

1 **Direct Comparative Analysis of 10X Genomics Chromium**  
2 **and Smart-seq2**

3 Xiliang Wang<sup>1#</sup>, Yao He<sup>2#</sup>, Qiming Zhang<sup>1</sup>, Xianwen Ren<sup>1,2</sup>, Zemin Zhang<sup>1,2\*</sup>

4

5 <sup>1</sup> BIOPIC, Beijing Advanced Innovation Center for Genomics, and School of  
6 Life Sciences, Peking University, Beijing 100871, China

7 <sup>2</sup> Peking-Tsinghua Center for Life Sciences, Academy for Advanced  
8 Interdisciplinary Studies, Peking University, Beijing 100871, China

9

10 # Equal contribution

11 \* Corresponding author

12 E-mail: zemin@pku.edu.cn (Zhang Z)

13

14 Running title: *Wang XL et al / Comparison of 10X and Smart-seq2*

15

16 There are 5870 words, 6 figures, 0 tables, 7 supplementary figures and 6  
17 supplementary tables.

18 **Abstract:**

19 Single cell RNA sequencing (scRNA-seq) is widely used for profiling  
20 transcriptomes of individual cells. The droplet-based 10X Genomics Chromium  
21 (10X) approach and the plate-based Smart-seq2 full-length method are two  
22 frequently-used scRNA-seq platforms, yet there are only a few thorough and  
23 systematic comparisons of their advantages and limitations. Here, by directly  
24 comparing the scRNA-seq data by the two platforms from the same samples of  
25 CD45<sup>+</sup> cells, we systematically evaluated their features using a wide spectrum  
26 of analysis. Smart-seq2 detected more genes in a cell, especially low  
27 abundance transcripts as well as alternatively spliced transcripts, but captured  
28 higher proportion of mitochondrial genes. The composite of Smart-seq2 data  
29 also resembled bulk RNA-seq data better. For 10X-based data, we observed  
30 higher noise for mRNA in the low expression level. Despite the poly(A)  
31 enrichment, approximately 10-30% of all detected transcripts by both platforms  
32 were from non-coding genes, with lncRNA accounting for a higher proportion in  
33 10X. 10X-based data displayed more severe dropout problem, especially for  
34 genes with lower expression levels. However, 10X-data can better detect rare  
35 cell types given its ability to cover a large number of cells. In addition, each  
36 platform detected different sets of differentially expressed genes between cell  
37 clusters, indicating the complementary nature of these technologies. Our  
38 comprehensive benchmark analysis offers the basis for selecting the optimal  
39 scRNA-seq strategy based on the objectives of each study.

40

41 **KEYWORDS:** Single cell RNA sequencing; 10X; Smart-Seq2; Comparison.

42

## 43 **Introduction**

44 Following the first single-cell RNA sequencing (scRNA-seq) method developed  
45 in 2009 [1], scRNA-seq has dramatically influenced many research fields  
46 ranging from cancer biology, stem cell biology to immunology [2-5]. Compared  
47 with RNA-seq of bulk tissues with millions of cells, scRNA-seq offers the  
48 opportunity to dissect the composition of tissues and the dynamic of  
49 transcriptional states, as well as to discover rare cell types. With the  
50 improvement of sequencing technologies, scRNA-seq is becoming robust and  
51 broadly accessible to perform transcriptome analysis [6].

52 Two scRNA-seq platforms are frequently used [7, 8]: Smart-seq2 [9] and  
53 10X (10X Genomics Chromium, 10X Genomics, Pleasanton, CA). Smart-seq2  
54 is based on microtiter plates [10, 11], where mRNA is isolated and reverse  
55 transcribed to cDNA for high-throughput sequencing for each cell [12]. Reads  
56 mapped to a gene are used to quantify its expression in each cell, and TPM  
57 (Transcripts Per Kilobase Million) is a common metric of expression  
58 normalization [13, 14]. By contrast, 10X is a droplet-based scRNA-seq method,  
59 allowing genome-wide expression profiling for thousands of cells at once. The  
60 UMI (unique molecular identifier) is used to directly quantify the expression  
61 level of each gene [15]. Both TPM (Smart-seq2) and normalized UMI (10X) is  
62 analyzed to detect HVGs (highly variable genes), which are often used for  
63 either cellular phenotype classification or new subpopulation identification [16].

