Realtime morphological characterization and sorting of unlabeled viable cells using deep learning

Mahyar Salek[1], Nianzhen Li[1], Hou-Pu Chou[1], Kiran Saini[1], Andreja Jovic[1], Kevin B. Jacobs[1], Chassidy Johnson[1], Esther J. Lee[1], Christina Chang[1], Phuc Nguyen[1], Jeanette Mei[1], Krishna P. Pant[1], Amy Y. Wong-Thai[1], Quillan F. Smith[1], Stephanie Huang[1], Ryan Chow[1], Janifer Cruz[1], Jeff Walker[1], Bryan Chan[1], Thomas J. Musci[1], Euan A. Ashley [1,2], Maddison (Mahdokht) Masaeli[1]

[1]Deepcell Inc; 4025 Bohannon Dr., Menlo Park, CA 94025, USA.

[2]Department of Medicine, Genetics, & Biomedical Data Science, Stanford University, Stanford, CA USA.

*Corresponding author. Email: yar@deepcellbio.com

Abstract

Phenotyping of single cells has dramatically lagged advances in molecular characterization and remains a manual, subjective, and destructive process. We introduce COSMOS, a platform for phenotyping and enrichment of cells based on deep learning interpretation of high-content morphology data in realtime. By training models on an atlas of >1.5 billion images, we demonstrate enrichment of unlabeled cells up to 33,000 fold. We apply COSMOS to multicellular tissue biopsy samples demonstrating that it can identify malignant cells with similar accuracy to molecular approaches while enriching viable cells for functional evaluation. We show high-dimensional embedding vectors of morphology generated by COSMOS without any need for complex sample pre-processing, gating, or bioinformatics capabilities, which enables discovery of cellular phenotypes, and integration of morphology into multi-dimensional analyses.

One sentence summary: A novel platform capable of high-throughput imaging and gently sorting cells using deep morphological assessment.

Technological advances in genomics and proteomics have enabled molecular profiling of single cells. Indeed, single cell characterization at the genomic, epigenomic, transcriptomic, and proteomic levels has been realized (Gawad, Koh, and Quake 2016; Schwartzman and Tanay 2015; Stegle, Teichmann, and Marioni 2015) and international collaborations are generating increasingly comprehensive cell atlases with exquisitely detailed molecular characterization of hundreds of cell types from multiple organisms (Rozenblatt-Rosen et al. 2017; Regev et al. 2017). In contrast, while cell morphology is often the gold standard for diagnosis and prognosis of many diseases and conditions, our conception of the physical form of single cells has changed little in centuries. Cell morphology characterization has not kept pace with advancements in molecular and functional characterization. This is largely due to the manual, time-sensitive, and subjective process of collecting cell morphology information and limited methods for sorting that do not perturb or damage the cells. Cytopathologists still classify cells stained with a limited number of chemical dyes using a small number of descriptive features (Alvarado-Kristensson and Rosselló 2019; Fischer 2020), and don't have access to further separate and assess cells based on their characteristics. Despite approaches like laser capture microdissection, molecular characterization of captured single cells is limited and sorting cells with high viability remains an unrealized goal. The invention of fluorescent activated cell sorting (FACS), and mass cytometry allowed high-throughput unidimensional or multidimensional classification and sorting of cells, albeit with the prerequisite of labeling with known markers, and at most a couple of features - side scatter and forward scatter of light in flow cytometry - to assess cell physical form. Additionally, these approaches alter cells making them non-ideal for downstream characterization (Bendall et al. 2011; Bendall et al. 2012). There have been recent efforts to improve upon our capability to isolate cells based on their morphological traits (Schraivogel et al. 2022), but these approaches still rely on staining cells with fluorescent markers, which alters them. Additionally, they are limited by the number of morphological traits that can be visualized simultaneously and require heavily involved processes to define a small number of features to quantify morphology. Finally, the feature engineering approach falls short of the human expert assessment in richness and complexity.

Application of machine intelligence has led to multiple approaches to classify pathology slide images on par with human experts, including the recapitulation of immunohistochemistry signals from light microscopy alone (Rivenson et al. 2019). One group combined shallow (six-layer) convolutional neural network (CNN) classification of single cells with a sorting device to identify a small number of cell types (Nitta et al. 2018, 2020). Despite this progress, machine learning approaches for single cell analysis have been based on small data sets.

A platform that can identify, classify, and sort living cells based on morphology could greatly empower our understanding of biology at the single cell level. Specifically, a method to facilitate molecular characterization approaches downstream of sorting and enrichment of minimally perturbed cells could redefine our understanding of cell type and state while at the same time considerably reducing costs by concentrating the cells of interest to the investigator. Complex multicellular tissues, such as the tumor microenvironment, could be deconvoluted prior to the application of molecular assays rather than the conventional post hoc analyses using single cell characterization techniques.

Here, we introduce the COSMOS platform, a novel microfluidic optical device capable of high-throughput cell imaging and sorting using morphological information (**Fig. 1**). The hardware is complemented by (i) a deep inference infrastructure, (ii) a machine learning assisted human image annotation tool, (iii) an atlas of expert-annotated images of single cells called Deep Cell Atlas (DCA), and (iv) a library of pre-trained machine learning models for specific biological applications. COSMOS yields populations of cells that are label-free, viable, and minimally perturbed, allowing sorted cells to be recovered and further characterized by molecular and functional assays. Additionally, cell images can be used to generate high-dimensional morphological profiles to reveal and explore previously unrecognized heterogeneous cell populations. We demonstrate several applications including enrichment of tumor cells, and gene expression analysis of sorted cells.

Hardware: A microfluidic cartridge allows for the input and flow of cells in suspension with confinement along a single lateral trajectory to obtain a narrow band of focus across the z-axis. Using a combination of hydrodynamic and inertial focusing, we collect high-speed bright-field images of cells (up to 20,000 frames per second) as they pass through the imaging zone of the microfluidic cartridge. Images capture subcellular and subnuclear features of the single cells in high contrast with each pixel representing an area of 0.044μm^2. An automated object detection module tracks the cells as they flow through the channel. The images are fed into a CNN for generation of high-dimensional morphological descriptors and classification in realtime. Based on the classification, pneumatic valves are used for sorting a cell into either the cell collection reservoir or waste outlet (**Fig. 1, A and D and fig. S1**). Sorted cells are then retrieved for downstream analysis. A laser-based tracking system identifies cells in realtime, to assist with imaging, sorting and to report on the purity and yield of the run. The instrument can automatically align the microfluidic chip within the camera's field of view, re-focus the optical z-plane, and adjust its operation based on sensors during instrument setup, imaging, and sorting.

Cell Annotation: Images of single cells are the input to the AI-assisted image annotation software (**Fig. 1B**), which uses an unsupervised learning approach to assign annotations to images to train machine learning models. Agglomerative clustering is used to cluster cell images, which can be viewed grouped by their focal plane. These cell groups are generated in 2 modes: 1) clusters that are formed based on morphological similarities deduced by an expressive unsupervised model, and 2) morphological proximity to cells annotated within the same session or prior sessions. This software enables a human expert to re-assign annotations to cells that are incorrectly annotated or partition morphologically distinct clusters into multiple cell annotations. Trained users have achieved annotation rates over 100 cells per second using this tool.

DCA: The DCA is an ever expanding database of expert-annotated images of single cells collected from a variety of immortalized cell lines, patient body fluids as well as tissue biopsies. At the time of this manuscript, DCA has amassed over 1.5 billion images of single cells. The annotations are structured based on a cell taxonomy which may allow a cell to be assigned multiple annotations on its lineage. The training pipeline extracts training and validation sets from DCA to train and evaluate neural net models aimed at identifying certain cell types and/or states. During training, one or more annotations may be selected for each cell image according to the architecture of the model (**Fig. 1C**).

Machine learning: A machine learning infrastructure capable of realtime analysis of cell images was developed to generate high-dimensional morphologic descriptors and classifications (**Fig. 1C**). Our model architecture is based on the InceptionV3 (Szegedy et al. 2016) CNN, modified for grayscale images and to output quantitative morphological descriptors (often called "embeddings" in the machine learning literature). This architecture consists of 48 layers and 24 million parameters. Features from cell images are summarized as an embedding from which cell class annotations are predicted. These embedding vectors are not generally interpretable in terms of conventional morphology metrics but can be used to perform cluster analysis to group morphologically similar cells and visualized using tools like Uniform Manifold Approximation and Projection (UMAP) (McInnes et al. 2018), and clustered heatmaps. This architecture runs in realtime on our instrument which allows images to be analyzed by previously trained models and generates classification and high-dimensional morphology descriptions for each imaged cell. If cell sorting is desired, the model outputs are used to determine whether to discard or retain each cell and, if retained, which collection well to route each cell.

Cells cluster in embedding space. To demonstrate that COSMOS can identify unique cell types, we applied it to cells that may circulate in body fluids, like fetal cells and cancer cells. Therefore, we created training and validation sets that included non-small cell lung cancer (NSCLC) cell lines, hepatocellular carcinoma (HCC) cell lines, fetal nucleated red blood cells (fnRBC), and adult peripheral blood mononuclear cells (PBMCs) (**Fig. 2B**), and measured the performance of COSMOS in identifying these different cell types (**Fig. 2, A and E)**. We generated low-dimensional projections of the embeddings from our trained model, using UMAP plots (**Fig. 2A**). We found a strong correlation between the dimensions of the embedding space and cell type, as illustrated using heatmap and UMAP representations (**Fig. 2, A and C**). The UMAP plot shows that distinct cell types are clustered separately from one another. Within the NSCLC and HCC cell line clusters, the three cell lines were clustered separately (e.g. A549, H522, H23). PBMCs show a large degree of variation consistent with being comprised of several morphologically distinct classes of cells. We then showed certain coordinates of the embedding space correlate with different cell classes by projecting the value of each coordinate in the embedding space onto the UMAP representation (**fig. S2**). Representative images of each of these four classes captured by COSMOS are shown in **Fig. 2D**.

Classification of cell types with low error. We next measured the accuracy of the model in classifying the four different cell types in a supervised fashion. For all the cell classes, the cell lines assessed in the validation dataset were distinct from those used for training. The validation dataset also included fnRBCs drawn from a pool of three fetal samples, and PBMCs extracted from the blood samples of three different donors, that were not used in the training dataset. **Fig. 2E** is the confusion matrix of classifier prediction correlations for each cell class against their true class. The data shows that the model's prediction for fNRBCs, HCCs, NSCLCs and PBMCs matches the actual class at 87%, 100%, 92% and 100%, respectively. The confusion matrix demonstrates that morphology alone can accurately differentiate and identify these cell types when compared against each other.

***In silico* evaluation of cell enrichment in contrived blood samples.** We assessed the ability of COSMOS to identify low abundance NSCLCs, HCCs and fnRBCs from a background of PBMCs. We considered two different strategies for evaluating performance of the supervised model: positive (selecting the target cell class: NSCLC or HCC) and negative selection (selecting all nucleated blood cells: PBMC). The classifier performance metrics for these cell lines yielded an area under curve (AUC) of 0.9842 for positive selection and 0.9996 for negative selection, respectively, for the NSCLC class, and an AUC of 0.9986 and 0.9999 for positive and negative selection, respectively, for the HCC class (**fig. S3, A and B**). In addition, we demonstrated low false positive rates for both modes of classification. Although the AUCs are superior in the negative selection strategy in both cases, the positive selection strategy in both cases would enable higher yields at low false positive rates (FPR < 0.0004). For fnRBCs, we assessed only the mode of positive selection which yielded an AUC of 0.97 (**fig. S3C**).

To better understand the model performance, different spike-in ratios were analyzed *in silico*. Estimated precision-recall curves at different proportions of target cells (NSCLC, HCC and fNRBCs) in a background of healthy donor PBMCs demonstrates that even at a dilution of 1:100,000, the model supports detection of target cells at >70% precision (positive predictive value or post-enrichment purity) and 50% recall (sensitivity) in both the fnRBC and HCC samples, while precision drops to 15% for NSCLC class (**Fig. 2F**). We also show the probability distribution for each of the classes as it relates to their identification against PBMCs for both positive selection (P_{NSCLC}, P_{HCC} and P_{fnRBC}) and negative selection (P_{PBMC}) (**fig. S4**).

Enrichment of target cells. To biologically validate our i*n silico* analysis, we performed simultaneous classification and enrichment experiments by spiking NSCLC cell lines (A549 and H522) into PBMCs at defined proportions ranging from 1:1,000 to 1:100,000. The fnRBC sample was spiked into PBMCs from matching maternal blood. Each spike-in mixture was then processed on COSMOS and cells identified as target cells (fnRBC or NSCLC) by the classifier were sorted in realtime and subsequently retrieved.

For each spike-in mixture, we assessed the purity of the sorted cells retrieved from our system by analyzing allele fractions of the spiked-in cell lines and the background cells in a panel of single nucleotide polymorphism (SNP) assays (**fig. S5**). By comparing the known spike-in mixture proportions and the final purity, we computed the degree of enrichment achieved for each of the samples analyzed. COSMOS was able to achieve similar enrichment and purity for A549 and H522 cells (**Fig. 3A**, **table S1**), even though the former was used to train the classifier and the latter was not. For the lowest spike-in ratio investigated (1:100,000), 20% and 30-33% purities corresponding to folds enrichments of 13,904x and 30,000x-32,500x were obtained for A549 and H522, respectively.

We also assayed for a frameshift mutation in *TP53* (c.572_572delC), for which the H522 cell line is homozygous and the A549 cell line is wildtype (Tate et al. 2019). The proportion of the total sequence reads that contain this frameshift mutation are shown in **Fig. 3B** and **table S2** and are consistent with purity estimates from the panel of SNPs depicted in **table S1**. Even at the lowest investigated spike-in ratio of 1:100,000, we found the mutation present at an allele fraction of 23% in the DNA extracted from the enriched cells, suggesting that functionally important cancer mutations may be detected even when the cells containing them are present at proportions significantly lower than 1:100,000.

Next, we spiked A549 cells into whole blood at concentrations of 40 cells/mL and 400 cells/mL and processed them as outlined in the methods. The purity and fold enrichment of the sorted cells was estimated by jointly analyzing allele fractions in a SNP panel for both the A549 cell line and the enriched cells (**Fig. 3C** and **fig. S6**). The sorted samples had final purities of 55% and 80% for the 400 cells/mL replicates (corresponding to an overall enrichment of >10,900 fold and >29,000 fold respectively) and purities of 43% and 35% for the 40 cells/mL replicates (corresponding to an overall enrichment of >33,500 fold and >27,800 fold, respectively) (**Fig. 3C** and **table S3**).

Compatibility of sorted cells with single cell RNA sequencing (scRNAseq). We tested if the cells sorted with COSMOS were viable and amenable to downstream scRNAseq analysis. We found COSMOS had minimal or no impact on cell viability across the cell lines and primary cells tested (**table S4**). We further compared the single cell gene expression profiles between unprocessed and COSMOS processed PBMCs by scRNAseq with a targeted immune response panel and whole transcriptome amplification (WTA). We found high correlation between the gene expression profiles using both targeted assays (R^2=0.97) and whole transcriptome (R^2=0.983), indicating that the cells processed through COSMOS are directly comparable with unprocessed cells, and are compatible with downstream single cell RNA analyses (**fig. S7, A, B and C**). Additionally, we compared the cell health states of neutrophils, a cell type that is known to be sensitive to cell processing (Alvarez-Larran et al. 2005), after various processing workflows (**fig. S7D**). With bulk RNA sequencing analysis, COSMOS-sorted cells showed fewer up- or down-regulated genes relative to control cells (**fig. S7E**) compared to FACS. They had less activation in genes involved in multiple immune cell activation pathways and neutrophil degranulation pathways (**fig. S7F**), suggesting COSMOS sorting was gentler to the cells.

Identification of malignant cells from dissociated solid tissue biopsies. We evaluated the accuracy of the model in identifying malignant cells from dissociated solid tissue biopsies, by running the model on three NSCLC dissociated tumor cell samples (DTC) with low, medium, and high percentages of malignant cells and comparing the model results to flow cytometry and scRNAseq analysis (**fig. S8**). Malignant cell frequencies determined by the model had high concordance to scRNAseq analysis of EpCAM expression for low (2.2% vs 4.6%), medium (12% vs 16.8%) and high (40% vs 46.7%) malignant cell purities (**fig. S8)**.

Enrichment of malignant cells from DTC samples. Finally, as proof that COSMOS can specifically distinguish and enrich malignant cells from tumor tissue, we sorted cancer cells from a DTC sample of a stage IIB NSCLC patient. To confirm the run-to-run consistency, the sample was split into two aliquots, and each aliquot was run on two COSMOS instruments. Sorted cells were split into multiple fractions for molecular analysis, including targeted DNA panel amplification for mutation analysis, whole genome amplification (WGA) followed by copy number variation (CNV) analysis, and scRNAseq for gene expression analysis (**Fig. 3D**). Our model predicted 1.2% ± 0.7% of malignant cell fraction on multiple runs, consistent with the EpCAM+ percentage reported by FACS (1.3%-1.5%; data not shown). Using a targeted lung cancer panel we found one KRAS and two different TP53 mutations and in sorted samples the allele frequency increased from <3% to 20% and 1-6% to 33-59%, respectively **(Fig. 3E)**. The two pre-sorted aliquots showed variations in the allele frequencies, possibly due to both tumor cell heterogeneity and technical noise in amplification of rare cells at 1-2% range. Nonetheless, we were able to enrich the mutations to 20-60%, suggesting COSMOS enrichment both captured the mutational heterogeneity of the pre-sorted sample and improved confidence in mutation calling beyond any technical noises for low tumor content samples. We then profiled the bulk copy numbers by WGA and were able to significantly increase the sensitivity of CNV detection (**Fig. 3, F, G and H)**. For example, chr8q was amplified (**Fig. 3H**), upon which the *MYC* and *PRDM14* oncogenes are located (Baykara et al. 2015).

We confirmed the identities of the sorted cells, their suitability for single cell gene expression analysis and compared the scRNA profiles to the pre-sorted sample using a WTA workflow. We found that 86-92% of the sorted cells overlapped with EpCAM+/CD45- populations from the pre-sorted cells, indicating a high degree of purity in the sorting capability **(Fig. 3, I and J and fig. S9a)**. The sorted and pre-sorted cells from the EpCAM+/CD45- cluster showed strong gene expression correlation ($R^2 = 0.98$), and overlapped in all subclusters, suggesting that COSMOS sorting was unbiased at least for the EpCAM+ population and did not change gene expression profile due to the gentle microfluidic flow (**Fig. 3K and fig. S9, B and C**). A close examination of 166 stress and apoptosis-related genes (a preloaded gene set from DataView software) also did not show differences in the sorted cells compared to the pre-sorted sample (**fig. S9, E, F and G**).

In conclusion, we present COSMOS, a novel technology platform for the characterization, classification, isolation, and enrichment of cells from living organisms based on high-dimensional morphology. Recent work has motivated morphology as an analyte in cell sorting (Schraivogel et al. 2022). Here we capture the power of deep neural networks in processing morphology by amassing an annotated atlas of greater than 1.5 billion single cell images and training deep models with the computational capacity to classify high resolution high content images. COSMOS offers deep interpretation of single cell phenotype in realtime, with no need for complex sample pre-processing, gating, feature engineering, or bioinformatics capabilities. Using its label-free unbiased approach, COSMOS provides a unique capability to analyze and enable discovery in cell populations with unknown phenotypic or molecular makeup. By enriching viable unaltered cells from tissue and the circulation, the platform enables the combination of morphologic and molecular characterization at the single cell scale, providing novel insights to advance our understanding of biology in basic, translational and clinical applications.

