Curated Single-Cell RNA-Seq for Machine Learning and Artificial Intelligence Applications in Pre-Clinical Biomarker Discovery and Validation in the Immune System

Jana Sponarova, PhD
- CSO and Head of Biocuration at NEBION, an Immunai company

de.NBI 3rd Annual Industrial Forum Meeting 2022
24 November 2022
NEBION Core Workflow

From public repositories to a curated search engine

1. The world’s expression data
2. Extensive curation by experts
3. Curated database and search engine
4. Powerful and user-friendly tools
Single-Cell RNA-Seq Data – Curation Workflow

Meta-data enrichment based on publications & suppl. materials

<table>
<thead>
<tr>
<th>Stimulus</th>
<th>Subject ID</th>
<th>age at assessment (HS)</th>
<th>sex</th>
<th>sample status</th>
<th>disease condition</th>
<th>disease duration</th>
<th>intestinal antinflammatory agent</th>
<th>Ulcerative colitis (Montreal classification)</th>
</tr>
</thead>
<tbody>
<tr>
<td>normal colonic mucosa cell sample</td>
<td>237</td>
<td>31 years (postnatal)</td>
<td>male</td>
<td>healthy</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>normal colonic mucosa cell sample</td>
<td>291</td>
<td>51 years (postnatal)</td>
<td>male</td>
<td>healthy</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>normal colonic mucosa cell sample</td>
<td>221</td>
<td>23 years (postnatal)</td>
<td>female</td>
<td>healthy</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>normal colonic mucosa cell sample</td>
<td>204</td>
<td>29 years (postnatal)</td>
<td>female</td>
<td>healthy</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>normal colonic mucosa cell sample</td>
<td>219</td>
<td>30 years (postnatal)</td>
<td>female</td>
<td>healthy</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>ulcerative colitis study 79</td>
<td>198</td>
<td>30 years (postnatal)</td>
<td>male</td>
<td>diseased</td>
<td>ulcerative colitis</td>
<td>0.1 years</td>
<td>aminosalicylic acid and similar</td>
<td>E3 (extensive UC, pancolitis)</td>
</tr>
<tr>
<td>ulcerative colitis study 79</td>
<td>238</td>
<td>77 years (postnatal)</td>
<td>male</td>
<td>diseased</td>
<td>ulcerative colitis</td>
<td>8.3 years</td>
<td>---</td>
<td>E2 (left-sided UC)</td>
</tr>
<tr>
<td>ulcerative colitis study 79</td>
<td>202</td>
<td>60 years (postnatal)</td>
<td>male</td>
<td>diseased</td>
<td>ulcerative colitis</td>
<td>13.6 years</td>
<td>aminosalicylic acid and similar</td>
<td>E2 (left-sided UC)</td>
</tr>
<tr>
<td>ulcerative colitis study 79</td>
<td>290</td>
<td>57 years (postnatal)</td>
<td>female</td>
<td>diseased</td>
<td>ulcerative colitis</td>
<td>42.2 years</td>
<td>aminosalicylic acid and similar</td>
<td>E3 (extensive UC, pancolitis)</td>
</tr>
</tbody>
</table>
Cell-Type/ State Identification, Verification and Harmonization

Core expertise!

Original publication

GV cell compendium

Cell-type reference

Identification & verification of cell types

Cell-type harmonization & identification of rare cell types

Cell type marker genes (Seurat)

Genevestigator®

Dataset 1

Dataset 2

Dataset 3

Dataset 4

Dataset 5
Process of Cell-Type Prediction using Tissue Reference

Global prediction

- CD4, CD8, NK
- Non-immune cells
- B-cell, plasma
- Myeloid, pDC, mast

TRM
- CD8 cytotoxic T-cells
- Epithelial cells
- Cancer cells

automatic prediction
manual prediction
Curated Studies as Harmonized Compendia

Delivery
- aggregate level
- single-cell level

Downstream analysis
- Tool of choice
GENEVESTIGATOR® Applications

- Characterize and prioritize targets & biomarkers
- Identify novel targets & biomarkers
- Analyze disease signatures
- Perform differential expression
- Examine cell signatures
- Analyze changes in cell populations
- Characterize and prioritize targets & biomarkers

- RNA expression data
- Surface proteins
- TCR sequence
- Cell annotations
- Patient metadata
- Response data
- Gene modules
- Gene signatures
- Disease indication
- Tissue type
- Treatment information
Use Case: Rare Cell Type Identification

**Aim:**
- Identification of monocytic and granulocytic MDSCs
- Derivation of MDSCs signatures
- Integration of GV myeloid cell compendium human and mouse
- Building 2 atlases:
  - Myeloid cell atlas (mouse)
  - Myeloid cell atlas (human)

**Application:**
- Identification of different MDSCs in de-novo sequenced mouse/human datasets
- Correlation of this population with disease progression

Mouse
120,752 cells, 18 datasets

Human
587,685 cells, 57 datasets
Summary

- We have built a manually **curated and globally normalized** scRNA-Seq compendium, mainly consisting of immune cells obtained from studies focused on immuno-oncology, autoimmune diseases, and other therapeutic areas.

- The **processing** of a study includes raw data mapping, standardized and strict quality control, data normalization, integration, and cell-type/-state identification.

- The **cell type annotation is synchronized** across all studies in the compendium.

- The pipeline outputs are enriched with **sample-level information** (e.g., patient-level data), and data are integrated into a user-friendly analysis software – GENEVESTIGATOR®, a high-performance visualization tool for gene expression data.

- This **deeply harmonized** compendium represents an important asset for downstream ML and AI applications in pre-clinical **biomarker discovery and validation**.
Thank you for your attention!
Aggregation From Cell to Cell-Type/State/Treatment Level

→ Storage of signal data in the aggregate format enables fast calculations