64 Although each platform has its own expected advantages and drawbacks  
65 based on the design of each method, there are only a few systematic  
66 comparisons of Smart-seq2 and 10X [17, 18]. Here, we applied these two  
67 technologies to the same set of samples, and directly compared the sensitivity  
68 (the probability to detect transcripts present in a single cell), precision  
69 (variation of the quantification), and power (subpopulation identification) of  
70 these two platforms.

71

## 72 **Results**

### 73 **Data generation and evaluation**

74 Our data were derived from two cancer patients. For the first patient,  
75 diagnosed to have hepatocellular carcinoma (HCC), we collected the liver  
76 tumor (LT) and its adjacent non-tumor tissue (NT). For the second patient,  
77 diagnosed to have rectal cancer with liver metastasis, we collected both the  
78 primary tumor (PT) and the metastasized tumor (MT). For each sample, we  
79 used fluorescence activated cell sorting (FACS) to obtain CD45<sup>-</sup> cells, and  
80 used both 10X and Smart-seq2 to perform scRNA-seq analysis. Following the  
81 standard experimental protocols, we obtained 10X data for 1,338, 1,305, 746,  
82 and 5,282 cells for LT, MT, NT, and PT tissues, respectively, and obtained  
83 Smart-seq2 data for 94, 183, 189, and 135 cells for the corresponding tissues  
84 (Table S1). Bulk RNA-seq data of those four samples were also generated.

85 We first examined the read counts for each cell derived from both platforms.  
86 The average total reads of each cell from Smart-seq2 were 6.2M, 1.7M, 6.3M,  
87 and 1.7M for LT, MT, NT, and PT, respectively, whereas 10X obtained relatively  
88 lower reads as followings: 59K, 34K, 92K, and 20K for the corresponding  
89 tissues respectively (**Figure 1A** and Figure S1A). For transcriptome analysis,  
90 we followed conventional practice and selected uniquely mapped reads in the  
91 genome for downstream analysis. The number of uniquely mapped reads was  
92 nearly 10-fold higher in Smart-seq2 (Figure S2A). Although, the 3' ends of  
93 genes have been reported to have higher homology than other parts of the  
94 genome, leading to increased level of multi-alignments [19], our results  
95 showed that the unique mapping ratios were similar, at approximately 80% for  
96 both datasets (Figure S2A).

97 As has been reported [20], damaged cells exhibited higher representation  
98 of genes in the “membrane” ontology category, but lower representation in the  
99 “extracellular region” and “cytoplasm” categories, when compared to  
100 high-quality cells. However, we did not observe obvious differences in term of



101 “extracellular region” category between those two scRNA-seq platforms  
102 (Figure 1B and Figure S1B). For Smart-seq2, the “membrane” category was  
103 over-represented (Figure 1C and Figure S1C) (all  $P < 1.0E-4$ , two-sided t-test)  
104 and “cytoplasm” category under-represented (Figure 1D and Figure S1D) (all  
105  $P < 1.0E-10$ , two-sided t-test), implying more complete lysis of membranes.

106 Cell cycle has a major impact on gene expression [21], and is an important  
107 confounding factor of cell subpopulation classification [22]. We used an  
108 established method [23] to classify cells into cell cycle phases based on gene  
109 expression (Figure S2B). The distributions of cells in G1, G2/M, and S phases  
110 were similar between the two platforms for all samples we studied (Figure 1E  
111 and Figure S1E).

### 112 **Higher proportion of mitochondrial genes for Smart-seq2 and** 113 **ribosome-related genes for 10X**

114 One metric we used to examine cell qualities is the proportion of reads  
115 mapped to genes in the mitochondrial genome [24]. High levels of  
116 mitochondrial reads are indicative of poor quality, likely resulting from  
117 increased apoptosis and/or loss of cytoplasmic RNA from lysed cells [20]. Most  
118 cells from 10X contained a much lower abundance of mitochondrial genes  
119 ranging from 0-15% of their total RNA. By contrast, the mitochondrial  
120 proportion from Smart-seq2 was 3.8-10.1 folds higher, at a level similar with  
121 bulk RNA-seq data (Figure 1F and Figure S1F). Such high proportions (an  
122 average of approximately 30%) by both Smart-seq2 and bulk RNA-seq were  
123 likely caused by more thorough disruption of organelle membranes by the  
124 Smart-seq2 and the standard bulk RNA-seq protocols than the relatively weak  
125 cell lysis procedure by 10X. Abnormally high proportion (such as  $> 50\%$ ) may  
126 reflect poor cell quality from Smart-seq2 in this study. However, caveats should  
127 be considered when examining mitochondrial genes, because naturally larger  
128 mitochondrial proportions can be expected from certain cells such as  
129 cardiomyocytes (58-86%) [25] or those in apoptosis [20].

