Supplementary Information

Long-read single-cell RNA sequencing analysis of key genes and isoforms during corneal wound healing in cynomolgus monkeys

Hu Chen[1,2,#], Qi Dai[2,#,*], Ming Zhou[1,#], Yuan-Xia Yang[1], Zhi-Bo Huang[3], Shi-Da Chen[1], Ying-Feng Zheng[1,*] and Zhuo-Xing Shi[1,*]

[1] *State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangdong Provincial Key Laboratory of Ophthalmology and Visual Science, Guangdong Provincial Clinical Research Center for Ocular Diseases, Guangzhou 510060, China*

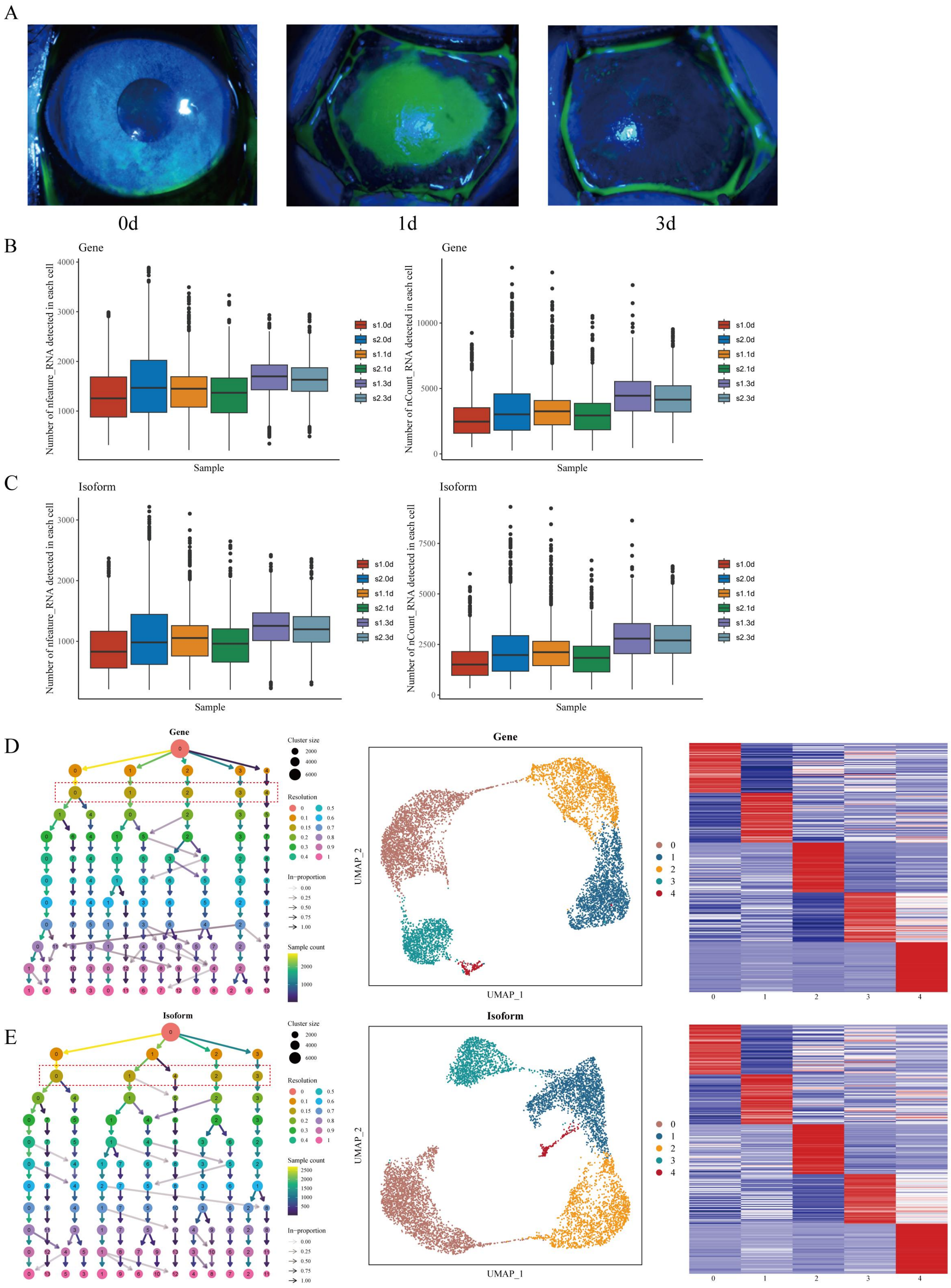
[2] *College of Life Science and medicine, Zhejiang Sci-Tech University, Hangzhou 310018, China*

[3] *South-Central Minzu University, Wuhan 430074, China*

[#] These authors equally contribute to this work.

