Machine Learning-based Approaches for Identifying Human Cells Harboring Fetal Chromatin Domain Ablations

Yi Li
The University of Texas at Dallas
Shadi Zaheri
The University of Texas at Dallas
Khai Nguyen
The University of Texas at Dallas
Li Liu
The University of Texas at Dallas
Fatemeh Hassanipour
The University of Texas at Dallas
Leonidas Bleris (bleris@utdallas.edu)
The University of Texas at Dallas

Research Article

Keywords: Y-globin, FCD (Fetal Chromatin Domain), deep learning, machine learning, CRISPR

Posted Date: September 22nd, 2021

DOI: https://doi.org/10.21203/rs.3.rs-906377/v1

Abstract

Two common hemoglobinopathies, sickle cell disease (SCD) and β-thalassemia, arise from genetic mutations within the β-globin gene. A 500-bp motif termed Fetal Chromatin Domain (FCD), upstream of human Y-globin locus, may function as a transcriptional regulatory element driving inhibition of the Y-globin gene. Here, we hypothesize that the removal of this motif using CRISPR technology may reactivate the expression of Y-globin and subsequently restore fetal hemoglobin functionality. In this work we present two different cell morphology-based machine learning approaches that can be used identify cells that harbor FCD genetic modifications. Three candidate models from the first, which uses multilayer perceptron algorithm (MLP 20–26, MLP26-18, and MLP 30 – 26) and flow cytometry-derived cellular data, yielded 0.83 precision, 0.80 recall, 0.82 accuracy, and 0.90 area under the ROC (receiver operating characteristic) curve when predicting the edited cells. In comparison, the candidate model from the second approach, which uses deep learning (T2D5) and DIC microscopy-derived imaging data, performed with less accuracy (0.80) and ROC AUC (0.87). We envision both assays could be valuable and complementary to currently available genotyping protocols for specific genetic modifications which result in morphological changes in human cells.

Introduction

The Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) genome editing technology, which is adapted from an immune system analog found in archaea and prokaryotes, has been applied to exceedingly broad scientific, industrial, and medical domains at an exceptional pace, since the first demonstration in cells[1–8]. One particularly exciting application is CRISPR-based therapeutics, which as of today, have been extended to at least five treatment areas: blood disorders, cancers, eye diseases, chronic infections, and protein-folding disorders[9].

Two most common hemoglobinopathies, sickle cell disease (SCD) and β-thalassemia, arise from genetic mutations within the β-globin gene. These mutations result in deficient or absent β-globin synthesis, which in turn lead to oxygen being disassociated from the hemoglobin and eventually conformational changes in red blood cells[10,11]. No cure is available for these disorders except bone marrow transplantation (BMT) when a suitable donor is available, and most treatments are mainly aimed at relieving symptoms and preventing complications. Recently, the CRISPR technology has been used to reactivate the expression of fetal hemoglobin, which can take the place of defective adult hemoglobin, and shown remarkable results in improving the quality of life in such patients[12].

Machine learning, which can yield models for pattern recognition, classification, and prediction from acquired data, has been widely used in biological studies ranging from protein folding prediction[13] to cancer prognosis[14]. There are two main types of machine learning methods: (1) supervised learning (e.g. random forest, support vector machine), which derive the relationship between a set of input variables (features) and a designated dependent variable (label) from training instances and subsequently can be used to predict on new instances, and (2) unsupervised learning (e.g. clustering), which infer patterns from data without known labels[15]. More recently, deep learning, a collection of new machine learning techniques extended from classical neural networks, has gained popularity due to its better performance compared to existing best-in-class machine learning algorithms across several fields including linguistics[16], high-energy physics[17], computational chemistry[18], and biology[19].

One area that has received particular attention in recent years is classification of different cell types (e.g. different blood cell types)[20–23], states (apoptotic and healthy cells)[24–27], and genotypes[28] using machine learning approaches to provide novel insights for biological systems. In one study[29], Suzuki and colleagues developed a convolutional neural network (CNN) that was at least 90% accurate in classifying whole blood cells, peripheral blood mononuclear cells, human colon cancer cells, and human T lymphoma cells, using imaging flow. Similarly, in our previous work[24], we have shown that machine learning can be an efficient and cost-effective approach in identifying live and apoptotic human cells. Suzuki and colleagues also demonstrated that, using label-free, brightfield (BF) microscopy images, machine learning models (logistical regression) can be used to predict cells harboring ubiquitin-proteasome system-related genetic mutations with reasonably good performance (ROC AUC = 0.773)[28].

Recently, we discovered a 500-bp motif upstream of human Y-globin locus[30] (named as Fetal Chromatin Domain, FCD, **Supplementary Materials/FCD sequence**), which harbors a DNase hypersensitive site and the histone 3 lysine 4, mono-methylated (H3K4Me1) enhancer mark. Our study showed that it is a transcriptional regulatory element of the Y-globin gene and the removal of

this motif using CRISPR system reactivates the expression of fetal hemoglobin. Herein, we explore cell morphology-based machine learning approaches to classify cells with or without such genetic modifications within the FCD domain.

Results

Generation and characterization of heterozygous FCD-deficient KU-812 cell model (FCD-HT)

The CRISPR/SpCas9 technology was used to generate the FCD-deficient model in KU-812 cells, which were established from the peripheral blood of a patient with chronic myelogenous leukemia[31]. Briefly, the parental cells were transiently transfected with CRISPR/SpCas9 complex which targets both left and right genomic regions flanking the FCD motif, along with a homologous recombination repair template containing a puromycin resistance gene transcript (**Figure 1A**). The polyclonal stable cell line was then established after 2 weeks of puromycin selection (2μg/mL). Subsequently, to avoid potential interference with the transcriptional activities of the globin genes, the puromycin resistance gene transcript (~2.2kb), which was flanked by flippase recognition target (FRT) sites, was removed using flippase. Finally, monoclonal stable cells were established using FACS single cell sorting.