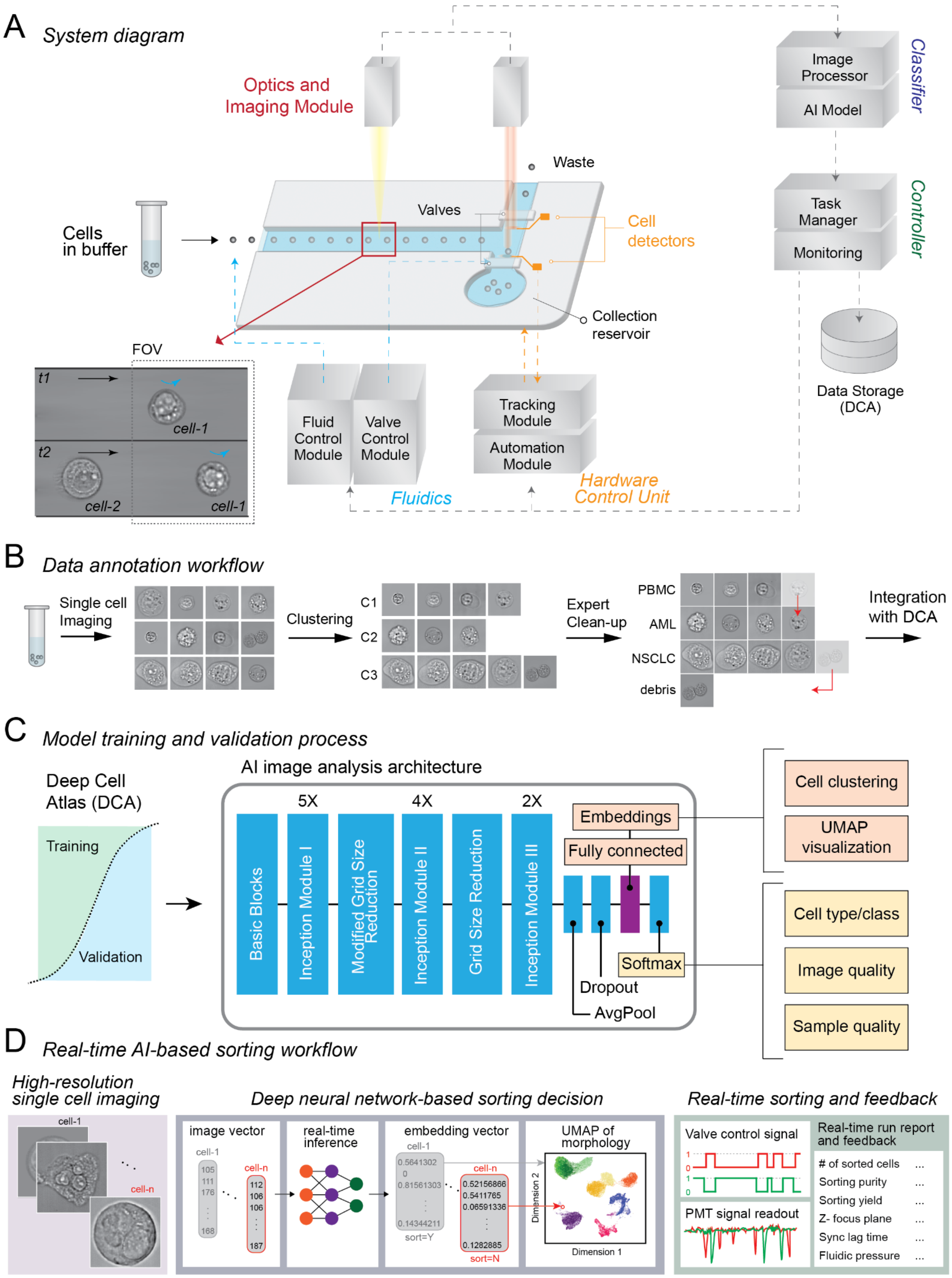

A
System diagram
Optics and Imaging Module
Waste
Valves
Cells in buffer
Cell detectors
Collection reservoir
FOV
t1
cell-1
t2
cell-2
cell-1
Fluid Control Module
Valve Control Module
Fluidics
Tracking Module
Automation Module
Hardware Control Unit
Image Processor
AI Model
Classifier
Task Manager
Monitoring
Controller
Data Storage (DCA)
B
Data annotation workflow
Single cell Imaging
Clustering
C1
C2
C3
Expert Clean-up
PBMC
AML
NSCLC
debris
Integration with DCA
C
Model training and validation process
Deep Cell Atlas (DCA)
Training
Validation
AI image analysis architecture
5X
4X
2X
Basic Blocks
Inception Module I
Modified Grid Size Reduction
Inception Module II
Grid Size Reduction
Inception Module III
Embeddings
Fully connected
Softmax
Dropout
AvgPool
Cell clustering
UMAP visualization
Cell type/class
Image quality
Sample quality
D
Real-time AI-based sorting workflow
High-resolution single cell imaging
cell-1
cell-n
Deep neural network-based sorting decision
image vector
cell-1
105
111
176
168
cell-n
112
106
106
187
real-time inference
embedding vector
cell-1
0.5641302
0
0.81561303
0.14344211
sort=Y
cell-n
0.52156866
0.5411765
0.06591336
0.1282885
sort=N
UMAP of morphology
Dimension 2
Dimension 1
Real-time sorting and feedback
Valve control signal
1
0
1
0
PMT signal readout
Real-time run report and feedback
# of sorted cells ...
Sorting purity ...
Sorting yield ...
Z- focus plane ...
Sync lag time ...
Fluidic pressure ...

Figure 1. COSMOS platform schematic. **(A)** System diagram: A portion of the microfluidic cartridge and interplay between different components of the software and hardware modules are shown. Cells in suspension are inserted into the cartridge. Cells are focused on a single z plane and lateral trajectory. Two images are collected per cell. The hardware includes: i. Fluidics (Fluid Control and Valve Control Modules), ii. Optics and Imaging Module and iii. Hardware Control Unit for auto-focusing and -alignment (Tracking and Automation Modules). The software includes Classifier, Controller, and Data Storage modules. **(B)** Data annotation workflow: High contrast, bright-field images of single cells are captured while flowing in the microfluidic chip. AI-assisted image annotation software is used to cluster individual cell images. A human expert uses the labeling tool to adjust and batch-label the cell clusters. In the example shown, one acute myeloid leukemia (AML) cell was mis-clustered with a group of PBMCs and an image showing debris was mis-clustered with a group of NSCLC cells. These errors are corrected by the "Expert Clean-up" step. The annotated cells are then integrated into DCA. **(C)** Model training and validation process: The DCA is split into training and validation image sets. The AI image analysis depicting the architecture of the Inception V3 model is shown. The fully connected layer of the architecture is used for cell clustering and UMAP visualization. The softmax layer generates per cell classification and the prediction probabilities. It also outputs the cell z-plane focus metrics, which are used to report on image quality. The model prediction for debris, doublets and cell clumps is used to report on sample quality. **(D)** Realtime AI-based sorting workflow: Images of single cells are converted to a vector, and a user-selected classifier assesses each cell. The embedding vector generated by the model is used to visualize sample profile (e.g. UMAP depiction is drawn based on the embeddings). Additionally the realtime inferences guide a sorting decision, based on user preferences. The sorting decision then translates into valve control signals. The laser tracking system detects cells as they arrive in different outlets, through evaluating two photomultiplier tube (PMT) signals. The system generates reports of the number and type of analyzed cells, number of sorted cells, sorting purity and yield, focus plane, synchronization signals, and the fluidic pressures and flow rates. The system uses this information in a feedback loop to adjust system parameters.

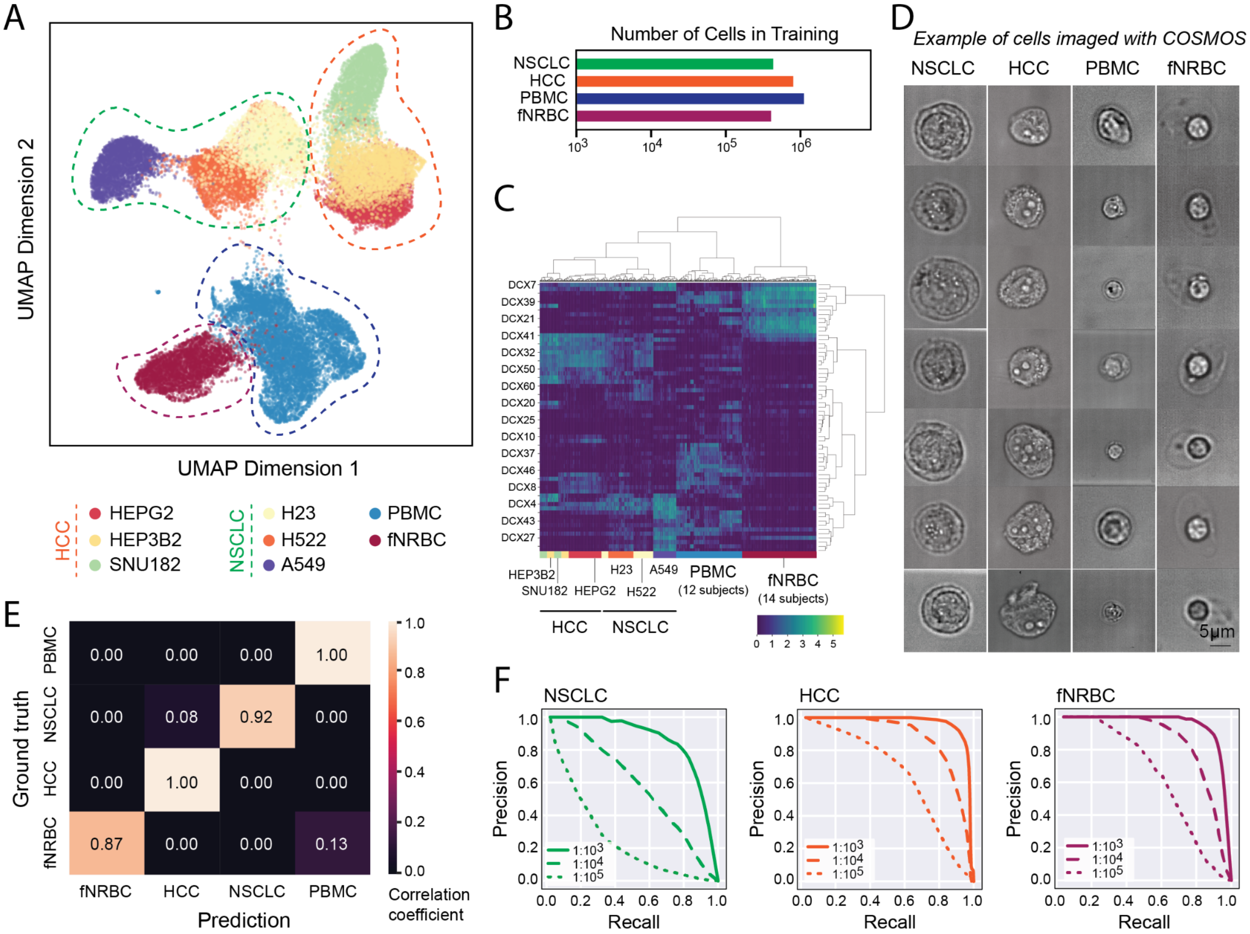
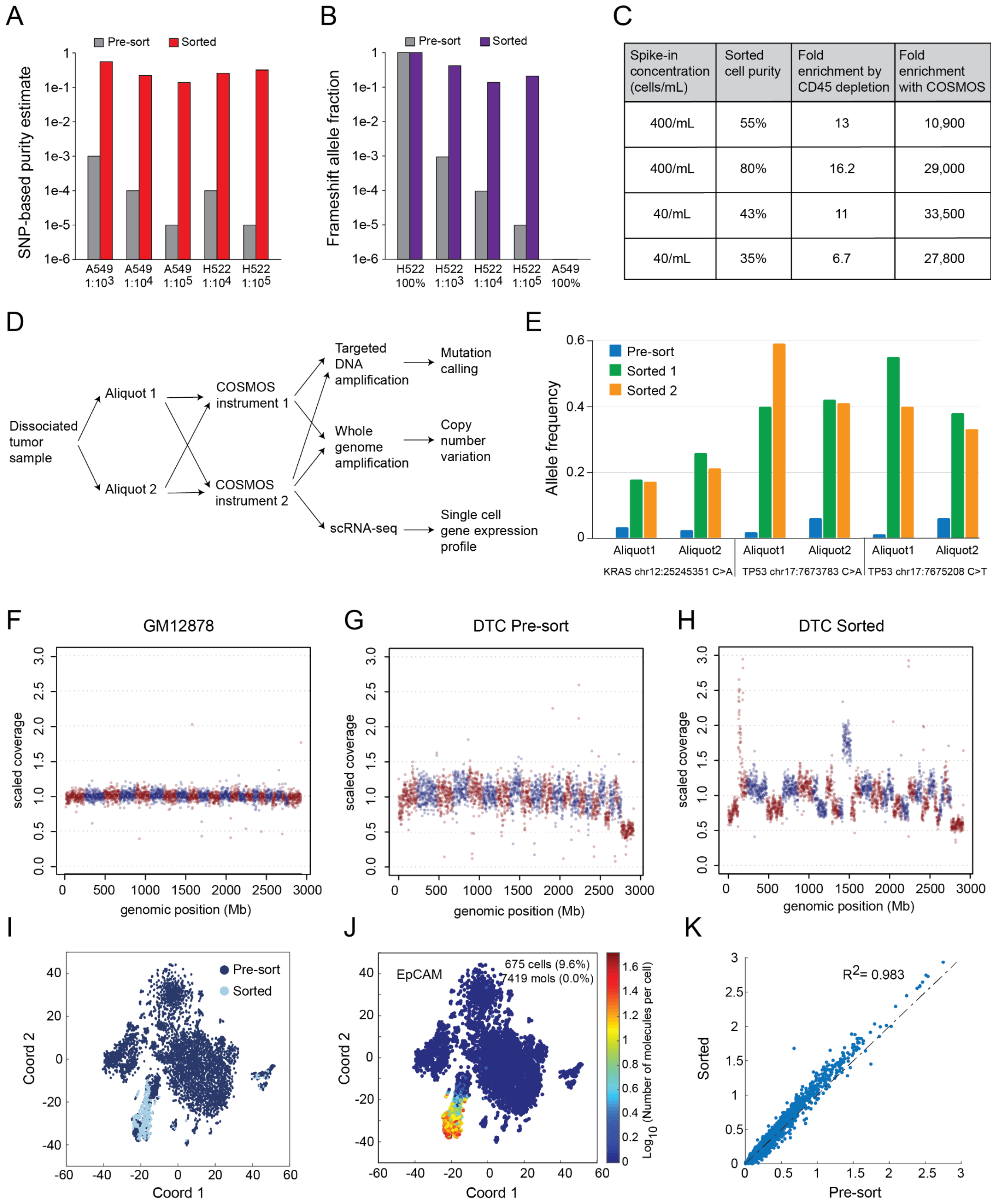


Figure 2. Quantitative morphological assessment of single cells, and performance of COSMOS in identifying cells. (**A**) UMAP projection of cell embeddings sampled from classes analyzed by the model. Each point represents a single cell. (**B**) The number of cells for each of the categories in the training set. (**C**) Heatmap representation of the embedding space. Each column is a single cell and each row is an embedding dimension. (**D**) Representative images of NSCLC, HCC, PBMC and fnRBC classes collected by COSMOS. (**E**) Confusion matrix representing the classifier's prediction accuracy (x axis) versus ground truth (y axis). (**F**) Estimated precision-recall curves at different proportions for positive selection of NSCLCs, HCCs and fNRBCs against a background of healthy donor PBMCs. Precision corresponds to the estimated purity and recall to the yield of the target cells. Three curves are shown for different target cell proportions: 1:1,000, 1:10,000 and 1:100,000.



Spike-in concentration (cells/mL)	Sorted cell purity	Fold enrichment by CD45 depletion	Fold enrichment with COSMOS
400/mL	55%	13	10,900
400/mL	80%	16.2	29,000
40/mL	43%	11	33,500
40/mL	35%	6.7	27,800

Figure 3. Performance of COSMOS in identifying and isolating target cells. A549 and H522 cell lines were spiked into donor PBMCs at the indicated ratios and processed on COSMOS for target cell identification and sorting. Purity of pre-sorted and sorted cells was estimated by comparing (**A**) allele fractions with a SNP panel to the known genotypes of both the cell lines and the donor samples that they were spiked into and (**B**) a frame-shift mutation assay in the *TP53* gene (c.572_572delC), for which the H522 cell line is homozygous and the A549 cell line is wild

type. (**C**) The indicated number of A549 cells were spiked into whole blood. CD45 depletion was performed and samples were processed on COSMOS for malignant cell identification and sorting. Purity of the sorted cells and fold enrichment were quantified by SNP analysis using known genotypes of both the A549 cell line and the blood samples that they were spiked into. (**D**) Workflow schematic of COSMOS sorting and downstream molecular analysis of DTCs. (**E**) A KRAS mutation (Chr12:25245351 C>A) and two TP53 mutations (chr17:7673783 C>A and chr17:7675208 C>T) were discovered in this sample and the allele frequency in pre-sorted and sorted samples is shown between sorting runs and aliquots. (**F-H**) WGA and CNV analysis of the pre-and post-sorted samples. Each data point represents 1Mb bin. Red and blue colors indicate different chromosomes. GM12878 genomic DNA was used as baseline control for copy number normalization. (**I**) scRNAseq was performed and a t-SNE plots of gene expression profiles using all 924 feature-selected genes for pre-sorted (dark blue) and post-sorted (light blue) is shown as an overlay and (**J**) the pseudo-color gene expression level of EPCAM (cancer cell marker) is shown. (**K**) Gene expression correlation plot of mean ($\log_{10}$(molecules per cell per gene)) for the sorted and the pre-sorted cells from the EPCAM^{+}/PTPRC(CD45)$^{-}$ cluster. Each data point is a gene. The gene expression correlation coefficient (R^2) was 0.98.

Methods and Materials

Microfluidics. Each cartridge design has a microfluidic channel height between 15 µm and 40 µm, chosen to be a few micrometers greater than the largest cells to be processed. A filter region at the input port prevents large particles, cells or cell aggregates from entering the flow channel. A buffer reagent (1X PBS) is introduced into the flow alongside the cell suspension on either side, achieving hydrodynamic focusing that keeps cells flowing at a consistent speed near the center of the flow horizontally. The flow rate used (~0.1 m/s) is also high enough that the effects of inertial focusing (Di Carlo et al. 2007) are realized, confining cells to the vicinity of two vertically separated planes close to the center of the flow channel.

Bright-field imaging of cells in flow. The microfluidic cartridge is mounted on a stage with lateral (horizontal) XY control and a fine Z control for focus. The objectives, camera, laser optics and fluidics components are all mounted on the same platform. After the microfluidic cartridge is loaded into COSMOS, it is automatically aligned and a focusing algorithm is used to bring the imaging region into the field of view. An LED illumination light (SOLA SE) is directed to the imaging region, and multiple images of each cell are captured as it flows through. Bright-field images are taken through objectives of high magnification (Leica 40X - 100X) and projected onto an ultra high-speed camera. To achieve higher accuracies and adjust for potential artifacts in the image, at least two images are captured from each cell as they flow downstream in the channel. These high-resolution cell images reveal not only the cell shape and size but also finer cellular structural features within the cytoplasm and the nucleus that are useful for discriminating cell types and states based on their morphology.

Computation. The COSMOS software workload is distributed over an Intel Xeon E-2146G central processing unit (CPU), a Xeon 4108 CPU, an Nvidia Quadro P2000 Graphical Processing Unit (GPU) and a custom microcontroller. The camera is periodically polled for the availability of new images. Image frames from the high speed bright-field camera are retrieved over a dedicated 1Gbps ethernet connection. Images are cropped to center cells within them, and the cropped images are sent to the GPU for classification by an optimized CNN that has been trained on relevant cell categories. The network architecture is based on the Inception V3 model architecture (Szegedy et al. 2016), is implemented using the TensorFlow v1.15 (Abadi et al. 2016), and is trained using cell images annotated with their corresponding cell categories. NVidia TensorRT is used to create an optimized model which is used for inference on the GPU. The classification inference from the models is sent to the microcontroller, which in turn sends switching signals to synchronize the toggling of valves with the arrival of the cell at the sorting location. To maximize throughput, image processing happens in a parallel pipeline such that multiple cells can be in different stages of the pipeline at the same time. The primary use of the GPU is to run the optimized CNN. Some basic image processing tasks such as cropping cells from the images are performed on the instrument CPU. The instrument CPU is also used to control all the hardware components and to read in sensor data for monitoring. The training and validation tasks are set up as recurring Apache Beam based data processing pipelines in Google Cloud Platform (GCP). Training and prediction jobs are orchestrated by Apache Airflow, and Google Cloud Dataflow is used to combine predictions, embeddings and annotations. Models are trained using TPUPodOperators on Google Cloud on version 3 of Google's Tensor Processing Units. PostgreSQL, Google Big Query, and Google Cloud Storage are used to store and query model predictions, embeddings, and run metadata.

