Quantitative Phenotyping and Enrichment of Live Single Cells via Deep Learning

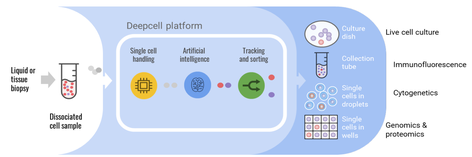
Euan Ashley[1], Mahyar Salek[2], Krishna P. Pant[2], Nianzhen Li[2], Christina Chang[2], Andreja Jovic[2], Esther Lee[2], Kiran Saini[2], Jeanette Mei[2], Thomas J. Musci[2], Maddison (Mahdokht) Masaeli[2]; 1. Stanford University, Stanford, CA; 2. Deepcell Inc. Mountain View, CA. *Contacting email:* euan@stanford.edu

Introduction

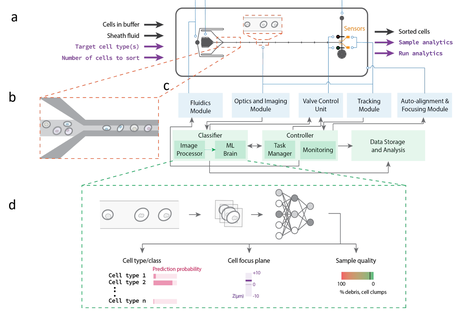
High-throughput single-cell multi-omic analysis has opened up a path to understand normal development and disease processes at cellular resolution. In order for the recent technological evolution to lead to major new insights about biological processes, these molecular-level data need to be coupled with phenotype and function- ideally also at the single cell level. Genotype-phenotype associations, while difficult to map, are critical for understanding how biological models function. Coming up with approaches to standardize and scale the phenotypic assessment at single cell resolution is at the heart of any solution to this challenge. Recent breakthroughs in machine intelligence and deep learning have demonstrated the possibility of achieving high levels of accuracy in identifying morphological features of cells on pathology slides. Here we introduce an artificial intelligence (AI)-powered morphological single cell analysis and sorting platform based on high-resolution imaging of cells in flow and a 48-layer deep Inception-based neural net. Despite the rich data and a deep network, the system is engineered to sort in real time. The sorting decisions are made purely based on fine cellular and nuclear morphological features, therefore allowing for label-free isolation and purification of populations of interest. The technology allows for the integration of cell morphology into the increasingly diverse set of molecular modalities now in use to characterize the function and state of individual cells.

Core technology and components

Core technology

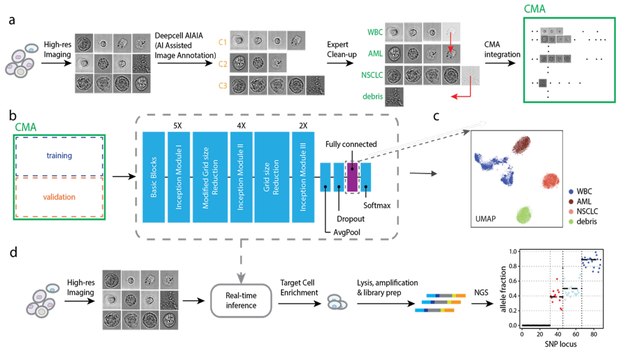


System Schematic



(a) The microfluidic chip and the inputs and outputs of the Deepcell sorter platform. (b) A combination of hydrodynamic focusing and inertial focusing is used to focus the cells on a single z plane and a single lateral trajectory. (c) Diagram shows the interplay between different software and hardware components. (d) The classifier is blown up to depict the process of image collection and real-time automated single cell classification.

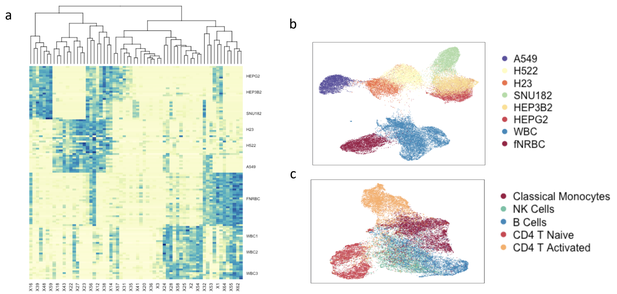
Labeling, AI model training, and intelligent sorting pipeline



(a) High resolution images of single cells in flow are stored. Deepcell **AIAIA** (Artificial Intelligence- Assisted Image Annotation) is used to cluster each individual cell into a morphologically similar group of cells. Cell clusters are reviewed manually and batch-labeled by an expert. These errors are corrected by the "Expert clean-up" step. The annotated cells are then integrated into Deepcell Cell Morphology Atlas (**CMA**). (b) The CMA is used to generate both training and validation sets for the training of the next generation of the models. (c) The last two layers of an Inception-based network are used to create a UMAP depiction of cell clusters and prediction probabilities. (d) During a sorting experiment, the pre-trained model shown in (b) is used to infer the cell type (class) in real-time. The enriched cells are retrieved from the chip and analyzed after cell lysis, amplification and library preparation.