To characterize the stable monoclonal cell line, genomic DNA was isolated using DNeasy Blood&Tissue Kit (Qiagen), and subsequently the transcript containing the FCD motif or FRT site was amplified using primers P13 and P14 (**Supplementary Table 1**). The PCR products were then subjected to gel electrophoresis and as shown in **Figure 1B**, the monoclonal cell line yielded two distinct bands corresponding to both the wild type (806 bp) and FCD-knockout (341 bp) alleles, confirming its heterozygous status (named as FCD-HT). Both bands were further extracted and subjected to Sanger sequencing, which confirmed that the FCD sequence was successfully removed in the FCD-Knockout allele (**Supplementary Figure 1**). To determine how the FCD-removal affects the Y-globin expression, total RNAs were extracted using the RNeasy Mini Kit from KU-812 and FCD-HT cells and the relative expression of Y-globin transcript was determined using quantitative reverse transcription-PCR (qRT-PCR). As shown in **Figure 1C**, the mRNA level of Y-globin significantly increased in FCD-HT cells (2.87-fold compared to its parental KU-812, named as HCT-WT), which is consistent with our hypothesis that the FCD motif may serve as a transcriptional repressor domain within human globin locus.

Flow cytometry-based data collection and visualization for FCD-WT and FCD-HT cells

To prepare cell morphology-based predictive models differentiating FCD-WT and FCD-HT cells, we first used flow cytometry assay to record 6 features (FSC-A, FSC-H, FSC-W, SSC-A, SSC-H, and SSC-W). In total, 192,772 FCD-WT cells (labeled as 0) and 185,544 FCD-HT cells (labeled as 1) were included (the ratio of labels 0 and 1 = 1.04, **Supplementary Table 2**). Next, this initial dataset was randomly split into training and testing datasets at a ratio of 80:20 (size of training dataset: size of testing dataset). Specifically, the training dataset contains 302,652 cells (label 0: 154,180 cells, label 1: 148,472 cells, **Supplementary Figure 2** and **Supplementary Table 3**), and the testing dataset contains 75,664 cells (label 0: 38,592 cells, label 1: 37,072 cells, **Supplementary Figure 3** and **Supplementary Table 4**).

We first compared the absolute readings among the 6 features using box plotting. As shown in **Figure 2A**, the means of these features varied significantly with the maximal ratio larger than 2.0-fold ($mean_{FSC\text{-}A}/mean_{SSC\text{-}H}$ = 2.47), indicating that standardization of the original training and testing datasets are required (standardized training and testing datasets in **Supplementary Table 5** and **6**, respectively). Subsequently, the standardized training dataset was subjected to two dimensionality reduction algorithms, principal component analysis (PCA) and t-distributed stochastic neighbor embedding t-SNE (t-SNE). As shown in **Figure 2B** (PCA) and **Figure 2C** (t-SNE), the two cell subpopulations (FCD-WT: green, FCD-HT: yellow) demonstrated distinct distributive patterns and were partially separable.

Cell morphology-based machine learning models using flow cytometry-derived data

A general workflow as described in our previous study was adopted to build and test various cell morphology-based machine learning models using flow cytometry-derived data[24]. In total, five (5) supervised learning algorithms (logistical regression, random forest, k-nearest neighbor, support-vector machine, and multilayer perceptron) were included (model hyperparameters in **Supplementary Table 7**).

First, using tenfold cross-validation, we screened all models with the standardized training dataset, and adopted the filtering conditions as (1) the mean accuracy > 0.80, and (2) the standard deviation of accuracy < 0.10. In total, one (1) logistic regression model (**Supplementary Table 8**), 94 random forest models (**Supplementary Table 9**), 96 k-nearest neighbor models (**Supplementary Table 10**), two (2) SVM models with linear kernel (**Supplementary Table 11**), 25 SVM models with Gaussian kernel (**Supplementary Table 12**), and 893 MLP models (**Supplementary Table 13**) were selected.

Next, all selected models (1,111) were trained using the training dataset, and subsequently applied to the standardized testing dataset and subjected to secondary filtering conditions as (1) precision when predicting FCD-HT cells > 0.80, and (2) recall when predicting FCD-HT cells > 0.80. As shown in **Supplementary Table 14**, only 533 MLP models survived this additional filter.

Finally, we chose three MLP models with largest AUC values (**Table 1**, MLP 20-26: number of nodes in the first layer: 20/number of nodes in the second layer: 26, MLP 26-18: number of nodes in the first layer: 26/number of nodes in the second layer: 18, and MLP 30-26: number of nodes in the first layer: 30/number of nodes in the second layer: 26), and plotted both the receiver operating characteristics (ROC, **Figure 3A**) and precision-recall curves (**Figure 3B**). The three models displayed essentially identical performance when predicting FCD-HT cells (precision: 0.83, recall: 0.80, accuracy: 0.82, and AUC: 0.90).

Cell morphology-based machine learning models using microscopy-derived data

In addition to flow cytometry, cell morphology information can also be directly assessed using imaging[28]. Using a Differential Interference Contrast (DIC) microscopy, we prepared 1,594 images of individual FCD-WT cells and 1,695 images of FCD-HT cells (**Supplementary Figure 4)**, the ratio of FCD-WT and FCD-HT = 0.94). Next, this starting dataset was randomly split into the training and testing datasets at a ratio of 90:10 (size of training dataset: size of testing dataset). The final training dataset contains 2,956 images (FCD-WT: 1,433 images, FCD-HT: 1,523 images), and the testing dataset contains 333 images (FCD-WT: 161 images, FCD-HT: 172 images).

Next, deep learning-based convolutional neural networks (CNNs) were used to construct genotype-predictive models. Two general CNN architectures were explored: (1) Type 1 (T1): $(\text{Conv-Conv-Pool})_n$, which was based on the VGG design[32], and (2) Type 2 (T2): $(\text{Conv-Pool})_n$, which contained a single convolution layer for each repeat. For each type, different number of convolution numbers were tested (two, four and six layers for T1, and two, three, four, five layers for T2) until the final feature map reaches a dimension of zero. Since our image inputs have a relatively small size (100 pixels by 100 pixels), we fixed the filter size at 3 and when applicable, the Maxpooling pool size at 2. The detailed architectural designs were included in **Supplementary Table 15**.

As an example, for Type 2/5 layers (T2D5, **Figure 4**), the numbers of layers at the feature extraction step were 32, 64, 92, 100 and 128 for each successive layer, and rectified linear unit (ReLU) was used as the activation function. Additionally, a Maxpooling layer was included after each convolution layer. Next, the outputs from convolutional layers were subjected to global average pooling and converted into a 1-dimensional vector (Flatten) for a fully connected layer (dense, 1028 nodes). Finally, a Softmax classifier, which applies a categorical cross-entropy loss function, was used, together with the adaptive moment estimation (ADAM) optimization algorithm.

