

1 **scCAT-seq: single-cell identification and quantification of mRNA isoforms by**
2 **cost-effective short-read sequencing of cap and tail**

3

4 Youjin Hu^{1,#,†}, Jiawei Zhong^{1#}, Yuhua Xiao¹, Zheng Xing³, Katherine Sheu⁴, Shuxin
5 Fan¹, Qin An², Yuanhui Qiu¹, Yingfeng Zheng¹, Xialin Liu¹, Guoping Fan², Yizhi Liu¹,

6 †

7 ¹ State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-Sen University,
8 Guangzhou, China

9 ² Department of Human Genetics, David Geffen School of Medicine, UCLA, Los Angeles, CA,
10 USA

11 ³ Earth, Planetary and Space Sciences, UCLA, Los Angeles, CA, USA

12 ⁴ Medical Scientist Training Program, David Geffen School of Medicine, UCLA, Los Angeles, CA,
13 USA

14 # These authors contributed equally to this work.

15 † Correspondence to Yizhi Liu (yzliu62@yahoo.com) or Youjin Hu (huyoujin@gzzoc.com).

16

17 **Abstract**

18 The differences in transcription start sites (TSS) and transcription end sites (TES) among gene isoforms
19 can affect the stability, localization, and translation efficiency of mRNA. Isoforms also allow a single
20 gene different functions across various tissues and cells. However, methods for efficient genome-wide
21 identification and quantification of RNA isoforms in single cells are still lacking. Here, we introduce
22 single cell Cap And Tail sequencing (scCAT-seq). In conjunction with a novel machine learning
23 algorithm developed for TSS/TES characterization, scCAT-seq can demarcate transcript boundaries of
24 RNA transcripts, providing an unprecedented way to identify and quantify single-cell full-length RNA
25 isoforms based on short-read sequencing. Compared with existing long-read sequencing methods,
26 scCAT-seq has higher efficiency with lower cost. Using scCAT-seq, we identified hundreds of
27 previously uncharacterized full-length transcripts and thousands of alternative transcripts for known
28 genes, quantitatively revealed cell-type specific isoforms with alternative TSSs/TEs in dorsal root
29 ganglion (DRG) neurons, mature oocytes and ageing oocytes, and generated the first atlas of the
30 non-human primate cornea. The approach described here can be widely adapted to other short-read or

31 long-read methods to improve accuracy and efficiency in assessing RNA isoform dynamics among
32 single cells.

33

34 **Background**

35 The extent of cellular heterogeneity across different tissues and cell types has become
36 increasingly apparent due to the development of genomics technology, especially
37 single-cell omics sequencing (1-3). With the launch of initiatives such as the human
38 single-cell atlas (4, 5), increased attention has been given to the regulatory
39 mechanisms of cell-specific gene transcription, including both transcript abundance
40 and alternative isoform usage, which can result in distinct protein sequences and
41 structures (6, 7). RNA isoform variability includes intron inclusion, exon skipping,
42 and alternative choice of transcription start sites (TSSs) (8) and transcription end sites
43 (TESs) (9, 10). Alternative TSSs and TESs account for the majority of tissue-specific
44 exon usage, are considered the principal drivers of transcript isoform diversity across
45 tissues, and underlie the majority of isoform-mediated, cell-type specific proteomes
46 (11). In addition, alternative TSS choices in the 5'-UTR, as well as alternative
47 polyadenylation (APA) in the 3'-UTR regions play key roles in mRNA stability,
48 translation, localization (9, 10, 12-14).

49 Previous studies have demonstrated the widespread heterogeneity of transcript
50 isoforms with alternative 5'-TSS or 3'-APA across different cell types, resulting in the
51 discovery of new transcripts with tissue- or cell-type specificity, and allowing updates
52 to transcript annotations of reference genomes (13, 15). Despite considerable success
53 in measurements made on bulk populations, current approaches for identifying RNA
54 isoforms and the dynamics of TSS/TEs choices in single cells are limited.
55 Fundamentally, there is currently no method for accurate, efficient, and quantitative
56 analysis of RNA isoforms of single cells genome-wide. Most single-cell transcriptome
57 approaches are based on single-ended quantification of RNA molecules (5' or 3')
58 which give partial information on one end but not the whole transcript (3, 16, 17),
59 resulting in loss of important information about the other end, especially for
60 transcripts regulated by UTR regions on both ends (13). Methods based on single-cell

61 full-length cDNA amplification such as Smart-seq2 can detect the full-length cDNA,
62 but its coverage at both ends is low, and it is not possible to accurately distinguish the
63 start and end positions of different transcript isoforms of the same gene (18, 19).
64 Recently, approaches based on long-read RNA sequencing technologies identified
65 RNA isoforms of thousands of cells, but challenges still remain. For example, the
66 sequencing depth needed to quantitatively assess the RNA isoform transcriptome
67 makes long-read sequencing too expensive, and the conventional approach has been
68 to first catalog isoforms using the long reads and then map short reads to the resulting
69 transcriptome references for quantification. In addition, the requirement of several
70 micrograms of cDNA input requires extensive PCR amplification from picograms of
71 mRNA of a single cell, which unavoidably results in higher PCR bias towards specific
72 isoforms.(13, 15, 20).

73 In order to address these problems, we developed a simple and efficient approach
74 based on well-established short-read sequencing platforms to explicitly exploit
75 transcription initiation and termination sites for the quantification of RNA isoforms in
76 single cells. When deployed in conjunction with optimized machine learning models,
77 scCAT-seq is more accurate and cost-effective, and has higher efficiency than existing
78 methods, making it suitable for quantitative and qualitative analysis of isoform
79 transcriptomes of single cells, and for analysis of RNA isoform dynamics in different
80 biological contexts.

81

82 **Results**

83 To develop scCAT-seq, we adopted a strategy to capture the boundaries of transcripts
84 at both 5' and 3' ends (21). Full-length cDNAs were first tagged with specific
85 sequences adjacent to both TSSs and TESs and further amplified, based on a modified
86 Smart-seq2 protocol (19). Segments of transcript ends with sequence tags were then
87 tagmented out by Tn5 transposases and captured by targeted PCR amplification.
88 Illumina sequencing adaptors were further added for standard Illumina sequencing. As
89 expected, the reads with tags are distributed at the terminal sides of transcripts
90 (**Supplementary Fig. 1a, b**). The analysis pipeline precisely determined TSSs by the

91 mapped position of reads with a head tag, along with the “GGG” signal added during
92 reverse transcription. TESs were determined from paired reads (R1 containing a tail
93 tag and R2 covering polyA sites) by mapping R2 sequences near polyA sites to the
94 genome (**Fig. 1a**). Peaks were called using the CAGEr package (22). Internal TES
95 peaks derived from the internal priming during reverse transcription of mRNA were
96 excluded.

97 To improve the accuracy in identification of real TSSs/TESs, we decided to employ
98 machine learning models. Based on the read distribution of scCAT-seq and
99 Smart-seq2 of the same single cell samples, we collected potential features that could
100 affect the identification of a peak as a TSS or TES peak (**Table 1**), and implemented
101 three widely used machine learning models: logistic regression classifier (LR),
102 random forest (RF), and support vector machine (SVM). The random forest model
103 indicated that “Slope_smart2_curve” and “Percentage” were the most important
104 features, while the logistic regression classifier and SVM put the highest weights on
105 “TPM_of_Dominant_Site” and “Trend_of_smart2_read_counts” (**Supplementary**
106 **Fig. 1d, e**). To derive the best predictions, we chose an ensemble learning strategy of
107 majority voting, integrating predictions from all models to systematically determine
108 the real TSSs/TESs (**Fig. 1b**). Our strategy resulted in perfect performance on an
109 independent test dataset of ERCC spike-in in single cells, with the true positive rate
110 improved by 3.7- (27% versus 100%) and 2.2-fold (43% versus 95%) for TSS and
111 TES, respectively (**Fig. 1c, d**), with sequencing depth of 4 million reads per sample
112 (**Supplementary Table 1**). Similarly, ERCC data from other methods, such as C1
113 CAGE (17), C1 STRT (23), and a method developed by Arguel et al. (24), also
114 showed high false positive rates for peaks identified as TSSs, (**Supplementary Fig.**
115 **1c**), but the accuracy was also improved to above 95% after using our machine
116 learning model (**Supplementary Fig. 1f**), indicating that our model can be applied to
117 other data sets that contain high false positive rates.

118

119 Table 1. Features used in the machine learning models

	Features	Description of the features
x ₁	TPM_of_peak	The total TPM value of the peak called by CAGEr.
x ₂	TPM_of_Dominant_Site	The highest TPM value of all sites within a peak.
x ₃	Gene_length	The length of the transcript annotated.
x ₄	Peak_width	The width of the peak called by CAGEr.
x ₅	Dominant_TPM_to_Smart2	The ratio of Dominant_TPM to the RPM value of the corresponding gene revealed by Smart-seq2.
x ₆	Slope_smart2_curve	The slope of Smart-seq2 coverage curve around the peaks
x ₇	Trend_of_smart2_reads	Calculated by dividing the number of reads increased/decreased within 50bp distance by 50
x ₈	Percentage	The percentage of read counts of a peak to the total counts of a transcript

120 Using the sequencing data from mouse dorsal root ganglion (DRG) neurons for
121 further benchmarking, we sequenced 18 DRG neurons with a mean of 2.4 million
122 reads per cell (**Supplementary Table 2**). As genomic sequence features can specify
123 the locations of TSSs and TESs, in addition to the eight features of read distribution,
124 we added an additional 650 and 150 features of motifs related to TESs and TSSs. To
125 train the TSS machine learning model, we used the data of neuron tissues from the
126 FANTOM5 database (25), and to train the TES model, we used the mouse polyA sites
127 peak from PolyA_DB (26). Using these databases, with 70% of the data for training
128 and 30% for testing, we found the prediction accuracy for TSS and TES to be 94.3%
129 and 94.2% respectively (**Supplementary Fig. 1g**). In total, after pooling all 18 cells
130 together and applying the machine learning model, we identified 11991 and 15481
131 peaks as TSSs and TESs, which were significantly enriched at annotated TSS and
132 TES regions, respectively (**Fig. 1e**). Over 93% of identified TSSs were located within
133 1 kb of annotated TSSs, and over 86% of identified TESs were within 1kb of
134 annotated TESs (**Supplementary Fig. 1h, i**). In summary, our results indicate that

135 scCAT-seq together with a machine learning model can identify TSSs and TESs of
136 transcripts with high accuracy, allowing demarcation of transcription boundaries of
137 full length isoforms.

138 Furthermore, we compared detected read counts with the known abundances of ERCC
139 mRNA molecules to assess quantification performance. The measured abundances
140 were highly concordant with the ground truth, with a Pearson's correlation coefficient
141 of 0.98 for both TSS and TES (**Fig. 1f, Supplementary Fig. 2a**). For the annotated
142 genes of the mouse genome, an internal comparison between random pools of 3 single
143 cells, each from the oocyte population, gave a correlation coefficient of 0.96 and 0.94
144 for the quantification of TSS and TES, respectively (**Fig. 1g, Supplementary Fig. 2b**).
145 Thus, the quantification of TSS and TES is reliable and provides an accurate and
146 reproducible measure of relative expression of transcript isoforms.