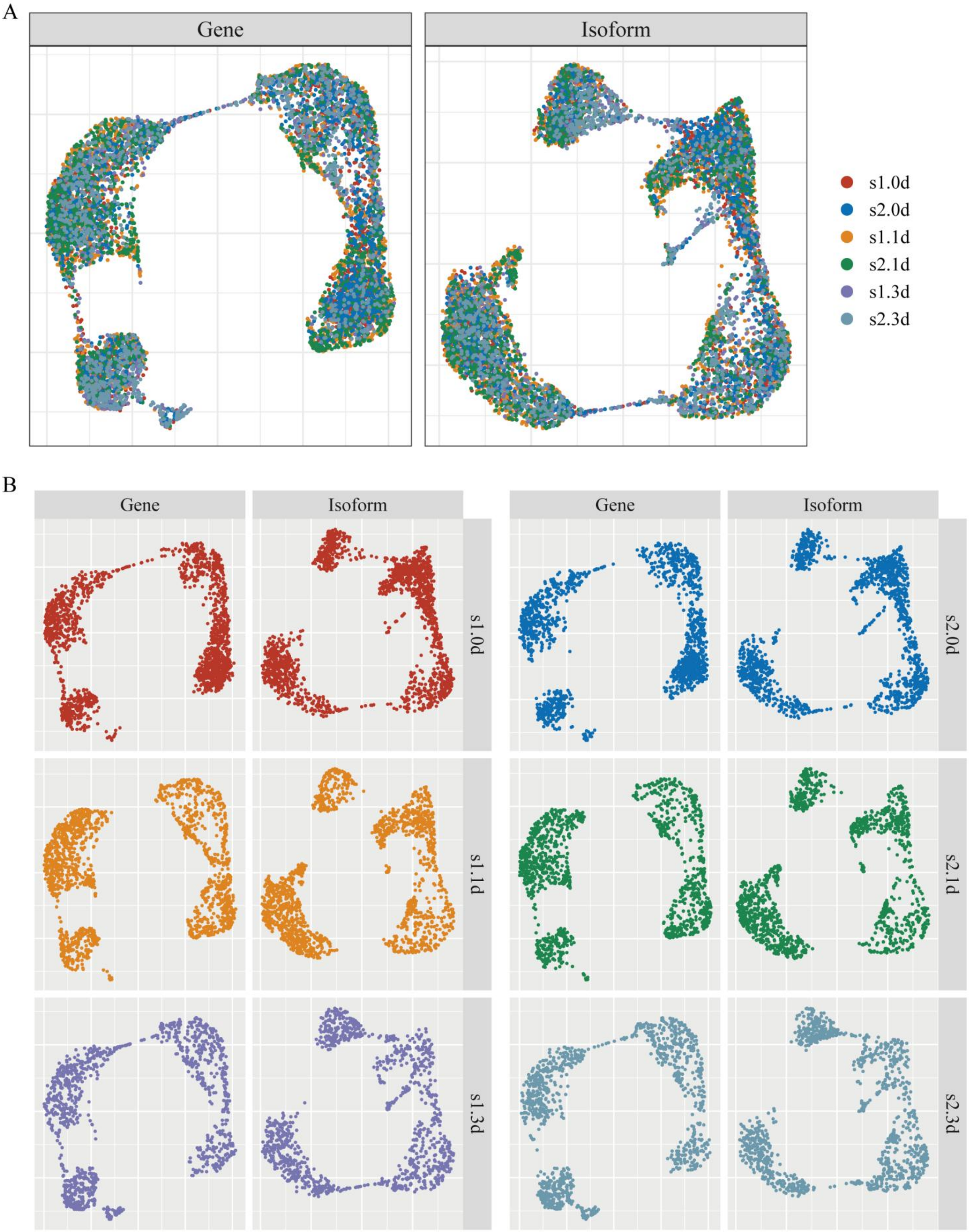
[*] Corresponding author: daiqi@zstu.edu.cn (Q. Dai), zhyfeng@mail.sysu.edu.cn (Y.-F Zheng), and zhuoxing1988@163.com (Z.-X Shi)