First, all 7 candidate architectures were subjected to tenfold cross-validation using the training dataset. As shown in **Supplementary Table 16**, models from Type 2 showed better performance compared to those from Type 1. Specifically, the best-performing model of Type 2 (T2D5) showed a mean of accuracies from 10 cross-validation tests at 67.3% (**Supplementary Figure 5**), while the best-performing model of Type 1 (T1D4) yielded a mean of accuracies at 58.3%.

We further trained models using all candidate architectures and the training dataset, and subsequently applied them to the testing dataset. As shown in **Table 2**, the architectures T2D5 displayed the best predictive outcome. Specifically, for FCD-HT cells, precision was 0.84, recall was 0.76, accuracy was 0.80 and AUC was 0.87 (**Figure 5**).

Discussion

In addition to five supervised machine learning algorithms, we subjected our flow cytometry-derived dataset (the standardized training dataset) to two unsupervised clustering algorithms (k-means clustering and Gaussian mixture clustering) in parallel[33]. As shown in **Supplementary Fig. 6** (SSC-A vs. FSC-A), the predicted distributive pattern for two subpopulations from k-means algorithm

differed drastically from real genotype labels (**Supplementary Fig. 2**). Specifically, even with the best-performing labeling schema (green: FCD-WT, red: FCD-HT), the model yielded poor predictive performance when predicting FCD-HT cells (precision: 0.52, recall: 0.66, accuracy: 0.53). Similarly, as shown in **Supplementary Fig. 7**, the predictive model derived from Gaussian mixture clustering performed poorly, with precision at 0.61, recall at 0.25 and accuracy at 0.55 when predicting FCD-HT cells (green: FCD-WT, red: FCD-HT). It is interesting to note that in our previous study, which focused on predicting cell states using cell morphological information, supervised learning also performed significantly better than unsupervised learning. These results may be because compared to supervised learning, unsupervised learning uses less information (unknown outputs/labels), and thus may be less accurate when applied to data from complex systems such as human cells.

In this study, we have investigated two alternative approaches in predicting cell genotypes: (1) numeric data based on flow cytometry assay, and (2) imaging data based on DIC microscopy. Our analysis showed that the best performing models from approach 1 (MLP 20–26, MLP 26 – 18, and MLP 30 – 26) yielded better prediction results compared to the best model from approach 2 (T2D5) (**Tables 1 and 2**, ROC AUC values for MLP 20–26, MLP 26 – 18, and MLP 30 – 26: 0.90, for T2D5: 0.87). Multiple factors maybe be involved in this observed discrepancy. For example, the resolution of our images was relatively low (100 pixels by 100 pixels) due to the instrumental limitation of our DIC microscopy. Additionally, compared to the training dataset from flow cytometry (302,652 cells), the size of imaging dataset was vastly smaller (3,289 images). To overcome this challenge, we resorted to data augmentation techniques by applying zooming (range: 0.5-1.5x), rotation (range: 90 degree), width shifting (range: -10 to 10 pixels), and height shifting (range: -10 to 10 pixels) to the original training dataset (**Supplementary Fig. 8**)[34]. Together with original samples, the augmented training dataset contained 26,604 images (FCD-WT: 12,897 images, FCD-HT: 13,707 images), which was subsequently subjected to deep learning modeling using the T2D5 architecture. As shown in **Supplementary Table 17**, the newly acquired model did not yield better predictive performance. As an example, for FCD-HT cells, precision was 0.77, recall was 0.74, accuracy was 0.75, and ROC AUC was 0.75, which were lower than those of the original model (precision: 0.84, recall: 0.76, accuracy: 0.80, ROC AUC: 0.87). These results showed that synthetic samples do not always enhance the model performance in deep learning.

Lastly, for our default T2D5 model, while the accuracy approached 1.0 for the training dataset with the progression of epochs, the accuracy of the testing dataset plateaued at a much lower value (~ 0.8). This discrepancy indicated a potential overfitting (**Supplementary Fig. 9**). Consequently, we investigated additional tunings on overfitting/underfitting-controlling hyperparameters: L2 regularization (Ridge regression) weights (parameter: kernel_regularizer), and dropout values[35]. As shown in **Supplementary Table 18**, of the parameter space that we have scanned (L2 regularization weight: [0.000001, 0.00001, 0.0001, 0.01], dropout value: [0, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.45]), none of the adjusted models (25) yielded better predictive performance compared to the original model (L2 regularization weight = 0, dropout value = 0), which indicated that additional factors may be involved in the divergency of accuracies between training and testing datasets.

Typically, to establish and confirm a mammalian stable cell line, polyclonal cells need to be sorted into single cells, commonly on 96-well plates. Next, the single cells are allowed to propagate until sufficient genomic material can be extracted and subjected to PCR-based genotyping assays (Fig. 1). While the protocol is well established, the full pipeline can become time-consuming (up to several weeks for cell propagation step dependent on the cell proliferation rates) and labor-intensive (hundreds of monoclones may be needed for acquiring desirable genotypes). On the other hand, sophisticated imaging flow cytometry techniques, which record extensive physically measurable quantities (features) of the cells, have been used to identify cell subpopulations with specific traits (e.g. cell types, apoptotic states)[26,29,36,37]. However, most of these instruments are complex, expensive, and may not be yet available to most research labs. In comparison, our flow cytometry and DIC microscopy-based machine learning approaches provide a unique balance between efficacy and availability, and theoretically can be applied to any engineered cells with genetic modifications known to introduce cell morphological changes.

In conclusion, we demonstrated the feasibility to use flow cytometry-based cellular information (FSC-A, FSC-H, FSC-W, SSC-A, SSC-H, and SSC-W) to predict specific cell genotypes using multilayer perceptron algorithms (MLP 20–26, MLP 26 – 18, and MLP 30 – 26). Additionally, we showed that deep learning framework (T2D5), when applied to DIC microscopy images, can also be indicative for certain genotyping purposes. We envision both assays could be valuable and complementary to currently available genotyping protocols.