147 The sensitivity and efficiency were first estimated with ERCC spike-ins. The lowest
148 detectable concentration was 4.4 molecules per million for both TSS and TES. In
149 other words, at a detection threshold of $TPM > 1$, at least 4.4 molecules are required to
150 get one detected read at sequencing depth of one million. Therefore, the sensitivity of
151 this method is estimated at roughly 22.7% ($1/4.4$) (**Fig. 1f**). This sensitivity is
152 approximately the same as the 22%-26% sensitivity previously reported for detection
153 of TSSs (24, 27, 28), but much higher than the 5.4% for TESs (29). In addition, the
154 number of TSSs detected genome-wide by scCAT-seq is highly dependent on the
155 number of reads mapped to the genome. Compared to existing methods which can
156 detect only a single end of transcripts (either the 5'-TSS or the 3'-TES), scCAT-seq
157 also has significantly better or comparable performance. When 1.28 million reads
158 were mapped to the mouse genome, around 8000 transcripts were detected by
159 scCAT-seq, comparable to the number for C1 CAGE (17) and the approach developed
160 by Arguel et al. (24), but much higher than STRT-seq (21) and Smart-seq2 (19), which
161 are the current single cell TSS profiling methods (**Supplementary Fig. 3a**). Similarly,
162 for TES detection with 1.28 million uniquely mapped reads, scCAT-seq can determine
163 TESs of more than 12000 transcripts, which is comparable to BAT-seq (29), and much
164 higher than Smart-seq2 (**Supplementary Fig. 3b**). Further, we compared the

165 performance of scCAT-seq to that of scISOr-seq (15, 20) which is the only method
166 available for profiling the full-length transcript of single cells. We sequenced 6 single
167 oocytes with the Pacbio sequel platform, with 54,000 circular consensus sequencing
168 (CCS) reads per single cell (**Supplementary Table 3**), which is much higher than that
169 of 270 reads per cell reported by Gupta et al. (15), and similar to that reported by
170 Byrne et al. on the Nanopore platform (20). By normalizing the sequencing depth to
171 the cost for both scCAT-seq and scISOr-seq, we found scCAT-seq had a much higher
172 efficiency in capturing both ends of full length isoforms than scISOr-seq, 3122 versus
173 919 genes for scCAT-seq versus scISOr-seq at the equal cost for 4 million PE150
174 short-reads from Illumina (**Fig. 1h, Supplementary Fig. 3c**). Around 15% of the
175 genes could be detected by both methods, with a higher overlapping ratio in highly
176 expressed genes (**Supplementary Fig. 3d, e**). In addition, for the number of
177 overlapping genes between single cells, scCAT-seq had a 2-fold higher overlapping
178 ratio than scISOr-seq (60% versus 30%), highlighting the high consistency of
179 scCAT-seq (**Fig. 1i, Supplementary Fig. 3f, g**). Comparison of the expression of the
180 transcripts detected by scISOr-seq and scCAT-seq showed that scISOr-seq mainly
181 detected the part of transcripts with the highest abundance (**Fig. 1j**), which only
182 account for 1/4 of those detected by scCAT-seq. Furthermore, for the same coverage,
183 our approach drastically reduces library preparation and sequencing cost. For instance,
184 scCAT-seq only requires 1/73 of the cost required by scISOr-seq for 1000 transcripts
185 covered (**Supplementary Fig. 2c**). These results indicate that scCAT-seq is a more
186 cost-effective and reliable approach for quantitatively detecting both start sites and
187 end sites of full-length transcripts at single cell level.

188

189 **Identification of novel transcripts with scCAT-seq**

190 Leveraging the capacity to demarcate the boundaries a transcript, we set out to
191 identify novel isoforms, both alternative TSSs/TESSs of annotated genes and novel
192 transcripts of unannotated genes (**Fig. 2a**). Data from mouse oocytes and DRG
193 neurons was used for benchmarking. For annotated genes, we identified both
194 alternative TSSs and TESSs events, as evidenced by 3102 novel TSSs and 5746 novel

195 TESSs in oocytes (**Fig. 2b**), and 2031 novel TSSs and 4693 novel TESSs in DRG
196 neurons (**Fig. 2c**). In addition, 71 and 107 novel, unannotated transcripts were
197 identified in DRG and oocytes respectively. Of note, many RNA isoforms identified
198 by scCAT-seq, and validated by Smart2-seq and Sanger sequencing, were drop-out by
199 scISOr-seq (**Fig. 2d, f, h**), indicating that scCAT-seq can identify novel transcripts
200 with higher efficiency than scISOr-seq.

201 Further, to characterize the full-length information of novel RNA isoforms, such as
202 alternatively spliced exons, full-length cDNAs were cloned with primers binding to
203 the terminal ends identified by scCAT-seq (**Fig. 2a**). Full-length transcripts were
204 sequenced by Sanger sequencing or scISOr-seq, and validated by Smart2-seq (**Fig.**
205 **2d-i**). For example, Figure 2f shows an example of novel gene with several isoforms,
206 which were identified by Sanger sequencing of full-length cDNAs. Three isoforms
207 differing in cDNA length have differential first exon choices (**Fig. 2f, g**), and
208 alternative splicing events between isoform 2 and isoform 3 were revealed, which
209 were also validated by Smart-seq2, including the exon not detected by scISOr-seq. In
210 total, 96% (68/71) of novel transcripts detected by scCAT-seq were validated by
211 Smart-seq2, while only 10% (7/71) of them were detected by scISOr-seq, indicating
212 high drop-out rate of full-length transcripts in scISOr-seq. Our data suggest that when
213 combined with targeted full-length sequencing, scCAT-seq can achieve higher
214 coverage to reveal different isoforms of individual genes. In summary, scCAT-seq can
215 accurately identify not only novel TSSs and TESSs, but also completely unannotated
216 full-length transcripts in single cells.

217

218 **scCAT-seq improves upon the performance of scISOr-seq for single cell RNA**
219 **isoform quantification.**

220 Due to the higher efficiency and lower cost of scCAT-seq compared to scISOr-seq for
221 identifying alternative isoforms, we hypothesized that scCAT-seq could also improve
222 upon performance of scISOr-seq for accurately quantifying alternative isoforms
223 (**Supplementary Fig. 4a**). It is currently too expensive to use scISOr-seq to obtain the
224 sequencing depth required for accurate isoform quantification of multiple samples,

225 especially at single cell level. Byrne et al. also tried to quantify isoforms with the
226 number of CCS reads, but the number of genes covered was very limited.
227 Concordantly, our data showed that the CCS readout for the majority of genes covered
228 was less than 3 even though the sequencing depth was 0.5M for one single cell
229 (**Supplementary Fig. 4b**). Although CCS read numbers are positively correlated with
230 the number of reads of scCAT-seq, much higher variation was observed for the former
231 with 10- to 1000-fold fewer read counts (**Supplementary Fig. 4c, d**). Intriguingly,
232 when using the scCAT-seq to quantify the isoforms identified by scISOr-seq, the
233 squared coefficient of variation (CV^2) was reduced at least 10-fold, making isoform
234 quantification much more accurate (**Supplementary Fig. 4d**). For example, two
235 alternative isoforms of *Ermp1* were quantified with a CCS number below 5 in both
236 DRG and oocytes, without sufficient power to differentiate the quantification of the
237 two isoforms (**Supplementary Fig. 4e, f**). However, when quantified with scCAT-seq,
238 with much lower variance, the longer isoform was found to be significantly higher
239 expressed in oocytes than in DRGs. In summary, scCAT-seq can be used to quantify
240 isoforms identified by scISOr-seq in single cells to improve accuracy with lower cost.

241

242 **Characterization and quantification of cell-type specific transcripts with** 243 **scCAT-seq**

244 To further assess differential gene expression between different cell types based on
245 quantified abundances of TSS and TES tag counts, we performed scCAT-seq on three
246 different cell types – mouse DRG, oocytes at Day 3, and oocytes at Day 4. Both TSS
247 and TES transcriptome data clearly discriminated different cell types from each other
248 (**Fig. 3a, Supplementary Fig. 5a**). In addition, because our method can identify both
249 ends of transcripts, we set out to identify cell type specific transcript isoforms.
250 Comparing DRG and oocyte cell-type specific isoforms, we identified 166 transcript
251 isoforms encompassing 83 genes that only differed in TSS choices, and 222 isoforms
252 encompassing 111 genes that only differed in TES choices (**Fig. 3b, Supplementary**
253 **Fig. 5b, c**). For example, *Tsc22d1* and *Grpel* had no difference in total gene
254 expression between DRG and oocytes, but the two isoforms of each gene were

255 expressed in a cell-type specific manner (**Fig. 3c, Supplementary Fig. 5d-f**).

256 We also used scCAT-seq to assess RNA dynamics during ageing of post-ovulatory

257 oocytes, and compared oocytes at day 3 post-ovulation (control) with oocytes at day 4

258 post-ovulation (post-ovulatory ageing oocytes). After assessing the 975 detectable

259 TSSs and TESs across the control and ageing oocytes, we found that TESs are more

260 prone to change positions, and the alternative choice of TESs is strongly associated

261 with TSSs invariability (**two-sided Fisher's exact test, P value = 3.0×10^{-53}**),

262 supporting the notion of interdependency between transcription initiation and

263 polyadenylation (**Fig. 3d**). Further, a change in the choice of major isoform from day

264 3 to day 4 oocytes is observed in 343 genes with alternative TSSs and 1612 genes

265 with alternative TESs, with a trend that shorter 5' UTRs (**Fig. 3e**) or longer 3' UTR

266 are preferred (**Fig. 3f**). Thus, using scCAT-seq we can observe that the dynamics of

267 major isoform choice during oocyte ageing is accomplished according to a general

268 rule, which is through degradation of the major isoform on day 3, and activation of

269 the minor isoform to switch to the alternative major isoform on day 4, as illustrated by

270 *Ska3* (**Supplementary Fig. 6a**). In addition, the observations made by RT-qPCR

271 validated our scCAT-seq data analysis (**Supplementary Fig. 6b**).

272

273 **Single cell atlas of non-human primate corneal epithelial based on RNA**

274 **expression and APA analysis.**

275 We next employed scCAT-seq to profile a much larger number of single cells. Taking

276 the non-human primate cornea as an example, we collected single cells and generated

277 multiplexed cDNA using the 10x genomics platform. scCAT-seq libraries were

278 subsequently generated and sequenced, and the 7848 single cells successfully

279 captured were clustered into 5 major groups. Hundreds of marker genes for each cell

280 type were identified (**Supplementary Fig. 7a**), with GO items relating to epithelial

281 development enriched in the genes up-regulated and those relating to cell adhesion

282 down-regulated (**Supplementary Fig. 7b, c**). Based on the RNA expression of the

283 known marker genes, the following subtypes were identified: corneal epithelial cells

284 (CEC) highly expressing KRT3 and KRT12, transient amplifying cells (TAC) highly

285 expressing KRT12 but not KRT3, and limbal epithelial cells (LEC) highly expressing
286 KRT19 (**Fig. 3g**). Pseudotime analysis on scCAT-seq data revealed the trajectory from
287 TAC to LEC and CEC (**Fig. 3h**). We next identified the cell-type specific isoforms of
288 the three major subtypes and assessed their dynamics. From TAC to LEC, we found
289 285 genes and 244 genes switched to proximal and distal APA sites, respectively (**Fig.**
290 **3i**). From TAC to CEC, we found 457 genes and 414 genes switched to proximal and
291 distal APA sites, respectively (**Supplementary Fig. 8a**). For example, the longer
292 isoform of UBE2B preferentially uses the distal TES in CEC, while the shorter
293 isoform preferentially uses the proximal TES in TAC (**Supplementary Fig. 8b**). We
294 also found that expression of genes with proximal APA sites was significantly higher
295 in TAC than CEC/LEC, while there was no significant difference in expression
296 between CEC and TAC for genes with distal APA sites in epithelial cells, suggesting a
297 potential role of proximal APA choices in gene regulation during differentiation of
298 epithelial cells from TACs (**Supplementary Fig. 8c-f**).

299

300 **Discussion**

301 The approach we introduce here is highly accurate for transcript demarcation and
302 isoform quantification in single cells. Through a machine learning algorithm that
303 employs a majority voting strategy, the noisy false positive peaks were filtered out,
304 enabling scCAT-seq to identify authentic terminal signatures with a true positive rate
305 of 95%. Previously, machine learning has been successfully used to predict
306 differential alternative splicing (30, 31), but none of them can be used to identify
307 authentic demarcation of RNA isoforms to elucidate the transcriptomic complexity of
308 single cells. The machine learning model developed here can also improve the
309 accuracy of other methods to 95% , as evidenced by the ERCC data from C1 CAGE
310 (17), C1 STRT (23), and Arguel et al., indicating that our model can be applied to
311 other data sets that contain previously unrecognized high false positive signals. In
312 addition to identification, the accuracy of our approach for quantification of the
313 alternative isoforms is also very high, as the measured abundances are highly
314 concordant with the ground truth, with a pearson's correlation coefficient of 0.98. The

315 high accuracy of both identification and quantification of isoforms provides an
316 unprecedented opportunity for detection of previously unannotated genes and
317 unidentified alternative TSSs and TESs, as well as for quantitation of cell-type
318 specific RNA isoforms.