130 Ribosome-related genes (genes in “ribosome” GO term) accounted for a

131 large portion of detected transcripts by 10X, 3.6-8.2 folds higher than  
132 Smart-seq2 data (Figure 1G and Figure S1G). Indeed, 10X detected genes  
133 were enriched in the “ribosome” GO term, rather than ribosomal DNA (rDNA).  
134 The proportion of sequencing reads assigned to rDNA were only 0.03-0.4% in  
135 10X, significantly lower than those by Smart-seq2 (10.2-28.0%). Few reads  
136 were uniquely mapped among those reads (Figure S1H), therefore removing  
137 non-uniquely mapped reads was essential to minimizing rDNA interference in  
138 Smart-seq2.

139 **10X detected a higher proportion of lncRNA and Smart-seq2 identified**  
140 **more lncRNA as highly variable genes**

141 Despite that both Smart-seq2 and 10X followed the poly-A enrichment strategy,  
142 approximately 10-30% of all detected transcripts were from non-coding genes  
143 (Figure 2A and Figure S3A), with lncRNA accounting for 2.9-3.8% in  
144 Smart-seq2 and relatively higher (6.5-9.6%) in 10X (Figure 2B and Figure  
145 S3B). In total, protein-coding genes and lncRNA accounted for 80.5-92.6% of  
146 all detected transcripts for Smart-seq2, and 77.4-99.2% for 10X. Other classes  
147 of RNAs and/or their precursor were also detected with a great variance  
148 among experiments. Among protein-coding genes, the proportions of  
149 house-keeping (HK) genes and transcriptional factors (TFs) were 1.7-2.5 and  
150 1.1-1.4 folds higher in 10X, respectively (Figure 2C-2D and Figure S3C-S3D).

151 One common method to cluster in scRNA-seq datasets was to identify  
152 highly variable gene (HVG) [26, 27], which assumed that large variation in  
153 gene expression across cells mainly come from biological difference rather  
154 than technical noise. We selected the top 1,000 HVGs, and found 333 HVGs  
155 shared between two platforms (Figure 2E). Smart-seq2 specific HVGs only  
156 enriched two KEGG pathways, while 10X specific HVGs enriched 34 pathways,  
157 including common pathways in cancer, such as “PI3K–Akt signaling pathway”  
158 (Figure S3E), suggesting that HVGs identified by 10X were more conducive to  
159 understanding biological difference among samples. Protein-coding genes  
160 accounted for 94.9%, 22.3%, and 92.8% of shared, Smart-seq2 specific, and

161 10X specific HVGs, respectively (Figure 2F). Huge differences in HVGs come  
162 from the lncRNA which has been previously shown to be expressed with  
163 biological function in scRNA-seq [19]. The enrichment of lncRNA in  
164 Smart-seq2-specific HVGs, which resulted in a few enriched KEGG pathways,  
165 may be caused by specific sub-populations which predominantly expressed  
166 those lncRNA [28, 29]. The possible reasons may lead to less lncRNA  
167 identified as HVGs in 10X as follows: lncRNA was detected at much lower  
168 levels than protein-coding genes [30, 31], and higher dropout ratio.