Materials And Methods

Cell culture

The KU-812 parental and derived cells were acquired from the American Type Culture Collection (ATCC, catalog number: CRL-1573) and maintained at 37°C, 100% humidity and 5% CO_2. The cells were grown in Dulbecco's modified Eagle's medium (DMEM, Invitrogen, catalog number: 11965–1181) supplemented with 10% Fetal Bovine Serum (FBS, Invitrogen, catalog number: 26140), 0.1 mM MEM non-essential amino acids (Invitrogen, catalog number: 11140–050), and 0.045 units/mL of Penicillin and 0.045 units/mL of Streptomycin (Penicillin-Streptomycin liquid, Invitrogen, catalog number: 15140). To pass the cells, the confluent cell culture was diluted in fresh medium at a ratio of 1:10. When applicable, 2 µg/mL puromycin (ThermoFisher Scientific, catalog number: A1113803) was added to the growth medium.

Generation of FCD-HT monoclonal stable cell line

To generate the FCD-HT monoclonal stable cell line, ~10 million of the human KU-812 cells were seeded onto a 10 cm petri dish. 16 hours later, the cells were transiently transfected with 4.5 µg of PCMV-SpCas9-U6-sgRNA-L, 4.5 µg of PCMV-SpCas9-U6-sgRNA-R, and 1 µg of the donor plasmid using the JetPEI reagent (Polyplus Transfection). 48 hours later, puromycin was added at the final concentration of 2 µg/mL. The selection lasted ~2 weeks, after which the surviving clones were pooled to generate the polyclonal stable cells. Next, to remove the puromycin resistance gene-T2A-mKate cassette, ~10 million of the polyclonal stable cells were seeded onto a 10 cm petri dish, and after 16 hours were transfected with 10 µg of EF1-Flpase (unpublished data) using the JetPEI reagent. 48 hours later, single cells were isolated using flow cytometry. The established monoclonal stable cell line was confirmed to be heterozygous by genotyping and further expanded and maintained in the complete growth medium.

Genotyping of FCD-HT monoclonal stable cell line

The genomic DNAs were isolated from FCD-HT monoclonal stable cells using DNeasy Blood&Tissue Kit (Qiagen). The transcripts containing the CRISPR-targeting region was amplified with primers P13 and P14. The PCR products were then subjected to gel electrophoresis and Sanger sequencing using primers P13 and P14.

Quantitative reverse transcription-PCR (qRT-PCR)

For measurement of mRNA levels of various human globin variants, total RNA was extracted using the RNeasy Mini Kit (Qiagen, #74104). First strand synthesis was performed using the QuantiTect Reverse Transcription Kit (Qiagen, #205311). Quantitative PCR was performed using the KAPA SYBR FAST Universal qPCR Kit (KAPABiosystems, #KK4601), with GAPDH levels used for normalization. Quantitative analysis was performed using the $2^{-\Delta\Delta Ct}$ method. Fold-change values are reported as mean with standard deviation. Primers used for Y-globin were (P15) 5'-GGCAACCTGTCCTCTGCCTC-3' and (P16) 5'-TAGGAAGCCATTTCTGCCTTG-3'. Primers used for GAPDH were (P17) 5'-AATCCCATCACCATCTTCCA-3' and (P18) 5'-TGGACTCCACGACGTACTCA-3'.

Flow cytometry

FCD-WT and FCD-HT cells from a 10-cm petri dish were washed with 5 mL PBS buffer, and subsequently trypsinized with 2 mL 0.25% Trypsin-EDTA at 37°C for 5 min. Trypsin-EDTA was then neutralized by adding 10 mL of complete medium. The cell suspension was centrifuged at 1,000 rpm for 5 min and after removal of supernatants, the cell pellets were re-suspended in 5 mL PBS buffer. The cells were analyzed on a BD Reforest flow analyzer. The voltages (V) for each channel were: 270 for FSC-A, 270 for FSC-H, 270 for FSC-W, 280 for SSC-A, 280 for SSC-H, and 280 for SSC-W.

Differential Interference Contrast (DIC) microscopy

Approximately 50,000 FCD-WT or FCD-HT cells were seeded on 12-well plates (Greiner Bio-One) in the complete medium. Cells were imaged using an Olympus IX81 microscope in a Precision Control environmental chamber. The images were captured using a Hamamatsu ORCA-03 Cooled monochrome digital camera. The filter set was Differential Interference Contrast (DIC) with magnification at 40X. After obtaining the images, Adobe Photoshop was used to isolate individual cells with fixed size at 100 pixels by 100 pixels.

Machine learning model training and testing

For flow cytometry-derived dataset, a Dell desktop computer (Intel Core i7-10700 CPU @ 2.90 GHz, Windows 10 enterprise 64-bit OS and 32 GB RAM) was used to conduct the machine learning modeling. Scikit-Learn, a free Python machine learning library, was used to conduct all model training and testing procedures. For DIC microscopy-derived dataset, a Lenovo Laptop (Intel Core i7-10510 CPU @ 1.80 GHz, Ubuntu 20.04 OS and 16 GB RAM) was used to conduct the deep learning modeling. The Keras library in TensorFlow was used to conduct all model training and testing procedures. Other Python libraries, including NumPy, Pandas, and Matplotlib, were also included for data analysis and presentation.

Performance metrics

Performance of different models was evaluated using threshold dependent and independent metrics, which include:

(1) precision: this parameter measures how accurate a model is when predicting cells being at live state.

Precision = TP/(TP + FP), where TP refers to correctly predicted live cells and FP refers to falsely predicted live cells.

(2) recall: this parameter measures the model's ability to correctly predict live cells from actual live cells.

Recall = TP/(TP + FN), where TP refers to correctly predicted live cells and FN refers to falsely predicted apoptotic cells.

(3) true positive rate (TPR): this parameter measures the model's ability to correctly predict live cells from actual live cells.

TPR = TP/(TP + FN), where TP refers to correctly predicted live cells and FN refers to falsely predicted apoptotic cells.

(4) false-positive rate (FPR): this parameter measures the model's level of falsely predicting live cells from actual apoptotic cells.

FPR = FP/(FP + TN), where FP refers to falsely predicted live cells and TN refers to correctly predicted apoptotic cells.

(5) accuracy: this parameter determines the success of correctly predict live and apoptotic cells from overall data.