319 Another clear advantage of scCAT-seq is its efficiency. Based on short-read
320 sequencing, scCAT-seq can identify TSSs and TESs simultaneously from sequencing
321 data derived from a single library, enabling investigation of transcription initiation and
322 polyadenylation in a large number of single cells. Compared with methods which
323 capture only single ends of RNA transcripts, either the TSS or TES, scCAT-seq is
324 demonstrably better for elucidating transcriptome complexity.

325 Compared with the recently developed long-read sequencing based method
326 scISOR-seq, which can profile full-length transcripts for a group of single cells (15,
327 20), our approach requires 1/73 of the cost to detect the same number of transcripts,
328 with higher efficiency. In addition, scISOR-seq requires at least 1 μ g of cDNA input,
329 necessitating extensive amplification of cDNA with unavoidable PCR bias due the
330 requirement for extra PCR cycles. This results in a decrease in the number of covered
331 transcripts (a few hundred per single cell) and a lower transcript overlap ratio among
332 single cells. In contrast, scCAT-seq only requires 0.1 ng of cDNA to achieve sufficient
333 coverage of thousands of genes. Most importantly, it is still challenging to use
334 scISOR-seq to quantify the isoforms differentially expressed between single cells, as
335 accurate quantification requires deep sequencing that is currently too expensive for
336 many labs. In contrast, our method can accurately quantify the transcripts ($r=0.98$) at
337 an affordable cost for most labs. Due to the high accuracy and efficiency of
338 scCAT-seq in identifying transcript ends, scCAT-seq also offers an efficient pipeline
339 for full-length characterization of novel isoforms after targeted construction of
340 full-length cDNA libraries, simply by PCR from the terminal sites identified by
341 scCAT-seq in single cells.

342 In summary, the performance of scCAT-seq is a significant improvement upon that of
343 scISOR-seq in terms of cost, efficiency, and accuracy of both identification and
344 quantification of RNA isoforms.

345 Like all technologies, scCAT-seq has its limitations. First, the initial accuracy of TSS
346 and TES identification is dependent on the effective cloning of full length cDNA.
347 Although we adapted a widely used method Smart-seq2 to obtain cDNA, other
348 protocols with better performance may be substituted in the future. Second, whereas
349 the information of full-length isoforms of novel genes can be revealed by PCR using
350 primers targeted to transcript ends identified by scCAT-seq, in this study we
351 multiplexed only small number of example genes. However, profiling full-length
352 transcripts with higher multiplexing can be done by complementing scISOr-seq
353 downstream of scCAT-seq, in order to efficiently profile the targeted amplified
354 full-length cDNA libraries. Including the scCAT-seq approach to initially identify
355 isoforms of interest will help increase the efficiency of scISOr-seq with lower cost.
356 In conclusion, we believe that this robust and cost-effective approach is an ideal
357 technology for comprehensive and systematic assessment of RNA isoform dynamics
358 across heterogeneous single cells and biological conditions. Not only can it help
359 define cell types with specific isoform expression patterns, but it can also establish a
360 multi-faceted mammalian cell atlas in conjunction with other methodologies to
361 identify tissue specific epigenetic elements, genotypes, and cis-elements. It can be
362 widely implemented and may play important roles in projects such as the Human
363 Single Cell Atlas.

364

365 **Methods**

366 **Single cell isolation.** The experiment was performed on 4-6 week old C57BL/6 mice
367 of both genders. Mice were maintained under standard conditions (12 h light and dark
368 cycles, with sufficient food and water). To obtain single DRG neurons, euthanasia was
369 performed by CO₂ and cervical dislocation, L4-L5 DRG from mice of both sides were
370 dissected and dissociated into single cells. Single DRG neurons were manually picked
371 by using a micro-capillary pipette. Single cells were incubated into a 0.2-ml thin-wall
372 PCR tube containing 4 µl Smart-seq2 lysis buffer according to the published
373 protocol(19, 32). To obtain postovulatory-aged oocytes, female mice were
374 administered intraperitoneal injections of 10 IU pregnant mare serum gonadotropin

375 and 10 IU human chorionic gonadotropin 48 hours later. Cumulus-oocyte-complexes
376 (COCs) were collected 24 h after human chorionic gonadotropin injections from the
377 oviductal ampullae. All cumulus cells were removed from the oocytes enzymatically
378 by trypsin treatment (Sigma-Aldrich) for 2 min and oocytes were subsequently
379 washed in DMEM medium containing 10% fetal bovine serum (FBS)
380 (Sigma-Aldrich). Oocytes were picked into a 0.2-ml thin-wall PCR tube contains 4 μ l
381 Smart-seq2 lysis buffer as described before.

382

383 **scCAT-seq library construction.** The full-length cDNA was generated through
384 reverse transcription with transcriptase III and the RT primer
385 (5'-AAGCAGTGGTATCAACGCAGAGTN4 [16bp of cell barcode] T30VN-3'),
386 followed by PCR amplification according to Smart-seq2 protocol(19) with minor
387 modification that Superscript II was replaced by superscript III to improve the yield of
388 cDNA. ERCC RNA spike-in Mix which contains 92 transcripts (Thermo Fisher) was
389 added and processed in parallel with poly-A RNA. After purification, 0.1 ng cDNA
390 was used for Nextera tagmentation and fragments of both ends of the cDNA were
391 selectively amplified by using the primers targeting TSO and Tn5 adaptors as shown
392 in Fig. 1a. Library are purified using $1.8 \times$ Agencourt AMPure XP beads
393 (BECKMAN COULTER), and then loaded on an E-Gel 2% SizeSelect, and fragments
394 of a length of 200-300bp bases were selected. Simultaneously, 0.1 ng of cDNA was
395 used for standard Smart-seq2 libraries. Library was assessed by using Agilent
396 Bioanalyzer 2100, and sequenced on Illumina Xten platform. The rest of the cDNA
397 were used for PacBio ISO-seq analysis.

398

399 **Single cell ISO-seq.** Single cell ISO-seq was performed on PacBio Sequel platform.
400 Full-length cDNA of eight single cells were mixed together to reach the total amount
401 of 2ug for each flowcell. PacBio library construction is done by using SMRTbell
402 Template Prep Kit (PacBio cat#100-991-900), and sequenced using SMRTcells
403 (PacBio cat#101-008-000), with eight single sample per SMRTcell.

404

405 **Single cell isolation of crab-eating monkey cornea epithelium and library**
406 **construction.** Whole eyes were dissected from a healthy crab-eating monkey. The
407 lens, retina, iris, and trabecular network were removed and most of the conjunctiva
408 was dissected and discarded. The corneal rims were subsequently treated with 1.5mL
409 of 10mg/mL Dispase II in PBS at 37°C for 2 hours and 0.25% trypsin and 1 mM
410 EDTA solution at 37°C for 15 minutes with gentle pipetting to yield single cells
411 suspension. The disassociated corneal epithelial cells were captured on the 10x
412 Genomics Chromium controller according to the Chromium Single Cell 3' Reagent
413 Kits V2 User Guide (10x Genomics PN-120237). Library was prepared using 0.3ng
414 cDNA from 10x Genomics following the scCAT-seq protocol as described above.

415

416 **Data processing of next generation sequencing data.** TSS and TES raw data were
417 extracted and processed separately. For TSS data, reads with the sequencing tag
418 5'-GTGGTATCAACGCAGAGTACATGGG-3' were selected, and TSO sequences
419 5'-GTGGTATCAACGCAGAGTACAT-3' were trimmed away with the "GGG" tag
420 retained. Then, these reads were aligned to mouse genome (mm10) with STAR
421 (version 2.6.1a) with parameters (--outFilterMultimapNmax 1
422 --outFilterScoreMinOverLread 0.6 --outFilterMatchNminOverLread 0.6). Uniquely
423 mapped reads were kept but discarded if the 5' GGG was mapped. Reads that aligned
424 to ribosomal RNA region were also discarded.

425 For the TES data, we first processed to remove 3' adaptor sequences with cutadapt
426 (version 1.18), and then extract pairing reads with R1 has 3' Tag and R2 contains at
427 least 10 polyA sequences at the 3' side. Poly A sequences in the end of R2 were
428 further trimmed with 5 A bases left at the 3' side. By using STAR with parameters
429 described above, reads were aligned to mouse genome (mm10). The reads with the
430 terminal 5 A bases not mapped to the genome were retained for downstream analysis
431 for polyadenylation sites. Reads mapped to multiple sites, with low quality alignment,
432 and aligned to mitochondrial or ribosomal RNA region were discarded.

433 For Smart-seq2 data, raw reads past quality control were aligned by STAR using
434 parameter as described above. Only reads that uniquely mapped to mm10 were

435 retained and read count on each gene in each sample was computed using
436 featureCounts (33). Differentially expressed gene analysis was performed using
437 SCDE (version 2.10.1) (34).

438 For comparison, we downloaded BAT-seq data (accession number: GSE60768), C1
439 STRT (accession number: GSE60361) data and data generated by Arguel et al.
440 (accession number: GSE79136) from the Gene Expression Omnibus database. C1
441 CAGE data were downloaded from DDBJ (Project ID: PRJDB5282). For the BAT-seq
442 data, we picked 192 mouse ES cells. For the C1 STRT data, 80 mouse cerebellum
443 cells from the single-cell dataset were randomly picked. Same strategies were used
444 with small modification to process C1 STRT data and BAT-seq data. For all data, we
445 converted bam files to bed files with BEDtools (version 2.27.1). For 5' end data, we
446 extract the 5' end from bed files for further analysis. Likewise, we extract the 3' end
447 from bed files for 3' end data.

448

449 **Data processing of scISO-Seq data**

450 Circular consensus reads (CCS) were obtained from the raw data of subreads Bam
451 files by using PacBio Sequel SMRT-Link 7.0 Soft, with the default setting of
452 parameters: minLength 10, maxLength 21000, minReadScore 0.75, minPasses 3.
453 Then, reads were considered FLNC if they contained 5' and 3' primers in addition to a
454 polyA tail. Primer and polyA tails were removed by cutadapt. Further, FLNC reads
455 were aligned to reference genome mm10 using Minimap2(35) (version 2.17) with
456 parameters (-t 30 -ax splice -uf --secondary=no -C5 -O6,24 -B4). CCS count on each
457 gene in each sample was computed using featureCounts. The output Sam files were
458 fed into Cupcake ToFU to collapse the mapped FLNC reads into unique transcripts.
459 Scripts are available at: https://github.com/Magdoll/cDNA_Cupcake. Eventually,
460 isoforms were identified and filtered using SQANTI2 against mm10 transcriptome
461 annotation.

462

463 **Peak calling.** To identify TSSs and TESs, we used CAGER (version 1.24.0) package
464 in R. Peaks were called using distclu (threshold = 5, nrPassThreshold = 1,

465 thresholdIsTpm = TRUE, removeSingletons = FALSE, keepSingletonsAbove = 10,
466 maxDist = 20). The position of dominant TSS/TES in each peak was set to represent
467 the position of peak. TSS and TES annotation reference was based on gencode
468 release_M18, and peaks mapped between 2kb upstream the annotated TSSs and 2k
469 downstream the annotated TESs were considered to belonging to the said gene. We
470 then extracted 5'-end and 3'-end of all annotated transcripts and converted to bed files
471 with a custom R script, and distance between the called peaks and the nearest
472 annotated TSS/TES was calculated by a custom script. We adopted the following
473 priority in calculating the distribution of TSS peaks mapped to genome features:
474 $TSS \pm 1000 > 5' \text{ UTR} > \text{first exon} > \text{first intron} > \text{other exon} > \text{other intron} > 3' \text{ UTR} >$
475 intergenic . Similarly, The priority in calculating distribution of TES peaks mapped to
476 genome features is $TES \pm 1000 > 3' \text{ UTR} > \text{last exon} > \text{last intron} > \text{other exon} > \text{other}$
477 $\text{intron} > 5' \text{ UTR} > \text{intergenic}$.

478

479

480 **Machine learning analysis.** To predict a peak is real or false TSS/TES, we employed
481 three widely used models, including logistic regression classifier, random forest and
482 support vector machine.