Table of Contents

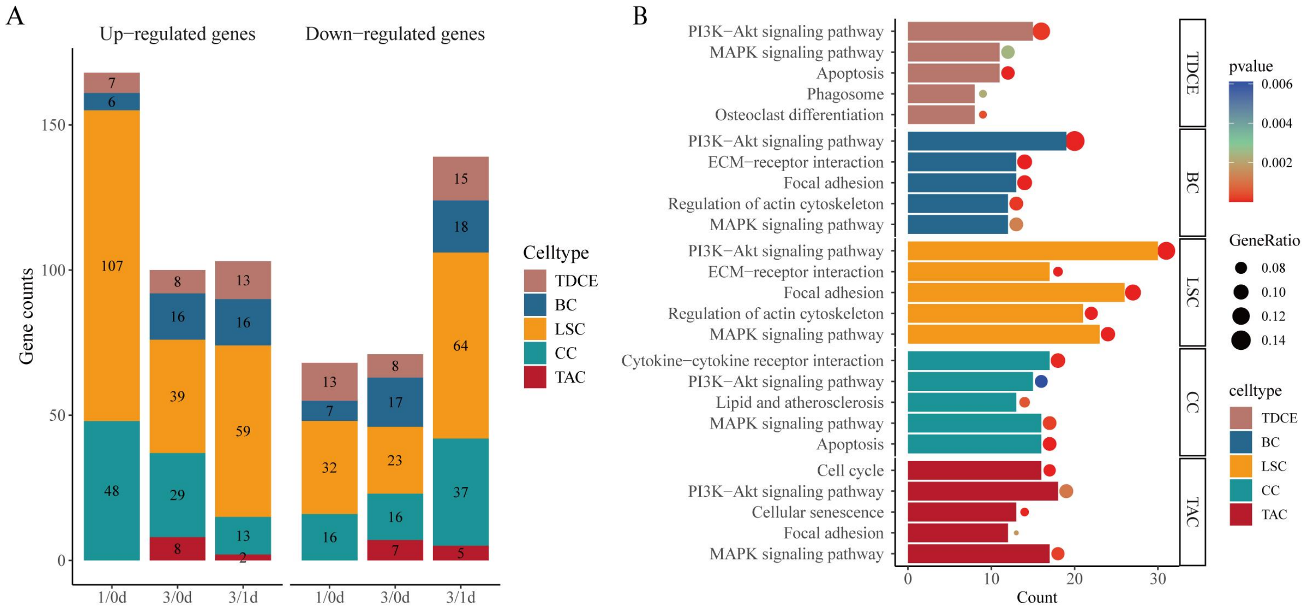


Supplementary Fig. S1: Quality control and cell annotation. (A). Corneal wound healing model. (B-C). Boxplots showing the distribution of nFeature_RNA and nCount_RNA across all cells in each sample, both at the gene (B) and isoform (C) levels. D-E. At the gene level (D) and isoform level (E), clustering trees (left) show clusterings at different resolutions; UMAP plots (middle) display clustering results; and heatmaps (right) illustrate the expression of the top 50 features in each cluster.

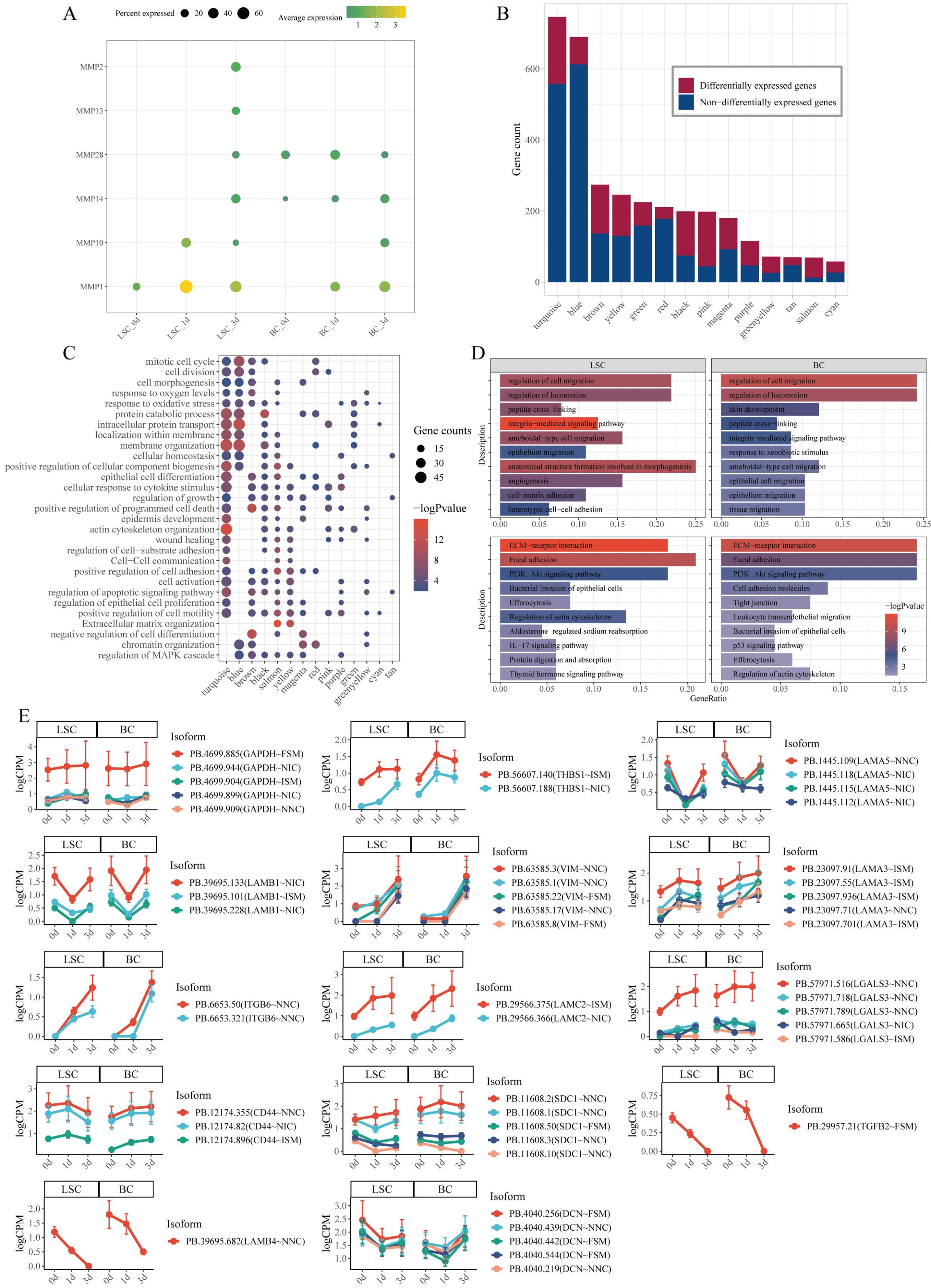
Supplementary Fig. S2: Feature plot of marker genes. (A-B). At the gene level (A) and isoform level (B), expression levels of well-known representative cell markers across clusters.



