Deep learning enables the identification and isolation of single cells of interest using high resolution images of non-labeled cells in flow

Mahyar Salek, Hou-Pu Chou, Prashast Khandelwal, Krishna P. Pant, Thomas J. Musci, Nianzhen Li, Esther Lee, Christina Chang, Andreja Jovic, Jeff Walker, Tariq Shafaat, Phuc Nguyen, Kiran Saini, Jeanette Mei, Quillan F. Smith, Himani P. Trivedi, Maddison (Mahdokht) Masaeli;
Deepcell Inc. Mountain View, CA

INTRODUCTION

- Cell morphology has long been considered a powerful phenotype to identify cell type and state in various clinical applications.
- The qualitative, laborious and complex nature of pathology makes it subjective, non-scalable, and therefore challenging to integrate into the emerging single cell assays and technologies.
- Applying machine learning techniques to microscopic tissue images has become a prolific area of research in digital pathology[1,2]
- Current cell sorting approaches have exploited either rudimentary physical characteristics or specific cell surface protein expression as a basis for sorting and isolation of cells.[5] These approaches can introduce biases that can significantly limit the ability to discover new molecular dimensions of the cell.
- Several research groups have recently presented cell sorting technologies that use a machine learning processor to make sorting decisions[3,4].
- These recent achievements invite the following question: is it possible to truly achieve a cyto-pathologist's level of accuracy, or beyond, to analyze and purify cellular samples at scale?
- To this end, we here introduce an AI-powered cell analysis & sorting platform based on high-resolution imaging of cells in flow. By developing a continuous labeling, training, and sorting pipeline, we show that we obtain near-perfect classification of various cell types (and cell states). We also demonstrate the power of morphology for clustering of various cell types. Finally, we show enrichment of cell types of interest against PBMC inspired by clinically actionable ratios in applications such as circulating tumor cell isolations and prenatal Dx.

METHODS

- **Platform Development:** We developed a novel platform for real time high-res brightfield imaging, cell tracking, analysis and sorting of cells in flow *(Description outside the scope of this presentation)*
- **Data Collection:** high-resolution images from roughly 54 million cells, including cells from normal adult blood, fetal blood, trophoblast cell lines, and multiple cell lines derived from non-small-cell lung adenocarcinoma (NSCLC), hepatocellular carcinoma (HCC), and other types of solid and liquid tumors were collected across 4 replicas of the platform to account for environmental variations.
- **AI Assisted Image Annotation (AIAIA):** We deployed a combination of techniques in self-supervised, unsupervised, and semi-supervised learning to facilitate cell annotation at scale.
- **Model training and validation:** We trained an inception-based convolutional neural network. We implemented several augmentation methods to generate altered replicas of the cell images used to train our classifier. In addition to standard augmentation techniques, we studied systematic variation in our image characteristics to develop custom augmentation algorithms that simulate environmental variabilities and sample-correlated imaging artifacts in our system. All validation experiments are carried out on samples that were not included in training.
- **Blood processing:** All blood samples were collected at external sites according to individual institutional review board (IRB) approved protocols and informed consent was obtained for each case. For adult control and maternal blood samples, white blood cells (WBCs) were isolated from whole blood by first centrifugation then the buffy coat was lysed. Fetal cells were isolated from fetal blood by directly lysing with the RBC lysis buffer then washed with PBS. Cells were then fixed in 4% paraformaldehyde (Electron Microscopy Sciences) and stored in PBS at 4°C for longer term usage. Cell lines were purchased from ATCC and cultured in a humidity and CO_2-controlled 37°C cell culture incubator according to ATCC recommended protocols.
- **Molecular analysis:** Cell lines and WBCs of individual blood donors were genotyped with Next Generation Sequencing using a targeted SampleID panel (Swift Biosciences) that includes 95 assays for exonic single nucleotide polymorphisms (SNPs) and 9 assays for gender ID.