### 169 **Smart-seq2 detected more genes and 10X identified more cell clusters**

170 We first assessed the gene-detection sensitivity, represented as the number of  
171 detected genes (TPM > 0 or UMI > 0) per cell [32]. Smart-seq2 had  
172 significantly higher sensitivity, capturing an average of 5,713, 4,761, 4,079,  
173 and 3,860 genes per cell for LT, MT, NT, and PT, respectively, compared to  
174 2,682, 1,853, 2,123, and 1,104 genes for 10X, respectively (Figure 3A and  
175 Figure S4A). In total, more than 25,000 genes were covered from each sample  
176 by Smart-seq2; however, despite a magnitude more cells captured by 10X,  
177 approximately 20% genes were still dropped out (Figure 3B and Figure S4B).  
178 For detected genes, Smart-seq2 data showed a unimodal distribution with few  
179 low-expressed genes detected in all cells. By contrast, 10X data showed an  
180 obvious bimodal distribution due to a large number of genes with near-zero  
181 expression (Figure 3C and Figure S4C), suggesting higher noise or random  
182 capture of mRNA at very low expression level.

183 To examine the expression dynamic ranges covered by each platform, we  
184 determined the expression levels reaching saturation. All genes were divided  
185 into four quartiles by expression values. While sequencing depths of all four  
186 quartiles were saturated for Smart-seq2, only upper two quartiles were  
187 adequate for 10X (Figure 3D and Figure S4D), suggesting that Smart-seq2  
188 has advantages in detecting genes at low expressed levels. Meanwhile, the  
189 top 10 most highly expressed genes accounted for 33.0-38.5% of total counts  
190 in Smart-seq2 and 18.4-33.0% in 10X (Figure 3E and Figure S4E). Those 10

191 genes were dominated by mitochondrial genes, especially in Smart-seq2.  
192 Moreover, bulk RNA-seq data showed strikingly similar results to Smart-seq2  
193 (Table S2).

194 We next determined if the two platforms covered different sets of genes.  
195 For any given sample, approximately 2/3 of genes present in the upper quartile  
196 were shared between the two platforms, leaving the remaining 1/3 genes  
197 distinct (Figure 3F and Figure S4F). Analysis of the distinct genes represented  
198 indicated that 5.6% of 10X detected genes had full KEGG annotation, whereas  
199 only 2.7% of Smart-seq2 detected genes were annotated (Table S3). Thus,  
200 Smart-seq2 is better equipped at finding genes with unknown functions. In  
201 addition, Smart-seq2 shared more genes with bulk RNA-seq (Figure 3F and  
202 Figure S4F). PCC of each gene between bulk RNA-seq and the averaged  
203 Smart-seq2 single cell output was higher (Figure 3G and Figure S4G), again  
204 showing more similarity between Smart-seq2 and bulk RNA-seq.

205 HVGs were used to cluster cells into putative subpopulations, which was  
206 one of the most common goals of an scRNA-seq experiment. 11 clusters were  
207 identified in 10X using Seurat (version 2.3.4) [33]. By applying conventional  
208 cell markers, those clusters were annotated as fibroblasts, epithelial cells,  
209 endothelial cells, and two special clusters: “hepatocyte” and “malignant cell”,  
210 which highly expressed their respective markers, such as, ALB and SERPINA1  
211 in hepatocyte, STMN1, H2AFZ, CKS1B, and TUBA1B in malignant cells [34,  
212 35] (Figure 4A). By contrast, only five clusters were identified in Smart-seq2  
213 due to limited cell number, these clusters were annotated as epithelial cells,  
214 endothelial cells and fibroblasts (Figure 4B). Four clusters of tumor fibroblasts  
215 were identified in 10X: cluster 0, cluster 2, cluster 5 and cluster 10 (Figure 4A).  
216 Cluster 0 cells showed fibroblasts signatures (RGS5 and NDUFA4L2), cluster  
217 2 cells had strong expression of CAF (cancer associated fibroblasts) cell  
218 markers (LUM, SFRP4, and COL1A1), cluster 5 cells expressed  
219 myofibroblasts markers (MYH11, TAGLN, and ACTA2). We also highlight a  
220 fibroblasts cluster (cluster 10) with a striking enrichment for mitochondrial

221 genes (MT-ND2, MT-CO3, and MT-CO2). Smart-seq2 only identified two  
222 fibroblasts subtypes, with cluster 2 cells expressing fibroblasts signatures  
223 (RGS5 and NDUFA4L2), and cluster 4 cells showing CAF markers (LUM, DCN,  
224 and FBLN1).

