Supplementary file 1

Predicting transcription factor binding in single

cells through deep learning

Laiyi Fu^{1,2}, Lihua Zhang^{3,5}, Emmanuel Dollinger^{3,4,5,6}, Qinke Peng¹, Qing Nie^{3,4,5,6*},

Xiaohui Xie^{2,5,6*}

1 Systems Engineering Institute, School of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an, Shannxi, 710049, China 2 Department of Computer Science, 3 Department of Mathematics, 4 Department of Developmental and Cell Biology, 5 NSF-Simons Center for Multiscale Cell Fate Research, 6 Center for Complex Biological Systems, University of California, Irvine, CA, 92697, USA

*: co-corresponding authors: <u>qnie@uci.edu</u>, xhx@uci.edu

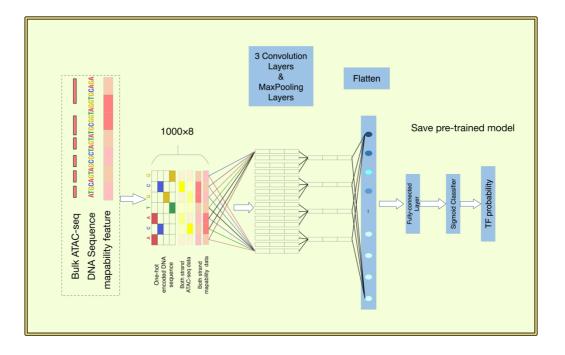


Figure S1: **Deep learning model in scFAN scheme.** The figure shows the overall scheme of the pre-trained model in scFAN. We adopt 3 convolution layers with corresponding Max-pooling layers to extract features, then concatenate the feature map with two fully-connected layers and finally get the output TF probability using Sigmoid function.

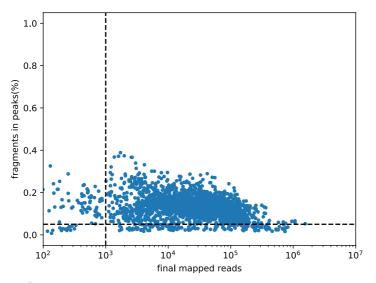


Figure S2: **Single Cell filtering according to read and fragment counts.** The dash lines show the threshold of filtering cells. We set the threshold of the fraction of total read counts per total number of peaks per cell to be 0.05 and also set the threshold of total read counts of each cell to be at least 1000.

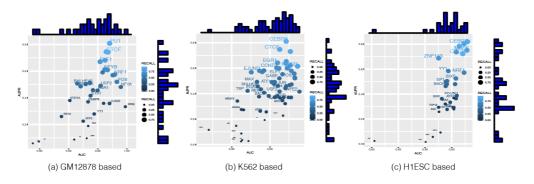


Figure S3: **TF binding prediction performance in the pre-trained model using bulk data.** The figure shows the TF prediction performance (AUC, AUPR, and RECALL values) of scFAN implemented by three cell lines based bulk data. X-axis is the AUC value, Y-axis is the AUPR value and the different RECALL value is shown in the lightness of color and the size of the dot. We also draw the histogram plot of the AUC and AUPR along the x-axis and y-axis.



Figure S4: **Each TF prediction performance of in all the three cell lines.** The figure shows each TF prediction performance(AUC, AUPR, and RECALL values) of three cell lines (GM12878/K562/H1ESC) using scFAN pre-trained model. The higher the bar is, the larger the value is