483

484 Firstly, we use eight features of read distribution to train the three machine learning
485 models and they are summarized in the table 1. To perform the analysis, we used two
486 independent data sets derived from ERCC spike-ins which can serve as a standard for
487 true TESs/TSSs determination, one is for training data, and the other is for test data.
488 The features were normalized, "TPM_of_peak", "TPM_of_Dominant_Site" are firstly
489 being taken a log and secondly normalized to be in the range of [0,1]. Training data
490 was generated by using scCAT-seq for ERCC spike-ins, and the test data was derived
491 from the ERCC spike-ins mixed in the single cells. The True positive and False
492 positive of TSSs and TESs predicted was calculated. Secondly, for the genomic data,
493 in addition to the eight features of read distribution, we added an additional 650 and
494 150 features of motifs related to TESs and TSSs, and used FANTOM5 database (25)

495 and PolyA_DB (26) to train the model for TSS and TES prediction respectively.
496 TSSs/TEs were predicted from the peaks of single cells based on scCAT-seq. We
497 utilize the popular open source python machine learning library scikit-learn to train
498 these models.

499 With a logistic regression model, the probability (π) of a peak with the given values of
500 the features ($x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8$) was determined as:

$$501 \quad \pi = p(y = 1|x; w) = \frac{1}{1+e^{-w^T x}} ;$$

$$w^T x = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_3 + w_4 x_4 + w_5 x_5 + w_6 x_6 + w_7 x_7 + w_8 x_8$$

502 Where the ($x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8$) are the observed value of the features
503 shown in **Table 1**, and $w_0, w_1, w_2, w_3, w_4, w_5, w_6, w_7, w_8$ are the coefficients
504 of the corresponding features of the training model. The decision was made based on
505 the following function:

$$y = \ln \frac{\pi}{1 - \pi}$$

506 We also applied L2 regularization and the coefficient is determined using cross
507 validation on the training set.

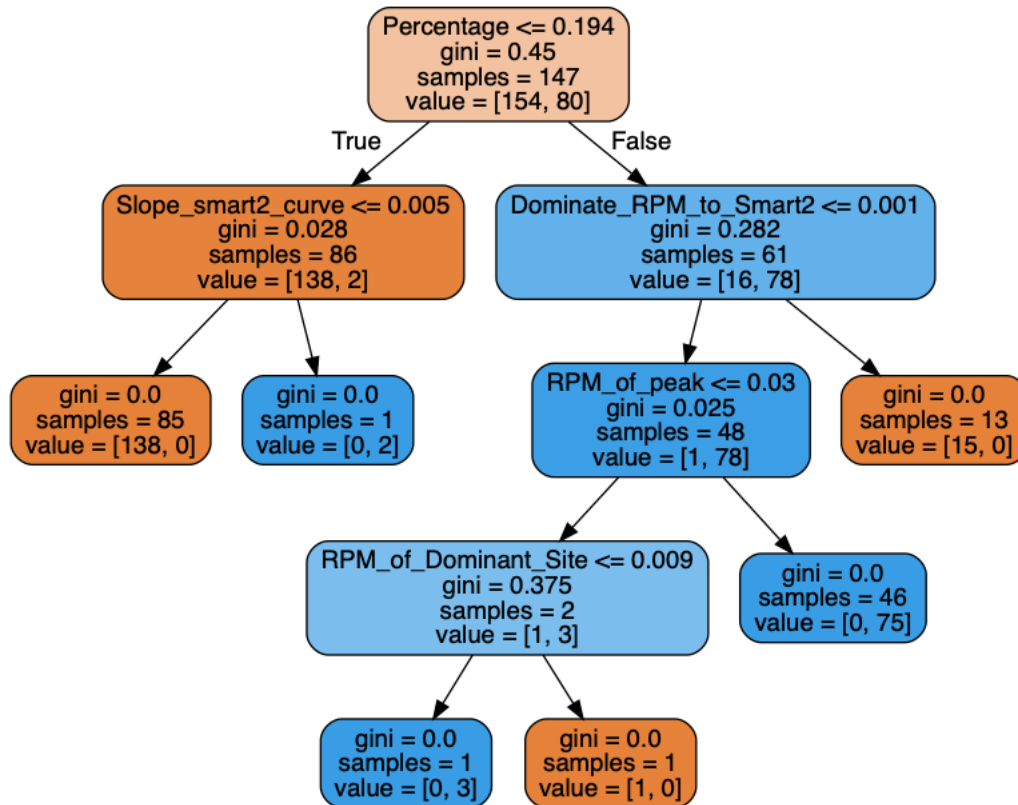
508

509 Random forest model(36) consists of a large number of individual decision trees that
510 operate as an ensemble. Every tree in the random forest makes its own class
511 prediction and the class with the most votes becomes the random forest model's
512 prediction. In random forest, each decision tree is independently trained using partial
513 features and bootstrap sampled training data. To generate a tree, it has to go over each
514 feature, and find the best one has the maximum gini index reduction(37) after splitting.
515 The gini index for each node is defined as:

$$\text{Gini}(D) = 1 - \sum_{k=1}^N p_k^2$$

516 Where D is a node in the tree, N is the number of different classes, and p_k is the
517 percentage of data in this node that is labeled class k. Conceptually, gini index reflect
518 how different they are if we randomly choose two samples from the node. The smaller
519 the gini index, the more pure the node is. After the split, if the child node still contains

520 more than one class, it will go through the search process again to split it. This
521 process generally ends when all the leaf nodes contain only one class samples. An
522 example of a decision tree learned is shown as below:



523

524

525 In every node of the tree in this plot, it first shows the selected feature and splitting
526 criteria to maximize gini index reduction. Then it shows the gini index for this node.
527 The “samples” represents number of distinct samples this node has. And the value has
528 two numbers, corresponding to the number of negative and positive training data
529 (some training data can have multiple copies since they are bootstrapped from the
530 original dataset). After we have learned a number of decision trees, we’ll do a majority
531 vote using all the trees’ predictions. In statistical theory, this step helps reduce model
532 variance.

533 The SVM(38) is another widely used supervised machine learning models for two
534 class classification (can be extended for multi-class classification and regression as
535 well.) The SVM algorithm tries to find a hyper plane in a mapped high dimensional
536 space (with kernel trick) that separates the two classes that achieves largest margin.

537 From any textbook, the SVM with soft margin and regularization is formularized as:

$$\begin{aligned} \min_{\{w,b,\xi_i\}} \quad & \frac{1}{2} \|w\|^2 + C \sum_{i=1}^m \xi_i \\ \text{s. t.} \quad & y_i(\theta^T x_i + b) \geq 1 - \xi_i, \\ & \xi_i \geq 0, i = 1, 2, \dots, m \end{aligned}$$

538 Where ξ_i is used to allow soft margin, and m is the number of training data you have.

539 The C controls the relative regularization and is determined using cross validation

540 method. And w is the vector of weights, and x is the feature vector.

541 Lastly, we try to further improve model performance by ensemble all three models.

542 Dietterich (39) indicated statistical, computational and representational benefits of

543 combining models. This theory is also validated here as the ensembled model

544 achieves better performance than any one of the three models alone, despite the fact

545 that the three models already achieve good performance on their own.

546

547 **Quantification of cell-type specific isoforms**

548 Expression values for each peak (TSS/TES) were quantified as tags per million (TPM)

549 generated by CAGER. To identify cell-type specific isoforms, the major TSS/TES

550 positions of genes co-expressed between the two types of cells are compared by

551 intersect the bed files of each with BEDtools (40). Genes with either alternative TSS

552 or alternative TES between the two were selected. Then, the differential expression

553 analysis on the TPM value of the major isoform of each cell type between the two was

554 performed with DESeq2. **Further**, we performed qRT-PCR with Unique AptamerTM

555 qPCR SYBR[®] Green Master Mix (Novogene) on the RocheLightCycler480 (Roche)

556 using the same samples used for next-generation sequencing to validate the alternative

557 TES in different cell types. All assays were run in triplicate for six individual samples.

558 The qRT-PCR conditions used were as follows: 5 min at 95°C, 45 cycles of 10 sec at

559 95°C and 30 sec at 60°C. The qRT-PCR primers sequences used were listed in

560 Supplementary Table 4. Gene body primers were used to quantify total gene

561 abundance. 3' UTR primers were used to quantify long 3' UTR isoform. Data were

562 analyzed using the $2^{-\Delta\Delta Ct}$ method.

563

564 **Sequencing full-length cDNA of target genes.** Primers were designed according to
565 the coordinates of TSS/TES identified by scCAT-seq. Full-length cDNA of all
566 isoforms of a target gene was amplified by PCR from the cDNA pool of single cells
567 generated with Smart2-seq. Briefly, 1 ng full length cDNA was used to perform
568 35-cycle PCR with Premix TaqTM (TaKaRa). PCR products were purified with
569 QIAquick Gel Extraction Kit (Qiagen) and Sanger-sequenced with corresponding
570 primers. All assays were performed for three individual single cell samples. PCR
571 primers used for novel genes are listed in Supplementary Table 5.

572

573 **Data processing of corneal single-cell data.** Each 10x droplet sequencing data was
574 processed using the Cell Ranger (version 2.1.1) pipeline from 10x Genomics. In brief,
575 reads was demultiplexed and aligned to the *Macaca fascicularis* genome. UMI counts
576 was quantified to generate a gene-barcode matrix. Cells were filtered to remove those
577 containing less than 500 genes. Genes that were detected in less than 3 cells were also
578 removed. Further analyses of these cells were performed using the Seurat (version
579 3.0.2) R packages, as described in the tutorials ("<https://satijalab.org/seurat/>")(41).
580 Briefly, cells were normalized using LogNormalize and multiplied by a scale factor of
581 10000. HVGs (high variable genes) were identified and used for further analysis.
582 Shared cell states were identified using integration procedure in Seurat.

583 Dimensionality reduction was performed using principal component analysis (PCA).
584 Statistically significant PCs were identified using the Jackstraw function. The score of
585 cells in those significant PCs were used to build a k-nearest neighbor (KNN) graph.
586 Louvain algorithm was used for identifying cell clusters in KNN graph (parameter
587 resolution=0.06). Uniform manifold approximation and projection (UMAP)
588 dimensionality reduction was used to project these populations in two dimensions.
589 Pseudotime analyses of CEC was performed using Monocle2 (42) (version 2.12.0) R
590 package. Differentially expressed genes among LEC, CEC and TAC were identified
591 using differentialGeneTest function and used as input for temporal ordering of those
592 cells along the differentiation trajectory.

593

594 **Code availability.** Custom computer code used in this study is freely available at
595 <https://github.com/huyoujinlab/scCAT-seq>.

596

597 **Availability of data and material**

598 All the related data can be downloaded from GEO with the accession number
599 **GSE134311**.

600

601 **Acknowledgement**

602 The work is supported by National Key R&D Program of China (2018YFA0108300,
603 2017YFC1001300); National Natural Science Foundation of China (31700900;
604 81530028; 81721003); Clinical Innovation Research Program of Guangzhou
605 Regenerative Medicine and Health Guangdong Laboratory (2018GZR0201001);
606 Local Innovative and Research Teams Project of Guangdong Pearl River Talents
607 Program; the State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center,
608 Sun Yat-sen University.