225 We next examined if the two platforms covered different sets of  
226 differentially expressed genes (DEGs). We first identified DEGs within each  
227 sample compared to all other samples (Figure 4C and Figure S5A). 10X  
228 detected more DEGs, and less than 50% of total DEGs were shared between  
229 two platforms, leaving the remaining genes distinct. For example, 864 DEGs  
230 were identified between LT and other samples using 10X, and 20 KEGG  
231 pathways were enriched. Such number were 638 DEGs and 22 pathways for  
232 Smart-seq2, respectively. Only 214 DEGs (Figure 4C) and 11 pathways  
233 (Figure 4D) were shared. Considering up-regulated DEGs and down-regulated  
234 genes separately, less than 50% DEGs were shared between two platforms as  
235 well (Figure S5B). Moreover, we observed a few DEGs with conflicting  
236 directions (Table S4). We furthermore identified DEGs within each cell type  
237 compared to all other cell types (Figure 4E and Figure S5C). The same  
238 tendency was also found with several conflicted DEGs (Table S5). Exemplified  
239 with fibroblasts, 876 DEGs were identified between fibroblasts and other type  
240 cells, and enriched in 30 KEGG pathways in 10X, whereas 776 DEGs  
241 identified and 23 pathways enriched in Smart-seq2. Only 352 DEGs (Figure 4E)  
242 and 11 pathways (Figure 4F) were shared. In summary, the concordance  
243 between DEGs and enriched KEGG pathways by Smart-seq2 and 10X was  
244 limited, suggesting that the choice of platform indeed have an impact on the  
245 results. Notably, the “Ribosome” pathway was spotted in 10X results (Figure  
246 1G, Figure 4D and 4F, Figure S3E), showing gene detection bias of 10X.

#### 247 **10X had higher dropout ratio than Smart-seq2**

248 Dropout events in scRNA-seq can result in many genes undetected and an  
249 excess of expression value of zero, leading to challenges in differential  
250 expression analysis [21, 36]. The average dropout ratios of majority genes in

251 10X were 1.3 to 1.4-fold higher for all samples tested (Figure 5A and Figure  
252 S6A). For example, the widely used HK gene ACTB had no dropout in  
253 Smart-seq2, whereas 2.8-5.9% dropout ratios were observed in 10X. Similarly,  
254 GAPDH had dropout ratios from 0-0.67% in Smart-seq2 but 4.2-18.8% in 10X  
255 (Figure 5B and Figure S6B).

256 The frequency of dropout events was correlated to gene expression levels,  
257 which can be fitted by a modified non-linear Michaelis-Menten equation  
258 introduced in the M3Drop package (<https://github.com/tallulandrews/M3Drop>).  
259 Genes with lower expression levels had higher dropout ratios (Figure 5C and  
260 Figure S6C), consistent with a previous report [37]. Mitochondrial genes were  
261 the least likely to be dropped out, especially in Smart-seq2 (Table S6). In both  
262 platforms, genes with lower abundance were detected in smaller number of  
263 cells, and those genes could lead to higher noise, especially in 10X (Figure 5D  
264 and Figure S6D). Because that genes with near-zero expression are noise  
265 without enough information for reliable statistical inference [38], removal of  
266 them may mitigate noise level and reduce the amount of computation without  
267 much loss of information.

268 We also found that the gene expression coefficient of variation (CV) across  
269 cells were associated with dropout ratios. 10X had more genes with large CV  
270 than Smart-seq2 (Figure 5E and Figure S6E). While genes with large CV  
271 generally had lower expression, especially for 10X (Figure 5F and Figure S6F),  
272 genes with larger CV also had higher dropout ratio (Figure S6G). For example,  
273 genes with CV larger than 800 had > 80% of dropout ratio in Smart-seq2, near  
274 100% of dropout in 10X (Figure 5G and Figure S6H).

### 275 **Difference in capture of gene structural information**

276 We finally evaluated how each of the two platforms capture the gene  
277 structural information. We first confirmed that the 10X reads showed a strong  
278 bias toward the 3' ends of mRNAs as expected, while Smart-seq2 reads  
279 were more uniformly distributed in the gene bodies (Figure 6A-6B and Figure  
280 S7A-S7B). For Smart-seq2, our sequencing depth was adequate for junction

281 detection, evidenced by the number of detected known junctions reaching a  
282 plateau (Figure 6C and Figure S7C). The 10X data were not equipped for  
283 alternative splicing analysis due to the 3'-bias (Figure 6C and Figure S7C).  
284 Nevertheless, 10X still detected non-negligible number of junctions, even  
285 though they only accounted for approximately 50% of those junctions detected  
286 by Smart-seq2. Although Smart-seq2 data were clearly much more suitable for  
287 alternative splicing studies [39, 40], the limited number of splicing junctions  
288 detected by 10X might be suitable for certain analyses that rely on  
289 junction-based characterization, such as the RNA velocity analysis [41].