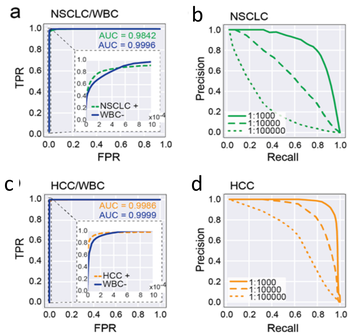
Results

Heatmap and UMAP depiction of cells represented by 64-node fully connected layer of the convolutional neural net



(a) Heatmap depicting representative data from the fully-connected layer trained on cells of 18 classes in validation. Each row represents a single cell, and each column represents one of the 64 dimensions of the embedding feature vector. The clustering of the 64 dimensions shows strongly correlated features that discriminate among different conditions (malignant vs not), the major classes (NSCLC vs Hepatocellular Carcinoma), and features that distinguish among individual cell lines (A549 vs H522). (b) UMAP depiction of the same validation data cell classes. Each point represents a single cell. (c) UMAP depiction of subtypes of immune cells combined with a set of activated CD4 T cells.

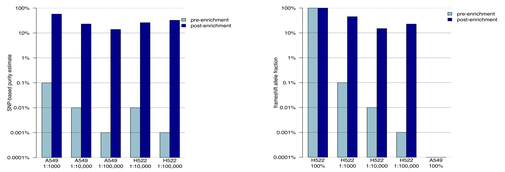
Model performance in classification of NSCLC, liver Carcinoma against WBC



(a)(c) Receiver operating characteristic (ROC) curves for the classification of NSCLC and HCC. Two ROC curves each are shown: one for the positive selection of each category, and one for negative selection, specifically for the selection of non-blood cells. Insets zoom into the upper left portions of the ROC curves where false positive rates are very low to highlight the differences between modes of classification. (b) (d)Estimated precision-recall curves at different proportions for each cell category.

Enrichment of cells at known ratio via Deepcell sorter based on the label-free morphological classifier

a) Spike into WBC



SNP-based purity and enrichment assessment

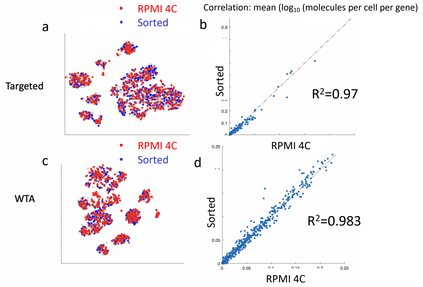
Detecting a frame-shift mutation in sorted H522 cells

b) Spike into whole blood

Cell source	Cell type	Spike-in cell concentration	# cells processed	Sorted cell purity	Fold enrichment by CD45 depletion	Overall fold enrichment
A549	NSCLC	400/ml	1,029,175	55%	13	10,900
A549	NSCLC	400/ml	932,665	80%	16.2	29,000
A549	NSCLC	40/ml	949,836	43%	11	33,500
A549	NSCLC	40/ml	1,012,315	35%	6.7	27,800

Cells from A549 and H522 cell lines were spiked into WBC (a) or whole blood (b) from an unrelated individual and sorted via Deepcell sorter. An additional CD45 depletion step was used to partly pre-deplete WBCs from the whole blood. Purity of enriched cells was estimated by comparing allele fractions for a SNP panel to the known genotypes of both the cell lines and the samples that they were spiked into.

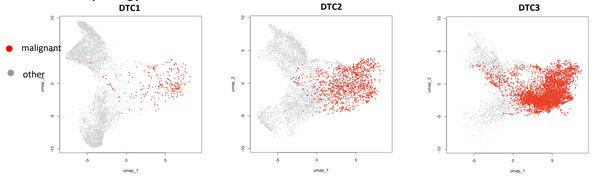
Label-free cell sorting yields live and healthy cells compatible with single cell RNA-seq



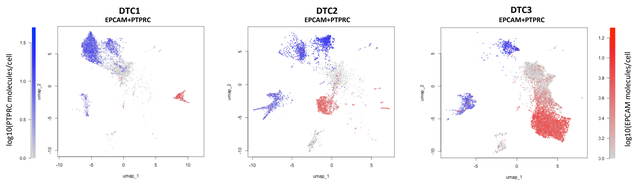
PBMCs flowed-through the Deepcell sorter were compared with control cells stored in RPMI medium using either single cell targeted immune panel or single cell whole transcriptome (WTA) with the BD Rhapsody system. (a)(c)The two samples overlapped with each other in UMAP (each dot is a cell) and (b)(d) gene expression levels showed high correlations (each dot is a gene), indicating no significant mRNA change after microfluidic flow.

Profiling of dissociated lung tumor and cancer cell enrichment

UMAP of Morphology



UMAP of scRNA-seq



Sample	Vendor reported EPCAM+%	In-house RNAseq EPCAM+%	Classifier NSCLC%
DTC1	1.3%	3.3%	2.2%
DTC2	11%	11.6%	12%
DTC3	69%	39%	40%

Three samples of NSCLC dissociated tumor cells (DTC) (from low to high tumor EPCAM load) were profiled using the Deepcell classifier / sorter and data compared with flow cytometry and scRNA-seq profiles of the same samples. The label-free Deepcell AI-based classification of NSCLCs demonstrated agreement with orthogonal methods. AI-based sorting of NSCLCs from these DTCs confirmed the identity of the cells and highly enriched the tumor fraction from the low EPCAM+ sample.

Conclusions

- Deep learning enables high accuracy of cell classification using morphology-alone.
- Rare cells from <=1:100,000 spike-in ratios can be enriched to a final purity of ~ 30%.
- Label-free sorted cells are compatible with downstream DNA/ scRNA-seq workflow.
- Real life DTC samples can be accurately profiled and tumor cells confirmed/enriched.
- Morphology-based deep classifiers holds promise for the development of a universal and standardized metric to understand and interpret morphology and for its adoption along with other -omics.