609

610 **References**

611

- 612 1. Trapnell C (2015) Defining cell types and states with single-cell genomics. *Genome Res*
613 25(10):1491-1498.
- 614 2. Wagner A, Regev A, & Yosef N (2016) Revealing the vectors of cellular identity with single-cell
615 genomics. *Nat Biotechnol* 34(11):1145-1160.
- 616 3. Tang FC, *et al.* (2009) mRNA-Seq whole-transcriptome analysis of a single cell. *Nat Methods*
617 6(5):377-U386.
- 618 4. Regev A, *et al.* (2017) The Human Cell Atlas. *Elife* 6.
- 619 5. Nosedà M & Harding SE (2018) Understanding dynamic tissue organization by studying the
620 human body one cell at a time: the human cell atlas (HCA) project. *Cardiovasc Res*
621 114(12):E93-E95.
- 622 6. Barash Y, *et al.* (2010) Deciphering the splicing code. *Nature* 465(7294):53-59.
- 623 7. Pan Q, Shai O, Lee LJ, Frey J, & Blencowe BJ (2008) Deep surveying of alternative splicing
624 complexity in the human transcriptome by high-throughput sequencing. *Nat Genet*
625 40(12):1413-1415.
- 626 8. Donczew R & Hahn S (2018) Mechanistic Differences in Transcription Initiation at TATA-Less
627 and TATA-Containing Promoters. *Mol Cell Biol* 38(1).
- 628 9. Di Giammartino DC, Nishida K, & Manley JL (2011) Mechanisms and consequences of
629 alternative polyadenylation. *Mol Cell* 43(6):853-866.
- 630 10. Moqtaderi Z, Geisberg JV, & Struhl K (2018) Extensive Structural Differences of Closely

- 631 Related 3' mRNA Isoforms: Links to Pab1 Binding and mRNA Stability. *Mol Cell* 72(5):849-861
632 e846.
- 633 11. Reyes A & Huber W (2018) Alternative start and termination sites of transcription drive most
634 transcript isoform differences across human tissues. *Nucleic Acids Res* 46(2):582-592.
- 635 12. Liu YL & Elliott DJ (2010) Coupling genetics and post-genomic approaches to decipher the
636 cellular splicing code at a systems-wide level. *Biochem Soc T* 38:237-241.
- 637 13. Anvar SY, *et al.* (2018) Full-length mRNA sequencing uncovers a widespread coupling
638 between transcription initiation and mRNA processing. *Genome Biol* 19.
- 639 14. Chen W, *et al.* (2017) Alternative Polyadenylation: Methods, Findings, and Impacts. *Genom*
640 *Proteom Bioinf* 15(5):287-300.
- 641 15. Gupta I, *et al.* (2018) Single-cell isoform RNA sequencing characterizes isoforms in thousands
642 of cerebellar cells. *Nat Biotechnol* 36(12):1197-+.
- 643 16. Hochgerner H, *et al.* (2017) STRT-seq-2i: dual-index 5' single cell and nucleus RNA-seq on an
644 addressable microwell array. *Sci Rep-Uk* 7.
- 645 17. Kouno T, *et al.* (2019) C1 CAGE detects transcription start sites and enhancer activity at
646 single-cell resolution. *Nat Commun* 10.
- 647 18. Goetz JJ & Trimarchi JM (2012) Transcriptome sequencing of single cells with Smart-Seq. *Nat*
648 *Biotechnol* 30(8):763-765.
- 649 19. Picelli S, *et al.* (2014) Full-length RNA-seq from single cells using Smart-seq2. *Nat Protoc*
650 9(1):171-181.
- 651 20. Byrne A, *et al.* (2017) Nanopore long-read RNAseq reveals widespread transcriptional
652 variation among the surface receptors of individual B cells. *Nat Commun* 8.
- 653 21. Ng P, *et al.* (2005) Gene identification signature (GIS) analysis for transcriptome
654 characterization and genome annotation. *Nat Methods* 2(2):105-111.
- 655 22. Haberle V, Forrest ARR, Hayashizaki Y, Carninci P, & Lenhard B (2015) CAGEr: precise TSS data
656 retrieval and high-resolution promoterome mining for integrative analyses. *Nucleic Acids Res*
657 43(8).
- 658 23. Islam S, *et al.* (2014) Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat*
659 *Methods* 11(2):163-+.
- 660 24. Arguel MJ, *et al.* (2017) A cost effective 5' selective single cell transcriptome profiling
661 approach with improved UMI design. *Nucleic Acids Res* 45(7).
- 662 25. Consortium F, *et al.* (2014) A promoter-level mammalian expression atlas. *Nature*
663 507(7493):462-470.
- 664 26. Wang R, Nambiar R, Zheng D, & Tian B (2018) PolyA_DB 3 catalogs cleavage and
665 polyadenylation sites identified by deep sequencing in multiple genomes. *Nucleic Acids Res*
666 46(D1):D315-D319.
- 667 27. Zeisel A, *et al.* (2015) Cell types in the mouse cortex and hippocampus revealed by single-cell
668 RNA-seq. *Science* 347(6226):1138-1142.
- 669 28. Hashimshony T, *et al.* (2016) CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq.
670 *Genome Biol* 17:77.
- 671 29. Velten L, *et al.* (2015) Single-cell polyadenylation site mapping reveals 3' isoform choice
672 variability. *Mol Syst Biol* 11(6).
- 673 30. Leung MK, Xiong HY, Lee LJ, & Frey BJ (2014) Deep learning of the tissue-regulated splicing
674 code. *Bioinformatics* 30(12):i121-129.

- 675 31. Qin Z, Stoilov P, Zhang X, & Xing Y (2018) SEASTAR: systematic evaluation of alternative
676 transcription start sites in RNA. *Nucleic Acids Res* 46(8):e45.
- 677 32. Hu YJ, *et al.* (2016) Simultaneous profiling of transcriptome and DNA methylome from a
678 single cell. *Genome Biol* 17.
- 679 33. Liao Y, Smyth GK, & Shi W (2014) featureCounts: an efficient general purpose program for
680 assigning sequence reads to genomic features. *Bioinformatics* 30(7):923-930.
- 681 34. Kharchenko PV, Silberstein L, & Scadden DT (2014) Bayesian approach to single-cell
682 differential expression analysis. *Nat Methods* 11(7):740-U184.
- 683 35. Li H (2018) Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*
684 34(18):3094-3100.
- 685 36. L. B (2001) Random forests. *Machine learning* 45(1):5-32.
- 686 37. Breiman L, Friedman, J.H., Olshen, R.A., and Stone, C.I. (1984) Classification and regression
687 trees. (Belmont, Calif.: Wadsworth).
- 688 38. Boser BE, Isabelle M. Guyon, and Vladimir N. Vapnik. (1992) A training algorithm for optimal
689 margin classifiers. *Proceedings of the fifth annual workshop on Computational learning*
690 *theory. ACM*, pp 144-152.
- 691 39. Dietterich TG (2000) Ensemble methods in machine learning. in *International workshop on*
692 *multiple classifier systems* (Springer, Berlin, Heidelberg).
- 693 40. Quinlan AR & Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic
694 features. *Bioinformatics* 26(6):841-842.
- 695 41. Stuart T, *et al.* (2019) Comprehensive Integration of Single-Cell Data. *Cell* 177(7):1888-1902
696 e1821.
- 697 42. Trapnell C, *et al.* (2014) The dynamics and regulators of cell fate decisions are revealed by
698 pseudotemporal ordering of single cells. *Nat Biotechnol* 32(4):381-386.

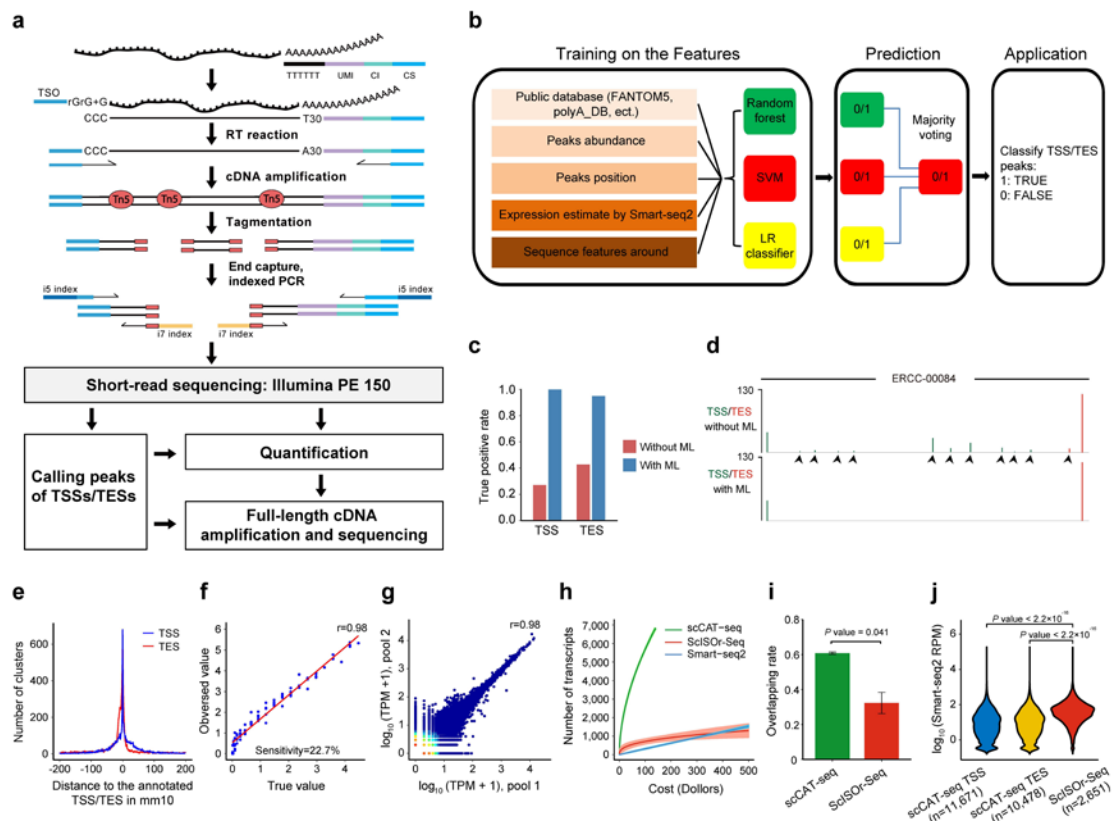
699

700

701

702

703 **Figure**



704

705 Figure 1. The scCAT-seq method and performance. **a**, Schematic of the scCAT-seq
 706 method. Template switching reverse transcription of full-length cDNA was performed
 707 with oligo-dT primer containing a unique molecular identifier (UMI), a cell identifier
 708 (CI), and a common sequence (CS). After PCR amplification, cDNA was tagged
 709 with Tn5 transposases. Both 5' and 3' ends of the cDNA were captured and amplified
 710 by PCR using primers binding to CS and TSO sequences, during which Illumina
 711 sequencing indexes were tagged. In addition, Smart-seq2 libraries are generated from
 712 cDNA of the same cell. Sequencing data was processed and transcription start sites
 713 (TSSs) and transcription end sites (TESs) were identified by machine learning models,
 714 following by quantification of transcript isoforms. **b**, Schematic of the machine
 715 learning model. Features were collected and three machine learning models were
 716 implemented. Predictions from all models were integrated by majority voting. **c**, True
 717 positive for identification of TSSs and TESs with and without optimization by the
 718 majority voting strategy based on machine learning models. **d**, Genome browser
 719 shows the example of TSS/ TES identification with or without machine learning (ML).

720 The false positive peaks filtered out by ML were indicated by arrows. **e**, Distance of
721 TSSs/TESSs identified by scCAT-seq in the genome to those annotated in mm10. **f**,
722 Scatter plot of observed transcript expression levels (y axis) and true abundance (x
723 axis) of ERCC spike-ins through 5'-end quantification (n = 92). Each point represents
724 a transcript. The Pearson's correlation coefficient is shown in the upper right corner.
725 The capture efficiency is estimated by the probability of an individual transcript could
726 be detected at sequencing depth of one million. **g**, Scatter plots shows the Pearson's
727 correlation of transcriptional level of isoforms between replicated pools of 3 single
728 cells. **h**, The number of transcripts with both ends captured using scCAT-seq,
729 Smart-seq2, or ScISOr-Seq, versus cost. The shaded regions represent 95%
730 confidence interval. **i**, Barplot shows the overlapping rate of genes detected among
731 single cells, by scCAT-seq versus ScISOr-Seq (n = 3 single cells). Significance was
732 computed using two sided t-test. Error bars represent standard error of the mean. **j**,
733 Violin plot for expression level comparison between genes detected by scCAT-seq and
734 ScISOr-Seq. Gene expression levels were quantified by Smart-seq RPM value.
735 Significance was computed using two-sided Wilcoxon test.