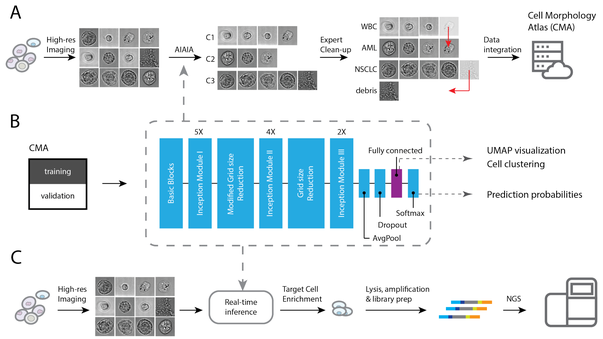

FIGURE 1. Continuous labeling, training, and sorting pipelines (A) High resolution images of single cells in flow are stored. Deepcell **AIAIA** (AI Assisted Image Annotation) is used to cluster each individual cell into a morphologically similar group of cells. Cell clusters are reviewed manually and batch-labeled by an expert. 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 next generation of the neural network models. The last two layers of an Inception-based network are used to create a UMAP depiction of cell clusters and prediction probabilities (C) During a sorting experiment, the pre-trained model shown in (b) is used to infer the cell type (class) for every single cell in real-time. The model prediction is used in real-time to enrich a target cell of interest. The enriched cells are retrieved from the Deepcell sorter and molecularly analyzed after cell lysis, amplification and library preparation.

RESULTS

- Area under the curve (AUC) of the Receiver Operating Characteristic (ROC) curves achieved for NSCLC are 0.9842 (positive selection) and 0.9996 (negative selection); (b) AUCs for HCC are 0.9986 (positive selection) and 0.9999 (negative selection); (c) the AUC for fNRBC is 0.97 (positive selection). (**Figure 2 A-C**)
- Precision-Recall plots show expected accuracies in identifying target cells at 3 hypothetical mixture ratios of 1:1000, 1:10,000 and 1:100,000. Precision corresponds to the estimated purity and recall to the yield of the target cells. (**Figure 2 D-F**)
- The heatmap depiction of an 18-class classifier's fully-connected layer shows clustering of the 64 dimensions. This shows strongly correlated features that discriminate among different conditions (malignant vs not), the major classes (NSCLC vs HCC), and also features that distinguish among individual cell lines (A549 vs H522). (**Figure 3A**)
- UMAP depiction of corresponding heatmap show another view of how the model discriminates various cell types. (**Figure 2B**)
- UMAP depiction of subtypes of immune cells shows separation between Monocytes, NK cells, B cells and CD4 T cells. UMAP depiction shows well separation between CD4 T Naïve and CD4 Activated T cells, demonstrating significant morphological changes associated with T cell activation. (**Figure 2C**)
- Purifying target cells in a variety of spike-in experiments shows enrichment up to 32,000 fold using morphology alone. (**Table 1**)

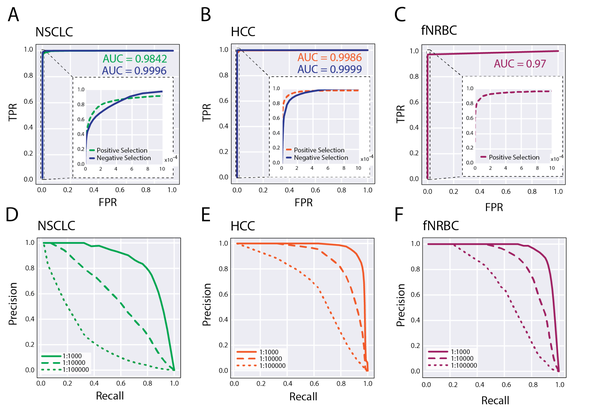

FIGURE 2. Model performance in classification of NSCLC, Liver Carcinoma and fNRBC cells against PBMC (A-C) ROC curves and estimated Precision Recall curves curves for the classification of three cell categories - NSCLC, HCC, and fNRBC. For each cancer cell lines, two ROC curves are shown: one for the 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. (D-F) Estimated precision-recall curves at different proportions for each cell category.