Data augmentation and model training. Several steps were taken to make the image classifier robust to imaging artifacts by systematically incorporating variation in cell image characteristics into our training data. Cells were imaged under a range of focus conditions to sample the effects of changes in focus during instrument runs. We gathered images across four of our instruments to sample instrument-to-instrument variation. We also implemented several augmentation methods to generate altered replicas of the cell images used to train our classifier. These included standard augmentation techniques such as horizontal and vertical flips of images, orthogonal rotation, gaussian noise, and contrast variation. We also added salt-and-pepper noise to images to mimic microscopic particles and pixel-level aberrations. Finally, we studied systematic variation in our image characteristics to develop custom augmentation algorithms that simulate chip variability and sample-correlated imaging artifacts on our microfluidic cartridge.

All cell images were resized to 299x299 pixels to make them compatible with the Inception architecture. We trained a model comprising cell types present in normal adult blood, cell types specific to fetal blood, trophoblast cell lines, and multiple cancer cell lines drawn from NSCLC, HCC, pancreatic carcinoma, acute lymphoblastic leukemia (ALL), AML. The model was also trained to detect out-of-focus images, both to use this information in auto-focusing during instrument runs and to exclude out-of-focus cell images from possible misclassification.

AI-assisted annotation of cell images. For the supervised model, we collected high-resolution images from 25.7 million cells, including cells from normal adult blood, fetal blood, trophoblast cell lines, and multiple cell lines derived from NSCLC, HCC, pancreatic carcinoma, ALL, and AML. Images were collected by an ultra high-speed bright-field camera as cell suspensions flowed through a narrow, straight channel in a microfluidics cartridge. We deployed a combination of techniques in self-supervised, unsupervised, and semi-supervised learning to facilitate cell annotation on this scale. First, we used subject and sample source data to restrict the set of class labels permitted for each cell; as an example, fetal cell class annotations were disallowed in cells drawn from non-pregnant adult subjects. Next, we extracted embedding vectors for each cell image in two pre-trained CNNs: one trained on the ImageNet dataset (Russakovsky et al. 2015) and the other on a subset of our own manually annotated cell images. We then used agglomerative clustering of these feature vectors to divide the dataset into morphologically similar clusters which were presented for manual annotation, thereby facilitating efficient cell annotation at scale.

To further enhance the accuracy of subsequent cell classification, we also selectively annotated false positive images identified from the predictions of previous trained models in an iterative manner. Finally, we balanced the classes that we wish to discriminate by feeding the harder examples of more abundant classes inspired by an active learning approach. The hard examples were identified as those that a model trained on a smaller training set had classified incorrectly (Settles 2010).

Training and validation sets. 57.4 million images were gathered to train and validate the classifier. A dataset of 25.7 million cells was imaged for the purpose of training our deep CNN in the model: PBMCs of 44 blood samples of normal adult individuals were collected which resulted in 22 million cell images. Additionally, 18 fetal blood samples were collected which yielded 2.8 million imaged cells. We imaged a total of 156,000 cells from four NSCLC cell lines, a total of 400,000 cells from four HCC cell lines, and another 440,000 cells from four cell lines of other types. A separate dataset of 25.1 million cells from 111 samples of the cell types above were gathered to validate the results of the classifier. We used the NCI-H522 (H522) cell line as the sample in validation for NSCLC and Hep 3B2.1-7 (HEP3B2) for HCC respectively.

Cell sorting. Cell sorting is performed using built-in pneumatic microvalves (Unger et al. 2000) on both the positive (targeted) and negative (waste) sides of the flow channel downstream of the bifurcation point. Valve timing is controlled by a DSP-based microcontroller circuit with 0.1ms time precision. When the model infers that a cell belongs to a targeted category, switching signals are timed to synchronize the toggling of valves with the arrival of the cell at the flow bifurcation point, and the cell flows into a reservoir on the microfluidic cartridge where targeted cells are collected (also called the positive well). If the model infers that a cell does not belong to a targeted category, the cell flows into a waste tube. Elliptical laser beams are focused onto both the positive and negative output channels downstream of the sorting flow bifurcation to detect passing cells and thereby monitor sorting performance in realtime.

Sample processing and cell culture. All human 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 (PBMCs) were isolated from whole blood by first centrifugation then the buffy coat was lysed with Red Blood Cell (RBC) Lysis Buffer (Roche) and then washed with PBS (Thermo Fisher Scientific). Fetal cells were isolated from fetal blood by directly lysing with the RBC lysis buffer then washed with PBS. Cells were then fixed with 4% paraformaldehyde (Electron Microscopy Sciences) and stored in PBS at 4ºC for longer term usage. A549, NCI-H1975, NCI-H23 (H23), NCI-H522 (H522), NCI-H810, Hep G2 (HEPG2), SNU-182, SNU-449, SNU-387, Hep 3B2.1-7 (HEP3B2), BxPC-3, PANC-1, Kasumi-1, Reh, and HTR-8/SVneo cell lines were purchased from ATCC and cultured in a humidity and CO_2-controlled 37ºC cell culture incubator according to ATCC recommended protocols.   GM12878 cell line was obtained from the NIGMS Human Genetic Cell Repository at the Coriell Institute for Medical Research and cultured according to their recommended protocols.

For neutrophil isolation and sorting, human neutrophils were isolated from whole blood using the EasySep Direct Human Neutrophil Isolation kit from Stemcell Technologies by immunomagnetic negative selection. When applicable, isolated neutrophils were labeled with a panel of primary antibodies (anti-CD3, anti-CD45, anti-CD19, anti-CD14, anti-CD66b, anti-CD15 from Biolegend) for 20 minutes at room temperature and washed twice. Propidium iodine was added to the cell mixture prior to acquisition and sorting on a BD FACSMelody instrument.

For spike-in experiments, cancer cell lines or fetal cells were first fixed with 4% paraformaldehyde and stored at 4ºC until mixing into PBMCs. For experiments in which cell lines were spiked into whole blood, live A549 cells were first stained with CellTracker Green CMFDA (Thermo Fisher Scientific), then spiked into whole blood (collected in EDTA tubes) at predefined ratios (e.g. 400 or 4000 cells in 10 mL blood), followed by buffy coat RBC lysis and fixation. Prior to loading into the sorter, the cell mixtures were pre-enriched by selective depletion of CD45 positive PBMC cells using magnetic beads (Miltenyi). Twenty percent of the samples were saved for flow cytometry analysis to estimate the number of total cells and cancer cells before and after CD45 depletion. Based on flow cytometry analysis, the CD45 magnetic bead depletion step resulted in 11-15 fold enrichment of A549 cells.

DTCs from NSCLC patients were purchased from Discovery Life Sciences. Cancer type and stage information and cell type composition report from flow cytometry were provided by the vendor. To account for possible cell type composition changes from the freeze-thaw process, after thawing the DTC aliquots, we split the samples to analyze some cells with flow cytometry and image and sort some cells on COSMOS. The panel used for flow cytometry includes markers: EpCAM, CD45, CD3, CD16, CD19. CD14, CD11b.

For cell viability assessment, pre-sorted or sorted cells were stained with either trypan blue or a Calbiochem live/dead double staining kit (Millepore Sigma) which uses a cell permeable green fluorescent Cyto-dye to stain live cells and propidium iodine to stain dead cells. Cells were then counted under a fluorescent microscope.

Molecular analyses

Single cell RNA sequencing. Cells were either directly loaded or retrieved from the positive wells of the microfluidic cartridge then loaded on a BD Rhapsody single cell analysis system (BD Biosciences). Single cells were then processed following either targeted RNA sequencing (human immune response panel) or whole transcriptome amplification protocols. The sequencing data were analyzed using BD DataView software.

Bulk RNA sequencing. Total RNA was extracted from cells using the RNeasy mini kit from Qiagen. cDNA synthesis, amplification and library preparation were performed with the Quantseq 3'm RNAseq library prep kit from Lexogen according to the manufacturer's protocol. The final libraries were sequenced on an Illumina Miniseq. Read QC, trimming, alignment and counting were performed with the Lexogen Quantseq analysis pipeline. Differential expression analysis was done using DESeq2 and iDEP (http://bioinformatics.sdstate.edu/idep/).

Genotyping. Cell lines and PBMCs 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. Briefly, genomic DNA was extracted from bulk cells using QIAGEN DNeasy Blood & Tissue Kit (Qiagen) and 1ng DNA was used as input to amplify the amplicon panels and prepare the sequencing library. For cancer cell lines, a 20-amplicon panel that covers full length of TP53 gene (Swift Biosciences) was pooled with the SampleID panel so cells were genotyped on both common SNPs and TP53 mutational status. From ATCC and COSMIC annotation, A549 cells are known to be TP53 wild type and NCI-H522 are known to carry a homozygous frameshift mutation (c.572_572delC). Our bulk genotyping results confirmed the relative mutation status for these two cell lines. For sorted cells from the COSMOS experiments, cells were retrieved from the positive outlet well of the microfluidic cartridge into a PCR tube, then directly lysed using Extracta DNA

Prep for PCR (Quanta Bio). Cell lysates were amplified with the Swift amplicon panels and followed by the same library preparation procedure for NGS.

Dissociated tumor cells from lung cancer patients. The cells before sorting and after sorting were profiled on targeted DNA mutations and copy number variations (CNV). For mutation analysis following direct lysis with Extracta DNA Prep for PCR (Quanta Bio) a 208-amplicon panel that includes 17 lung cancer genes (Swift Biosciences) were used. For CNV analysis, after direct lysis, genomic DNA was amplified using ResolveDNA Whole Genome Amplification Kit (BioSkryb Genomics) and then libraries were prepared for sequencing (Kapa Hyperplus Kit, Roche). All libraries were sequenced on either an Illumina MiniSeq or NextSeq instrument (Illumina) using 2x150 bp kit (DNA) or 2x75 bp kit (RNA).

Primary sequencing analysis and QC. Sequencing reads were aligned to the reference genome using the BWA-MEM aligner. SNP allele counts were summarized using bcftools. SNP data were subjected to quality control checks: each sample was required to have a mean coverage per SNP of > 200; each SNP locus needed to have a median coverage across all samples > 0.1x the median SNP to be considered; each individual SNP assay for a sample needed to have a depth of coverage > 50. 89 SNP assays were selected on this basis for further use in mixture analysis. Samples and individual SNP assays that failed QC were excluded from genotyping and the estimation of mixture proportions.

Mixture proportion estimation by SNP analysis. Pure diploid samples that formed the base of each mixture for spike-in experiments were clustered into the three diploid genotypes (AA, AB, BB) for each SNP using a maximum likelihood estimation that incorporated an internal estimate of error within homozygous SN. The mixture proportion of the component of interest (tumor cell line or fetal sample) was determined using maximum likelihood estimation (MLE), in which all discrete mixture fractions in increments of 0.005 were considered (0.0, 0.005, 0.01, …, 1.0). For each possible mixture proportion, expected allele fractions at each SNP were determined by linearly combining the allele fractions in the two mixture components. A binomial log likelihood corresponding to each individual sample-SNP combination was computed using the expected allele fraction and an effective number of independent reads N per SNP estimated from the variance of allele fraction in mixture SNPs at which the base genotype is heterozygous (AB) and the spike-in component genotype is homozygous (AA or BB). By estimating N from the mixture data directly and using SNPs expected to have a shared allele fraction, the procedure is robust to low input for which the number of reads might exceed the number of independent molecules sampled. The overall log likelihood for each possible mixture proportion is computed as the sum of contributions from each SNP, and the mixture proportion is estimated as that at which the highest overall log likelihood is obtained. The accuracy of the procedure was verified on DNA mixtures with known composition (**fig. S5**). Each composite sample contained 250 pg of DNA and the mixture proportion of DNA from the second individual was set at 5%, 10%, 20%, 30%, 40%, 60%, 80% and 90%.

Joint Estimation of Genotypes and Sample Purity. In two cases, genotypes and mixture fraction were jointly estimated from the allele fractions ϕ of SNPs in the mixture: (i) to genotype the fetal sample Fet1, which included some maternal cells in addition to fetal cells (ii) for the spike-in of A549 cells into whole blood. In each case, genotypes for one of the mixture components, designated $\boldsymbol{G_0}$, were obtained from a pure sample (from maternal DNA for the former, and from the pure A549 cell line for the latter), while the genotypes of the other sample, designated $\boldsymbol{G}$ (corresponding to the fetal sample in the former case and to the unrelated blood sample for the latter) were estimated from the data. The maternal sample was genotyped as diploid, but for pure A549, the allowed allele fractions for genotypes were 0, ⅓, ½, ⅔ and 1, in keeping with the known hypotriploidy of that cell line. An expectation maximization (EM) procedure was then used to jointly estimate the purity and missing genotypes. Briefly, given $\boldsymbol{G_0}$ and a current estimate of purity f, a binomial likelihood was estimated for each allowed missing genotype, and a maximum likelihood estimate was used to update $\boldsymbol{G}$. Given $\boldsymbol{G}$, a revised estimate of f was obtained by linear regression, using the expected linear relationship between the observed allele fraction ϕ and $\boldsymbol{G_0}$ over SNPs of identical $\boldsymbol{G}$. The procedure incorporated an error rate estimate drawn from the SNPs where both components are identically homozygous. The procedure was iterated until convergence, defined as changes in the purity estimate < 0.0001. Results of the EM procedure for A549 cells enriched from a starting concentration of 40 cells/mL are shown in **fig. S6**. The three dotted lines depict the linear regression used to estimate the purity given the genotypes; their slope is equal to the final purity estimate of 0.43. **Fig. 6D** also shows well-separated clusters corresponding to each of the inferred genotypes in the blood sample.

Mutation and CNV analysis. Mutation allele fractions in sorted and enriched samples (**fig. 3, B and E**) were estimated from targeted amplicon sequencing data. In each sample, the mutation allele fraction was estimated as the fraction of high-quality read alignments overlapping the mutation locus that contained the variant allele. For spike-in samples at concentrations < 1 % (1:1000, 1:10,000 or 1:100,000 in **fig. 3I**), the depicted pre-enrichment allele fraction is the experimental spike-in fraction.

Six aliquots from the GM12878 cell line, consisting of 100, 50, 25, 10, 5 and 1 cell(s) respectively, were used as a normalization cohort for copy number estimation in dissociated tumor cells before and after enrichment by sorting. Read coverage was first aggregated over 1 Mb genomic intervals across the genome within the dissociated tumor sample and each of the GM12878 normalization aliquots. The coverage within each sample was then scaled by the mean coverage per Mb over the entire genome for that sample. Next, the median assay bias and median absolute deviation (MAD) of the scaled coverage for each 1 Mb interval across the genome were computed from data from the normalization cohort. Genomic intervals for which the MAD across normalization samples exceeded 20% of the median were excluded from further analysis. Finally, the coverage values within the dissociated tumor sample before and after enrichment were further scaled by the median assay bias estimated from the normalization cohort. The resulting scaled coverage data reveal several large-scale aneuploidies in the dissociated tumor cells after sorting but not prior to the sort (**fig. 5, C, D and E**), and thereby provide strong evidence for an enrichment of tumor cells by sorting.

Supplementary Data and Figures

Cell Source	Primary Cell Class	Target Spike-in Ratio	Cells Imaged	Classifier Positive Rate	Sorted Cell Purity	Fold Enrichment
Fet1	fnRBC	1:1,304	999,978	0.017%	74%	965
A549	NSCLC	1:1,000	69,611	0.150%	62%	348
A549	NSCLC	1:1,000	101,180	0.170%	67%	380
A549	NSCLC	1:10,000	1,105,997	0.060%	27%	1,978
A549	NSCLC	1:10,000	876,421	0.099%	17%	1,201
A549	NSCLC	1:10,000	1,107,669	0.025%	31%	2,305
A549	NSCLC	1:10,000	1,063,745	0.0083%	42%	4,200
A549	NSCLC	1:10,000	1,169,744	0.0094%	33%	3,300
A549	NSCLC	1:10,000	719,499	0.028%	33%	3,300
H522	NSCLC	1:10,000	1,050,036	0.030%	26%	2,550
A549	NSCLC	1:100,000	1,342,632	0.003%	20%	13,904
H522	NSCLC	1:100,000	1,514,263	0.005%	30%	30,000
H522	NSCLC	1:100,000	1,561,847	0.006%	33%	32,500

Table S1. Enrichment of cells spiked into PBMCs. Fet1 is a fetal blood sample spiked into cells from the corresponding maternal sample. Cells from the A549 and H522 cell lines were spiked into PBMCs from a healthy donor. Cell mixtures were flown through, imaged and target cell sorted via COSMOS system. In some experiments, actual spike-in ratios of the mixtures were estimated and confirmed by pre-staining cancer cells with a fluorescent Cell Tracker dye before mixing into PBMC or whole blood and then analyzing a portion of cells using flow cytometry. Classifier Positive Rate is the percentage of the classifier identified positive cells against all imaged cells. Sorted Cell purity was estimated by comparing allele fractions using a SNP panel to the known genotypes of both the cell lines and the samples that they were spiked into and normalized by copy numbers of the cell lines. Fold enrichment was calculated by the SNP-estimated purity in sorted cells divided by target purity or flow cytometry-based estimation of pre-sort cells.

Cell Line	Proportion in Mixture	Proportion of TP53 frameshift c.572_572delC	Fold enrichment of c.572_572delC
A549	100%	0%	--
NCI-H522	100%	100%	--
NCI-H522	0.1% (1:1,000)	45%	451
NCI-H522	0.01% (1:10,000)	15%	1,499
NCI-H522	0.001% (1:100,000)	23%	22,770

Table S2. Detection and enrichment of a known frame-shift mutation in the TP53 gene for which the NCI-H522 cell line is homozygous. The indicated cell lines were spiked into healthy donor PBMCs 0.1% (1:1,000), 0.01% (1:10,000) and 0.001% (1:100,000). Each of these mixtures was then enriched using COSMOS. DNA from the enriched cells was assayed for the frame-shift mutation. In each case, the mutation was detected with an allele fraction of 15% or more. For the 1:100,000 spike-in mixture, an enrichment of 22,770x was achieved.

Spike-in Cell Concentration	Percentage of A549 after RBC lysis	Percentage of A549 after CD45 depletion	Fold Enrichment by CD45 depletion	Cells Imaged	Classifier Positive Rate	Sorted Cell Purity	Overall Fold Enrichment
400/mL	0.004%	0.06%	13	1,029,175	0.019%	55%	10,900
400/mL	0.003%	0.06%	16.2	932,665	0.018%	80%	29,000
40/mL	0.001%	0.01%	11	949,836	0.007%	43%	33,500
40/mL	0.001%	0.01%	6.7	1,012,315	0.009%	35%	27,800

Table S3. Enrichment with COSMOS.cA549 cells spiked into healthy donor whole blood at concentrations of 400 cells/mL or 40 cells/mL and processed and sorted with COSMOS. An additional CD45 depletion step was used to partly enrich the A549 cells prior to COSMOS sorting. A549 cells were pre-stained with a fluorescent cell tracker dye before spike-in. A portion of cells after RBC lysis and after CD45 depletion were analyzed with flow cytometry to estimate the fraction of A549 cells in the mixtures after each step. A549 cell purities were estimated from SNP analysis of sorted cells.

Sample and cell type	Number of runs	% Viability of pre-sorted cells (average ± SD)	% Viability of sorted or flown-through cells (average ± SD)
PBMC	13	94.7 ± 2.5%	94.2 ± 3.8%
B-lymphoblastoid cell line (GM12878)	10	90.6 ± 3.1%	91.0 ± 5.4%
Cancer cell lines (H522 and A549)	13	96.1 ± 2.0%	96.9 ± 2.2%

Dissociated NSCLC tissue	3	71.0 ± 4.4%	65.0 ± 1.0%

Table S4. Cell viability after flowing through or sorting on COSMOS.