736

737

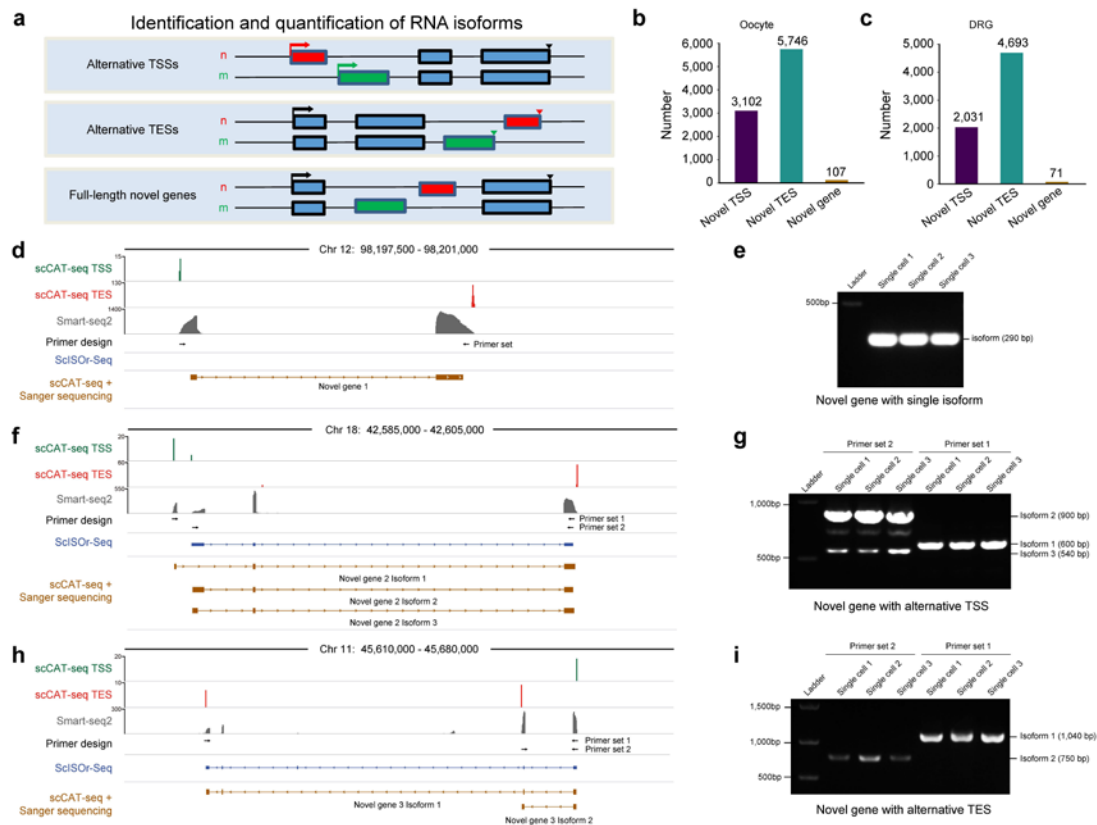
738

739

740

741

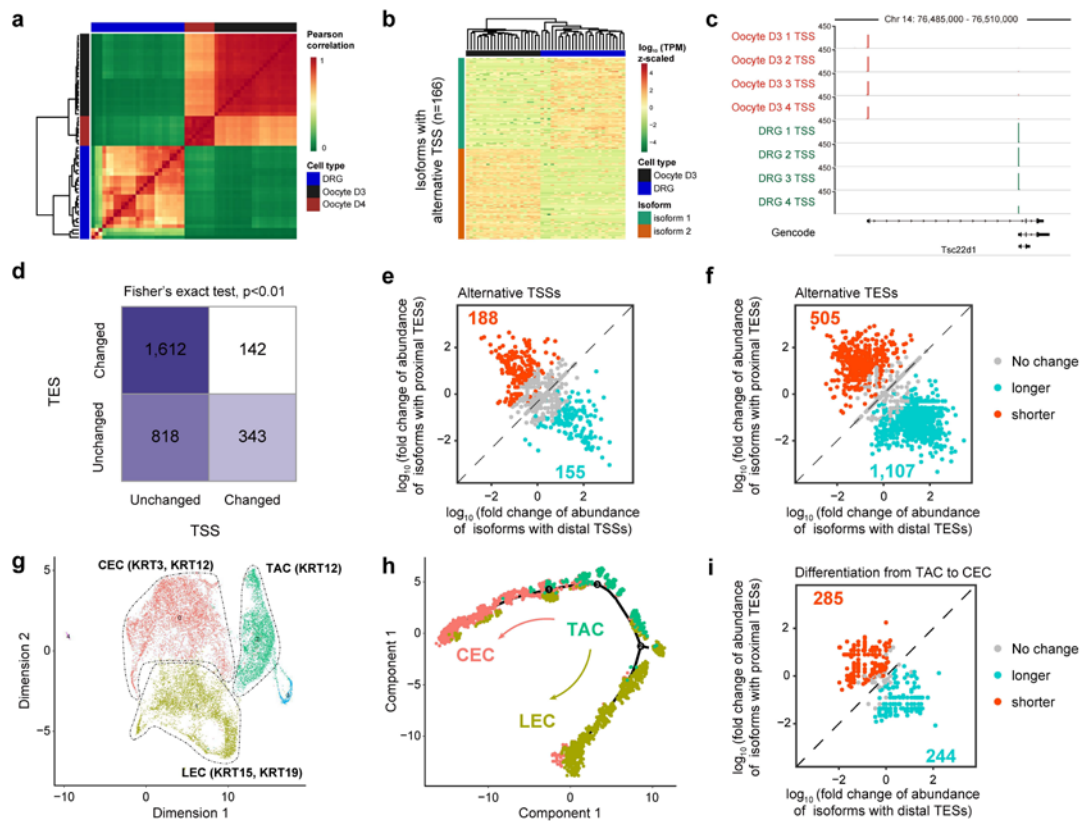
742



743

744 Figure 2. Characterization of novel transcripts and isoforms of single cells with
 745 scCAT-seq. **a**, Schematic of the functions of scCAT-seq. **b**, Barplot showing the
 746 number of novel isoforms of annotated genes and novel, unannotated transcripts in
 747 DRG neurons. The number of transcripts for each category is indicated above the box.
 748 **c**, Barplot showing the number of novel isoforms of annotated genes and novel,
 749 unannotated transcripts in oocytes. **d**, Genome browser track for an example of novel
 750 genes with single isoform. **e**, Gel image showing validation result of novel gene in **d**. **f**,
 751 Genome browser track for an example of novel genes with alternative TSSs on a
 752 different exon. **g**, Gel image showing validation result of novel gene in **f**. **h**, Genome
 753 browser track for an example of novel genes with alternative polyadenylation sites on
 754 a different exon. **i**, Gel image showing validation result of novel gene in **h**.

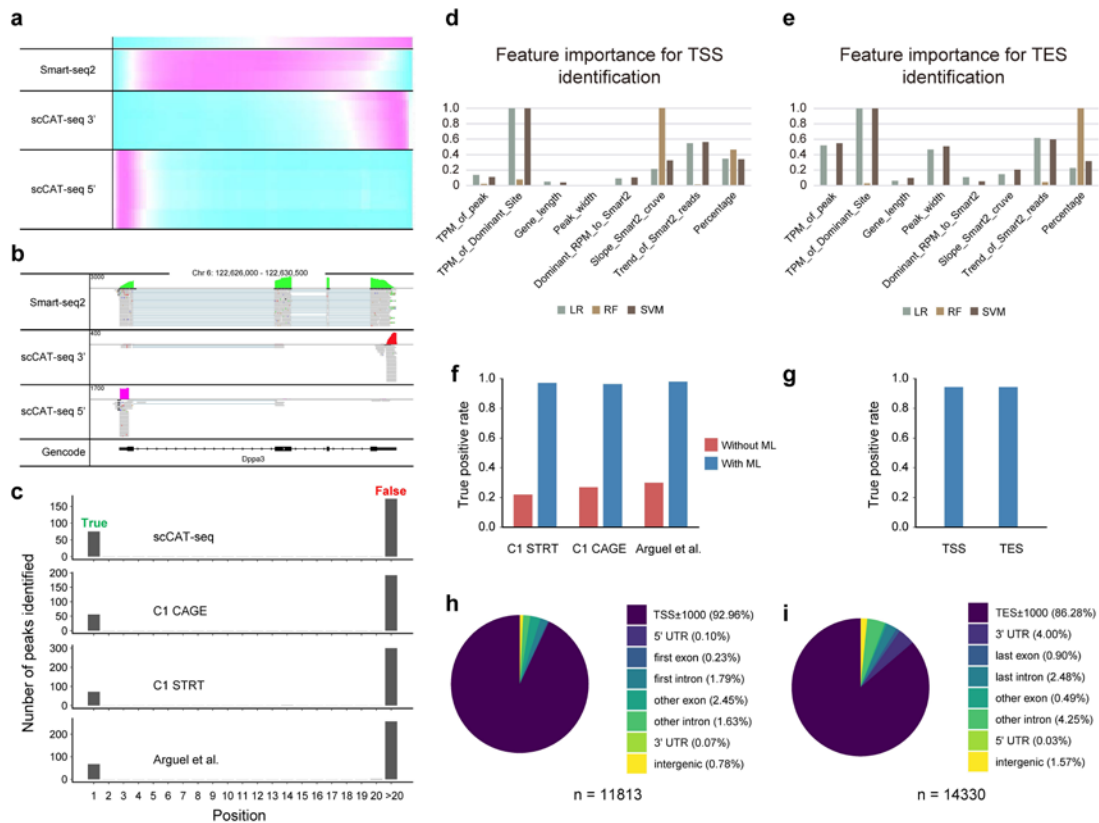
755



756

757 Figure 3. Quantification of cell specific isoforms discriminate cell types and illustrates
 758 the dynamics of isoform choices during oocyte ageing and corneal epithelial
 759 regeneration. **a**, Heatmap for Pearson's correlation coefficient of transcriptomes of
 760 DRG neuron and oocytes, based on 5'-end quantification of RNA isoforms. **b**,
 761 Heatmap showing RNA isoforms of alternative TSS choices with cell type specificity.
 762 The major isoforms either in oocytes or in DRG neurons are shown ($n = 166$
 763 isoforms). **c**, Genome browser tracks showing the alternative choices of TSS of
 764 *Tse22d1* between oocytes and DRG neurons. **d**, Heatmap showing the number of
 765 transcripts with or without TSSs/TESS changes during oocyte post-ovulatory ageing. **e**,
 766 Expression data with isoform specificity reveals isoform expression dynamics during
 767 oocyte post-ovulatory ageing ($n = 1,161$ genes). **f**, Expression data with isoform
 768 specificity reveals TES changes and isoform expression dynamics during oocyte
 769 post-ovulatory ageing ($n = 1,754$ genes). **g**, UMAP plot depicting cell clusters
 770 identified with scCAT-seq, including corneal epithelial cell (CEC), limbal corneal
 771 epithelial cell (LEC), and transient amplifying cell (TAC), and their specific marker
 772 genes ($n = 7,848$ cells). **h**, A pseudotime trajectory of single cells constructed using

773 Monocle. Indicated in color are the three presumptive states corresponding to CEC,
774 TAC, and LEC. **i**, Expression data with isoform specificity reveals TES changes and
775 isoform expression dynamics during differentiation of CEC from TAC (n = 584
776 genes).
777



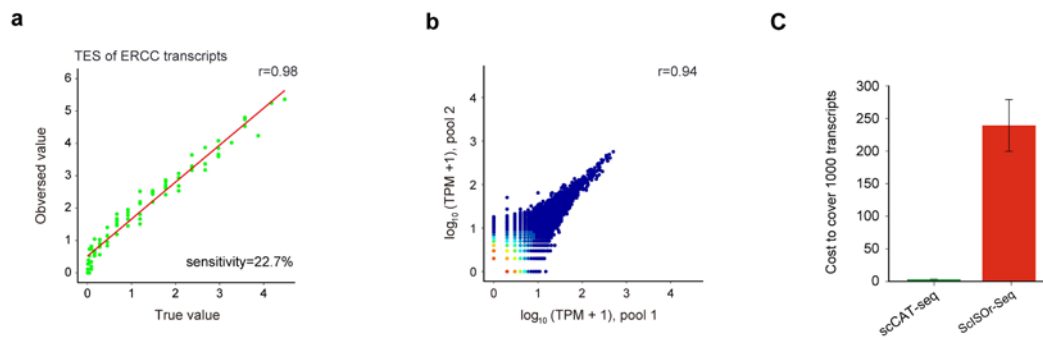
778

779 Supplementary figure 1. Machine learning improves accuracy of scCAT-seq
 780 demarcated isoform boundaries in single cells. **a**, Distribution of sequencing reads
 781 along the genes from head to tail from Smart-seq2 and scCAT-seq. **b**, *Dppa3* as an
 782 example gene, showing the distribution of sequencing reads of Smart-seq2 and
 783 scCAT-seq. **c**, TSS peaks identified in the data of scCAT-seq, as well as public data
 784 sets of ERCC for C1 CAGE, C1 STRT and Arguel et al. True positive peaks located
 785 around the annotated TSSs and false positive TSS peaks located elsewhere are
 786 indicated respectively. **d**, Relative feature importance of the eight features for TSS
 787 identification in random forest model (RF), support vector machine (SVM), and
 788 logistic regression classifier (LR). **e**, Relative feature importance of the eight features
 789 for TES identification in the three machine learning models. The value of importance
 790 for SVM and LR are first transformed to absolute value and normalized to the highest
 791 value of the eight features. **f**, True positive for identification of TSSs for the public
 792 data sets with and without optimization by the majority voting strategy based on
 793 machine learning models. **g**, True positive for identification of TSSs and TESs for the
 794 pooled single DRG neurons data sets with optimization by the majority voting

795 strategy based on machine learning models. **h**, Pie chart with the genomic distribution
796 of the identified TSSs. The total number of TSS peaks identified after optimization by
797 the machine learning models is indicated under the pie chart. **i**, Pie chart with the
798 genomic distribution of the identified TESs.