ANA 17 2800.00 1	Ruler			and the second		5.2			
HIN1772662.bx	SBB1779662 htt		178234000	178235000		178236000	17823700	0	178238000
	3HH1779062.0W								
AR1773678-W 1-9 - <	SRR1779660.bw	0.0]							
AR1773678-W 1-9 - <	SBB1770600 htt								
bit 100 or 33806000 00 33807000 00 33807000 00 0 </td <td>5HH177 5000.0W</td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td>	5HH177 5000.0W								
Link 4Y1 33809000 33809000 33809000 33809000 38910000 HR175866.bv 10 10 10 10 10 10 HR175866.bv 10 10 10 10 10 10 HR175866.bv 10 1731000 1731000 1731000 1731000 HR175866.bv 10 1731000 1731000 1731000 1731000 HR175866.bv 10 10 10 10 10 HR175866.bv 10 1731000 1731000 1731000 1731000 HR175866.bv 10 10 10 10 10 HR175866.bv 10 10 10 10 10 HR175866.bv 10 17300V 17300V 10 10 HR175866.bv 10 10 10 10 10 HR175866.bv 10 10 10 10 10	SRR1779678.bw	1.0							
etci 33800000 33807000 33800000 33800000 RR177560.2w a a a a a a a a a a a a a a a a a a a a a a a a a a <td></td> <td>0.0</td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td>		0.0							
etci 33800000 33807000 33800000 33800000 RR177560.2w a a a a a a a a a a a a a a a a a a a a a a a a a a <td></td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td>									
AR1773662.bw 10 AR1773660.bw 0 AR1773660.bw 1731000 AR1773660.bw 0 AR1773660.bw 17300K AR1773660.bw 10	Ruler								
RR177960.0w Image: set of the set of	SRR1779662.bw		33806000	33807000	33808000		33809000	334	310000
AR177500.bw AR1775		0.0							
AR1172602.bw 40 AR1172602.bw 40 AR1172600.bw 40 AR1172	SRR1779660.bw								
0.0 0.0 1000 17312000 17314000 17315000 173173 1.0 17315000 17314000 17314000 17315000 173173 1.0 0 17315000 17314000 17315000 173173 1.0 0 17315000 17314000 17315000 173173 1.0 0 0 17314000 17315000 173173 1.0 0 0 0 17314000 17315000 173173 1.0 0 0 0 0 0 0 0 1.0 0 0 0 0 0 0 0 1.0 0 0 0 0 0 0 0 0 1.0 0 0 0 0 0 0 0 0 0 1.0 0 0 0 0 0 0 0 0 0 0 1.0 0 0 0 0 0 0 0 0 0 1.0 0 0 0 0 0 0 0 0 0 0 1.0 0 0 0 0 0<	SRR1779600.bw								
0.0 0.0 1000 17312000 17314000 17315000 173173 1.0 17315000 17314000 17314000 17315000 173173 1.0 0 17315000 17314000 17315000 173173 1.0 0 17315000 17314000 17315000 173173 1.0 0 0 17314000 17315000 173173 1.0 0 0 0 17314000 17315000 173173 1.0 0 0 0 0 0 0 0 1.0 0 0 0 0 0 0 0 1.0 0 0 0 0 0 0 0 0 1.0 0 0 0 0 0 0 0 0 0 1.0 0 0 0 0 0 0 0 0 0 0 1.0 0 0 0 0 0 0 0 0 0 1.0 0 0 0 0 0 0 0 0 0 0 1.0 0 0 0 0 0<	0001720678	107							
UMW P36.13 1731000 17315000 17317000 RINT75660.2W 0 17315000 1731000 1731000 RINT75600.2W 0 0 0 0 RINT75600.2W 10 17200K 17300K 17400K RINT75600.2W 10 17200K 17300K 17400K	SHH1779678.Dw								
ev:1000 17312000 17313000 17314000 17315000 17315000 ARI1775800.2wv 0.0 17314000 17314000 17315000 17314000 17315000 ARI1775800.2wv 0.0 17315000 17314000 17314000 17315000 17314000 17315000 ARI1775800.2wv 0.0 10.0		0.0 _							
ev:1000 17312000 17313000 17314000 17315000 17315000 ARI1775800.2wv 0.0 17314000 17314000 17315000 17314000 17315000 ARI1775800.2wv 0.0 17315000 17314000 17314000 17315000 17314000 17315000 ARI1775800.2wv 0.0 10.0	Ruler								
AR1775600.bw 00 00 00 00 00 00 00 00 00 00 00 00 00		chr11000	173120	00 173			14000	17315000	173
NR1779602.0w Image: Second S	SRR1779662.bw	1.0				1.1			
AR11778078.bw 10 00 4V117100K 15 00 4V117100K 17200K 17200K 17200K 17200K 17200K 1740K 1740K 17	SRR1779660.bw	0.0							
AR11778078.bw 10 00 4V117100K 15 00 4V117100K 17200K 17200K 17200K 17200K 17200K 1740K 1740K 17									
0.0	SRR1779600.bw								
International and the first state of the first state	SRR1779678.bw	1.0 7				1.00			
drv117100K 1 17200K 17300K 17400K 1.5 0.0 1 1 1 1 10,0000 Juw 1.9 1.9 1.9 1.9 1.9 1173000 Juw 1.9 1.9 1.9 1.9 1.9 1.9 1173000 Juw 1.9 <td></td> <td>0.0</td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td>		0.0							
drv117100K 1 17200K 17300K 17400K 1.5 0.0 1 1 1 1 10,0000 Juw 1.9 1.9 1.9 1.9 1.9 1173000 Juw 1.9 1.9 1.9 1.9 1.9 1.9 1173000 Juw 1.9 <td></td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td>									
drv117100K 1 17200K 17300K 17400K 1.5 0.0 1 1 1 1 10,0000 Juw 1.9 1.9 1.9 1.9 1.9 1173000 Juw 1.9 1.9 1.9 1.9 1.9 1.9 1173000 Juw 1.9 <td>Ruler</td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td>	Ruler								
RR 1778600.0w		chr117100K		17200K	, p36.			17400K	
RR177600.0w	SRR1779662.bw	1.5							
RH177800.bw	SRR1779660.bw	0.0							
RR1779678.bw 1.0									
	SRR1779600.bw								
	SRR1779678.bw	1.0 7							

Figure S5: **The raw scATAC-seq signal across four H1ESC cells in chromosome 1.** The figure shows the appearance raw signal of scATAC-seq data in the predicted regions (from four different regions in chromosome 1) that are bind with CTCF across four H1ESC cells. Each blue panel (the first row and the last row) indicate one H1ESC cell that shows more active score (more scATAC-seq signal) than the white panels (the second row and the third row), which results in the separation on the clustering of H1ESC cells.