Adjusted p-value	Number of Genes	Pathways	Genes
1.54E-04	47	Symbiont process	LTF THOC5 EIF3D CALM2 CALM3 CAMP PCBP1 JUN APOBEC3G CCL3 MPO PSMB1 MRE11 EIF2AK2 WAPL TNPO1 KPNA3 SGTA FKBP8 DDX58 VPS29 LTBR NFE2L2 CD46 VPS4B IRF3 BECN1 ILF3 TRIM22 TANK TAF1 ZYX ITGA5 SYK LMBRD1 RAB6A SATB1 KPNA4 NUCKS1 RAB1B PTBP1 SERPINB9 AP2S1 RPL21 RPS4Y1 NMT2 POLR2A
1.41E-03	41	Viral process	EIF3D PCBP1 JUN APOBEC3G CCL3 PSMB1 MRE11 EIF2AK2 WAPL TNPO1 KPNA3 SGTA FKBP8 DDX58 VPS29 LTBR NFE2L2 CD46 VPS4B IRF3 BECN1 ILF3 TRIM22 TANK TAF1 ZYX ITGA5 SYK LMBRD1 RAB6A SATB1 KPNA4 NUCKS1 RAB1B PTBP1 LTF AP2S1 RPL21 RPS4Y1 NMT2 POLR2A
1.41E-03	75	Cellular response to stress	MRE11 MAP2K4 UBE2A ACD SGTA LTBR NPRL2 NFE2L2 PRDX6 RPA2 BECN1 HSPA2 RBM17 ERCC5 HUS1 OXSR1 TMEM259 UBE2G2 ZNF652 TAF1 SIRT6 GSK3B SUSD6 GADD45G POLN RHOB OPA1 SMC5 MPO BID EIF2AK2 TSPO CINP CCDC47 COL4A3BP SLC12A4 MAX DYSF TANK PARP9 PEA15 SYK ATMIN TERF2IP DDIT4 NABP1 JUN CALR UPP1 ERO1A IRAK4 AGER SIPA1 CCDC88C PDK3 NUCKS1 THOC5 CBX3 MCM7 FEM1B PTTG1IP KMT5A CHUK ARHGEF6 PSMB1 TAB2 SH2D3C CAMKK2 IRF3 CNOT6L TPP1 DNAJC7 YIF1A POLR2A GFPT1
1.52E-03	93	Protein localization	TMSB10 TBC1D22A ZC3H11A WAPL TNPO3 RABL2B TNPO1 SUN2 THOC5 TM9SF1 ACD STX10 SGTA KDELR1 VPS29 TM9SF2 BECN1 KIF13A SRP14 SRP9 IMMP1L COMMD1 RAB1B ZDHHC14 RAB6A SVBP EXOC3 PTTG1IP IPO9 BID BTN3A1 TOLLIP GSK3B TESC DVL1 SRSF3 RPA2 APPL2 FLCN CALR CD24 MYCBP2 LSG1 AP2S1 DGKD BIRC5 SNAP29 TSPO KPNA3 CORO1C VPS4B FFAR1 IRF3 KTN1 YWHAQ DYSF CTDSPL2 ARL3 ARL8A RHOB FRMD4A AHCTF1 GHRL FCHO2 ZFAND2B CALM3 SYK GAPVD1 FNTA YIF1A SMAD2 RABEP2 KPNA4 AGER SNX2 DENND1B SFT2D2 PARP9 SCIN YBX1 CD33 DDX58 TRIM22 TERF2IP UBE2G2 CHUK CCL3 GNAS CASC3 RPL21 RPS4Y1 HAX1 ATP6V1B2
1.52E-03	73	Protein transport	TBC1D22A ZC3H11A TNPO3 RABL2B TNPO1 THOC5 STX10 SGTA VPS29 KIF13A SRP14 SRP9 IMMP1L RAB1B ZDHHC14 RAB6A SVBP PTTG1IP IPO9 BTN3A1 GSK3B SRSF3 APPL2 CALR CD24 BID LSG1 AP2S1 DGKD TESC BIRC5 SNAP29 KPNA3 KDELR1 VPS4B FFAR1 IRF3 KTN1 YWHAQ DYSF CTDSPL2 ARL3 ARL8A RHOB FRMD4A AHCTF1 GHRL ZFAND2B SYK GAPVD1 COMMD1 YIF1A SMAD2 RABEP2 EXOC3 KPNA4 AGER SNX2 DENND1B SFT2D2 ACD CD33 DDX58 UBE2G2 CHUK CCL3 GNAS TSPO CASC3 RPL21 RPS4Y1 HAX1 ATP6V1B2
1.52E-03	75	Amide transport	TBC1D22A ZC3H11A TNPO3 RABL2B TNPO1 THOC5 STX10 SGTA VPS29 COL4A3BP KIF13A SRP14 SRP9 IMMP1L GHRL SLC19A1 RAB1B ZDHHC14 RAB6A SVBP PTTG1IP IPO9 BTN3A1 GSK3B SRSF3 APPL2 CALR CD24 BID LSG1 AP2S1 DGKD TESC BIRC5 SNAP29 KPNA3 KDELR1 VPS4B FFAR1 IRF3 KTN1 YWHAQ DYSF CTDSPL2 ARL3 ARL8A RHOB FRMD4A AHCTF1 ZFAND2B SYK GAPVD1 COMMD1 YIF1A SMAD2 RABEP2 EXOC3 KPNA4 AGER SNX2 DENND1B SFT2D2 ACD CD33 DDX58 UBE2G2 CHUK CCL3 GNAS TSPO CASC3 RPL21 RPS4Y1 HAX1 ATP6V1B2
1.60E-03	67	Intracellular transport	GSK3B CCDC88C TBC1D22A ZC3H11A TNPO3 RABL2B TNPO1 SNAP29 THOC5 STX10 SGTA KDELR1 VPS29 COL4A3BP BECN1 KIF1C KIF13A SRP14 SRP9 IMMP1L SYK RAB1B ZDHHC14 RAB6A PTTG1IP IPO9 SNX2 DENND1B ATP5MG MT-ATP8 LSG1 CORO1C SRSF3 VPS4B APPL2 RHOB TMCC1 CALR CD24 BID AP2S1 SUN2 KPNA3 DVL1 TFG RAP1B YWHAQ CTDSPL2 ARL3 ARL8A ZFAND2B CALM3 KPNA4 SRSF10 ACD UBE2G2 MYO10 GNAS HMGXB4 TSPO CASC3 RPL21 WDR60 RPS4Y1 ACTR10 HAX1 OPA1
2.05E-03	73	Peptide transport	TBC1D22A ZC3H11A TNPO3 RABL2B TNPO1 THOC5 STX10 SGTA VPS29 KIF13A SRP14 SRP9 IMMP1L GHRL RAB1B ZDHHC14 RAB6A SVBP PTTG1IP IPO9 BTN3A1 GSK3B SRSF3 APPL2 CALR CD24 BID LSG1 AP2S1 DGKD TESC BIRC5 SNAP29 KPNA3 KDELR1 VPS4B FFAR1 IRF3 KTN1 YWHAQ DYSF CTDSPL2 ARL3 ARL8A RHOB FRMD4A AHCTF1 ZFAND2B SYK GAPVD1 COMMD1 YIF1A SMAD2 RABEP2 EXOC3 KPNA4 AGER SNX2 DENND1B SFT2D2 ACD CD33 DDX58 UBE2G2 CHUK CCL3 GNAS TSPO CASC3 RPL21 RPS4Y1 HAX1 ATP6V1B2
2.05E-03	75	Establishment of protein localization	TBC1D22A ZC3H11A TNPO3 RABL2B TNPO1 THOC5 STX10 SGTA VPS29 KIF13A SRP14 SRP9 IMMP1L RAB1B ZDHHC14 RAB6A SVBP PTTG1IP IPO9 BID BTN3A1 GSK3B SRSF3 APPL2 CALR CD24 LSG1 AP2S1 DGKD TESC BIRC5 SNAP29 TSPO KPNA3 KDELR1 CORO1C VPS4B FFAR1 IRF3 KTN1 YWHAQ DYSF CTDSPL2 ARL3 ARL8A RHOB FRMD4A AHCTF1 GHRL ZFAND2B CALM3 SYK GAPVD1 COMMD1 YIF1A SMAD2 RABEP2 EXOC3 KPNA4 AGER SNX2 DENND1B SFT2D2 ACD CD33 DDX58 UBE2G2 CHUK CCL3 GNAS CASC3 RPL21 RPS4Y1 HAX1 ATP6V1B2
2.08E-03	29	Small GTPase mediated signal transduction	RABL2B GNA13 RHOF RHOB RAB1B RAB6A ARL3 JUN CCDC125 DENND4B RFXANK SH2D3C SOS2 DOCK8 RALGPS2 RAP1B ARPP19 ARHGEF6 YWHAQ BNIP2 VAV1 TIAM2 DOCK11 KCTD13 CHUK SIPA1 FLCN ARHGAP25 SIAH2
2.11E-03	76	Establishment of localization in cell	GSK3B CCDC88C TBC1D22A ZC3H11A TNPO3 RABL2B TNPO1 SNAP29 THOC5 STX10 SGTA KDELR1 VPS29 COL4A3BP BECN1 KIF1C KIF13A PLCB2 SRP14 SRP9 IMMP1L FCHO2 SYK RAB1B ZDHHC14 RAB6A PTTG1IP IPO9 SNX2 DENND1B CALM2 CALM3 ATP5MG MT-ATP8 LSG1 CORO1C SRSF3 VPS4B APPL2 RHOB TMCC1 CALR CD24 CCL3 BID AP2S1 BIRC5 SUN2 KPNA3 DVL1 TFG RAP1B YWHAQ CTDSPL2 ARL3 ARL8A GHRL ZFAND2B MEI1 KPNA4 SRSF10 ERO1A ACD CEP19 UBE2G2 MYO10 GNAS HMGXB4 TSPO CASC3 RPL21 WDR60 RPS4Y1 ACTR10 HAX1 OPA1
2.54E-03	96	Negative regulation of metabolic process	HIVEP1 PHF12 ATAD2B ARPP19 FXR2 CARHSP1 REL SERPINB9 JUN SVBP PTBP1 LTF EIF2AK2 SMARCA2 AES SGTA DVL1 TRIM37 CSRNP2 NPRL2 CD46 TSHZ3 CBX3 MAX ILF3 YWHAQ TUT4 TAF1 MGAT5 FLCN PHF6 GHRL TERF2IP ZBTB33 CALR KMT5A SRSF10 ARID5A CCL3 MYCBP2 YBX1 SIRT6 TCF7 TESC BIRC5 MAEA TSPO CHD8 PHF14 CASC3 BIRC6 NMI IRF3 BECN1 MLLT1 TRIM22 AKIRIN2 MICAL1 DYSF APPL2 HUS1 CNOT6L SRP9 YY1AP1 RYBP DDIT4 ESRRA SMAD2 CEP295NL SIAH2 RNF41 SATB1 RBM10 CEBPD ACD MRE11 WAPL FKBP8 CORO1C PARP9 FAM192A ANGEL2 UBE2G2 TFE3 GSK3B JAK3 PSMB1 TAB2 TNPO1 RPL21 RPS4Y1 CALM2 CALM3 RBBP4 S100A11 POLR2A
2.54E-03	30	Myeloid leukocyte mediated immunity	SYK CCL3 DDX58 SERPINB9 IRAK4 MPO PSMB1 LTF CAPN1 ATP11A TOLLIP SNAP29 ASAH1 CD33 PDAP1 BST1 PRDX6 CYSTM1 RAP1B ACTR10 RHOF SRP14 C1orf35 ARL8A SURF4 EEF1A1 S100A11 CAMP OSCAR RAB6A
2.71E-03	66	Regulation of intracellular signal transduction	TFG BID MAP2K4 LTBR NPRL2 GADD45G IRAK4 CCL3 LTF GNAS TRIM22 CALM2 MGAT5 CALM3 RNF41 CCDC125 DENND4B CD24 EIF2AK2 SOS2 RALGPS2 IRF3 BECN1 RAP1B ARPP19 ARHGEF6 BNIP2 VAV1 TIAM2 FLCN SYK TERF2IP DDIT4 KCTD13 JUN CALR AGER SIPA1 TAB2 REL OPA1 PTBP1 GSK3B BST1 NFE2L2 TANK PARP9 HAX1 GHRL ARHGAP25 PTTG1IP KMT5A CHUK MRE11 TAF2 PPP2R5C JAK3 RPA2 GNA13 HUS1 RHOF RHOB TAF1 RBBP4 PEA15 TAF15
2.79E-11	83	Intracellular transport	NPEPPS PICALM RAB7A SAR1A RAB2A SEC61B IFT20 MAK SNX17 TMED5 SPCS3 NAPG BLOC1S1 STX11 RBM26 RHOT2 DOP1B HSPD1 RAB5A GOLGA7 RABGAP1L RANBP2 ZDHHC7 STX5 ZDHHC3 TBC1D10C FES BCAP31 POM121 CHMP1B EPS15 MAVS ATXN1 ZC3H12A RNF139 IFIT1 CHP1 DDX39B PDCD6 LAMP2 CD74 CDC42 CBL ERBIN HSPA9 RAP1A TXN MAPK8IP3 LAMTOR1 RILPL2 PAN3 SRP68 CREB3L2 HGS FAM160A2 GSK3A AUP1 PEX16 OSBPL2 AKT1 DPY30 TSC1 LGALS9 ARHGAP1 PTPN1 TAP2 VAMP2 ACAA1 CAPZB MAPK3 RPL34 PPP6C RPL36 HMGA1 EIF4A3 DYNC1LI1 FIP1L1 RPS3A DDX19B ACOX1 RPL9 RPS26 RPL23A
2.79E-11	93	Establishment of localization in cell	NPEPPS PICALM RAB7A SAR1A RAB2A SEC61B IFT20 MAK SNX17 TMED5 SPCS3 NAPG BLOC1S1 STX11 KIAA1109 RBM26 RHOT2 DOP1B HSPD1 RAB5A GOLGA7 RABGAP1L RANBP2 ZDHHC7 STX5 ZDHHC3 TBC1D10C FES BCAP31 POM121 CHMP1B EPS15 MAVS ATXN1 ZC3H12A RNF139 IFIT1 CHP1 DDX39B PDCD6 LAMP2 CD74 BAK1 CDC42 CBL ERBIN HSPA9 ITGA4 RAP1A CANX PPT1 TXN CDK5RAP2 MAPK8IP3 LAMTOR1 RILPL2 PAN3 SRP68 SLC30A1 CREB3L2 ACTG1 HGS NSFL1C UBXN2B FAM160A2 GSK3A AUP1 PEX16 OSBPL2 AKT1 DPY30 TSC1 LGALS9 ARHGAP1 PTPN1 TAP2 VAMP2 ACAA1 CAPZB MAPK3 RPL34 PPP6C RPL36 HMGA1 EIF4A3 DYNC1LI1 FIP1L1 RPS3A DDX19B ACOX1 RPL9 RPS26 RPL23A
5.61E-10	86	Vesicle-mediated transport	CDC42 PICALM RAB7A SAR1A SNX17 TMED5 PPT1 BLOC1S1 STX11 KIAA1109 MAPK8IP3 DOP1B RAB5A GOLGA7 C9orf72 VPS8 XKR8 STX5 FES BCAP31 CHMP1B EPS15 CEACAM1 CSK RNF139 PTPN1 PYCARD RAB2A RAPGEF1 CBL ITGA4 RAP1A MAPK8IP1 CANX NAPG LAMTOR1 RABGAP1L PI4K2A ELMO1 CDC42SE2 CD14 CREB3L2 ACTG1 CHP1 AIF1 VAMP2 FAM160A2 NOTCH1 DPY30 LGALS9 ARHGAP1 TBC1D10C HGS PDCD6 GABARAPL2 LAMP2 CD44 ACAA1 CAPZB MAGT1 MAPK3 MAN2B1 GCA QPCT PPP6C ARAP3 RAP2C ARPC1B JCHAIN CHI3L1 PRCP ADAM10 DYNC1LI1 FAM49B CDA APEH FPR1 LRG1 ATP6AP2 PSMD13 ATP6V0C XRCC6 TAP2 TMSB4X PLEKHO2 MGAM
4.05E-09	56	Intracellular protein transport	NPEPPS RAB7A SAR1A RAB2A SEC61B SNX17 TMED5 SPCS3 NAPG STX11 HSPD1 RAB5A GOLGA7 RABGAP1L RANBP2 ZDHHC7 STX5 ZDHHC3 TBC1D10C POM121 MAVS ZC3H12A IFIT1 BCAP31 CHP1 DDX39B PDCD6 LAMP2 CD74 CDC42 ERBIN HSPA9 TXN MAPK8IP3 RILPL2 PAN3 SRP68 HGS GSK3A AUP1 PEX16 AKT1 TSC1 PTPN1 VAMP2 ACAA1 RPL34 RPL36 EIF4A3 FIP1L1 RPS3A DDX19B ACOX1 RPL9 RPS26 RPL23A
4.05E-09	119	Positive regulation of metabolic process	CD44 SBNO2 KAT6A MAP3K1 RBM23 RBM3 MED26 CHCHD2 CAMTA2 VDR DPF2 MAPK8IP3 ASXL2 LAMTOR1 ING1 ZNF296 CAMTA1 MAP3K5 APOPT1 TXN CD74 BAK1 SPI1 PICALM MAVS FUS MUL1 PYCARD LYL1 CDKN1B STAT1 ATF6 JCHAIN NOTCH2 UBQLN1 CDK5RAP2 PELI2 AKT1 HSPD1 OGT NOTCH1 YTHDF1 FLI1 KLF10 BTG2 ZC3H12A BRD7 LGALS9 ATF7IP CEBPB ZNF24 MLXIP AURKAIP1 TBL1XR1 ST20 CREB3L2 MAFF SP1 YTHDF3 BCAP31 SEMA4D PTPN1 XRCC6 GPATCH3 TMSB4X PDCD6 HIPK2 TAF11 TNFRSF1A MNT CDC42 EPS15 MAPK3 CSK NFKBIB USF2 BUD31 RAPGEF1 PDLIM1 MAK KCTD20 RAP1A MAPK8IP1 RAP2C CHI3L1 RANBP2 NPTN ADIPOR1 TSC1 CCS CHP1 DDX39B AIF1 DDT GSK3A C9orf72 RAB7A PPP2CA CCT4 RPS6KA1 CLN6 FBH1 HMGA1 EIF4A3 CACUL1 STX5 PIP4K2C RNF187 IRF2 SIN3A RNF139 HGS LILRA5 SIRT7 PRKAR1A SF3B1 TAB3 FPR1 HMGN1