290 To evaluate whether gene lengths would introduce any bias in either of the  
291 platforms, we examined the correlation between the two platforms in terms of  
292 gene length and expression level. All calculated PCCs were near perfect for all  
293 tested samples (Figure 6D and Figure S7D), demonstrating that mRNA  
294 molecular quantification was not influenced by either full-length or 3' capture  
295 strategies.

296

## 297 **Discussion**

298 Here we comprehensively evaluated two scRNA-seq platforms: Smart-seq2  
299 was more sensitive for gene detection, and 10X had more noise and higher  
300 dropout ratio. 10X could detect rare cell populations due to high cell throughput.  
301 Both platforms had similar results in unique mapping ratio and assigning cells  
302 into different cell cycle phase. Smart-seq2 had better performance in detection  
303 of genes with low expression levels and of splicing junction. In terms of  
304 defining HVGs and detection DEGs, each platform showed unique strength  
305 with limited overlap and they could provide complementary information.  
306 However, there are some limitations that should be acknowledged in our study.  
307 Firstly, the analysis of dropout rates was influenced by the large difference in  
308 sequencing depth of those two platforms. Considering an intrinsic property of  
309 the two methods, we did not perform downsample to equal sequencing  
310 coverage. Secondly, we only sequenced 94-189 cells per sample with the



311 Smart-seq2 protocol, which may reduce the power to detect groups of cells. As  
312 has been previously shown, Smart-seq2 libraries should contain ~70 cells per  
313 cluster to achieve decent power [42]. Lastly, UMI counts and read counts have  
314 different mean distributions, namely the negative binomial model is a good  
315 approximation for UMI counts, and zero-inflated negative binomial for read  
316 counts [43], which may impair the CV measure because that CV is linked to  
317 the mean gene expression levels.

318 The advantage of scRNA-seq crucially depends on two parameters: cell  
319 number and sample complexity. These two parameters can be designed and  
320 chosen based on study objectives. The number of cells is a key determinant  
321 for profiling the cell composition. In this study, several hundreds of cells could  
322 capture abundant, but not rare, cell types using Smart-seq2. Thousands of  
323 cells or more could capture unique cell subtypes in both Smart-seq2 and 10X.  
324 Thus, we believe that the range of sample sizes in our comparisons are  
325 relevant for other study. In a heterogenous population where the cellular states  
326 are transcriptionally distinct and equally distributed, 1,000-2,000 single cells  
327 could be sufficient for de novo clustering of the different cell states [44].

328 However, the cost is still prohibitive for studies that involve hundreds of  
329 thousands of cells even at low sequencing depths [7]. It seems a now standard  
330 practice to investigate tens of thousands of cells in a published paper. The cost  
331 is certainly an important factor for the optimal selection of the cell number.  
332 Smart-seq2 is not restricted by cell size, shape, homogeneity, and cell number,  
333 and thus is an efficient method to uncover an in-depth characteristic of a rare  
334 cell population such as germ cells. However, its overall cost is very high, and  
335 the laborious nature and technical variability can be intimidating because the  
336 reactions are carried out in individual wells for Smart-seq2 [42]. The huge  
337 advantage of 10X is the low cost and high throughput, making it better for  
338 complex experiments such as multiple treatments. Although many cells of  
339 each sample were added to each channel for 10X in our study, we just  
340 obtained 746, 1,305, 1,338, and 5,282 cells by CellRanger (version 2.2,



341 <http://www.10xgenomics.com/>). 10X cannot guarantee the yield of cells, and  
342 cell number may fluctuate wildly among experiments. For example, 60-4,930  
343 cells among 68 samples [45], and 1,052-7,247 cells among 25 samples [46]  
344 were obtained in two reports, respectively. The huge variability may come from  
345 tissue/cell types, inaccurate estimation of input cell number, or poor conditions  
346 and death of cells during experiments. Dataset from a small number of cells is  
347 not adequate to reflect fully the biological image [47]. Therefore, the trade-off  
348 between Smart-seq2 and 10X should be carefully assessed depending on  
349 data throughput and ultimate study objectives.