Ruler		1	I		1	1	1 03	8 13 I		1	1	1		1			1
	chr1)K	17260K	17270K	17280K	17290K	17300K	17310K	17320K		7330K	17340K	17350K		17360		17370K	
SRR1779678.bw	1.0																H1ESC cell 5
SRR1779678_impu ted.bw	0.0			1.111	l di di su		1			d at a		1	I	L L	T.		H1ESC cell 5 imputed
SRR1779600.bw																	H1ESC cell 82
SRR1779600_impu ted.bw	0.0			1.111			al.			1 II II		1	1	П	T		H1ESC cell 82 imputed
SRR1779660.bw																	H1ESC cell 22
SRR1779660_impu ted.bw	^{2.0}						1			d at a		1	1	П	T		H1ESC cell 22 imputed
SRR1779662.bw	1.0 0.0																H1ESC cell 21
SRR1779662_impu ted.bw	0.0			1.111	t ni ni tra	n n daa	al.		11	d at a		1	1	L L	T.		H1ESC cell 21 imputed

Figure S6: **The imputed ATAC-seq and scATAC-seq signal from the same regions.** Four cells are chosen. The data in blue panel is raw scATAC-seq data and the white panel is the imputed scATAC-seq data. After the imputation, those regions that are originally missing some signal were recovered, which lets the model able to predict TF in these regions more accurately and further clusters these cells together.

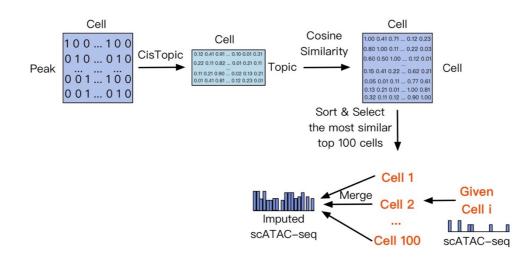


Figure S7: **The imputation procedure of scATAC-seq data.** The figure shows the procedure of implementing scATAC-seq imputation using binarized peak-cell count matrix. We used cisTopic to reduce the original high-dimensional matrix into low-dimensional latent space(topic-cell) matrix, which can be further used to calculate similarity matrix between each cell. Given a single cell i, we sorted the similarity matrix and select the most similar 100 cells to cell i, and merge their signal into one file using bigWigMerge tool to get the final imputed scATAC-seq data with respect to cell i.

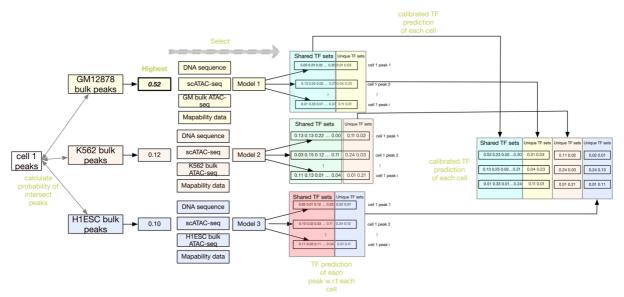


Figure S8: **The calibration of TF binding prediction in one single cell.** The figure shows the calibration of TF binding prediction from three different models. All the peaks in one single cell were intersected with bulk peaks from three cell line. The probability of the intersecting number in single cell was calculated and the highest one was selected as the most likely matched cell line and chosen as the representation of all the shared TFs by three models. All the unique TFs in three models were kept separately. After that, the calibrated TF prediction of each cell was obtained.

Table S1: **Partial summay of scFAN cross-cell performance on bulk data.** From the table we could see that our model could capture different TFs from different cell lines.

TF name	Cell line	AUC	AUPR	RECALL
CTCF	GM12878	0.968	0.742	0.919
ELK1	GM12878	0.954	0.670	0.765
PU1	GM12878	0.970	0.808	0.911
RAD21	GM12878	0.965	0.745	0.912
SMC3	GM12878	0.963	0.682	0.845
CTCF	K562	0.963	0.748	0.855
CEBPB	K562	0.971	0.811	0.904
MAFF	K562	0.977	0.731	0.848
RAD21	K562	0.986	0.615	0.829
EGR1	K562	0.960	0.642	0.762
JUND	H1ESC	0.941	0.479	0.463
CEBPB	H1ESC	0.973	0.804	0.865
CTCF	H1ESC	0.971	0.723	0.937
MAFK	H1ESC	0.978	0.728	0.840
USF1	H1ESC	0.979	0.820	0.925