4.05E-09	82	Protein transport	NPEPPS RAB7A SAR1A RAB2A SEC61B SNX17 TMED5 SPCS3 NAPG STX11 HSPD1 RAB5A GOLGA7 RABGAP1L RANBP2 ZDHHC7 STX5 ZDHHC3 CD14 TBC1D10C POM121 CHMP1B MAVS PYCARD CSK RAP1A ZC3H12A LGALS9 IFIT1 BCAP31 LILRA5 CHP1 DDX39B AIF1 TMSB4X PDCD6 LAMP2 CD74 GABARAPL2 FAM160A2 CYB5R4 CDC42 EPS15 MAPK3 IFT20 ERBIN HSPA9 CHI3L1 TXN MAPK8IP3 DOP1B RILPL2 PAN3 VPS8 DDX19B SRP68 HGS TAP2 NOTCH1 GSK3A AUP1 PEX16 PPT1 AKT1 TSC1 ARHGAP1 PTPN1 HCAR2 VAMP2 ACAA1 RPL34 CANX RPL36 ATP6V1G1 EIF4A3 FIP1L1 RPS3A ACOX1 RPL9 ATP6V0C RPS26 RPL23A
4.05E-09	49	Viral process	USF2 CD74 TAF11 MAVS TRIM25 ZC3H12A LGALS9 SP1 IFIT1 MCTS1 CHMP1B SPEN TNFRSF1A EPS15 MAPK3 SELPLG RNGTT SKP1 STAT1 HSPD1 DYNC1LI1 OGT F11R PSMB4 ATF7IP BCAP31 ATP6V0C ISG15 MAP3K5 PSMB8 TAP2 CDC42 RAB7A TOP2B RAB5A NOTCH1 ISG20 HIPK2 RPL34 RPL36 HMGA1 RPS3A RANBP2 ELMO1 RPL9 POM121 XRCC6 RPS26 RPL23A
4.05E-09	51	Symbiont process	USF2 CD74 TAF11 MAVS TRIM25 ZC3H12A LGALS9 SP1 IFIT1 DDX39B MCTS1 CHMP1B SPEN TNFRSF1A EPS15 MAPK3 SELPLG RNGTT SKP1 STAT1 HSPD1 DYNC1LI1 OGT F11R PSMB4 ATF7IP BCAP31 ATP6V0C ISG15 MAP3K5 PSMB8 TAP2 CDC42 RAB7A TOP2B RAB5A NOTCH1 ISG20 HIPK2 RPL34 CBL RPL36 HMGA1 RPS3A RANBP2 ELMO1 RPL9 POM121 XRCC6 RPS26 RPL23A
4.05E-09	104	Cellular localization	RAP1A CCT4 NPEPPS PICALM RAB7A SAR1A RAB2A SEC61B IFT20 MAK SNX17 TMED5 SPCS3 NAPG BLOC1S1 STX11 KIAA1109 RBM26 RHOT2 DOP1B HSPD1 RAB5A GOLGA7 LAMTOR1 RABGAP1L RANBP2 ZDHHC7 STX5 ZDHHC3 TBC1D10C FES BCAP31 POM121 TMSB4X CHMP1B EPS15 MAVS CSK PEX16 ATXN1 ZC3H12A RNF139 IFIT1 CHP1 DDX39B PDCD6 LAMP2 CD74 BAK1 CDC42 CBL ERBIN HSPA9 DBN1 ITGA4 CANX PPT1 TXN CDK5RAP2 ADAM10 MAPK8IP3 AKT1 RILPL2 PAN3 NPTN SRP68 SIN3A SLC30A1 CREB3L2 ACTG1 HGS NSFL1C UBXN2B FAM160A2 TNFRSF1A MUL1 GSK3A AUP1 OSBPL2 F11R DPY30 TSC1 LGALS9 ARHGAP1 PTPN1 TAP2 SKP1 VAMP2 ACAA1 CAPZB MAPK3 RPL34 PPP6C RPL36 HMGA1 EIF4A3 DYNC1LI1 FIP1L1 RPS3A DDX19B ACOX1 RPL9 RPS26 RPL23A
4.39E-09	112	Positive regulation of cellular metabolic process	CD44 SBNO2 KAT6A MAP3K1 RBM23 RBM3 CHCHD2 CAMTA2 VDR DPF2 MAPK8IP3 ASXL2 LAMTOR1 ING1 ZNF296 CAMTA1 MAP3K5 APOPT1 TXN CD74 BAK1 SPI1 PICALM MAVS FUS MUL1 PYCARD LYL1 STAT1 ATF6 NOTCH2 UBQLN1 CDK5RAP2 PELI2 AKT1 HSPD1 OGT NOTCH1 YTHDF1 FLI1 KLF10 BTG2 ZC3H12A BRD7 ATF7IP CEBPB ZNF24 MLXIP AURKAIP1 TBL1XR1 ST20 CREB3L2 MAFF SP1 YTHDF3 BCAP31 SEMA4D PTPN1 XRCC6 GPATCH3 TMSB4X PDCD6 HIPK2 TAF11 TNFRSF1A MNT CDC42 EPS15 MAPK3 CSK NFKBIB MED26 USF2 BUD31 RAPGEF1 PDLIM1 CDKN1B MAK KCTD20 RAP1A MAPK8IP1 RAP2C CHI3L1 RANBP2 NPTN TSC1 CHP1 DDX39B AIF1 DDT GSK3A C9orf72 PPP2CA CCT4 RPS6KA1 CLN6 FBH1 HMGA1 EIF4A3 CACUL1 PIP4K2C RNF187 IRF2 LGALS9 SIN3A RNF139 LILRA5 SIRT7 PRKAR1A TAB3 FPR1 HMGN1
7.76E-09	82	Peptide transport	NPEPPS RAB7A SAR1A RAB2A SEC61B SNX17 TMED5 SPCS3 NAPG STX11 HSPD1 RAB5A GOLGA7 RABGAP1L RANBP2 ZDHHC7 STX5 ZDHHC3 CD14 TBC1D10C POM121 TAP2 CHMP1B CD74 MAVS PYCARD CSK RAP1A ZC3H12A LGALS9 IFIT1 BCAP31 LILRA5 CHP1 DDX39B AIF1 TMSB4X PDCD6 LAMP2 GABARAPL2 FAM160A2 CYB5R4 CDC42 EPS15 MAPK3 IFT20 ERBIN HSPA9 CHI3L1 TXN MAPK8IP3 DOP1B RILPL2 PAN3 VPS8 DDX19B SRP68 HGS NOTCH1 GSK3A AUP1 PEX16 PPT1 AKT1 TSC1 ARHGAP1 PTPN1 HCAR2 VAMP2 ACAA1 RPL34 CANX RPL36 ATP6V1G1 EIF4A3 FIP1L1 RPS3A ACOX1 RPL9 ATP6V0C RPS26 RPL23A
7.76E-09	84	Establishment of protein localization	NPEPPS RAB7A SAR1A RAB2A SEC61B SNX17 TMED5 SPCS3 NAPG STX11 HSPD1 RAB5A GOLGA7 RABGAP1L RANBP2 ZDHHC7 STX5 ZDHHC3 CD14 TBC1D10C POM121 CHMP1B MAVS PYCARD CSK RAP1A ZC3H12A LGALS9 IFIT1 BCAP31 LILRA5 CHP1 DDX39B AIF1 TMSB4X PDCD6 LAMP2 CD74 GABARAPL2 FAM160A2 CYB5R4 CDC42 EPS15 MAPK3 IFT20 ERBIN HSPA9 CHI3L1 TXN MAPK8IP3 DOP1B RILPL2 PAN3 VPS8 DDX19B SRP68 HGS TAP2 NOTCH1 GSK3A AUP1 CCT4 PEX16 PPT1 AKT1 TSC1 ARHGAP1 PTP4A3 PTPN1 HCAR2 VAMP2 ACAA1 RPL34 CANX RPL36 ATP6V1G1 EIF4A3 FIP1L1 RPS3A ACOX1 RPL9 ATP6V0C RPS26 RPL23A
8.16E-09	108	Positive regulation of nitrogen compound metabolic process	CD44 SBNO2 KAT6A MAP3K1 RBM23 RBM3 CHCHD2 CAMTA2 VDR DPF2 MAPK8IP3 ASXL2 LAMTOR1 ING1 ZNF296 CAMTA1 MAP3K5 APOPT1 TXN CD74 BAK1 SPI1 PICALM MAVS FUS MUL1 PYCARD LYL1 CDKN1B STAT1 ATF6 NOTCH2 UBQLN1 CDK5RAP2 PELI2 AKT1 HSPD1 OGT NOTCH1 YTHDF1 FLI1 KLF10 BTG2 ZC3H12A BRD7 ATF7IP CEBPB ZNF24 MLXIP AURKAIP1 TBL1XR1 ST20 CREB3L2 MAFF SP1 YTHDF3 BCAP31 SEMA4D PTPN1 XRCC6 GPATCH3 TMSB4X PDCD6 HIPK2 TAF11 TNFRSF1A MNT CDC42 MAPK3 CSK NFKBIB MED26 USF2 BUD31 RAPGEF1 PDLIM1 MAK RAP1A MAPK8IP1 RAP2C CHI3L1 NPTN CHP1 DDX39B AIF1 DDT GSK3A RAB7A PPP2CA CCT4 RPS6KA1 CLN6 FBH1 HMGA1 EIF4A3 CACUL1 STX5 RNF187 IRF2 LGALS9 SIN3A RNF139 LILRA5 SIRT7 PRKAR1A TAB3 FPR1 HMGN1
1.29E-08	102	Intracellular signal transduction	MAVS MUL1 LGALS9 CD44 BAK1 MAPK6 CDC42 RAB7A MAP3K1 MAPK3 PYCARD RAB2A PRKAR1A MAK MAPK8IP3 RHOT2 AKT1 RAB5A LAMTOR1 TSC1 CD14 TNFAIP8L1 GNG2 MAP3K5 ARHGAP19 APOPT1 TXN CD74 WWC3 HIPK2 STAT1 TRIM25 NOTCH2 PELI2 RIT1 OGT NOTCH1 ZC3H12A BAG5 ARHGAP1 ATP6AP2 PTP4A3 HGS CHP1 PDCD6 TNFRSF1A POLB CSK RHEB RAPGEF1 CBL ERBIN RAP1A RPS6KA1 MAPK8IP1 RAP2C PSD4 CHI3L1 C9orf72 PIK3AP1 NPTN CDC42SE2 ADIPOR1 DUSP18 CEBPB PTPN1 GPATCH3 AIF1 DDT ELMO1 TMSB4X PPIF FBH1 UBQLN1 TAB3 CAMTA1 BCAP31 LILRA5 SEMA4D CEACAM1 GSK3A PPP2CA G3BP2 CARD19 TAF11 KAT6A CORO2A CDKN1B CNOT2 SKP1 ARAP3 LY96 PSMD4 PSMB4 BTG2 TOPBP1 BRD7 TAF10 PIP4K2C FPR1 PSMD13 PSMB8
1.07E-14	172	Response to organic substance	FOXRED2 TSC2 ZNF106 IL10RA MAPKAPK3 SERP1 MED1 ANXA1 ABHD2 NLRC5 IFITM3 IFNAR1 MARCH6 MAPKAPK2 TGFBR2 RAB8A SGTB UPF1 RUNX3 HERPUD1 ATP2B4 CYLD REST ICAM1 SNW1 PPM1A CARD8 PADI2 FFAR2 SNX6 DSTYK MDM2 TLR2 PAK1 SPATA2 ABCG1 CTNNB1 UBE2O MX2 CIB1 SELL DAXX ATF6B STRAP DGAT2 DNM2 ATRX SNX10 RFX2 GRAMD1A RANGAP1 YY1 TRIB3 EHD4 GAB1 KLF3 CHKA ARPC3 PDCD10 GNAI2 EIF4G1 STXBP3 RAD23B DNMT3A IFIT2 PTK2B HERPUD2 ITPR2 MBD4 LDLR KRAS TOR1A ACTR2 DUSP6 TOB1 ARHGDIA MTF2 ZEB1 EIF4EBP2 VPS26B CDK19 KAT6B EIF4A2 ZBTB7B RBM15 MNDA IFI16 NFIL3 INPPL1 TET2 PTAFR SYAP1 STX8 AKAP13 PRPF8 RPS6KA3 TAF7 PDE4B IRF7 STRN3 MYO5A NUB1 ST3GAL6 PSEN1 RTF2 ZFAND6 EZR RPS6KA5 RIOK3 EMD ZNF451 KLF7 PTPN12 CDKN1C LILRB2 LAMTOR5 KAT7 FLOT1 TRIM41 EPG5 UBXN1 H2AFZ KMT2D STAT2 BCL9L CD47 GET4 LILRA2 ACAP2 TCF7L2 GNAQ PIAS1 SHOC2 GAB2 PSMA4 ATP6V1H UBE2D1 ACADVL ADD1 DNAJB11 PSMD8 HNRNPM ATP6V1D PSME2 ARFGAP1 HCK XIAP POLR2C TYK2 PSMD3 UBE2D3 OAS3 TRIM38 ATP6V1A SUMO1 TNFSF14 RBCK1 PSME3 TLN1 MBTPS1 ARF1 HNRNPDL MX1 TNFRSF14 ATP6V0D1 UBB TALDO1 SRPRA HLA-DRA DCTN1 PSMB9 PSMB3
1.07E-14	151	Cellular response to organic substance	TSC2 ZNF106 IL10RA SERP1 MED1 ANXA1 ABHD2 NLRC5 IFITM3 IFNAR1 MAPKAPK2 TGFBR2 RAB8A UPF1 ATP2B4 CYLD REST ICAM1 SNW1 PPM1A CARD8 PADI2 FFAR2 SNX6 DSTYK MDM2 TLR2 PAK1 SPATA2 CTNNB1 UBE2O CIB1 DAXX ATF6B STRAP HERPUD1 DGAT2 DNM2 SNX10 RFX2 GRAMD1A RANGAP1 YY1 TRIB3 EHD4 KLF3 ARPC3 PDCD10 GNAI2 STXBP3 RAD23B DNMT3A IFIT2 PTK2B ITPR2 LDLR KRAS TOR1A ACTR2 TOB1 ARHGDIA MTF2 ZEB1 EIF4EBP2 VPS26B CDK19 KAT6B EIF4A2 ZBTB7B RBM15 MNDA IFI16 NFIL3 PTAFR SYAP1 STX8 AKAP13 PRPF8 TAF7 PDE4B IRF7 STRN3 MYO5A ST3GAL6 PSEN1 ZFAND6 EZR RPS6KA5 RIOK3 EMD GAB1 ZNF451 KLF7 PTPN12 CDKN1C LILRB2 LAMTOR5 FLOT1 TRIM41 EPG5 H2AFZ KMT2D STAT2 BCL9L CD47 LILRA2 ACAP2 GNAQ PIAS1 SHOC2 GAB2 PSMA4 ATP6V1H UBE2D1 ACADVL ADD1 DNAJB11 PSMD8 HNRNPM ATP6V1D PSME2 ARFGAP1 HCK XIAP POLR2C TYK2 PSMD3 UBE2D3 OAS3 TRIM38 ATP6V1A SUMO1 TNFSF14 RBCK1 PSME3 TLN1 MBTPS1 ARF1 HNRNPDL MX1 TNFRSF14 ATP6V0D1 INPPL1 UBB TALDO1 SRPRA MX2 HLA-DRA DCTN1 PSMB9 PSMB3
5.39E-13	166	Cellular response to chemical stimulus	TSC2 ZNF106 IL10RA SERP1 MED1 ANXA1 ABHD2 NLRC5 IFITM3 IFNAR1 MAPKAPK2 TGFBR2 RAB8A UPF1 ATP2B4 CYLD REST ADD1 ICAM1 SNW1 PPM1A CARD8 SUMO1 PADI2 FFAR2 SNX6 DSTYK MDM2 TLR2 PAK1 SPATA2 PINK1 S100A12 CTNNB1 UBE2O CIB1 DAXX ATF6B STRAP HERPUD1 DGAT2 DNM2 ATRX SNX10 RFX2 GRAMD1A MCOLN1 RANGAP1 YY1 TRIB3 EHD4 UBE2D3 KLF3 ARPC3 PDCD10 GNAI2 STXBP3 RAD23B DNMT3A IFIT2 KDM3B PTK2B ITPR2 TNFSF14 LDLR KDM6B KRAS TOR1A SLC31A1 ACTR2 TOB1 ARHGDIA MTF2 ZEB1 EIF4EBP2 VPS26B CDK19 KAT6B EIF4A2 ZBTB7B RBM15 MNDA IFI16 NFIL3 PTAFR SYAP1 STX8 AKAP13 PRPF8 ZFAND2A TAF7 CAMK1D PDE4B IRF7 STRN3 MYO5A ST3GAL6 PSEN1 BNIP3L RTF2 DAPK2 ZFAND6 EZR RPS6KA5 RIOK3 EMD GAB1 ZNF451 ATP6V1A KMT2A KLF7 CHD6 PTPN12 CDKN1C LILRB2 LAMTOR5 FLOT1 TRIM41 EPG5 ATP6V0D1 H2AFZ KMT2D STAT2 BCL9L CD47 LILRA2 ACAP2 GNAQ PIAS1 SHOC2 GAB2 PSMA4 ATP6V1H UBE2D1 ACADVL DNAJB11 PSMD8 HNRNPM ATP6V1D PSME2 ARFGAP1 HCK XIAP POLR2C TYK2 PSMD3 OAS3 TRIM38 RBCK1 PSME3 EGLN1 TLN1 MBTPS1 ARF1 HNRNPDL MX1 TNFRSF14 INPPL1 UBB TALDO1 SRPRA MX2 HLA-DRA DCTN1 PSMB9 PSMB3
2.33E-12	192	Cellular protein modification process	SPAG9 UBE3C RB1CC1 RNF19A USP36 ST3GAL6 SLK UBE2D1 ITCH CYLD USP48 RPS6KA5 STK4 TRIB3 HCK TNKS2 SHOC2 CCNY UNC119 UBE2D3 MAPKAPK3 MOGS USP34 AAK1 SUMO1 PADI2 KDM3B SERP1 PTK2B RBCK1 HAT1 MGAT1 NAA35 KAT7 DUSP6 ESCO1 PIK3R5 MTF2 PAK1 TRIP12 MPPE1 CDK19 UBE2L6 KAT6B TNFRSF14 PINK1 MAPKAPK2 RNF181 UBB QARS UBE2F PPP1R2 PIGN MIER1 PJA2 CTNNB1 CDC27 BAZ1B RUNX3 PIAS1 DAPK2 ATP2B4 REST B4GALT1 SRPK1 PPM1A RNF125 XIAP SLC9A3R1 FBXO9 ZNF451 PDCD10 KMT2A FBXL5 MED1 TGM3 PTPN12 ANXA1 MDM2 EGLN1 TLR2 PPIG WARS MARCH6 TLK2 PTPRJ SMG1 DYRK1A SPATA2 FBXW5 VPS28 UBXN1 TGFBR2 ATXN7 ANAPC16 TET2 ALK UBE2O TAF7 PRKAG1 ZFP91 PHF2 GRK5 CHML DAXX ST3GAL1 HERPUD1 PSEN1 ICAM1 PPP2R3C RIOK3 MGRN1 CTCF EHD4 TYK2 KMT5B FBXO3 TRIM38 GNAI2 CNPPD1 SET CDKN1C KDM6B DSTYK KRAS FKBP11 AGAP2 ARRDC4 TRIM41 TRIM11 ZBTB7B SPSB3 MTMR14 PIGX RNF169 FBXO22 KMT2D SETD5 PXK ZEB2 SYAP1 AKAP13 PPP1CA RPS6KA3 GAK CAMK1D CIB1 STK40 GRK6 MAN1A2 PHACTR4 LCP2 NUB1 STRAP ATRX EZR SNW1 H2AFY EIF4G1 TOR1A FLOT1 JTB GNAQ CTBP1 CAMLG CEP295 STAT2 TSC2 LILRB2 EMC10 SNX6 PSMA4 SEH1L NUP50 PSMD8 APOL1 RANGAP1 YY1 PSME2 FBXL20 PSMD3 RAD23B PSME3 DBNL MBTPS1 S100A12 MBD6 RAB8A SERPINA1 PSMB9 PSMB3
2.88E-12	68	Regulation of cellular catabolic process	USP36 ABHD5 RNF19A SMG6 HNRNPM PSME2 TRIB3 PSME3 UPF1 HERPUD1 ATP2B4 IL10RA SUMO1 RAD23B PTK2B MDM2 TOB1 UBXN1 MAPKAPK2 UBB DCP2 GIGYF2 PIAS1 PSEN1 BNIP3L FBXL5 LDLR PINK1 FBXO22 PPP1CA ZFAND2A IFI16 EZR TSC2 NUB1 ITCH EIF4G2 EIF4G1 TENT4B VPS13C TLK2 CAMLG RAB8A PAFAH1B2 SUPT5H ATP6V1H ATP6V1D ATP6V1A SNX6 SERBP1 ATP6V0D1 RB1CC1 DAPK2 PSMA4 DIS3 SEH1L NUP50 PSMD8 PSMD3 EXOC8 SET LAMTOR5 VPS26B PFKFB3 PRKAG1 PSMB9 PSMB3
3.53E-12	198	Macromolecule modification	SPAG9 UBE3C RB1CC1 RNF19A USP36 ST3GAL6 SLK BUD23 UBE2D1 ITCH CYLD USP48 RPS6KA5 STK4 TRIB3 HCK TNKS2 SHOC2 CCNY UNC119 UBE2D3 MAPKAPK3 MOGS USP34 AAK1 SUMO1 PADI2 KDM3B SERP1 PTK2B RBCK1 HAT1 MGAT1 NAA35 KAT7 DUSP6 ESCO1 PIK3R5 MTF2 PAK1 TRIP12 MPPE1 CDK19 UBE2L6 KAT6B TNFRSF14 PINK1 MAPKAPK2 TET2 RNF181 UBB QARS UBE2F PPP1R2 PIGN MIER1 PJA2 CTNNB1 CDC27 BAZ1B RUNX3 PIAS1 DAPK2 ATP2B4 REST B4GALT1 SRPK1 PPM1A RNF125 XIAP SLC9A3R1 FBXO9 ZNF451 PDCD10 KMT2A FBXL5 DNMT3A MED1 TGM3 PTPN12 ANXA1 MDM2 EGLN1 TLR2 PPIG WARS MARCH6 TLK2 PTPRJ SMG1 DYRK1A SPATA2 FBXW5 VPS28 UBXN1 RBM15 TGFBR2 ATXN7 ANAPC16 ALK UBE2O TAF7 PRKAG1 ZFP91 PHF2 GRK5 CHML DAXX ST3GAL1 HERPUD1 PSEN1 ICAM1 PPP2R3C RIOK3 MGRN1 CTCF EHD4 TYK2 KMT5B FBXO3 TRIM38 GNAI2 CNPPD1 SET CDKN1C KDM6B DSTYK KRAS FKBP11 AGAP2 ARRDC4 TRIM41 TRIM11 ZBTB7B SPSB3 MTMR14 PIGX RNF169 FBXO22 KMT2D SETD5 PXK ZEB2 SYAP1 AKAP13 PPP1CA RPS6KA3 GAK CAMK1D CIB1 STK40 GRK6 MAN1A2 PHACTR4 LCP2 NUB1 STRAP ATRX EZR SNW1 H2AFY EIF4G1 TOR1A FLOT1 JTB GNAQ CTBP1 CAMLG CEP295 STAT2 TSC2 LILRB2 EMC10 SNX6 TSR3 PSMA4 SEH1L NUP50 PSMD8 APOL1 RANGAP1 YY1 PSME2 FBXL20 PSMD3 RAD23B MBD4 PSME3 DBNL MBTPS1 S100A12 MBD6 RAB8A GATAD2A SERPINA1 PSMB9 PSMB3
4.53E-12	180	Organelle organization	SMARCD2 USP36 UBE2D3 PINK1 RB1CC1 ATG2B SMG6 NDE1 PCM1 DNM2 PDS5B ADD1 PPP2R3C SNAP23 ATP6V1D STAG2 BNIP3L TNKS2 CEP164 ARPC3 H2AFY GORASP2 WIPF1 PADI2 KDM3B SEPT7 HIP1 HAT1 CAP1 ANXA1 WASHC4 KAT7 MAP1LC3B ESCO1 MTF2 ARFGAP2 VPS51 ASAP1 TRIP12 KAT6B ATAD2 MX1 ARPC5 CDC42EP3 H2AFZ BRWD3 CEP295 RAB8A ATXN2L LMAN2 TAPT1 STX8 MX2 TUBB4B MCMBP MIER1 CAPZA2 ATG9A GET4 LILRB2 CTNNB1 UPF1 BAZ1B REST ATRX EZR RPS6KA5 HCK CTCF MTPN SLC25A36 KMT2A RAD23B KATNBL1 ACTR2 SERBP1 TLK2 PAK1 ATXN7 TAF7 CIB1 BLOC1S2 DAXX DCTN1 FARP2 BUD23 SMC1A PSEN1 CYLD ZFAND6 RFX2 ICAM1 YY1 SLC9A3R1 SLAIN2 KMT5B SUMO1 SET DNMT3A PTK2B CHD6 CDKN1C KDM6B KRAS CEP350 TMEM127 DBNL TOR1A TLN1 ARF1 DYRK1A FBXW5 ZBTB7B KMT2D SETD5 TET2 ZEB2 BRD3 UBB AKAP13 AGFG1 IKZF1 PHF2 MSRB1 PHACTR4 ODF2 CDC27 HEBP2 SLK AP1M1 SEH1L SNX10 SNW1 GAB1 ZNF451 SUB1 PDCD10 EIF4G1 EXOC8 TENT4B ZBTB1 VPS13C CHMP2A CLIP1 MBTPS1 ARHGDIA STRIP1 POGZ SH3KBP1 TCF7L2 SMG1 CTBP1 ARHGAP12 INPPL1 PDZD8 STAT2 DCP2 GAK ARAP1 PRPF40A FGD4 ATP6V0D1 CD47 TSC2 RNF19A ATP6V1H UBE2D1 SRPK1 CTSZ EMD ATP6V1A IFIT2 NFE2 ATP5F1E ALDOA VPS28 TRAPPC1 TRAPPC5 SERPINA1 DYNLL2
8.18E-12	72	Regulation of catabolic process	USP36 ABHD5 RNF19A SMG6 ITCH HNRNPM PSME2 TRIB3 PSME3 UPF1 HERPUD1 ATP2B4 IL10RA SUMO1 RAD23B PTK2B AGAP2 MDM2 TOB1 UBXN1 MAPKAPK2 UBB DCP2 GIGYF2 PIAS1 PSEN1 BNIP3L FBXL20 PSMD3 FBXL5 LDLR EGLN1 PINK1 FBXO22 PPP1CA ZFAND2A IFI16 EZR TSC2 NUB1 EIF4G2 EIF4G1 TENT4B VPS13C TLK2 VPS28 CAMLG RAB8A PAFAH1B2 SUPT5H ATP6V1H ATP6V1D ATP6V1A SNX6 SERBP1 ATP6V0D1 RB1CC1 DAPK2 PSMA4 DIS3 SEH1L NUP50 PSMD8 EXOC8 SET LAMTOR5 VPS26B PFKFB3 PRKAG1 PSMB9 PSMB3

