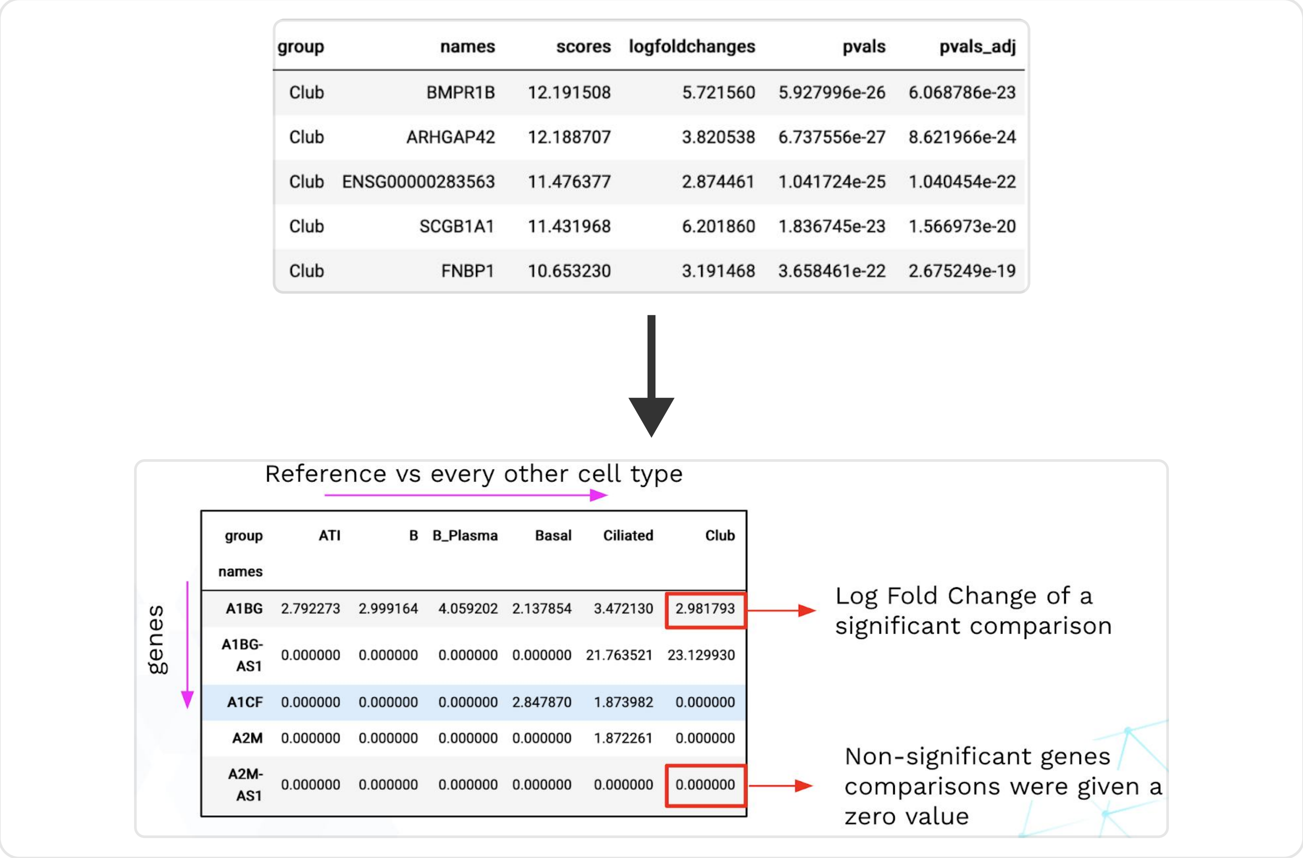
Using the same method, it also generated expression profiles for all genes. This enabled the pipeline to directly compare a gene to a reference gene by calculating the cosine similarity between their expression profiles. The pipeline utilized multiple statistical methods, including Wilcoxon's rank-sum test, the standard T-test, and T-test, to calculate both the log fold change and the corresponding p-values for each gene. The application of different statistical methods ensures the robustness of the pipeline workflow, thereby enhancing the reliability of the results.


