Poster
Presented at the
AACR
April 2024

Integrating public single-cell transcriptomics and patient profiles to guide clinical development

Jack Russella Pollard[1], Joshua Whitener[1], Jordan Byck[1], Monika Manne[1], Hani Alostaz[1], Marianna Elia[1], Kamil Krukowski[1], Gabrielle Wong[1], Francisco Adrian[1], Liang Schweizer[1], Robert HI Andtbacka[1], & Christos Hatzis[1]
[1]HiFiBiO Therapeutics, Cambridge, MA USA

BACKGROUND

Single-cell technologies provide invaluable insights into disease biology and inform drug development by revealing complex interactions among different cell types within patients. However, harnessing the potential of publicly available single-cell data remains challenging due to the lack of integrated data across diverse single cell platforms.

To maximize the potential of single cell insights, we have created an AI/ML powered curation and data integration process within our Drug Intelligence Science (DIS®) platform. This automated process identifies relevant published studies and integrates single cell transcriptomic data from publicly available sources with in-house generated datasets from our ex vivo translational efforts and our clinical programs.

MATERIALS AND METHODS

Dataset Search and Selection

Peer reviewed published datasets covering a variety of human tissues in the healthy and diseased settings (cancer and autoimmunity) are identified with a custom Python-based dataset crawler that utilizes keywords to surface datasets and publications from various public repositories.

Dataset acquisition is aimed at improving indication selection and combination strategies for our clinical programs. For example, data from approximately 300 patients who received standard of care (SOC) treatment including immune checkpoint inhibitor (IO) therapy were selected for this purpose.

After filtering datasets for the availability of raw count data and vetting data quality via the recovery of the key findings from the publications, our current dataset composition is described in Figure 1.

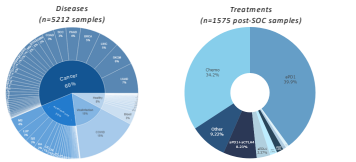

Figure 1. DATASET COMPOSITION. Over 198 datasets were extracted from 186 publications (58 on autoimmune disorders including 866 patients across 30 indications such as Crohn's and Sjogren's and 121 on cancer including 1476 patients across 39 indications such as lung adenocarcinoma and renal cell carcinoma).

Metadata Extraction and Standardization

To structure the cell level data and render it suitable for integration and analysis, we extract metadata from each dataset/publication according to the guidelines for single cell data reporting (1). Briefly, attributes detailing these aspects are collected including:

- Data source (e.g., title, DOI, and number of patients), sample specific-attributes (e.g., indication, biopsy location, treatment, timepoint, response, etc.)
- Sample processing attributes (e.g., tissue dissociation, cell type enrichment, etc.)
- Cell-level attributes (e.g., author annotated cell type/state, gene expression level, activation scores, clonotype, etc.)

These metadata are standardized where possible by referencing appropriate external ontologies such as the Disease Oncology (DO) for indications (2).

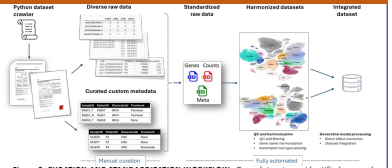

Figure 2. CURATION AND STANDARDIZATION WORKFLOW. Once datasets are identified, manual curation is used to extract both the raw data and the metadata in a standardized form. This standardized form is then used to process and integrate the datasets in a fully automated fashion.

Generative Modeling for Single Cell Analysis Tasks

A key advantage and differentiator for our approach is the use of generative modeling for a variety of single cell tasks such as integration, normalization and cell type/state annotation. Typically, these tasks are done with best in breed-tools (3) that then require rank-based or meta-analysis to compare results across datasets. Generative modeling has the advantage of using a single, integrated model for all these tasks in an automated fashion and enables direct comparisons across datasets for machine learning applications.

For example, data integration and annotation harmonization of diverse cell types/states across studies is a computational challenge (4) that is ripe for generative approaches. Briefly, using the autoimmune and cancer datasets, we pre-trained a model to annotate 31 immune and 4 non-immune cell types/states in new data sets in an automated fashion. In the latent space, new unlabeled data is integrated with data from the pre-trained model. Transfer learning then assigns the cell type/state annotation.

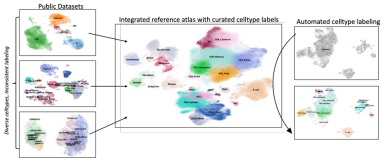

Figure 3. PRETRAINED MODEL FOR CELL TYPE / STATE CALLING. By integrating more than 377k cells drawn from 19 immune cell rich studies encompassing more than 411 samples from 225 cancer and autoimmune patients and healthy controls, we have created an algorithm to label cells from published and internal studies in an automated fashion.

RESULTS

Integrated Single Cell Database from Generative Modeling

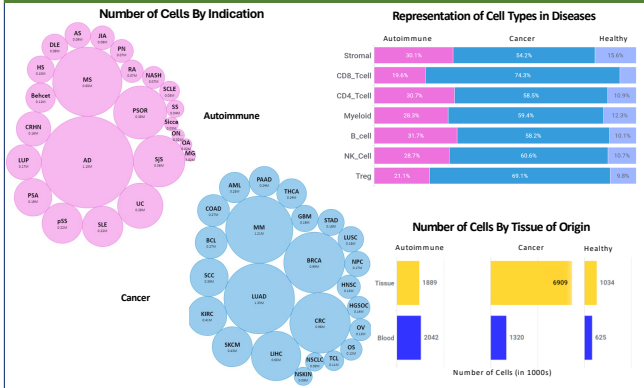

Figure 4. CELL COMPOSITION. The current single-cell database comprises > 15 million cells analyzed in >5,000 samples collected from >2,500 patients across 160 curated unique human studies, spanning oncology (53%), autoimmune disease (AID, 24%), and viral infection (14%), with the remaining being from reference healthy tissues (9%). Only indications exceeding 0.09M cells are shown here.