2.82E-11	125	Cellular catabolic process	USP36 UPF1 UBE3C ABHD5 RB1CC1 RNF19A PSMA4 ATG2B SMG6 UBE2D1 ITCH DIS3 PSMD8 HNRNPM FOXRED2 PSME2 CTSZ TRIB3 GDPD3 FBXL20 PSMD3 UBE2D3 FBXO9 FBXL5 RAD23B TENT4B RBCK1 PSME3 ABHD2 MAP1LC3B MARCH6 EPG5 TRIP12 UBE2L6 SMG1 PINK1 VPS28 SPSB3 UBAP1 TET2 UBB DCP2 SGTB ATG9A PSMB9 RBM8A PSMB3 HERPUD1 ATP2B4 ACADVL XRN2 RNF125 IL10RA SUMO1 PTK2B MDM2 TOB1 FBXW5 UBXN1 MAPKAPK2 CTNNB1 GIGYF2 ABHD16A RNPS1 PIAS1 PSEN1 CYLD MCOLN1 USP48 BNIP3L PDE4C TRIM38 USP34 NAGK LDLR MGAT1 TOR1A ALDOA IFI16 FBXO22 RAB8A RAB24 PPP1CA ZFAND2A PDE4B DDAH2 PLEKHM1 EZR TSC2 NUB1 EIF4G2 EIF4G1 VPS13C TLK2 VPS51 CAMLG PAFAH1B2 SUPT5H GET4 ATP6V1H ATP6V1D ATP6V1A SNX6 SERBP1 ATP6V0D1 CDC27 DAPK2 SEH1L NUP50 RPL19 EXOC8 SET MBD4 CHMP2A LAMTOR5 VPS26B MTMR14 PGM2L1 ANAPC16 PFKFB3 PRKAG1 RPL41 DYNLL2 RPL17
7.44E-11	114	Vesicle-mediated transport	SPAG9 DNM2 SNAP23 ARFGAP1 UNC119 WIPF1 EXOC8 HIP1 CHMP2A ARF3 ANXA1 WASHC4 ARF1 ARFGAP2 VPS51 VPS26B RER1 VPS28 RAB8A LMAN2 TRAPPC1 STX8 GAK TRAPPC5 BLOC1S2 CD47 AP5B1 CUX1 ATP6V1H IL10RA AAK1 GDI2 AP1M1 PSEN1 SNX10 LAT2 EZR UNC13D HCK EHD4 FBXL20 GNAI2 MAPKAPK3 STXBP3 LDLR CAP1 NUMB DBNL TOR1A FLOT1 SH3KBP1 PAK1 EPG5 MPPE1 MAPKAPK2 TGFBR2 CTNNB1 PTAFR MYO5A SYNRG MGRN1 SNX6 PINK1 ARHGAP12 INPPL1 UBE2O CAMK1D LAMP1 DCTN1 GAB2 REST TSC2 ACAP2 TLR2 PTPRJ DNAJC5 FCGRT CD93 SERPINB1 HEBP2 B4GALT1 APOL1 ATP6V1D CTSZ PSMD3 CYTH1 ARPC3 SCAMP3 PADI2 CTSD NPC2 FGL2 VPS13C LILRB2 TMBIM1 TLN1 ACTR2 SLC15A4 SERPING1 ALDOA ARPC5 S100A12 MNDA UBAP1 CLEC4D PAFAH1B2 RAB24 UBB DGAT1 TUBB4B SELL SERPINA1 CAPZA2 DYNLL2
1.26E-10	139	Negative regulation of macromolecule metabolic process	UPF1 SERPINB1 SMG6 HNRNPM XIAP TENT4B KAT7 DUSP6 NFIC EIF4EBP2 SERPING1 TRIP12 KAT6B ATAD2 SMG1 GATAD2A QARS DCP2 SERPINA1 MIER1 CUX1 RBM8A DNMT3A CYLD REST SNW1 RPS6KA5 YY1 TRIB3 NKAP CTCF CARD8 SLC9A3R1 EIF4G1 SUMO1 SET ZBTB1 SNX6 EAPP CDKN1C LAMTOR5 AGAP2 MDM2 WARS TOB1 ZEB1 TCF7L2 PTPRJ EIF4A2 DIP2A VPS28 UBXN1 MAPKAPK2 IFI16 ZNF148 RNF169 CTNNB1 ZEB2 CNBP RPS6KA3 TAF7 SUPT5H GIGYF2 DAXX RNPS1 STRAP PIAS1 HERPUD1 ITCH PSEN1 ATRX KLF3 H2AFY GNAI2 GON4L KHDRBS1 TTF1 MED1 TNFSF14 HAT1 LDLR MTF2 TNFAIP8 TRIM11 DYRK1A PINK1 CTBP1 ZBTB7B NFIL3 INPPL1 PPP1R2 CIB1 STRN3 PHACTR4 SMC1A EZR CTSZ STAG2 TNKS2 ZNF451 PDCD10 KMT2A KLF7 TLK2 GNAQ CAMLG PHF2 RUNX3 SFSWAP TSC2 LILRB2 KMT2D IKZF1 NACA ATP2B4 SERBP1 PSMA4 UBE2D1 DIS3 SEH1L NUP50 PSMD8 PPM1A PSME2 POLR2C RPL19 PSMD3 UBE2D3 TSNAX PSME3 UBB IRF7 RPL41 HSBP1 PSMB9 RPL17 PSMB3
1.26E-10	104	Intracellular transport	USP36 UBE2D3 PINK1 NDE1 PCM1 DNM2 SNAP23 NUP50 AFTPH VPS13C CHMP2A ARF3 COX5B WASHC4 ARF1 ARFGAP2 VPS51 VPS26B VPS28 RAB8A LMAN2 RAB24 TRAPPC1 STX8 TRAPPC5 SRPRA BLOC1S2 SGTB AP5B1 ATP5F1E SPAG9 RANGAP1 NPC2 KHDRBS1 UBB AP1M1 PSEN1 ZFAND6 LAT2 PPM1A EHD4 BNIP3L FBXL20 MED1 SNX6 CLIP1 MDM2 TOR1A ACTR2 MBTPS1 PAK1 EPG5 MPPE1 ABCG1 FBXO22 CTNNB1 AKAP13 PRKAG1 MYO5A CHML DCTN1 EZR HERPUD1 SNX13 UNC13D EMD MGRN1 SLC9A3R1 OSBP PDCD10 LDLR TCF7L2 UBE2O GAK MX2 CIB1 LAMP1 GAB2 TSC2 ACAP2 UPF1 SMG6 UBE2D1 SEH1L DNAJC5 CTSZ ARFGAP1 RPL19 HIP1 SMG1 U2AF1 SSR2 PSIP1 UBAP1 CDC40 AGFG1 SERPINA1 CAPZA2 RNPS1 RPL41 DYNLL2 RBM8A RPL17 SYNRG
1.04E-09	144	Negative regulation of metabolic process	UPF1 SERPINB1 SMG6 HNRNPM XIAP TENT4B KAT7 DUSP6 NFIC EIF4EBP2 SERPING1 TRIP12 KAT6B ATAD2 SMG1 GATAD2A QARS DCP2 SERPINA1 MIER1 CUX1 RBM8A DNMT3A ATP2B4 CYLD REST SNW1 RPS6KA5 YY1 TRIB3 NKAP CTCF CARD8 SLC9A3R1 IL10RA EIF4G1 SUMO1 SET ZBTB1 SNX6 EAPP CDKN1C LAMTOR5 AGAP2 MDM2 WARS TOB1 ZEB1 TCF7L2 PTPRJ EIF4A2 DIP2A VPS28 UBXN1 MAPKAPK2 IFI16 ZNF148 RNF169 CTNNB1 ZEB2 CNBP RPS6KA3 TAF7 SUPT5H GIGYF2 DAXX RNPS1 STRAP PIAS1 HERPUD1 DGAT2 ACADVL ITCH PSEN1 ATRX KLF3 H2AFY GNAI2 GON4L KHDRBS1 TTF1 MED1 TNFSF14 HAT1 LDLR MTF2 TNFAIP8 TRIM11 DYRK1A PINK1 CTBP1 ZBTB7B NFIL3 INPPL1 PPP1R2 CIB1 STRN3 PHACTR4 USP36 SMC1A EZR CTSZ STAG2 TNKS2 EIF4G2 ZNF451 PDCD10 KMT2A KLF7 TLK2 GNAQ CAMLG PHF2 RUNX3 SFSWAP TSC2 LILRB2 KMT2D IKZF1 NACA SERBP1 PSMA4 UBE2D1 DIS3 SEH1L NUP50 PSMD8 PPM1A PSME2 POLR2C RPL19 PSMD3 UBE2D3 TSNAX PSME3 UBB IRF7 RPL41 HSBP1 PSMB9 RPL17 PSMB3
1.04E-09	154	Immune system process	RUNX3 PPP2R3C BNIP3L FCGRT BTN3A3 MAPKAPK3 IFIT2 ANKHD1 ANXA1 TLR2 NLRC5 IFITM3 MX1 TNFRSF14 MAPKAPK2 TET2 MX2 IRF7 SRPK1 LILRB2 LCP2 CYLD LAT2 APOL1 RNF125 OAS3 PADI2 TNFSF14 RBCK1 TAPBPL IFNAR1 SERPING1 PTPRJ RBM15 S100A12 TGFBR2 IFI16 CD47 HLA-DRA LILRA2 FARP2 DAPK2 ROGDI APBB1IP ITCH PSEN1 REST B4GALT1 SNX10 ADD1 ICAM1 MCOLN1 UNC13D STK4 HCK RIOK3 NKAP ZMIZ1 FBXO9 TRIM38 STXBP3 GON4L PTK2B MED1 FFAR2 ZBTB1 CDKN1C LDLR KDM6B DBNL ACTR2 SECTM1 TMEM91 ZEB1 VPS26B TRIM11 ZBTB7B MNDA NFIL3 INPPL1 CLEC4D CTNNB1 PTAFR STX8 STAT2 RPS6KA3 CYBC1 CAMK1D PDE4B CIB1 SELL MSRB1 PSMB9 NUB1 GAB2 SNAP23 EZR FLOT1 EPG5 IKZF1 LAMP1 PJA2 CD93 TUBB4B SERPINB1 PIAS1 PSMA4 HEBP2 GDI2 UBE2D1 AP1M1 DNM2 PSMD8 ATP6V1D RPS6KA5 PSME2 DNAJC5 CTSZ XIAP TYK2 PSMD3 UBE2D3 ELF2 ARPC3 HRH2 WIPF1 SUMO1 CTSD KMT2A NPC2 NFE2 S1PR4 FGL2 CAP1 PSME3 KRAS TMBIM1 SLC15A4 IGSF6 ARF1 PAK1 ALDOA ARPC5 KMT2D PAFAH1B2 RAB24 TRAPPC1 UBB DGAT1 SERPINA1 CAPZA2 DCTN1 DYNLL2 PSMB3
2.81E-07	120	Organelle organization	SMARCA4 SH3GLB1 RAC2 UBE2J2 BRCA1 KDM5A XRCC5 SUPT16H RAB18 MAST3 FBXO7 DOT1L NEDD9 PHF1 CCND3 PHF13 CAPZA1 KDM5B NRDE2 RPL5 SIN3B MACF1 TUBGCP2 HAUS8 NASP DYNC1LI2 GCC2 PARVG WHAMM UQCRB PADI4 ALKBH4 MAP1A TPM4 NCKAP5L CHAF1A ATG4B TADA3 SAMD9L PLEC RFLNB SPG7 FCHSD1 RPL12 ATXN2 XRN1 FMR1 KAT8 HPS1 UBE2S ASH1L TNFSF10 CHMP1A PPHLN1 SMARCA5 XPC USP16 H3F3A DTX3L ABCA1 CRK PRELID1 ASXL1 CTDNEP1 MTHFR SETD2 C6orf89 NAP1L4 PUM2 HMG20B NCOA1 AKAP8 CDC37 PHACTR1 RO60 CREB1 PANK2 MKLN1 PPFIA1 KMT5C DOCK2 AGTPBP1 ARRB1 GINS4 MYSM1 MID1IP1 NSD1 ING2 KIF5B PAX5 SIPA1L1 SMURF1 NELFE RAD1 CHMP4A AKAP8L PNKP ARFGEF1 MAP2K7 TFEB IL1B IFI6 TMEM165 CLTC SCLT1 RPGR SYNJ1 TAOK1 ZNF274 SEPT9 PPP1R10 CORO7 FGD3 BIN3 GABPB1 ATP6V1E1 MT-ND6 F5 CSNK1E
7.41E-07	31	Covalent chromatin modification	BRCA1 PHF1 SIN3B PADI4 TADA3 KDM5A FMR1 KAT8 DOT1L ASH1L KDM5B PPHLN1 USP16 DTX3L ASXL1 MTHFR SETD2 C6orf89 NCOA1 KMT5C ARRB1 MYSM1 NSD1 ING2 PAX5 NELFE AKAP8L AKAP8 ZNF274 IL1B
7.41E-07	30	Histone modification	BRCA1 PHF1 SIN3B PADI4 TADA3 KDM5A FMR1 KAT8 DOT1L ASH1L KDM5B USP16 DTX3L ASXL1 MTHFR SETD2 C6orf89 NCOA1 KMT5C ARRB1 MYSM1 NSD1 ING2 PAX5 NELFE AKAP8L AKAP8 ZNF274 IL1B
7.41E-07	104	Positive regulation of cellular metabolic process	DBF4 BRCA1 PSMC4 CYFIP2 TP53BP1 MAP2K7 NCOA1 SUPT16H FBXO7 FMR1 PGK1 KAT8 GABPB1 MED25 PHF1 IL1B CCNT1 IL1RN CASP1 IGF1R RBMX TAOK1 CDK12 MAP4K2 ING2 DHRSX ASXL1 CASP4 NCOA6 CSNK1E ZNF350 ID2 AKAP8L BIRC3 CYBA KDM5A ARAF NRDC XRCC5 RBM22 PRPF6 RPRD1B DOT1L UBE2S NFKB1 NEDD9 TFEB CCND3 CEBPZ MEF2D ZBTB17 CREB1 ELF1 TNFSF10 RPL5 SMARCA4 HELZ2 TNNI2 F12 PPHLN1 MED4 ARRB1 WDFY2 ETS2 NSD1 RBPJ PRELID1 TADA3 DDIT3 C6orf89 SH3GLB1 CRKL GTPBP1 CST3 VRK3 CDC37 NCL ASH1L PUM1 SMARCA5 USP16 MID1IP1 NFATC2IP ARID3B ZBTB20 PAX5 SMURF1 NELFE DTX3L CTDNEP1 PNKP ARFGEF1 AKAP8 VHL MYSM1 PARP14 PPP1R10 ABCA7 EDF1 PRKAR2A CXCR4 ADCY4 ADM CRK
7.51E-07	41	Chromatin organization	SMARCA4 BRCA1 KDM5A SUPT16H DOT1L PHF1 KDM5B NRDE2 SIN3B NASP PADI4 CHAF1A TADA3 FMR1 KAT8 ASH1L PPHLN1 SMARCA5 USP16 H3F3A DTX3L ASXL1 MTHFR SETD2 C6orf89 NAP1L4 HMG20B NCOA1 PHF13 KMT5C ARRB1 MYSM1 NSD1 ING2 PAX5 NELFE AKAP8L AKAP8 ZNF274 IL1B
1.03E-06	109	Positive regulation of metabolic process	DBF4 BRCA1 PSMC4 CYFIP2 TP53BP1 MAP2K7 NCOA1 SUPT16H FBXO7 FMR1 PGK1 KAT8 GABPB1 MED25 PHF1 IL1B CCNT1 IL1RN CASP1 IGF1R RBMX TAOK1 CDK12 MAP4K2 ING2 DHRSX ASXL1 CASP4 NCOA6 SMURF1 CSNK1E ZNF350 ID2 AKAP8L BIRC3 CYBA KDM5A ARAF NRDC XRCC5 RBM22 PRPF6 RPRD1B DOT1L UBE2S NFKB1 NEDD9 TFEB CCND3 CEBPZ MEF2D ZBTB17 CREB1 ELF1 TNFSF10 RPL5 SMARCA4 HELZ2 TNNI2 F12 PPHLN1 MED4 ARRB1 WDFY2 ETS2 NSD1 RBPJ PRELID1 TADA3 DDIT3 C6orf89 EGLN2 MYLIP SH3GLB1 CRKL GTPBP1 CST3 VRK3 CDC37 NCL ASH1L KDM5B PUM1 SMARCA5 USP16 MID1IP1 ATG4B SLC50A1 NFATC2IP ARID3B ZBTB20 PAX5 NELFE DTX3L CTDNEP1 PNKP ARFGEF1 AKAP8 VHL MYSM1 PARP14 PPP1R10 ABCA7 CXCR4 EDF1 PRKAR2A ADCY4 ADM CRK
1.03E-06	92	Negative regulation of macromolecule metabolic process	BRCA1 BIRC3 FMR1 KAT8 DOT1L NRDE2 IFI6 SIN3B HELZ2 RBPJ MAF1 CNOT10 ZNF350 NFKB1 XRN1 PUM2 TP53BP1 CST3 PRKAR2A ID2 KDM5B CREB1 RPL5 SRSF6 SMARCA4 CHMP1A VHL PPHLN1 PUM1 ARRB1 IGF1R CLTC HDGF SAP18 ETS2 DDIT3 CD55 AP2A1 CARD16 HMG20B KDM5A SRRT FBXO7 GTPBP1 PPP1R37 VRK3 EIF3A HIPK3 PHACTR1 CCND3 ASH1L PSPC1 SLPI IL1B NONO SMARCA5 ZNF274 ZBTB20 SIGIRR PAX5 NELFE PPP1R10 LRPAP1 PNKP XRCC5 IGBP1 CRKL MED25 PHF1 NCL GBP1 CCNT1 OS9 DTX3L PIP5KL1 USP47 PARP14 ATXN2 ABCA7 RBMX NSD1 MAP1A N4BP2L2 ZBED6 CLN8 PSMC4 PSMA7 H3F3A PSMD1 PSMD2 RPL12
1.03E-06	16	Regulation of histone modification	BRCA1 PHF1 FMR1 MTHFR C6orf89 KDM5A ING2 TADA3 PAX5 NELFE AKAP8L AKAP8 ARRB1 NSD1 ZNF274 IL1B
1.03E-06	86	Negative regulation of cellular metabolic process	BRCA1 BIRC3 FMR1 KAT8 DOT1L NRDE2 IFI6 SIN3B RBPJ MAF1 CNOT10 ZNF350 NFKB1 XRN1 TP53BP1 CST3 PPP1R37 PRKAR2A ID2 KDM5B CREB1 RPL5 SRSF6 MPHOSPH10 SMARCA4 CHMP1A VHL PPHLN1 ARRB1 IGF1R CLTC HDGF SAP18 ETS2 DDIT3 CD55 AP2A1 CARD16 HMG20B KDM5A FBXO7 HERC1 VRK3 EIF3A HIPK3 PHACTR1 CCND3 ASH1L PSPC1 SLPI IL1B PUM1 NONO SMARCA5 ZNF274 ZBTB20 SIGIRR PAX5 NELFE PPP1R10 LRPAP1 PNKP XRCC5 IGBP1 CRKL MED25 PHF1 NCL GBP1 CCNT1 OS9 DTX3L PIP5KL1 USP47 PARP14 ATXN2 CHMP4A ABCA7 RBMX NSD1 MAP1A N4BP2L2 ZBED6 CLN8 H3F3A
1.03E-06	87	Regulation of cellular protein metabolic process	BRCA1 BIRC3 CYFIP2 MAP2K7 FMR1 PHF1 CCND3 IL1B IFI6 IL1RN CASP1 IGF1R TAOK1 MAP4K2 CNOT10 CASP4 CSNK1E NFKB1 ARAF NRDC XRCC5 CST3 UBE2S NEDD9 PRKAR2A TNFSF10 RPL5 F12 ARRB1 WDFY2 PRELID1 MTHFR CD55 C6orf89 CARD16 DBF4 MYLIP PSMC4 PUM2 HMG20B KDM5A MKNK1 IGBP1 CRKL FBXO7 MTG2 PPP1R37 VRK3 CDC37 EIF3A HIPK3 PHACTR1 XRN1 ASH1L SLPI RAC2 CCNT1 PUM1 ING2 TADA3 PAX5 NELFE PPP1R10 AKAP8L PNKP ARFGEF1 PUM3 AKAP8 NCL GBP1 OS9 USP16 DTX3L NSD1 PIP5KL1 CDK12 USP47 ZNF274 PARP14 SMURF1 ABCA7 MAP1A CLN8 ATXN2 CXCR4 ADCY4 CRK
1.07E-06	92	Regulation of protein metabolic process	BRCA1 BIRC3 CYFIP2 MAP2K7 FMR1 PHF1 CCND3 IL1B IFI6 IL1RN CASP1 IGF1R TAOK1 MAP4K2 CNOT10 CASP4 SMURF1 CSNK1E NFKB1 ARAF NRDC XRCC5 CST3 UBE2S NEDD9 PRKAR2A TNFSF10 RPL5 F12 ARRB1 WDFY2 PRELID1 MTHFR CD55 C6orf89 CARD16 EGLN2 DBF4 MYLIP PSMC4 PUM2 HMG20B KDM5A MKNK1 IGBP1 CRKL FBXO7 MTG2 PPP1R37 VRK3 CDC37 EIF3A HIPK3 PHACTR1 XRN1 ASH1L SLPI RAC2 CCNT1 PUM1 ATG4B ING2 TADA3 PSMD1 PSMD2 SIGIRR PAX5 NELFE PPP1R10 AKAP8L PNKP ABCA7 ARFGEF1 PUM3 AKAP8 NCL GBP1 OS9 USP16 DTX3L NSD1 PIP5KL1 CDK12 USP47 ZNF274 PARP14 MAP1A CLN8 ATXN2 CXCR4 ADCY4 CRK