799

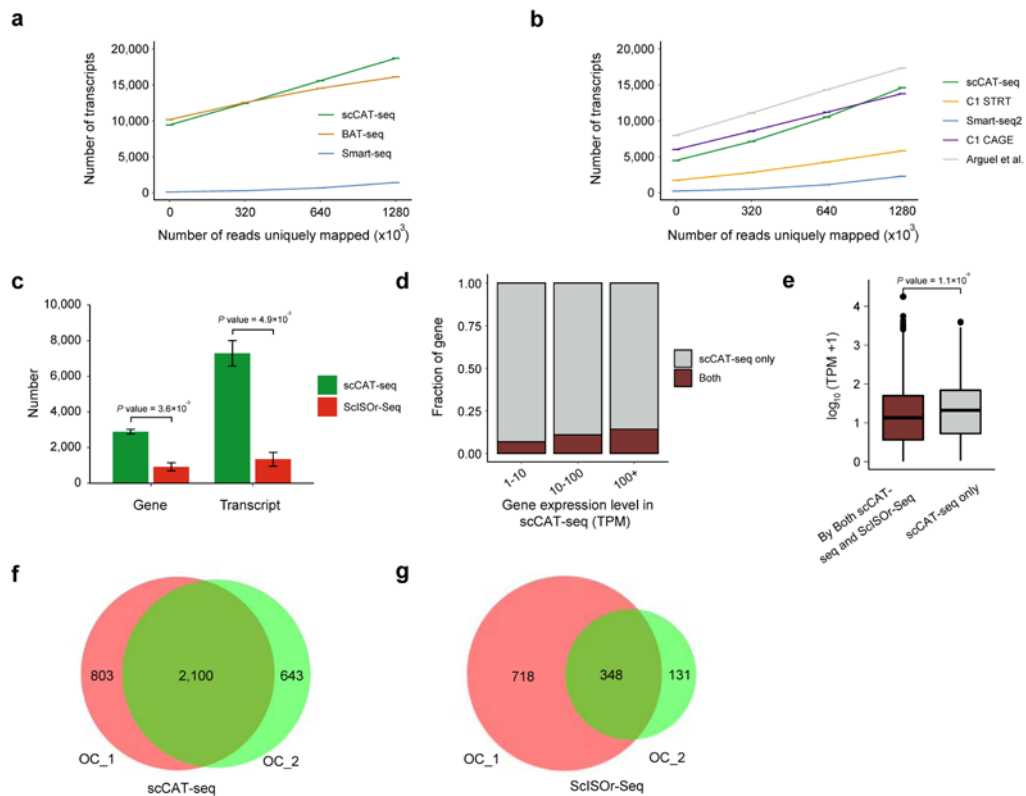
800



801

802 Supplementary figure 2. Accuracy and consistency of scCAT-seq performance for
803 isoform quantification. **a**, Scatter plot of observed transcript expression levels (y axis)
804 and true abundance (x axis) of ERCC spike-ins through 3'-end quantification. **b**,
805 Scatter plots showing the correlation of transcriptional level of isoforms between
806 replicated samples (3 cells pooled) based on 3'-end quantification. **c**, Comparison of
807 the cost for the same number of transcripts (1,000) between PacBio ScISO-seq,
808 scCAT-seq. The price is estimated based on the market price in China. Error bars
809 represent 95% confidence interval.

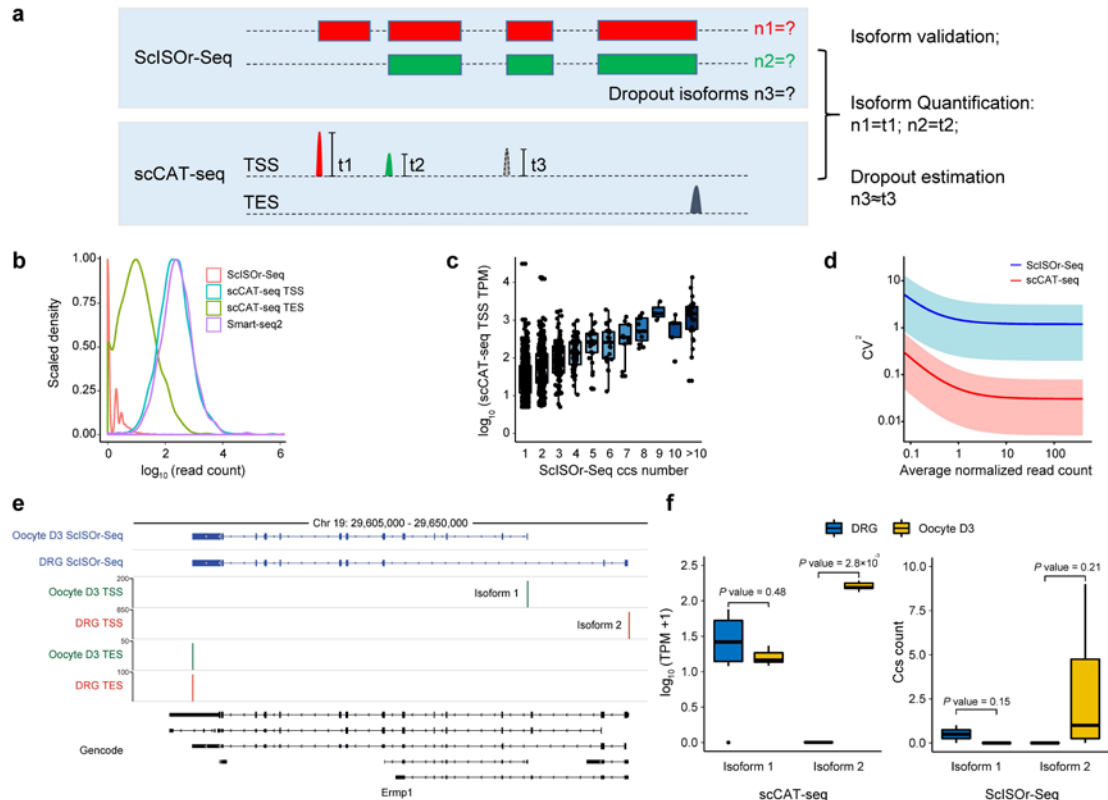
810



811

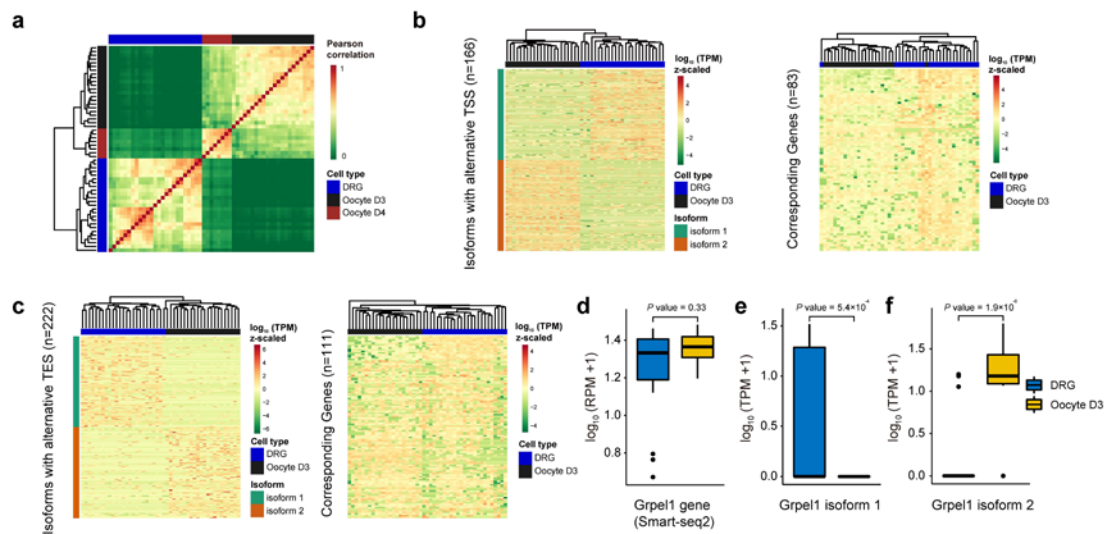
812 Supplementary figure 3. The number of genes covered by scCAT-seq. **a**, The number
 813 of transcripts with 3' tail detected by scCAT-seq, BAT-seq, and Smart-seq2 at variable
 814 sequencing depth. Error bars represents standard error of the mean. **b**, The number of
 815 transcripts with 5' head detected by scCAT-seq, C1 STRT, Smart-seq2, Arguel et al.,
 816 and C1 CAGE at variable sequencing depth. Error bars represents standard error of
 817 the mean. **c**, Number of genes and transcripts covered by scCAT-seq and ScISOr-Seq
 818 respectively (n = 3). The number of reads for scCAT-seq was 4 million per single cell
 819 and the CCS number for ScISOr-Seq is 50,000 per cell. Significance was computed
 820 using two sided t-test. Error bars represents standard error of the mean. **d**, Stacked
 821 barplots showing the number of genes with different expression levels detected in
 822 oocytes by scCAT-seq and ScISOr-Seq. **e**, Boxplot for expression level comparison
 823 between genes detected by scCAT-seq only and by both scCAT-seq and ScISOr-Seq (n
 824 = 9,626). Significance was computed using two-sided Wilcoxon test. **f**, Venn diagram
 825 for genes detected concordantly among single cells by scCAT-seq. **g**, Venn diagram
 826 for genes detected concordantly among single cells by ScISOr-Seq.

827



828

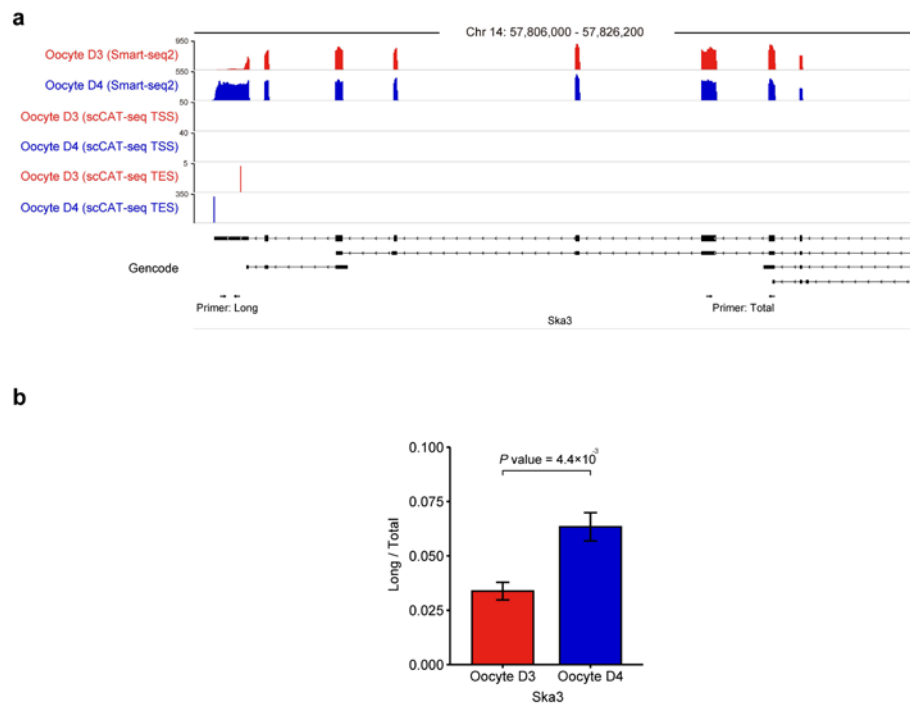
829 Supplementary figure 4. scCAT-seq improves upon the performance of scISOr-seq for
 830 single cell RNA isoform quantification. **a**, Schematic showing performance
 831 improvement of ScISOr-Seq via scCAT-seq quantification. **b**, Density plot showing
 832 the comparison of genes read count values among ScISOr-Seq, scCAT-seq and
 833 Smart-seq2. **c**, Boxplot for the expression level comparison among genes with
 834 different CCS numbers detected by ScISOr-Seq. **d**, Squared coefficients of variation
 835 of scCAT-seq and ScISOr-Seq, versus the means of normalized read counts. The
 836 shaded regions represent 95% confidence interval. **e**, Genome browser track showing
 837 an example of *Ermp1*, which has two isoforms detected by both scCAT-seq and
 838 ScISOr-Seq. **f**, Boxplot for the example gene *Ermp1*, which has two isoforms
 839 differentially expressed in oocytes or in DRG neurons, while the expression value
 840 assessed by ScISOr-Seq is not differential between the two cell types. Significance
 841 was computed using two-sided Wilcoxon test.