Accuracy = (TP + TN)/(TP + FP + TN + FN), where TP refers to correctly predicted live cells, FP refers to falsely predicted live cells, FN refers to falsely predicted apoptotic cells, and TN refers to correctly predicted apoptotic cells.

Declarations

Acknowledgements

This work was funded by the US National Science Foundation (NSF) grant 2029121, a Cecil H. and Ida Green Endowment, and the University of Texas at Dallas.

Contributions

Y.L., S.Z. and K.N. developed and performed the computational analysis. Y.L., K.N. and L.L. performed the experiments. Y.L., S.Z., K.N., F.H., and L.B. wrote the paper. L.B. supervised the project.

Corresponding author

Correspondence to Leonidas Bleris.

Competing interests

The authors declare no competing interests.

References

1. Wang, H. *et al.* One-Step Generation of Mice Carrying Mutations in Multiple Genes by CRISPR/Cas-Mediated Genome Engineering., **153** (4), 910–918 https://doi.org/10.1016/j.cell.2013.04.025 (2013).
2. Cong, L. *et al.* Multiplex genome engineering using CRISPR/Cas systems. *Science (80-)*, **339** (6121), 819–823 https://doi.org/10.1126/science.1231143 (2013).
3. Hsu, P. D., Lander, E. S. & Zhang, F. Development and Applications of CRISPR-Cas9 for Genome Engineering. *Cell Press*, **Vol 157**, 1262–1278 https://doi.org/10.1016/j.cell.2014.05.010 (2014).
4. Ran, F. A. *et al.* In vivo genome editing using Staphylococcus aureus Cas9. *Nature*, **520** (7546), 186–191 https://doi.org/10.1038/nature14299 (2015).
5. Jinek, M. *et al.* A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science (80-)*, **337** (6096), 816–821 https://doi.org/10.1126/science.1225829 (2012).
6. Mali, P. *et al.* RNA-guided human genome engineering via Cas9. *Science (80-)*, **339** (6121), 823–826 https://doi.org/10.1126/science.1232033 (2013).
7. Moore, R. *et al.* CRISPR-based self-cleaving mechanism for controllable gene delivery in human cells. *Nucleic Acids Res*, **43** (2), 1297–1303 https://doi.org/10.1093/nar/gku1326 (2015).
8. Li, Y., Nowak, C. M., Withers, D., Pertsemlidis, A. & Bleris, L. CRISPR-Based Editing Reveals Edge-Specific Effects in Biological Networks. *Cris J*, **1** (4), 286–293 (2018).
9. Luthra, R., Kaur, S. & Bhandari, K. Applications of CRISPR as a potential therapeutic. *Life Sci. Published online August*, **25**, 119908 https://doi.org/10.1016/J.LFS.2021.119908 (2021).
10. Asano, H., Li, X. S. & Stamatoyannopoulos, G. FKLF, a Novel Krüppel-Like Factor That Activates Human Embryonic and Fetal β-Like Globin Genes. *Mol Cell Biol*, **19** (5), 3571–3579 https://doi.org/10.1128/mcb.19.5.3571 (1999).
11. Li, B., Ding, L., Li, W., Story, M. D. & Pace, B. S. Characterization of the transcriptome profiles related to globin gene switching during in vitro erythroid maturation. *BMC Genomics*, **13** (1), 153 https://doi.org/10.1186/1471-2164-13-153 (2012).
12. Frangoul, H. *et al.* CRISPR-Cas9 Gene Editing for Sickle Cell Disease and β-Thalassemia. *N Engl J Med*, **384** (3), 252–260 https://doi.org/10.1056/nejmoa2031054 (2021).
13. Senior, A. W. *et al.* Improved protein structure prediction using potentials from deep learning. *Nature*, **577** (7792), 706–710 https://doi.org/10.1038/S41586-019-1923-7 (2020).
14. Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V. & Fotiadis, D. I. Machine learning applications in cancer prognosis and prediction. *Comput Struct Biotechnol J*, **13**, 8–17 https://doi.org/10.1016/J.CSBJ.2014.11.005 (2015).
15. Giger, M. L. Machine Learning in Medical Imaging. *J Am Coll Radiol*, **15** (3), 512–520 https://doi.org/10.1016/j.jacr.2017.12.028 (2018).
16. Lakretz, Y. *et al.* Mechanisms for handling nested dependencies in neural-network language models and humans., **213**, 104699 https://doi.org/10.1016/J.COGNITION.2021.104699 (2021).
17. Azimi, S. A. *et al.* Classification of radioxenon spectra with deep learning algorithm. *J Environ Radioact*, **237**, 106718 https://doi.org/10.1016/J.JENVRAD.2021.106718 (2021).
18. Townshend, R. J. L. *et al.* Geometric deep learning of RNA structure. *Science (80-)*, **373** (6558), 1047–1051 (2021). https://science.sciencemag.org/content/373/6558/1047%0Ahttps://science.sciencemag.org/content/373/6558/1047.abstract
19. Nabwire, S. *et al.* Application of artificial intelligence in phenomics. *Sensors*, **21** (13), 1–19 https://doi.org/10.3390/s21134363 (2021).
20. Habibzadeh Motlagh, M., Jannesari, M., Rezaei, Z., Totonchi, M. & Baharvand, H. Automatic white blood cell classification using pre-trained deep learning models. *ResNet and Inception*, **1069612** (April 2018), 105 https://doi.org/10.1117/12.2311282 (2018).
21. Huang, X. *et al.* Deep-learning based label-free classification of activated and inactivated neutrophils for rapid immune state monitoring. *Sensors (Switzerland)*, **21** (2), 1–14 https://doi.org/10.3390/s21020512 (2021).
22. Nassar, M. *et al.* Label-Free Identification of White Blood Cells Using Machine Learning. *Cytom Part A*, **95** (8), 836–842 https://doi.org/10.1002/cyto.a.23794 (2019).