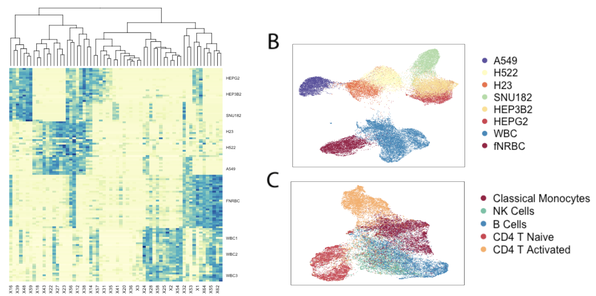

FIGURE 3. Heatmap and UMAP depiction of cells represented by 64-node fully connected layer of the convolutional neural net (A) Heatmap depicting data from the fully-connected layer trained on cells of 18 classes in validation. H522, H23 and A549 are lung cancer cell lines; HEPG2, SNU182 and HEP3B2 are liver cancer cell lines. Also depicted are labeled fetal nucleated RBCs (fNRBC) from three sample not used in training, and WBCs from another three adult subjects not used in training. Each row represents a single cell and each column represents one of the 64 dimensions of the embedding feature vector. (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.

A. Spiked into PBMC

Cell source	Cell type	Spike-in ratio	# cells processed	Sorted cell purity	Fold enrichment
Fet1	fNRBC	1:1,300	999,978	74%	965
A549	NSCLC	1:1,000	101,180	67%	380
A549	NSCLC	1:10,000	1,107,669	31%	2,305
A549	NSCLC	1:100,000	1,342,632	20%	13,904
H522	NSCLC	1:10,000	1,050,036	26%	2,550
H522	NSCLC	1:100,000	1,561,847	33%	32,500

B. Spiked into whole blood

Cell source	Cell type	Spike-in cell concentration	# cells processed	Sorted cell purity	Fold enrichment by CD45 depletion	Total fold enrichment after Deepcell sort
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

TABLE 1. Enrichment of cells at known ratio via Deepcell sorter based on the label-free morphological classifier Fet1 is a fetal blood sample spiked into cells from the corresponding maternal sample. Cells from A549 and H522 cell lines were spiked into (A) PBMC or (B) whole blood from an unrelated individual. An additional CD45 depletion step was used to partly enrich cells spiked into 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 (or the fetal sample) and the samples that they were spiked into.

CONCLUSIONS & FUTURE DIRECTIONS

- Multiple cancer cell lines and fetal nucleated red blood cells are shown to be discriminable against PBMC at near perfect accuracy based on morphology alone.
- Separability of various cell types, subtypes (e.g., Monocyte, NK, B, T) of immune cells, and states (CD4 T naïve vs activated) is demonstrated via clustering of values derived from a late layer of the convolutional neural network.
- Samples with extreme spike-in ratios simulating rare cell applications are sorted and enriched up to five orders of magnitude resulting in limit of detection of up to 1:100,000
- Characterizing cell types and cell states using morphology-based deep classification holds promise for the development of a universal and standardized lexicon to understand and interpret morphology and for its integration with other 'omic' analyses.
- Cell morphology at this resolution and distinction could serve as a powerful phenotypical compliment to single cell multi-omic data.
- Further studies are needed to demonstrate the feasibility of transcriptomic analysis following the isolation and culturing of rare cells of interest from a diverse background.

References

1 Prediction of early recurrence of hepatocellular carcinoma after resection using digital pathology images assessed by machine learning Saito A, Toyoda H, Kobayashi M, Koiwa Y, Fujii H, Fujita K, Maeda A *et al. Mod. Pathol.*, 2020
2 Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer Ehteshami Bejnordi B, Veta M, Johannes van Diest P, van Ginneken B *et al. JAMA*, 2017
3 Intelligent Image-Activated Cell Sorting Nitta N, Sugimura T, Isozaki A, Mikami H, Hiraki K, Sakuma S, Iino T *et al. Cell*, 2018
4 Intelligent image-activated cell sorting 2.0 Isozaki A, Mikami H, Tezuka H, Matsumura H, Huang K, Akamine M *et al. Lab Chip*, 2020
5 Circulating tumor cell technologies Ferreira MM, Ramani VC, Jeffrey SS *Mol. Oncol.*, 2016

Acknowledgements

The authors would like to thank Professor Euan Ashley for the insightful discussions. The authors gratefully acknowledge ABR, iSpecimen Inc, and ATCC for the support of this research by providing biological samples and all corresponding anonymous contributors.

Corresponding author

Mahyar Salek, yar@deepcellbio.com