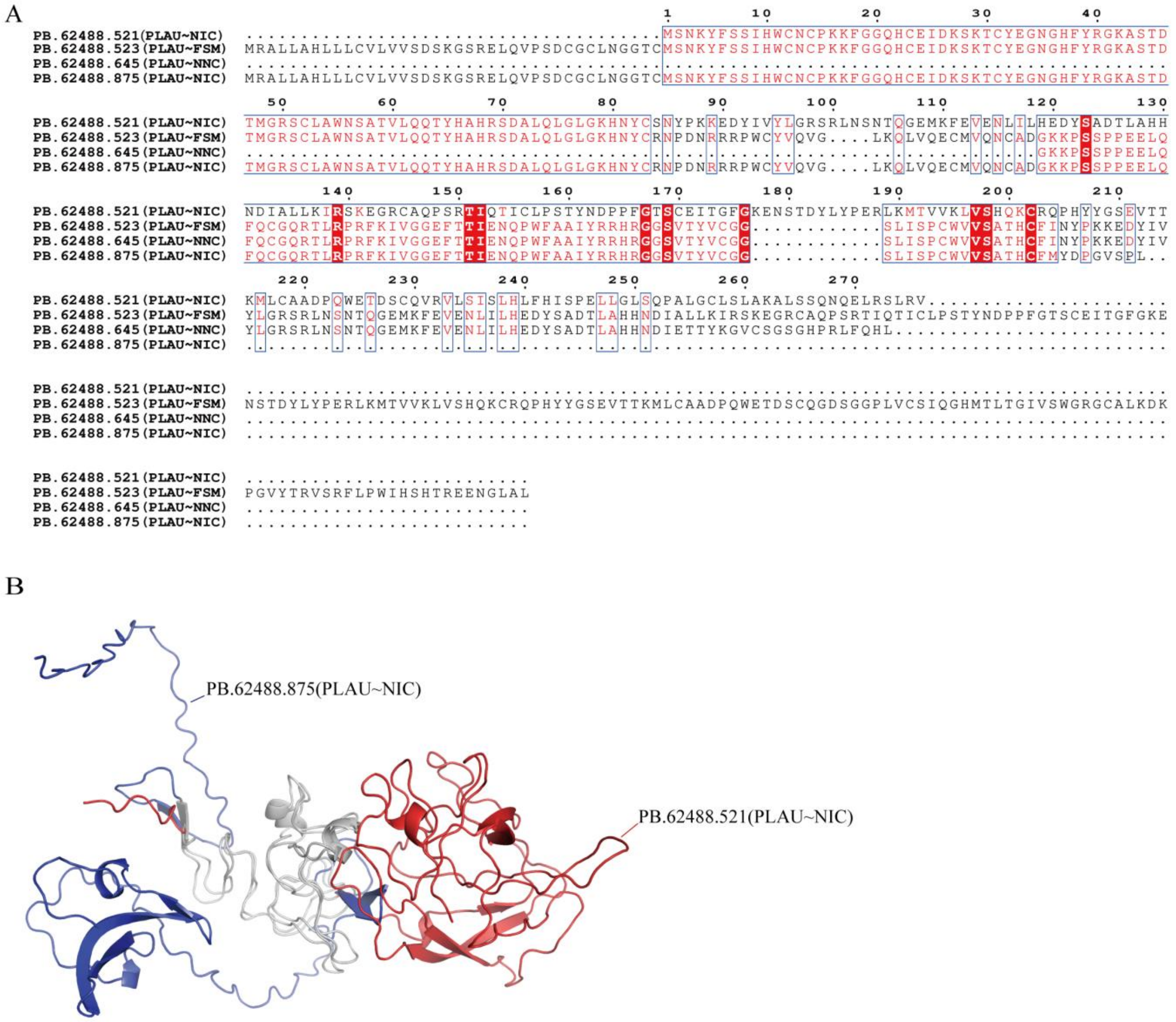
Supplementary Fig. S3: UMAP Visualization of Batch Effects. (A-B). UMAP plots showing the distribution of cells from different samples.



Supplementary Fig. S4: Differentially Expressed Genes. (A). Distribution of differentially expressed genes between pairs of time points across different cell types. (B). KEGG pathway enrichment analysis of differentially expressed genes across various cell types.