842

843 Supplementary figure 5. Identification and quantification of cell-type specific
844 transcript isoforms. **a**, Heatmap for Pearson's correlation coefficient of transcriptomes
845 of DRG neuron and oocytes, based on 3'-end quantification of RNA isoforms. **b**,
846 Heatmap showing RNA isoforms of alternative TSS choices with cell type specificity
847 (left panel), and the expression of corresponding genes assessed by Smart-seq2 (right
848 panel). **c**, Heatmap showing RNA isoforms of alternative TES choices with cell type
849 specificity (left panel), and the expression of corresponding genes assessed by
850 Smart-seq2 (right panel). **d-f**, Boxplot for the example gene *Grpel1*, which has two
851 isoforms differentially expressed in oocytes or in DRG neurons (e, f), while the
852 overall gene expression assessed by Smart-seq2 is not differential between the two
853 cell types (d). For **d-f**, significance was computed using two-sided Wilcoxon test (n =
854 35).

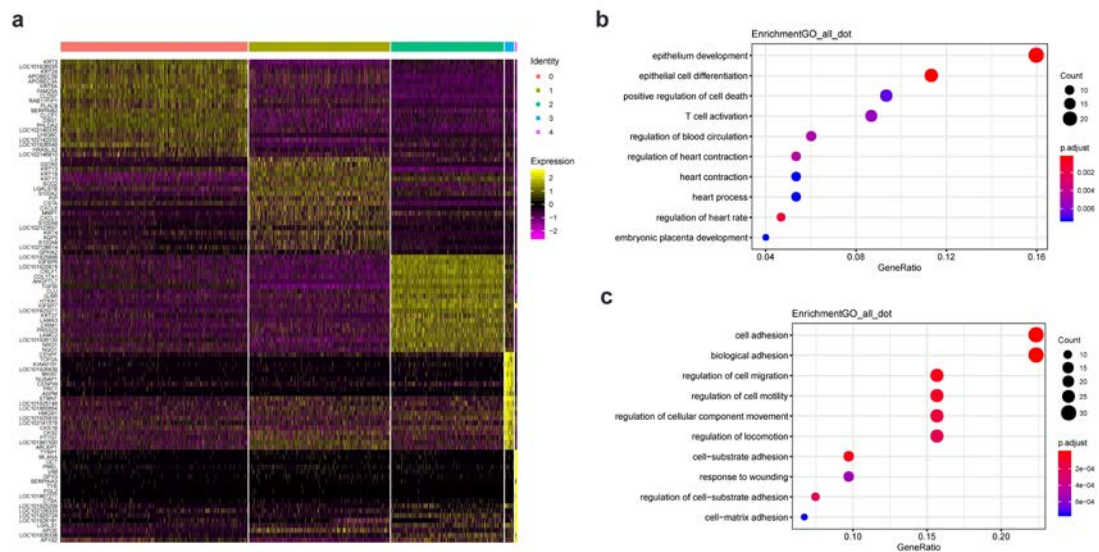
855



856

857 Supplementary figure 6. Examples of full-length isoforms with alternative TES during
858 oocyte post-ovulatory ageing. **a**, Genome browser track showing the TSS dynamic
859 choices during oocytes post-ovulatory ageing. **b**, Fold change in expression of the
860 *Ska3* long 3' UTR isoform (long) relative to total *Ska3* expression (total) between
861 oocyte D3 and oocyte D4 single cells, measured by RT-qPCR. Error bars represent
862 standard error of the mean (n = 6 biological replicates). Significance was computed
863 using two-sided t-test.

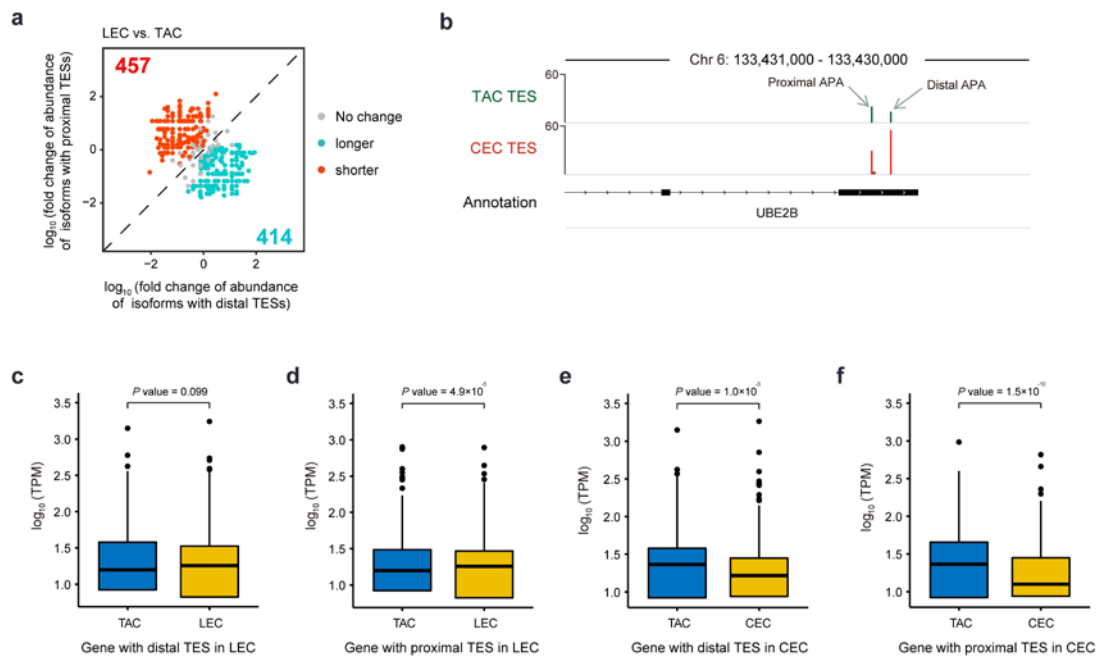
864



865

866 Supplementary figure 7. Cell-type and marker-gene identification in the crab-eating
867 monkey cornea. **a**, Heatmap shows the top 20 marker genes expressed in LEC, CEC
868 and TAC respectively. Color bars on the top are used to discriminate different cell
869 types. **b**, GO items enriched in the genes up-regulated in CEC compared to TAC. **c**,
870 GO items enriched in the genes down-regulated in CEC compared to TAC.

871



872

873 Supplementary figure 8. Differences between LEC and TAC in terms of RNA
874 expression and APA choices. **a**, Expression data with isoform specificity reveals
875 isoform expression differences between LEC and TAC (n = 956 genes). **b**, Genome
876 browser track shows an example of APA choices for the gene *UBE2B* during
877 differentiation of CEC from TAC (n = 414 genes). **c**, Boxplot comparing expression
878 of genes in LEC and TAC, which have distal TESs in LEC (n = 414 genes). **d**,
879 Boxplot comparing expression of genes in LEC and TAC, which have proximal TESs
880 in LEC (n = 457 genes). **e**, Boxplot comparing expression of genes in CEC and TAC,
881 which have distal TESs in CEC (n = 244 genes). **f**, Boxplot comparing expression of
882 genes in CEC and TAC, which have proximal TESs in CEC (n = 285 genes). For **c-f**,
883 significance was computed using two-sided Wilcoxon test.

884

885

886 Supplementary table 1.

Sample	Sequencing depth
ERCC_01	1,052,270
ERCC_02	1,571,949
ERCC_03	3,799,426
ERCC_04	3,989,246
ERCC_05	2,291,544
ERCC_06	3,835,792
ERCC_07	3,986,186
ERCC_08	7,964,775
ERCC_09	7,645,733
ERCC_10	5,223,629
Average	4,136,055

887

888 Sequencing depth for each ERCC spike-in libraries generated by scCAT-seq methods

889 are listed.

890

891 Supplementary table 2.

Library name	Sequencing Depth	TSS number	TES number	Gene number
D41_71	2,493,578	11,155	4,867	3,917
D44_52	3,100,138	12,906	5,667	4,594
D44_72	1,612,099	10,693	4,601	3,558
D45_52	2,723,321	12,858	4,866	4,066
D45_72	2,080,059	11,355	4,352	3,570
D46_71	3,955,396	13,570	5,547	4,674
D47_52	1,628,466	10,786	5,433	4,295
D47_72	1,056,585	9,567	4,276	3,232
D48_52	1,484,463	10,165	4,180	3,346
D48_72	1,869,191	10,710	4,407	3,501
D49_72	1,330,893	9,587	3,671	2,902
D50_52	1,518,348	10,201	4,288	3,474
D50_71	2,249,770	10,802	4,104	3,343
D50_72	2,717,055	11,725	5,136	4,262
D51_52	3,733,854	12,258	5,437	4,600
D51_72	2,474,653	11,314	3,860	3,253
D52_52	2,624,766	12,142	5,692	4,689
D52_72	4,015,130	13,347	5,869	4,998
Average	2,370,431	11,397	4,792	3,904

892

893 Information about 18 single DRG neurons generated by scCAT-seq method for further
894 benchmarking in this study are listed in this table. A gene is detected only if there are
895 TSS peaks and TES peaks mapped to the said gene.

896

897 Supplementary table 3. ScISOr-Seq information in this study

Sample	Subreads base (G)	Number of Subreads	Average subreads length	Number of CCS	Number of FLNC
DRG_1	0.32	214,295	1,487	19,947	4,289
DRG_2	0.35	237,777	1,481	23,070	3,835
OC_1	0.89	612,741	1,446	47,152	8,171
OC_2	0.24	165,171	1,436	13,305	780
OC_3	1.03	703,398	1,462	54,258	22,549
OC_4	3.52	2492,563	1,414	193,637	13,932
OC_5	0.22	145,585	1,483	14,857	1,009
OC_6	1.31	912,567	1,441	68,471	27,559

898

899 Information about 6 single oocytes and 2 single DRG neurons generated by scISO-seq
900 are listed.

901

902

903 Supplementary table 4. Cloning primers used in this study

Target gene	Sequences (5' → 3')
Novel gene 1 F	CTGCATCAGCTTCTGTTTCCT
Novel gene 1 R	GCTTAACAGTTTCGGAGGGT
Novel gene 2-1 F	CACTCCTCCACGGCCTC
Novel gene 2-1 R	TTCTTTACAGATATTTAAGGCACCC
Novel gene 2-2 F	GCTGGTCACGGTTGTACCTT
Novel gene 2-2 R	ATCATGGGAAGGGCATGAGC
Novel gene 3-1 F	TTACATGCTCTGACTTGGGCT
Novel gene 3-1 R	GTGTGCTCTGGCTTGCCATT
Novel gene 3-2 F	AGCCA ACTCTAAGATGGCACC
Novel gene 3-2 R	CTGAGCTTCGGTTTGGTGTG

904

905 Primer sequences used to clone full-length novel genes are listed.

906

907