23. Lin, Y-H., Liao, K. Y. K. & Sung, K-B. Automatic detection and characterization of quantitative phase images of thalassemic red blood cells using a mask region-based convolutional neural network. *J Biomed Opt*, **25** (11), 1–14 https://doi.org/10.1117/1.jbo.25.11.116502 (2020).
24. Li, Y., Nowak, C. M., Pham, U., Nguyen, K. & Bleris, L. Cell morphology-based machine learning models for human cell state classification. *npj Syst Biol Appl*, **7** (1), 1–9 https://doi.org/10.1038/s41540-021-00180-y (2021).
25. Pischel, D., Buchbinder, J. H., Sundmacher, K., Lavrik, I. N. & Flassig, R. J. A guide to automated apoptosis detection: How to make sense of imaging flow cytometry data. Rishi A, ed*PLoS One*. 2018;13(5):e0197208. doi:10.1371/journal.pone.0197208
26. Feng, J. *et al.* Feasibility study of stain-free classification of cell apoptosis based on diffraction imaging flow cytometry and supervised machine learning techniques., **23** (5–6), 290–298 https://doi.org/10.1007/s10495-018-1454-y (2018).
27. Vicar, T., Raudenska, M., Gumulec, J. & Masarik, M. Detection and characterization of apoptotic and necrotic cell death by time-lapse quantitative phase image analysis. *bioRxiv*, (March), 1–21 https://doi.org/10.1101/589697 (2019).
28. Suzuki, G. *et al.* Machine learning approach for discrimination of genotypes based on bright-field cellular images. *npj Syst Biol Appl*, **7** (1), 1–8 https://doi.org/10.1038/s41540-021-00190-w (2021).
29. Suzuki, Y. *et al.* Label-free chemical imaging flow cytometry by high-speed multicolor stimulated Raman scattering. *Proc Natl Acad Sci U S A*, **116** (32), 15842–15848 https://doi.org/10.1073/pnas.1902322116 (2019).
30. Liu, L. *et al.* A chromatin region regulates transcription of γ-Globin genes. manuscript in preparation
31. Nakazawa, M. *et al.* KU 812: a pluripotent human cell line with spontaneous erythroid terminal maturation., **73** (7), 2003–2013 https://doi.org/10.1182/blood.v73.7.2003.2003 (1989).
32. Younis, M. C. Evaluation of deep learning approaches for identification of different corona-virus species and time series prediction. *Comput Med Imaging Graph*, **90**, 101921 https://doi.org/10.1016/J.COMPMEDIMAG.2021.101921 (2021).
33. Lin, M. *et al.* Artificial intelligence in tumor subregion analysis based on medical imaging: A review. *J Appl Clin Med Phys*, **22** (7), 10–26 https://doi.org/10.1002/acm2.13321 (2021).
34. Moses, D. A. Deep learning applied to automatic disease detection using chest X-rays. *J Med Imaging Radiat Oncol*, **65** (5), 498–517 https://doi.org/10.1111/1754-9485.13273 (2021).
35. Li, H., Weng, J., Mao, Y. & Wang, Y. *on Biological Principles*, **32** (9), 1–10 (2021).
36. Shir, O. M., Raz, V., Dirks, R. W. & Bäck, T.. Classification of cell fates with support vector machine learning. In: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Vol 4447 LNCS.Springer, Berlin, Heidelberg; 2007:258–269. doi:10.1007/978-3-540-71783-6_25
37. Lee, K. C. M. *et al.* Multi-ATOM: Ultrahigh-throughput single-cell quantitative phase imaging with subcellular resolution. *J Biophotonics*, **12** (7), https://doi.org/10.1002/jbio.201800479 (2019).

Figures

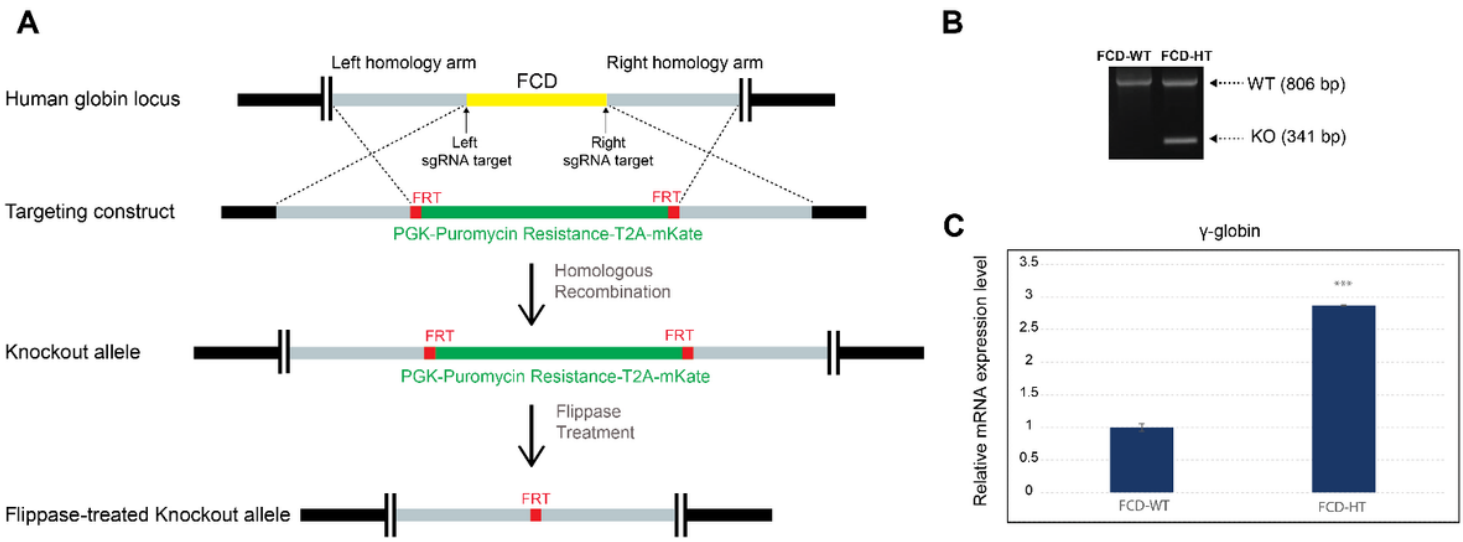

Figure 1

Generation and characterization of heterozygous FCD-deficient KU-812 cell model (FCD-HT). (A) Schematic illustration of the CRISPR/SpCas9 homologous recombination process to remove the FCD motif within human globin locus. The polyclonal stable cell line was established using 2 weeks of puromycin selection (2 μg/mL). Subsequently, the puromycin resistance gene transcript, which was flanked by flippase recognition target (FRT) sites, was removed using flippase. Finally, monoclonal stable cells were established using FACS single cell sorting. (B) Genotyping of FCD-HT monoclonal stable cell. Genomic DNAs were harvested from FCD-WT and FCD-HT cells and the transcript containing the FCD motif or FRT site was PCR amplified and subsequently subjected to gel electrophoresis. The stable cell line yielded two bands corresponding to both the wild type (806 bp) and FCD-knockout (341 bp) alleles, confirming its heterozygous status. (C) Quantitative reverse transcription-PCR (qRT-PCR) assay showed that compared to FCD-WT cells, the mRNA level of Y-globin significantly increased in FCD-HT cells (2.87-fold). *** denotes p-value < 0.001.