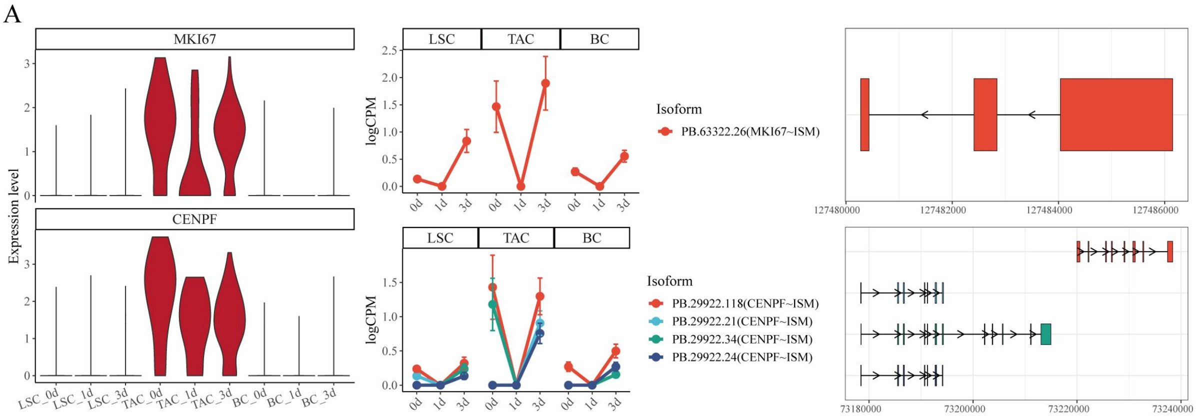


Supplementary Fig. S5: Cell migration-related genes. (A). Expression levels of the MMP family at three time points in LSC and BC. (B). Distribution of differentially expressed genes across various gene modules. (C). GO term enrichment analysis of different gene modules. (D). GO term enrichment analysis of functional genes related to cell migration in LSC and BC. (E). Expression trends of corresponding isoforms of cell migration-related genes.