The post-SOC resource was used to identify treatment settings where these mechanisms are enriched and markers co-expressed with targets of interest (e.g., PD1, CTLA4, OX40, BTLA, TNFR2, etc.) in specific cell populations to inform combination strategies.

Cellect™: Multiuser Visualization and Interrogation Interface

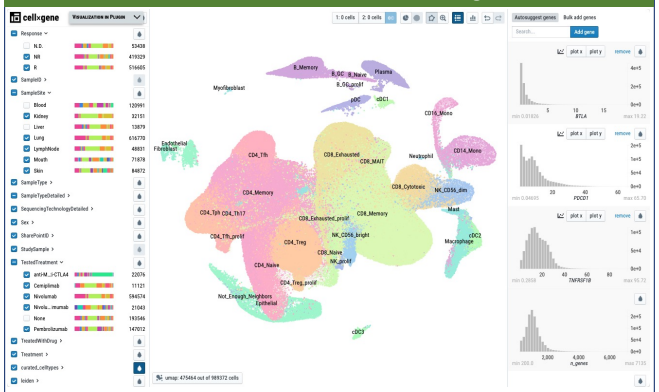

Figure 5. CELLECT. To facilitate multiuser visualization and interrogation of the curated and integrated datasets, we have developed Cellect, a customized version of the CellxGene Gateway (5). The system has a suite of tools that help both bench and computational scientists find, explore, and analyze our curated single cell datasets. It includes several tools with features to engage with the data and aids in identifying cell types, clusters, and marker genes. Shown in the figure is an integrated dataset of about 1-million cells from cancer patients treated with checkpoint inhibitors across a variety of indications. Our generative modeling was able to standardize and integrate the data making comparisons across studies and treatments possible. This integrated dataset was used to identify combination strategies for our clinical programs.

APPLICATIONS

Post IO Combination Strategies Informed by Single Cell Data

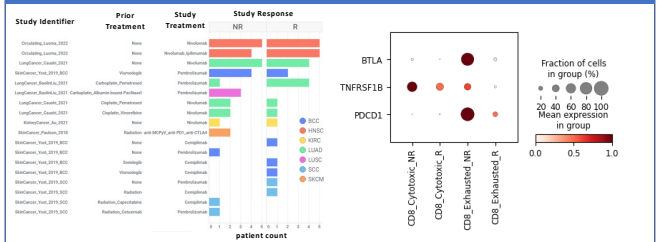

Figure 6. IDENTIFYING COMBINATIONS AND TARGET CELL TYPES. To explore the cell type/states associated to IO response and resistance, a collection of about 475k cells from the primary tumor site of post IO treated patients was selected for analysis. Analysis across CD8 T-cells subsets that express several surface targets revealed a distinct pattern of expression with respect to the validated PD1 target (PDCD1). Specifically, the B- and T-lymphocyte attenuator (BTLA) and the Tumor necrosis factor (TNF) receptor 2 (TNFRSF1b) show patterns suggesting they maybe targets to activate distinct subsets of CD8 T-cells.

Post IO HFB200301 & HFB200603 May Target Distinct CD8 Subsets

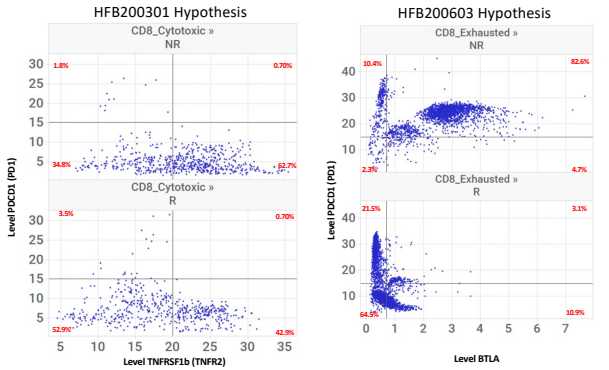

Figure 7. HFB200301 & HFB200603 MAY TARGET DISTINCT CD8 SUBSETS. TNFR2 shows a pattern of expression consistent with it being an activating receptor expressed on an alterative set of CD8 T-cells not expressing PD1, that if agonized with a compound such as HFB200301 these cells maybe able to attack tumor cells. Alternatively, BTLA shows a pattern consistent with it being a coinhibitory receptor expressed with PD1 on CD8 cells in the post IO treatment setting, that if antagonized with HFB200603 in combination with anti-PD1 might activate those cells to attack tumor cells.

CONCLUSIONS

We have presented an AI/ML guided approach to address the key challenge of integrating single-cell data across platforms and demonstrated that relevant disease biology is retained upon integration. We outline a path for deploying this solution at scale for bench and computational scientists to guide target as well as indication selection, as was done for our ongoing clinical programs, including our first-in-class TNFR2 agonist (HFB200301, NCT05238883) and our BTLA antagonist (HFB200603, NCT05238883).

References

1. Nature Biotechnology volume 38, pages 1384–1386.
2. Nucleic Acids Res. 2024 Jan 5; 52(D1): D1305–D1314.
3. Nat Rev Drug Discov. 2023 Jun;22(6):496-520.
4. Genome Biol 20, 194 (2019).
5. bioRxiv 2021.04.05; doi: https://doi.org/10.1101/2021.04.05.438318.

For additional information, please email contact@hifibio.com or visit Hifibio.com