350 Samples generally contain a mixture of cells at different phases. However,  
351 effects of cell cycle cannot be avoided by simply removing cell cycle marker  
352 genes, as the cell cycle can affect many other genes [48, 49]. To date, our  
353 results demonstrated that Smart-seq2 and 10X have similar power in  
354 assigning cells into different cyclic phases.

355 The scRNA-seq provides biological resolution that cannot be attained by  
356 bulk RNA-seq, at a cost of increased noise [50]. Reliable capture of transcripts  
357 into cDNA for sequencing is difficult for the low abundance genes in a single  
358 cell, which increases the frequency of dropout events. This was more  
359 noticeable in 10X (Figure 5C). Moreover, 10X may capture some ambient  
360 transcript molecules that float in droplet due to cell lysis or cell death [19],  
361 which also results in noise, however, increased capture single cells could  
362 compensate the inefficacy brought by noise and provide a more robust  
363 clustering. By contrast, Smart-seq2 had less noise and higher sensitivity but  
364 high cost, therefore the sample size attribute in Smart-seq2 and 10X should be  
365 established on rigorous design and well-defined rationale.

366

## 367 **Conclusions**

368 Here we comprehensively evaluated two scRNA-seq platforms from the  
369 aspects of sensitivity, precision and power: Smart-seq2 was more sensitive for  
370 gene detection, and 10X had more noise and higher dropout ratio. 10X could

371 detect rare cell populations due to high cell throughput. Both platforms had  
372 similar results in unique mapping ratio and assigning cells into different cell  
373 cycle phase. Smart-seq2 had better performance in detection of genes with  
374 low expression levels and of splicing junction. In terms of defining HVGs and  
375 detection DEGs, each platform showed unique strength with limited overlap  
376 and they could provide complementary information.

377

## 378 **Materials and methods**

### 379 **Sample collection and single cell processing**

380 Tumor tissue of two donors were obtained from about 2cm far from tumor  
381 edge, and adjacent normal liver tissues (donor 20170608) were located at  
382 least 2cm far from the matched tumor tissue. Those fresh tissue were cut  
383 into pieces about 1mm<sup>3</sup> and digested with MACS tumor dissociation kit for  
384 30min. Suspended cells were filtered with 70µm Cell-Strainer (BD) in the  
385 RPMI-1640 medium (Invitrogen), then centrifuged at 400g for 5min, and the  
386 supernatant was removed. To lyse red blood cells, pelleted cells were  
387 suspended in red blood cell lysis buffer (Solarbio) and incubated on ice for  
388 2min. Finally, cell pellets were resuspended in sorting buffer after washed  
389 twice with 1x PBS.

### 390 **Single cell RNA-seq**

391 Based on fluorescence activated cell sorting (FACS) analysis (BD Aria III  
392 instrument), CD45 (eBioscience, cat. no. 11-0459) was used to separate  
393 CD45+ and CD45- cells. Cells were sorted into 1.5mL low binding tubes  
394 (Eppendorf) with 50mL sorting buffer, and into wells of 96-well plates  
395 (Axygen) with lysis buffer, which contained 1µL 10mM dNTP mix  
396 (Fermentas), 1µL 10µM Oligo(dT) primer, 1.9µL 1% Triton X-100 (Sigma)  
397 plus 0.1µL 40U/µL RNase Inhibitor (Takara).

398 For 10X, single cells were processed with the GemCode Single Cell  
399 Platform using the GemCode Gel Bead, Chip and Library Kits (10x

400 Genomics, Pleasanton) following the manufacturer's protocol. Samples were  
401 processed using kits pertaining to the V2 barcoding chemistry of 10x  
402 Genomics. Estimated 10,000 cells were added to each channel with the  
403 average recovery rate 2,000 cells. Libraries were sequenced on Hiseq 4000  
404 (Illumina).