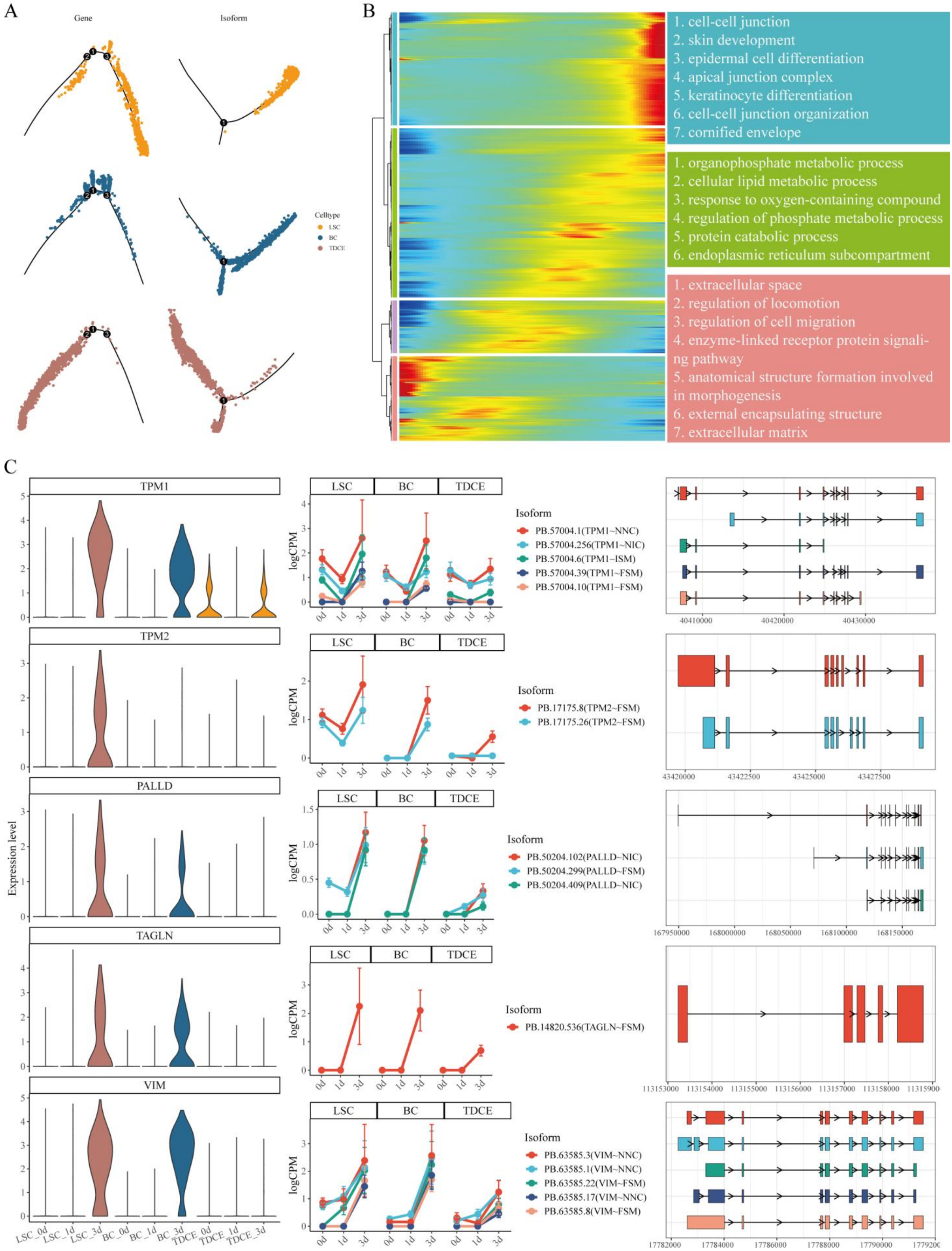


```
                                                             1       10        20        30        40
PB.62488.521(PLAU~NIC) ......................................MSNKYFSSIHWCNCPKKFGGQHCEIDKSKTCYEGNGHFYRGKASTD
PB.62488.523(PLAU~FSM) MRALLAHLLLCVLVVSDSKGSRELQVPSDCGCLNGGTCMSNKYFSSIHWCNCPKKFGGQHCEIDKSKTCYEGNGHFYRGKASTD
PB.62488.645(PLAU~NNC) ....................................................................................
PB.62488.875(PLAU~NIC) MRALLAHLLLCVLVVSDSKGSRELQVPSDCGCLNGGTCMSNKYFSSIHWCNCPKKFGGQHCEIDKSKTCYEGNGHFYRGKASTD

                         50        60        70        80        90       100       110       120       130
PB.62488.521(PLAU~NIC) TMGRSCLAWNSATVLQQTYHAHRSDALQLGLGKHNYCSNYPKKEDYIVYLGRSRLNSNTQGEMKFEVENLILHEDYSADTLAHH
PB.62488.523(PLAU~FSM) TMGRSCLAWNSATVLQQTYHAHRSDALQLGLGKHNYCRNPDNRRRPWCYVQVG....LKQLVQECMVQNCADGKKPSSPPEELQ
PB.62488.645(PLAU~NNC) ........................................................................GKKPSSPPEELQ
PB.62488.875(PLAU~NIC) TMGRSCLAWNSATVLQQTYHAHRSDALQLGLGKHNYCRNPDNRRRPWCYVQVG....LKQLVQECMVQNCADGKKPSSPPEELQ

                              140       150       160       170       180       190       200       210
PB.62488.521(PLAU~NIC) NDIALLKIRSKEGRCAQPSRTIQTICLPSTYNDPPFGTSCEITGFGKENSTDYLYPERLKMTVVKLVSHQKCRQPHYYGSEVTT
PB.62488.523(PLAU~FSM) FQCGQRTLRPRFKIVGGEFTTIENQPWFAAIYRRHRGGSVTYVCGG............SLISPCWVVSATHCFINYPKKEDYIV
PB.62488.645(PLAU~NNC) FQCGQRTLRPRFKIVGGEFTTIENQPWFAAIYRRHRGGSVTYVCGG............SLISPCWVVSATHCFINYPKKEDYIV
PB.62488.875(PLAU~NIC) FQCGQRTLRPRFKIVGGEFTTIENQPWFAAIYRRHRGGSVTYVCGG............SLISPCWVVSATHCFMYDPGVSPL..

                          220       230       240       250       260       270
PB.62488.521(PLAU~NIC) KMLCAADPQWETDSCQVRVLSISLHLFHISPELLGLSQPALGCLSLAKALSSQNQELRSLRV......................
PB.62488.523(PLAU~FSM) YLGRSRLNSNTQGEMKFEVENLILHEDYSADTLAHHNDIALLKIRSKEGRCAQPSRTIQTICLPSTYNDPPFGTSCEITGFGKE
PB.62488.645(PLAU~NNC) YLGRSRLNSNTQGEMKFEVENLILHEDYSADTLAHHNDIETTYKGVCSGSGHPRLFQHL.........................
PB.62488.875(PLAU~NIC) ....................................................................................

PB.62488.521(PLAU~NIC) ....................................................................................
PB.62488.523(PLAU~FSM) NSTDYLYPERLKMTVVKLVSHQKCRQPHYYGSEVTTKMLCAADPQWETDSCQGDSGGPLVCSIQGHMTLTGIVSWGRGCALKDK
PB.62488.645(PLAU~NNC) ....................................................................................
PB.62488.875(PLAU~NIC) ....................................................................................

PB.62488.521(PLAU~NIC) ..........................
PB.62488.523(PLAU~FSM) PGVYTRVSRFLPWIHSHTREENGLAL
PB.62488.645(PLAU~NNC) ..........................
PB.62488.875(PLAU~NIC) ..........................
```

Supplementary Fig. S6: PLAU Sequence and Protein Structure. (A). Multiple sequence alignment of the protein sequences corresponding to *PLAU* isoforms. (B). Protein structures of PB.62488.521(*PLAU*~NIC) and PB.62488.875(*PLAU*~NIC).



Supplementary Fig. S7: Cell proliferation-related genes. MKI67 and CENPF expression violin plots (left), isoform expression trend fold plots (middle), and isoform structure plots (right).



Supplementary Fig. S8: Cell differentiation-related genes. (A). Distribution of different cell types along the differentiation trajectory at both gene and isoform levels. (B). At the isoform level, heatmap showing the expression of differentially expressed genes (vertical columns) sorted by pseudotime (horizontal columns) that may be key genes driving cell differentiation. The right panel displays GO annotation entries for the corresponding clusters. (C). Expression violin plots of cell differentiation-related genes (left), isoform expression trend fold plots (middle), and isoform structure plots (right).