GSE136831
Normalized data from Control Samples
A refrence vs each other cell type
T-test
Wilcoxon
T-test with overestimation of variance
We generated data for the 3 tests. DE Tables were generated for further analysis
DE Table
DE Table

group	names	scores	logfoldchanges	pvals	pvals_adj
Club	BMPR1B	12.191508	5.721560	5.927996e-26	6.068786e-23
Club	ARHGAP42	12.188707	3.820538	6.737556e-27	8.621966e-24
Club	ENSG00000283563	11.476377	2.874461	1.041724e-25	1.040454e-22
Club	SCGB1A1	11.431968	6.201860	1.836745e-23	1.566973e-20
Club	FNBP1	10.653230	3.191468	3.658461e-22	2.675249e-19
...	...	...	...	...	...
Monocyte	PARD3	-74.502600	-6.840659	0.000000e+00	0.000000e+00

Pipeline workflow

Through the analysis conducted by the AI-enabled pipeline, we improved precision in identifying genes of interest and elucidating potential therapeutic targets, thus facilitating more effective development of RNAi drugs.



group	names	scores	logfoldchanges	pvals	pvals_adj
Club	BMPR1B	12.191508	5.721560	5.927996e-26	6.068786e-23
Club	ARHGAP42	12.188707	3.820538	6.737556e-27	8.621966e-24
Club	ENSG00000283563	11.476377	2.874461	1.041724e-25	1.040454e-22
Club	SCGB1A1	11.431968	6.201860	1.836745e-23	1.566973e-20
Club	FNBP1	10.653230	3.191468	3.658461e-22	2.675249e-19

Reference vs every other cell type

group	ATI	B	B_Plasma	Basal	Ciliated	Club
names						
A1BG	2.792273	2.999164	4.059202	2.137854	3.472130	2.981793
A1BG-AS1	0.000000	0.000000	0.000000	0.000000	21.763521	23.129930
A1CF	0.000000	0.000000	0.000000	2.847870	1.873982	0.000000
A2M	0.000000	0.000000	0.000000	0.000000	1.872261	0.000000
A2M-AS1	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000

Log Fold Change of a significant comparison

Non-significant genes comparisons were given a zero value

How does cosine similarity score in the pipeline work

The reannotation pipeline was a key enabler, providing **high-quality insights for 20 muscle and 25 kidney cell types**. In addition, we harmonized 43 datasets across five tissues, comprising **1.8 million cells**, which significantly reduced validation time from months to weeks. This allowed us to generate **ranked gene lists** for approximately **996 genes per target across 19 target genes** and **four tissues.**

01 Lung Analysis

Ranked gene list (920 genes @ ~60%) similar to target gene ITGB6 delivered

2 high potential genes validated for downstream research - **AGER, MUC5B**

02 Muscle Analysis

Ranked gene list (649 genes @ ~60%) similar to 5 target genes delivered

03 Kidney & Adrenal

Ranked gene list (649 genes @ ~60%) similar to 5 target genes delivered

Key Highlights

Outcomes

This successful partnership over 1 year has helped the customer:

1. **Access High-Quality Single-cell Datasets from Public Repositories:** We delivered **high-quality, AI-ready** datasets with ~1 million cells and 5000 single-cell samples relevant to rare diseases.
2. **Access Comprehensive Cell-type Annotated Single-cell Datasets:** We delivered single cells that were consistently processed, curated, and annotated for cell types based on their reference markers, with human-in-the-loop validation.
3. **Expedite their Research Objectives:** With our high-quality and AI-analyzed datasets, the customer accelerated the identification of genes of interest for the rare diseases associated with kidneys and lungs. Additionally, these analyses done by our pipeline elucidated more potential therapeutic targets and facilitated the effective development of RNAi drugs.

Category	Value Created
Finding relevant public datasets and enriching them	Finding **1 million cells and ~5000 relevant single-cell samples** from public datasets and annotating metadata from the source would take **~500 hours**
Cell-type annotation and re-annotation	Annotation and re-annotation of **over 1 million cells** with reference markers would take **~1000 hours**
Pipeline for analysis	**Reducing internal costs:** Without Elucidata, the customer would have needed to hire an additional full-time bioinformatician, an AI engineer, and a full-time data engineer to handle tasks such as harmonizing single-cell datasets, as well as developing and deploying pipelines for re-annotation and analysis.

Impact

2X faster identification of ranked potential target genes

1.8M cells harmonized, 5 tissues, 3 organisms, 43 datasets

~1500 hours saved on sourcing datasets, metadata annotation, and cell-type annotation from the public domain

High Quality Insights using **Re-annotation across 20 muscle & 25 kidney cell types**

4X faster analysis of single-cell datasets

About Elucidata

Elucidata optimizes data quality for pre-clinical drug discovery by harmonizing multi-omics and assay data into ML-ready formats. Elucidata's powerful data harmonization engine is utilized by trained experts to curate diverse data types, annotate metadata, and ensure consistent processing at affordable costs. These ML-ready data are stored on an Atlas on the platform, robust data stores ideal for analysis and management.

Our engine's advanced technology accommodates 25+ R&D data types, catering to teams in pre-clinical drug discovery and diagnostics R&D. Elucidata is trusted by over 25 research organizations, including 4 of the top 10 pharma companies, to accelerate their discovery programs.

Book a Demo

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

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