Figure 2

General statistics and visualization of the training dataset. (A) Box plot of the training dataset. The means of the six features (FSC-A, SSC-A, FSC-H, SSC-H, FSC-W, and SSC-W) varied significantly with the maximal ratio larger than 2.0-fold (meanFSC-A/meanSSC-H = 2.47), indicating that standardization of the original training and testing datasets are required. (B) Visualization of the standardized training dataset using PCA (green: FCD-WT, yellow: FCD-HT). (C) Visualization of the standardized training dataset using t-SNE (green: FCD-WT, yellow: FCD-HT).

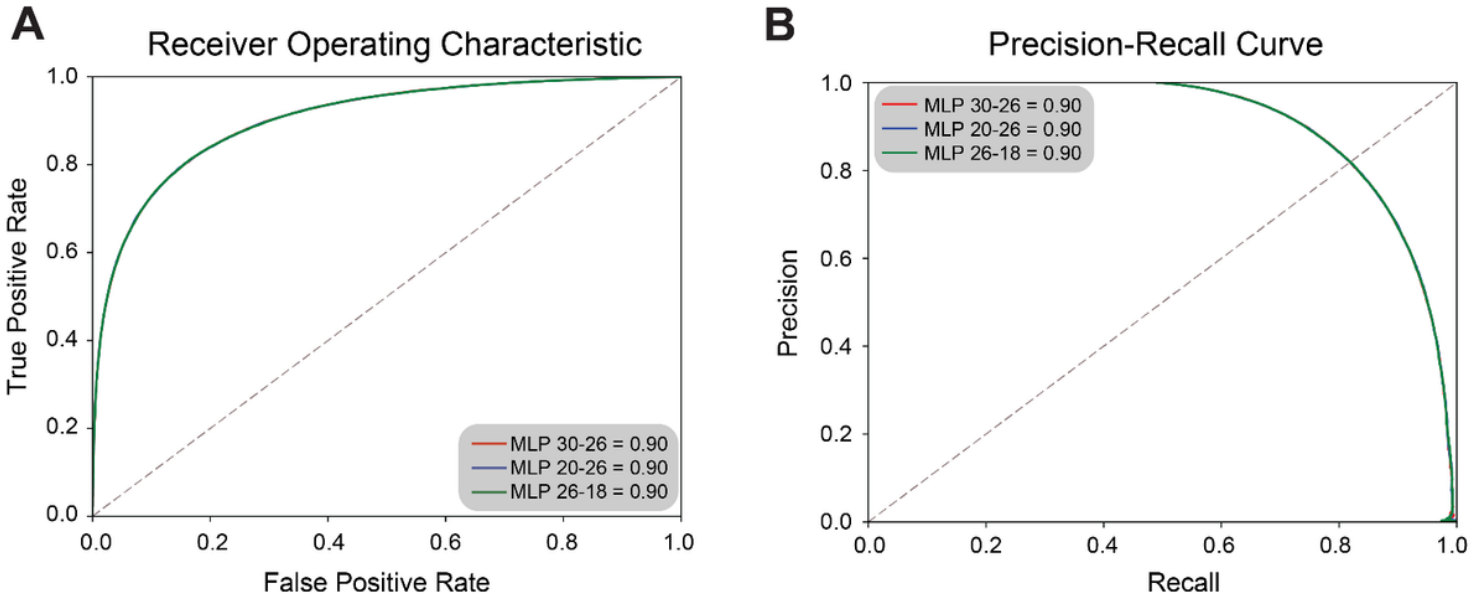

Figure 3

Predictive performances of the three candidate MLP models on FCD-HT cells. Both (A) ROC curve and (B) Precision-Recall curve showed that the three MLP models (MLP 20-26, MLP 26-18, and MLP 30-26) can predict the FCD-HT cells with relatively high precisions and recalls (precision: 0.83, recall: 0.80, accuracy: 0.82, and AUC: 0.90).