	0d_s1	**0d_s2**	**1d_s1**	**1d_s2**	**3d_s1**	**3d_s2**
CCS Reads	3427900	4226969	4609698	4389883	4684184	4190376
CCS Yield (GB)	16.75	21.62	21.08	19.01	21.3	17.74
CCS MaxLen	27647	27379	27073	27435	28089	26896
CCS MeanLen	4887.5	5114.62	4573.57	4330.86	4546.41	4233.92
CCS N50Len	5729	6065	5498	5291	5502	5094
CCS Mean Passes	23	20	12	13	12	12
CCS MeanQV	0.95	0.95	0.93	0.93	0.92	0.93

Supplementary Table 1. Quality control reports for CCS reads from the HIT-scISOseq dataset. CCS Reads : total CCS reads count of each sample; CCS Yield (GB): sum of all CCS reads bases (giga base) in each sample; CCS MaxLen (bp): the max read length of all CCS reads; CCS MeanLen (bp): the mean read length of all CCS reads; CCS N50Len (bp): the N50 read length of all CCS reads; CCS Mean Passes: the mean passes of all CCS reads; Avg. CCS MeanQV: the mean QV of all CCS reads.

	0d_s1	**0d_s2**	**1d_s1**	**1d_s2**	**3d_s1**	**3d_s2**
All Paire	11637865	14837342	13994773	10819299	12889308	11974218
FLNC	10472456	13227218	12550076	9631647	11464563	10807930
NFL	283104	394147	389367	317775	380252	318033
Unknow	882305	1215977	1055330	869877	1044493	848255
FLNC(%)	89.99	89.15	89.68	89.02	88.95	90.26
NFL(%)	2.43	2.66	2.78	2.94	2.95	2.66
Unknow(%)	7.58	8.20	7.54	8.04	8.10	7.08

Supplementary Table 2. Quality control reports for FLNC reads from the HIT-scISOseq dataset. All Paire: total linked cDNAs (which is defined as linked cDNA in each CCS reads) count in each sample; FLNC: total full-length non-concatemer (FLNC) reads count in each sample; NFL: total non-full length (NFL) reads count in each sample; Unknow: total artifact cDNA count in each sample; FLNC(%): percentage of FLNC in linked cDNAs of each sample; NFL(%): percentage of NFL in linked cDNAs of each sample; Unknow(%): the percentage of artifact cDNAs in linked cDNAs of each sample.

	0d_s1	**0d_s2**	**1d_s1**	**1d_s2**	**3d_s1**	**3d_s2**
Reads Count	10472456	13227218	12550076	9631647	11464563	10807930
Mapped	10341384	13048692	12374890	9514271	11302113	10681176
Mapped(%)	98.75	98.65	98.60	98.78	98.58	98.83
Uniq Mapped	9959307	12578311	12121594	9305690	11102031	10496253
Uniq Mapped(%)	95.10	95.09	96.59	96.62	96.84	97.12
UnMapped	131072	178526	175186	117376	162450	126754
UnMapped(%)	1.25	1.35	1.40	1.22	1.42	1.17
MutilMapped	382077	470381	253296	208581	200082	184923
MutilMapped(%)	3.65	3.56	2.02	2.17	1.75	1.71
MeanMapCoverage(%)	98.90	98.83	98.74	98.80	98.70	98.81
MeanMapIdentity(%)	97.74	97.59	96.75	96.69	96.62	96.80

Supplementary Table 3. Quality control reports for genome mapping of FLNC reads from the HIT-scISOseq dataset. Reads Count: Total count of full-length non-concatemer (FLNC) reads for each sample; Mapped: Total count of FLNC reads mapped to the genome for each sample; Mapped(%): Percentage of genome-mapped FLNC reads in each sample; UniqMapped: Total count of unique genome-mapped FLNC reads for each sample; UniqMapped(%): Percentage of unique genome-mapped FLNC reads in each sample; UnMapped: Total count of unmapped FLNC reads for each sample; UnMapped(%): Percentage of unmapped FLNC reads in each sample; MutilMapped: Total count of FLNC reads mapped to multiple locations in the genome for each sample; MutilMapped(%): Percentage of multiple mapped FLNC reads in each sample; MeanMapCoverage(%): Average mapping coverage of uniquely mapped FLNC reads in each sample; MeanMapIdentity(%): Average mapping identity of uniquely mapped FLNC reads in each sample.