1.31E-06	99	Positive regulation of nitrogen compound metabolic process	DBF4 BRCA1 PSMC4 CYFIP2 TP53BP1 MAP2K7 NCOA1 SUPT16H FMR1 PGK1 KAT8 GABPB1 MED25 PHF1 IL1B CCNT1 IL1RN CASP1 IGF1R RBMX TAOK1 CDK12 MAP4K2 ING2 ASXL1 CASP4 NCOA6 SMURF1 CSNK1E ZNF350 ID2 AKAP8L BIRC3 KDM5A ARAF NRDC XRCC5 RBM22 PRPF6 RPRD1B DOT1L UBE2S NFKB1 NEDD9 TFEB CCND3 CEBPZ MEF2D ZBTB17 CREB1 ELF1 TNFSF10 RPL5 SMARCA4 HELZ2 TNNI2 F12 PPHLN1 MED4 ARRB1 WDFY2 ETS2 NSD1 RBPJ PRELID1 TADA3 DDIT3 C6orf89 EGLN2 MYLIP CRKL GTPBP1 CST3 VRK3 NCL ASH1L PUM1 SMARCA5 USP16 ATG4B NFATC2IP ARID3B ZBTB20 PAX5 NELFE DTX3L PNKP ARFGEF1 AKAP8 VHL MYSM1 PARP14 PPP1R10 ABCA7 EDF1 PRKAR2A CXCR4 ADCY4 CRK
1.51E-06	96	Negative regulation of metabolic process	BRCA1 BIRC3 FMR1 KAT8 DOT1L NRDE2 IFI6 SIN3B HELZ2 RBPJ MAF1 CNOT10 ZNF350 NFKB1 XRN1 PUM2 TP53BP1 CST3 PPP1R37 PRKAR2A ID2 KDM5B CREB1 RPL5 SRSF6 MPHOSPH10 IL1B SMARCA4 CHMP1A VHL PPHLN1 PUM1 ARRB1 IGF1R CLTC HDGF SAP18 ETS2 DDIT3 CD55 AP2A1 CARD16 HMG20B KDM5A SRRT FBXO7 GTPBP1 HERC1 VRK3 EIF3A HIPK3 PHACTR1 CCND3 ASH1L PSPC1 SLPI NONO SMARCA5 ZNF274 ZBTB20 SIGIRR PAX5 NELFE PPP1R10 LRPAP1 PNKP XRCC5 IGBP1 CRKL MED25 PHF1 NCL GBP1 CCNT1 OS9 DTX3L PIP5KL1 USP47 PARP14 ATXN2 CHMP4A ABCA7 NRDC RBMX NSD1 MAP1A N4BP2L2 ZBED6 CLN8 PSMC4 PSMA7 H3F3A PSMD1 PSMD2 RPL12
2.33E-06	81	Negative regulation of nitrogen compound metabolic process	BRCA1 BIRC3 FMR1 KAT8 DOT1L NRDE2 IFI6 SIN3B RBPJ MAF1 CNOT10 ZNF350 NFKB1 XRN1 TP53BP1 CST3 PRKAR2A ID2 KDM5B CREB1 RPL5 SRSF6 SMARCA4 CHMP1A VHL PPHLN1 ARRB1 IGF1R CLTC HDGF SAP18 ETS2 DDIT3 CD55 AP2A1 CARD16 HMG20B KDM5A FBXO7 PPP1R37 VRK3 EIF3A HIPK3 PHACTR1 CCND3 ASH1L PSPC1 SLPI IL1B PUM1 NONO SMARCA5 ZNF274 ZBTB20 SIGIRR PAX5 NELFE PPP1R10 PNKP XRCC5 IGBP1 CRKL MED25 PHF1 NCL GBP1 CCNT1 OS9 DTX3L PIP5KL1 USP47 PARP14 ABCA7 RBMX NSD1 MAP1A N4BP2L2 ZBED6 CLN8 H3F3A
3.47E-06	17	Regulation of chromatin organization	BRCA1 PHF1 FMR1 PPHLN1 MTHFR C6orf89 KDM5A ING2 TADA3 PAX5 NELFE AKAP8L AKAP8 ARRB1 NSD1 ZNF274 IL1B

Table S5. Full list of enriched pathways in FACS-sorted vs COSMOS-sorted neutrophils.

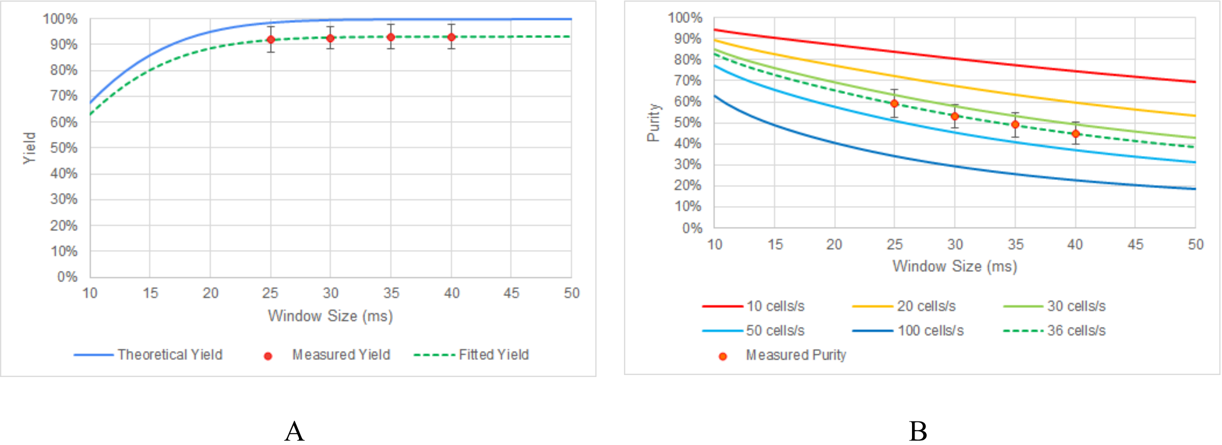


Figure S1. Performance of 0.5% random sorting of PBMC samples using different window sizes (25, 30, 35 and 40 milliseconds). A total of 341 experiments were run across four window sizes in 21 microfluidic devices (three chips each from seven photoresist mold sets) and on two hardware systems. (**A**) Yield: The theoretical curve assumes a normal distribution of cell arrival time with a standard deviation of 5 ms; fitted curve adds a limit of detection level at 93%. (**B**) Purity: Solid and dotted lines are theoretical values at various cell throughput; ±3 ms exclusion zone is assumed around each cell to match measured values with the theoretical values. The error bars in both graphs represent one standard deviation (2σ total) of the raw experimental data in each window size.

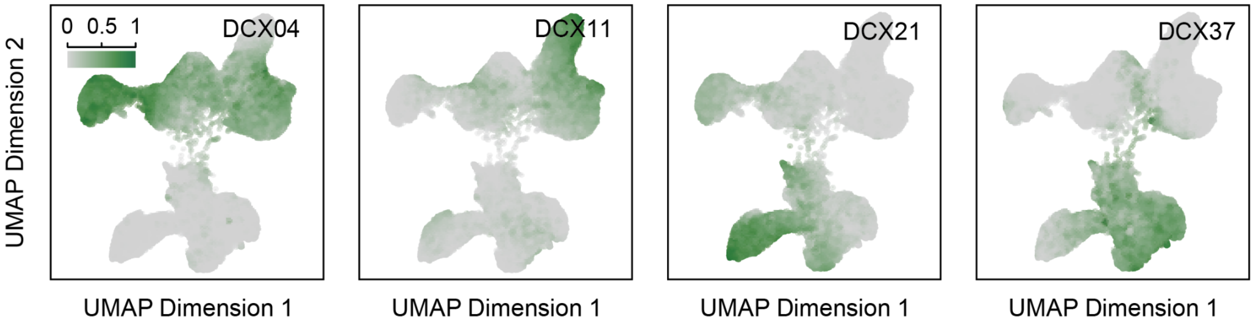


Figure S2. The UMAP projection in (**fig. 2A**), colored by the value of each coordinate in the embedding space of the model, demonstrating the contribution of that coordinate in identifying cells that are highlighted. For example, the leftmost plot demonstrates the value of coordinate number 4 (DCX04) and it shows that this coordinate "encodes" for the NSCLC and HCC (malignant) cells, whereas coordinates DCX11, DCX21 and DCX37 correspond to the HCC, fnRBC and PBMC classes, respectively.

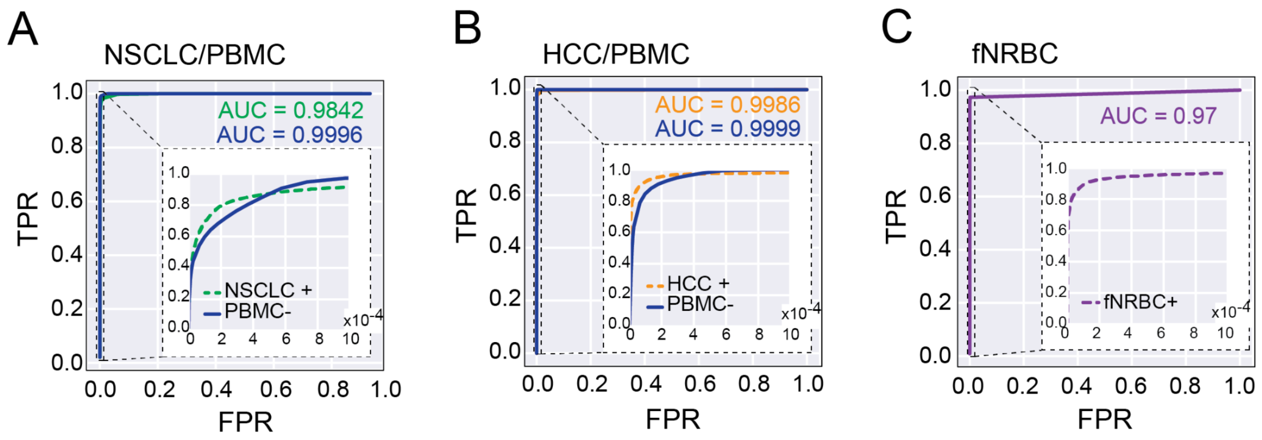


Figure S3. (**A and B**) Receiver operating characteristic (ROC) curves for the classification of (**A)** NSCLCs and (**B**) HCCs. Two ROC curves each are shown: one for the positive selection of each category, and one for negative selection, specifically for the selection of non-blood cells. Area Under Curves (AUCs) achieved for NSCLC are 0.9842 (positive selection) and 0.9996 (negative selection) and for HCC are 0.9986 (positive selection) and 0.9999 (negative selection). (**C**) ROC curves for the classification of fnRBCs and the AUC is 0.97 (positive selection). 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.

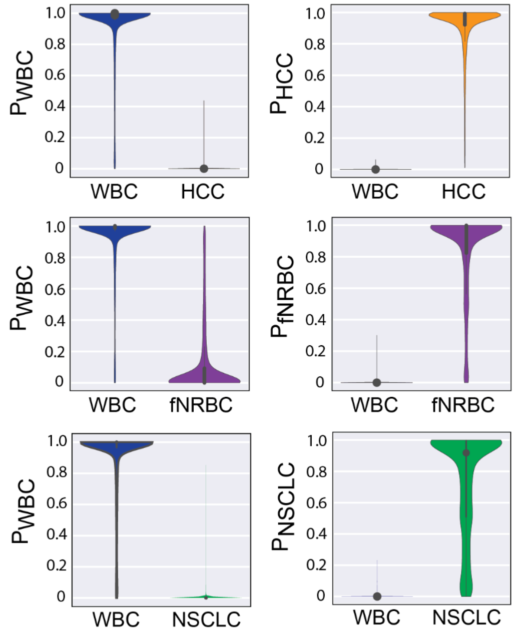


Figure S4. Violin plots showing the predicted probabilities of assigning cells in each category to its appropriate class. The plot on the left shows the probability distribution of PBMCs as well as NSCLCs being classified as PBMCs (P_{PBMC}) and the plot on the right shows the probability distribution of PBMCs as well as NSCLCs being classified as NSCLCs (P_{NSCLC}).

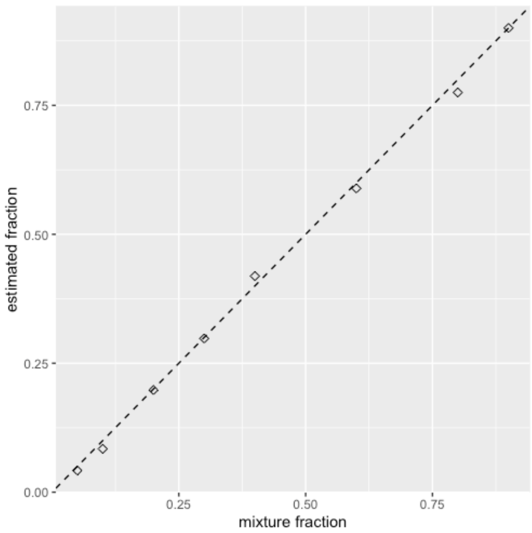


Figure S5. Accuracy of SNP-based mixture fraction estimates in control DNA mixtures. Each composite sample contained 250 pg of bulk DNA drawn from two individuals and the mixture proportion of DNA from the second individual was set at 5%, 10%, 20%, 30%, 40%, 60%, 80% and 90%. A close correspondence was found between the known and estimated mixture proportions.