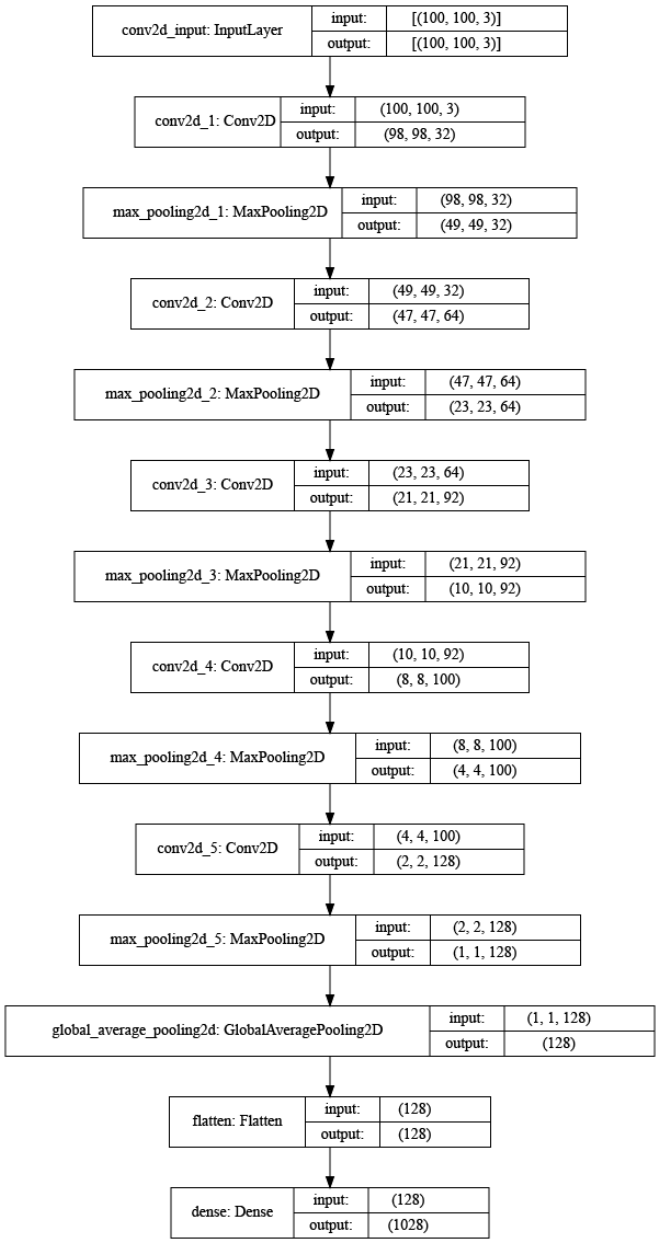

Figure 4

The T2D5 deep learning architecture. The model contained five convolutional layers (the numbers of each layer: 32, 64, 92, 100 and 128). Additionally, a Maxpooling layer was included after each convolution layer. Next, the outputs from convolutional layers were subjected to global average pooling and flattened for a fully connected layer (1028 nodes). Finally, a Softmax classifier, which applies a categorical cross-entropy loss function, was used.

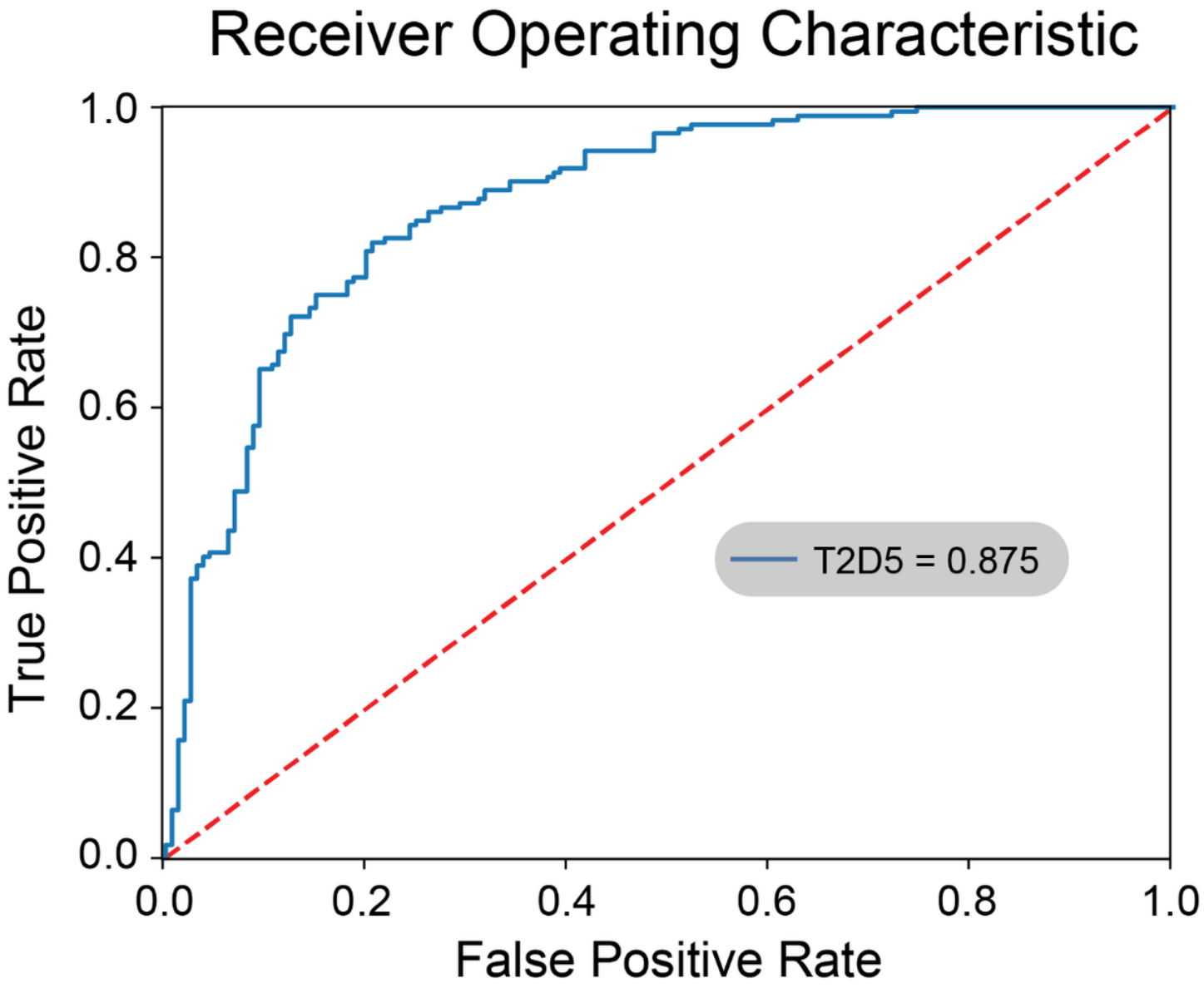

Figure 5

The ROC (Receiver Operating Characteristic) curve of the candidate deep learning model T2D5. The model showed relatively good performance when predicting FCD-HT cells (precision: 0.84, recall: 0.76, accuracy: 0.80, and AUC: 0.87).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Table1.xlsx
- Table2.xlsx
- SupplementaryTable1.xlsx
- SupplementaryTable2.xlsx
- SupplementaryTable3.xlsx
- SupplementaryTable4.xlsx
- SupplementaryTable5.xlsx

- SupplementaryTable6.xlsx
- SupplementaryTable7.xlsx
- SupplementaryTable8.xlsx
- SupplementaryTable9.xlsx
- SupplementaryTable10.xlsx
- SupplementaryTable11.xlsx
- SupplementaryTable12.xlsx
- SupplementaryTable13.xlsx
- SupplementaryTable14.xlsx
- SupplementaryTable15.xlsx
- SupplementaryTable16.xlsx
- SupplementaryTable17.xlsx
- SupplementaryTable18.xlsx
- supplementarymaterials.docx