			0d_s1	**0d_s2**	**1d_s1**	**1d_s2**	**3d_s1**	**3d_s2**
cellBC	All	Total FLNC	10472456	13227218	12550076	9631647	11464563	10807930
		CB in Whitelist	8823984	11051565	9739201	7436281	8736796	8327534
		CB in Whitelist(%)	84.26	83.55	77.60	77.21	76.21	77.05
		CB Correction	248384	324201	492090	390769	480879	454075
		CB Correction(%)	2.37	2.45	3.92	4.06	4.19	4.20
		Total Corrected CB	9072368	11375766	10231291	7827050	9217675	8781609
		Total Corrected CB(%)	86.63	86.00	81.52	81.26	80.40	81.25
	QV>=0.95(Passed)	Total FLNC	9375747	11729087	10194383	7831481	9222694	8726993
		CB in Whitelist	8648780	10820554	9363001	7150056	8396439	7982488
		CB in Whitelist(%)	92.25	92.25	91.84	91.30	91.04	91.47
		CB Correction	125123	159534	211427	176266	215593	195292
		CB Correction(%)	1.33	1.36	2.07	2.25	2.34	2.24
		Total Corrected CB	8773903	10980088	9574428	7326322	8612032	8177780
		Total Corrected CB(%)	93.58	93.61	93.92	93.55	93.38	93.71
UMI	cellBC Passed && UMI Passed	Total FLNC	8773903	10980088	9574428	7326322	8612032	8177780
		UMI Discarded	19011	30719	29564	21077	20503	18787
		UMI Discarded(%)	0.22	0.28	0.31	0.29	0.24	0.23
		UMI Uncorrected	8743058	10931494	9525729	7289218	8579686	8148661
		UMI Uncorrected(%)	99.65	99.56	99.49	99.49	99.62	99.64
		UMI Corrected	11834	17875	19135	16027	11843	10332
		UMI Corrected(%)	0.13	0.16	0.20	0.22	0.14	0.13
		Total Passed UMI	8754892	10949369	9544864	7305245	8591529	8158993
		Total Passed UMI(%)	99.78	99.72	99.69	99.71	99.76	99.77

Supplementary Table 4. Reports on cell barcode and UMI corrections from the HIT-scISOseq dataset. Total FLNC: number of Full-Length Non-Concatemer reads; For cell barcode correction, “All” means CB have no QV filter, “CB QV>=0.95 (passed)” means have QV filter conditions before CB correction. CB in whitelist: cell barcode can be found directly in the 10× Genomics barcode whitelists; CB in whitelist(%): the percentage of “CB in whitelist” in FLNC read count; CB Corrected: cell barcode have 1- Hamming distance with whitelists; CB Corrected(%): the percentage of “Corrected CB” in FLNC read count; Total Corrected CB: sum of the “CB in whitelist” and “Corrected CB”; Total Corrected CB(%): the percentage of “Total number of corrected CB” in FLNC read count; For UMI correction, “CB passed and UMI passed” means have QV filter conditions before UMI correction that CB QV must >=0.95 and have UMI filter conditions after correction that UMI must not “Discarded UMI”; UMI Discarded: means UMI not passed basic quality filtering; UMI Discarded(%): the percentage of “Discarded UMI” in FLNC read count; UMI Uncorrected: means UMI have passed the basic quality filtering but no need correction; UMI Uncorrected(%): the percentage of “Uncorrected UMI” in FLNC read count; UMI Corrected: meet the UMI correction conditions and have correction; UMI Corrected(%): the percentage of “Corrected UMI” in FLNC read count; Total Passed UMI: sum of the “Uncorrected UMI” and “Corrected UMI”; Total Passed UMI(%): the percentage of “Total number of passed UMI” in FLNC read count.

	0d_s1	**1d_s1**	**3d_s1**	**0d_s2**	**1d_s2**	**3d_s2**
Nunber of cell	1565	1482	1096	1417	1267	1068
Mean UMI per cell	2703.78	3325.31	4465.49	3393.07	2946.68	4292.30
Median UMI per cell	2462	3258	4447	3023	2931	4145
Mean genes per cell	1320.34	1409.17	1677.16	1544.64	1312.69	1634.21
Median genes per cell	1255	1451	1699	1467	1368	1631.5
Total genes detected	13077	12603	13037	13370	12185	12935

Supplementary Table 5. The gene and UMI count in HIT-scISOseq dataset. Number of cell: total number of cell in gene matrix; Mean UMI count per cell: mean UMI detection of each cell in gene matrix; Median UMI count per cell: median UMI detection of each cell in gene matrix; Mean gene per cell: mean of gene detection of each cell in gene matrix; Median gene per cell: median of gene detection of each cell in gene matrix; Total detected: sum of gene detection of all cell in gene matrix.

	0d_s1	**1d_s1**	**3d_s1**	**0d_s2**	**1d_s2**	**3d_s2**
Nunber of cell	1565	1482	1096	1417	1267	1068
Mean UMI per cell	1655.591	2157.602	2834.746	2189.615	1824.537	2822.137
Median UMI per cell	1502	2106.5	2795.5	1970	1800	2704.5
Mean isoforms per cell	892.1348	1030.419	1245.797	1069.456	928.2557	1207.896
Median isoforms per cell	829	1054	1259.5	982	947	1204.5
Total isoforms detected	29379	31338	29726	31780	26973	28786

Supplementary Table 6. The isoform and UMI count in HIT-scISOseq dataset. Number of cell: total number of cell in isoform matrix; Mean UMI count per cell: mean UMI detection of each cell in isoform matrix; Median UMI count per cell: median UMI detection of each cell in isoform matrix; Mean isoform per cell: mean type of isoform detection of each cell in isoform matrix; Median isoform per cell: median type of isoform detection of each cell in isoform matrix; Total isoforms detected: sum of isoform type detection of all cell in isoform matrix.