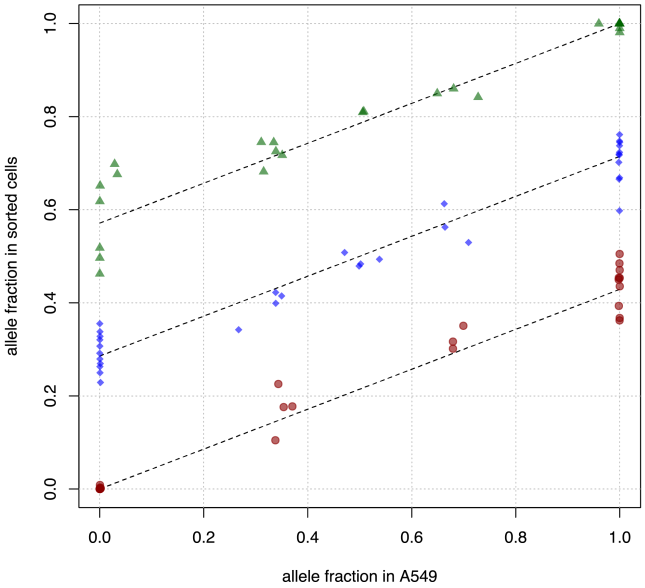
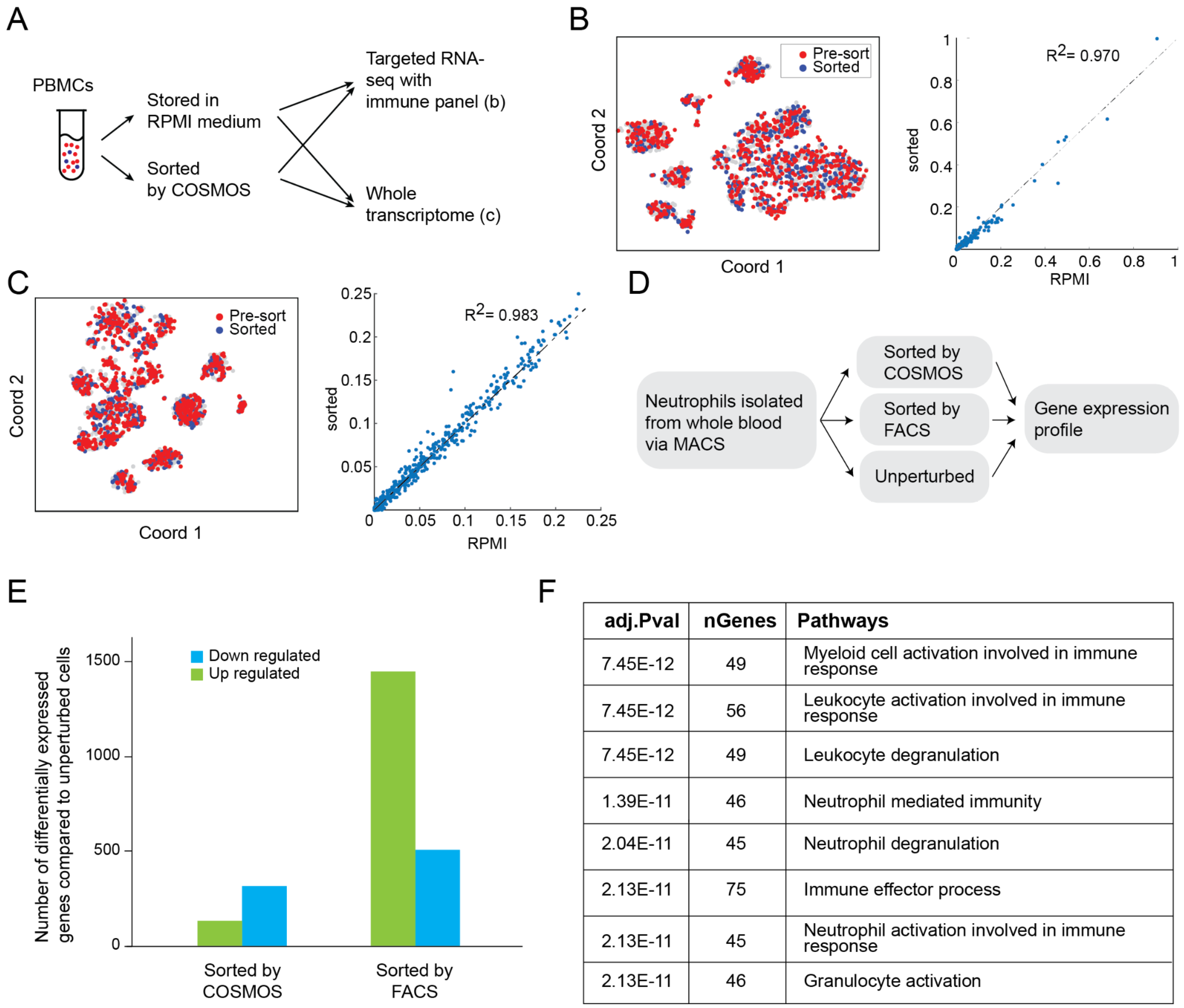


Figure S6. A549 purity in cells enriched using COSMOS from a 40 cells/mL spike-in into healthy donor whole blood. The purity and blood sample genotypes were estimated with an expectation-maximization (EM) algorithm. Green circles, blue diamonds and red triangles denote AA, AB and BB genotypes respectively in the blood sample used as a base for the spike-in mixture; dotted lines represent the expected allele fractions for the three blood genotypes at the inferred purity of 43% (95% confidence interval 0.40 - 0.45) which is also the slope of the lines.



adj.Pval	nGenes	Pathways
7.45E-12	49	Myeloid cell activation involved in immune response
7.45E-12	56	Leukocyte activation involved in immune response
7.45E-12	49	Leukocyte degranulation
1.39E-11	46	Neutrophil mediated immunity
2.04E-11	45	Neutrophil degranulation
2.13E-11	75	Immune effector process
2.13E-11	45	Neutrophil activation involved in immune response
2.13E-11	46	Granulocyte activation

Figure S7. Cell health and quality after COSMOS sorting. (**A**) Workflow schematics: PBMCs flowed-through COSMOS were compared with control cells stored in RPMI medium using either a single cell targeted immune panel (**b**) or single cell whole transcriptome (WTA) workflow (**c**) with the BD Rhapsody^{TM} system. (**B and C**) Left is a t-SNE plot of gene expression profiles of the pre-sort and sorted cells, each point is a cell. Right is a correlation plot of mean (log$_{10}$(molecules per cell per gene)) for the two conditions, each point is a gene. The two samples overlapped with each other in t-SNE plot and gene expression levels showed high correlations (R^2 equaled 0.97 and 0.98 respectively for targeted panel and WTA), indicating no significant gene expression change after sorting. (**D-G**) COSMOS sorting of unlabeled neutrophils yielded healthier cells compared to stained and FACS sorted cells. (**D**) Workflow schematics: human neutrophils were first isolated from whole blood by immunomagnetic negative selection then split into multiple aliquots for four conditions: unperturbed, stained and flow-sorted by FACS, unstained/unlabeled and sorted by COSMOS. Pre-sorted and sorted cells were lysed for bulk gene expression profiling by RNAseq. (**E and F**) Bulk RNAseq gene expression analysis of the cells in different groups. (**E**) Number of up- and down- regulated genes compared to unperturbed cells, confirmed that COSMOS-sorted cells had minimal gene expression differences compared to unperturbed cells, much fewer than FACS-sorted cells did. (**F**) Upregulated pathways in FACS-sorted cells compared to COSMOS-sorted cells, suggests that FACS induced upregulation of pathways in neutrophil activation and degranulation.

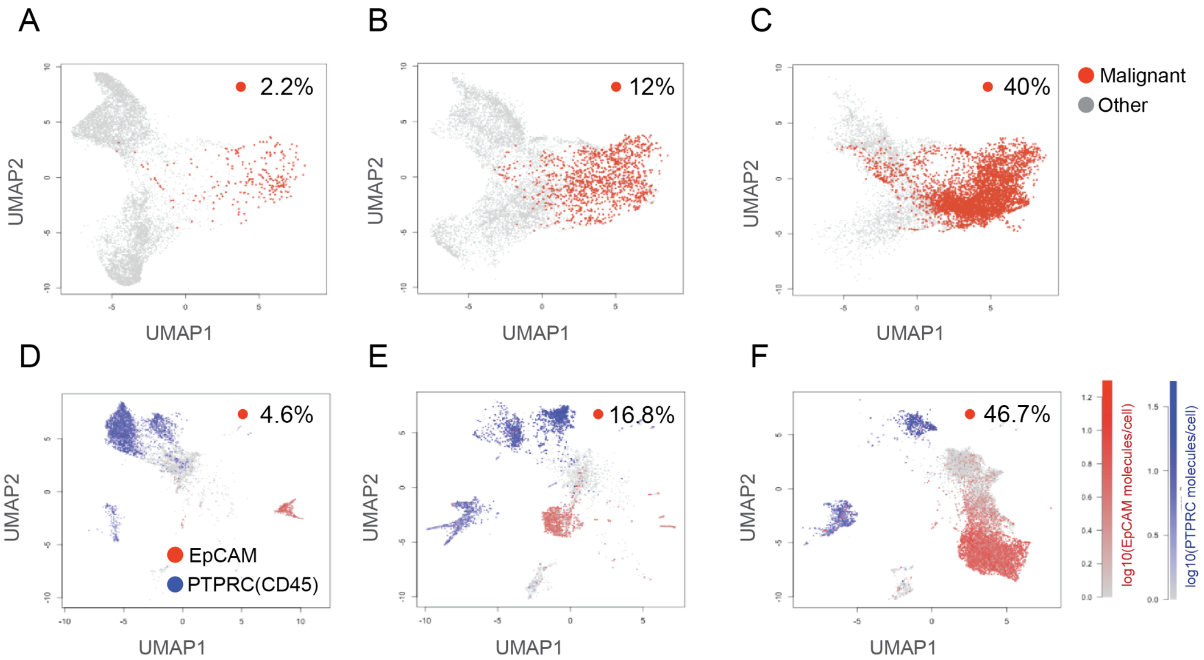


Figure S8. UMAP of morphology embeddings vs scRNAseq gene expression of three different dissociated tumor cell (DTC) samples. Samples from patients with lung adenocarcinoma containing low (**A, D**), medium (**B, E**) and high (**C, F**) percentage of malignant cells were tested. (**A-C**) UMAP of morphological embeddings: each data point is a cell; the predicted malignant cells are colored red and non-malignant cells colored gray. The plot labels indicate the fraction of malignant cells predicted by the model. (**D-E**) UMAP of single cell RNA gene expression profiles from all genes. The red and blue color gradients indicate the expression levels of EpCAM (tumor cell marker) and PTPRC (CD45, immune cell marker) respectively. The plot labels indicate the fraction of EpCAM+/PTRPC- cells. Overall, the morphology-based model predicted a similar fraction of malignant vs nonmalignant cells and UMAPs have similar resolution and separation of malignant vs nonmalignant cells for all three samples.

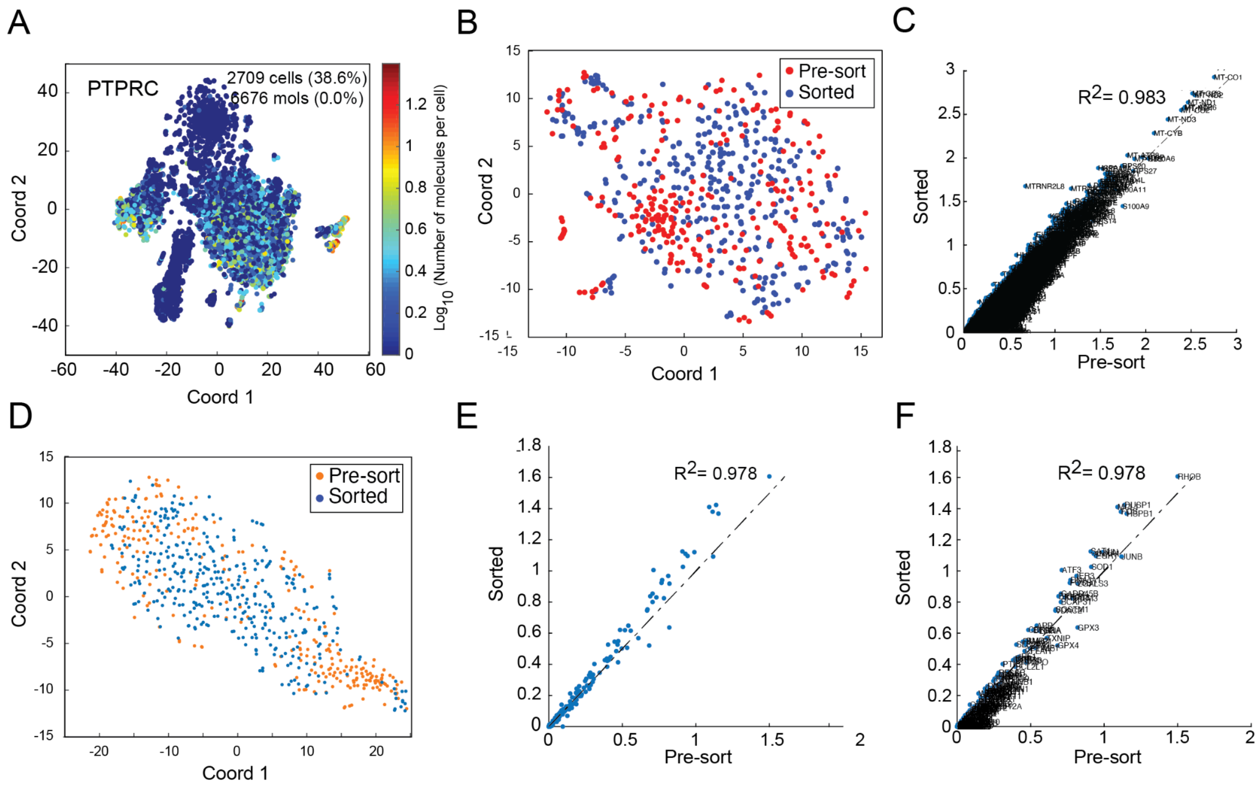


Figure S9. (**A**) Pseudo-color gene expression level of PTPRC (CD45, immune cell marker) in the DTC t-SNE plot. In combination with **fig. 3J** it confirmed that the Sorted cells were mostly in the EPCAM^{+}/PTPRC(CD45)$^{-}$ cluster. (**B**) Further sub-clustering of the EPCAM^{+}/PTPRC(CD45)$^{-}$ cluster showed that sorted cells almost completely overlapped with pre-sorted cells for all subclusters and (**C**) the gene expression profiles are highly correlated (gene correlation plot, each dot is a gene with the gene names annotated). (**D-F**) Stress and apoptosis related gene expression profile comparison of the pre-sorted and sorted cells in the EPCAM+/CD45- subpopulation. (**D**) t-SNE plot of the EpCAM^{+}/CD45^{-} cluster from **fig. 3I** using only the 166 stress and apoptosis genes, showing sorted cells overlap with pre-sort cells in all subclusters. (**E and F**) Gene expression profiles were highly corrected between sorted and pre-sort EpCAM+/CD45 cells; the correlation coefficient was 0.978 (each data point is a gene, **c** has gene names annotated), suggesting that COSMOS sorting did not cause additional cell stress.

References

Abadi, Martín, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. 2016. “Tensorflow: A System for Large-Scale Machine Learning.” In *12th ${USENIX} Symposium on Operating Systems Design and Implementation ({OSDI}$ 16)*, 265–83.

Alvarado-Kristensson, Maria, and Catalina Ana Rosselló. 2019. “The Biology of the Nuclear Envelope and Its Implications in Cancer Biology.” *International Journal of Molecular Sciences* 20 (10). https://doi.org/10.3390/ijms20102586.

Alvarez-Larran, A., T. Toll, S. Rives, and J. Estella. 2005. “Assessment of Neutrophil Activation in Whole Blood by Flow Cytometry.” *Clinical and Laboratory Haematology*. https://doi.org/10.1111/j.1365-2257.2004.00661.x.

Baykara, Onur, Burak Bakir, Nur Buyru, Kamil Kaynak, and Nejat Dalay. 2015. “Amplification of Chromosome 8 Genes in Lung Cancer.” *Journal of Cancer* 6 (3): 270–75.

Di Carlo, Dino, Daniel Irimia, Ronald G. Tompkins, and Mehmet Toner. 2007. “Continuous Inertial Focusing, Ordering, and Separation of Particles in Microchannels.” *Proceedings of the National Academy of Sciences of the United States of America* 104 (48): 18892–97.

Fischer, Edgar G. 2020. “Nuclear Morphology and the Biology of Cancer Cells.” *Acta Cytologica*, June, 1–9.

Gawad, Charles, Winston Koh, and Stephen R. Quake. 2016. “Single-Cell Genome Sequencing: Current State of the Science.” *Nature Reviews. Genetics* 17 (3): 175–88.

Hou, Le, Dimitris Samaras, Tahsin M. Kurc, Yi Gao, James E. Davis, and Joel H. Saltz. 2016. “Patch-Based Convolutional Neural Network for Whole Slide Tissue Image Classification.” *Proceedings / CVPR, IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 2016 (June): 2424–33.

Ma, Zhichao, Yinning Zhou, David J. Collins, and Ye Ai. 2017. “Fluorescence Activated Cell Sorting via a Focused Traveling Surface Acoustic Beam.” *Lab on a Chip* 17 (18): 3176–85.

McInnes, Leland, John Healy, and James Melville. 2018. “UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction.” *arXiv [stat.ML]*. arXiv. http://arxiv.org/abs/1802.03426.

Nitta, Nao, Takanori Iino, Akihiro Isozaki, Mai Yamagishi, Yasutaka Kitahama, Shinya Sakuma, Yuta Suzuki, et al. 2020. “Raman Image-Activated Cell Sorting.” *Nature Communications* 11 (1): 3452.

Nitta, Nao, Takeaki Sugimura, Akihiro Isozaki, Hideharu Mikami, Kei Hiraki, Shinya Sakuma, Takanori Iino, et al. 2018. “Intelligent Image-Activated Cell Sorting.” *Cell* 175 (1): 266–76.e13.

Rozenblatt-Rosen, Orit, Michael J. T. Stubbington, Aviv Regev, and Sarah A. Teichmann. 2017. “The Human Cell Atlas: From Vision to Reality.” *Nature* 550 (7677): 451–53.

Russakovsky, Olga, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, et al. 2015. “ImageNet Large Scale Visual Recognition Challenge.” *International Journal of Computer Vision* 115 (3): 211–52.

Schwartzman, Omer, and Amos Tanay. 2015. “Single-Cell Epigenomics: Techniques and Emerging Applications.” *Nature Reviews. Genetics* 16 (12): 716–26.

Stegle, Oliver, Sarah A. Teichmann, and John C. Marioni. 2015. “Computational and Analytical Challenges in Single-Cell Transcriptomics.” *Nature Reviews. Genetics* 16 (3): 133–45.

Schraivogel, Daniel, et al. "High-speed fluorescence image–enabled cell sorting." *Science* 375.6578 (2022): 315-320.

Szegedy, Christian, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. “Rethinking the Inception Architecture for Computer Vision.” In *Proceedings of the IEEE*

Conference on Computer Vision and Pattern Recognition, 2818–26.

Tate, John G., Sally Bamford, Harry C. Jubb, Zbyslaw Sondka, David M. Beare, Nidhi Bindal, Harry Boutselakis, et al. 2019. “COSMIC: The Catalogue Of Somatic Mutations In Cancer.” *Nucleic Acids Research* 47 (D1): D941–47.

Unger, M. A., H. P. Chou, T. Thorsen, A. Scherer, and S. R. Quake. 2000. “Monolithic Microfabricated Valves and Pumps by Multilayer Soft Lithography.” *Science* 288 (5463): 113–16.

Acknowledgements*:* We thank Catherine Blish for input and suggestions on the neutrophil study design and the manuscript and Prashast Knandelwal and Tariq Sharaat for early contributions to the COSMOS system development.

Funding: This work was supported by private funding.

Author contributions:

Research design and study concept: MS, MM, NL, HPC, EAA, KPP, TJM

Design studies: MS, MM, NL, HPC, KPP, KBJ, KS, AJ, EJL, CC, PN, SH

Perform experiments and analysis: KS, AJ, EJL, CC, PN, SH, RC, JM, KPP, AYWT, QFS, JC, JW, BC

Figures: MM, KPP, KBJ, CC, KS

Drafted manuscript: MM, MS, KBJ, NL, TJM, KS, AJ, CC, HPC, KPP, CJ

Review & editing manuscript: all authors

Competing interests

All authors are current or former employees at or are affiliated with Deepcell Inc.