405 For Smart-seq2, transcripts reverse transcription and amplification were  
406 performed according to Smart-seq2's protocol. We purified the amplified  
407 cDNA products with 1x Agencourt XP DNA beads (Beckman), then  
408 performed quantification of cDNA of every single cell with qPCR of GAPDH,  
409 and fragment analysis using fragment analyzer AATI. To eliminate short  
410 fragments (less than 500 bp), cDNA products with high quality were further  
411 cleaned using 0.5x Agencourt XP DNA beads (Beckman). The concentration  
412 of each sample was quantified using Qubit HsDNA kits (Invitrogen). Libraries  
413 were constructed with the TruePrep DNA Library Prep Kit V2 (Vazyme  
414 Biotech), and sequenced on Hiseq 4000 (Illumina) in paired-end 150bp.

#### 415 **Bulk RNA isolation and sequencing**

416 After surgical resection, tissue was firstly stored in RNAlater RNA  
417 stabilization reagent (QIAGEN) and kept on ice. Total RNA was extracted  
418 using the RNeasy Mini Kit (QIAGEN) according to the manufacturer's  
419 instructions. Concentration of RNA was quantified using the NanoDrop  
420 instrument (Thermo), and quality of RNA was evaluated with fragment  
421 analyzer (AATI). Libraries were constructed using NEBNext Poly(A) mRNA  
422 Magnetic Isolation Module kit (NEB) and NEBNext Ultra RNA Library Prep  
423 Kit (NEB), and sequenced on Hiseq 4000 (Illumina) in paired-end 150bp.

#### 424 **Data reference**

425 We used the GRCH38 human genome assembly as reference, which was  
426 downloaded from the Ensembl database (Ensembl 88)  
427 (<http://asia.ensembl.org>). The protein coding genes and lncRNA were  
428 categorized based on an Ensembl annotation file in the GTF format. Among  
429 those non-coding genes, rRNAs, tRNAs, miRNAs, snoRNAs, snRNA and

430 other known classes of RNAs were excluded, and lncRNA were defined as  
431 all non-coding genes longer than 200 nucleotides and not belonging to other  
432 RNA categories.

433 We retrieved the signature genes (extracellular region, cytoplasm,  
434 mitochondrion, ribosome, apoptotic process, metabolic process, membrane,  
435 and cell cycle) from the gene ontology database (GO:0005576, GO:0005737,  
436 GO:0005739, GO:0005840, GO:0006915, GO:0008152, GO:0016020, and  
437 GO:0007049, respectively) (<http://geneontology.org/>). A list of human TFs  
438 was obtained from the “Animal Transcription Factor Database”  
439 (<http://bioinfo.life.hust.edu.cn/AnimalTFDB/>).

#### 440 **Quality control for scRNA**

441 For Smart-seq2, sequenced reads were mapped to GRCH38 using the  
442 STAR aligner (version 2.6.0a) with the default parameters. These uniquely  
443 mapped reads in the genome were used, and reads aligned to more than  
444 one locus were discarded. The expression level of gene was quantified by  
445 the TPM value. Genes expressed (TPM > 0) in less than 10 cells were  
446 filtered out. Cells were removed according to the following criteria: (1) cells  
447 that had fewer than 800 genes and (2) cells that had over 50% reads  
448 mapped to mitochondrial genes.

449 For 10X, an expression matrix of each sample was obtained using the  
450 Cell Ranger toolkit (version 2.2, <https://www.10xgenomics.com/>). Genes  
451 presented (UMI > 0) in less than 10 cells were filtered out. Cells were  
452 removed according to the following criteria: (1) cells that had fewer than 500  
453 genes; (2) cells that had fewer than 900 UMI or over 8000 UMI; and (3) cells  
454 that had more than 20% of mitochondrial UMI counts.

#### 455 **CV**

456 The coefficient of variation (CV) is a standardized measure of dispersion of a  
457 probability distribution or frequency distribution. It is defined as the ratio of  
458 the standard deviation (SD) to the mean, namely  $CV = 100 \cdot SD / \text{mean}$

#### 459 **Cell cycle**

460 We used the reported method [23] to classify cells into cell cycle phases based  
461 on gene expression. Cells were classified as being in G1 phase if the G1 score  
462 is above 0.5 and greater than the G2/M score; in G2/M phase if the G2/M  
463 score is above 0.5 and greater than the G1 score; and in S phase if neither  
464 score is above 0.5 [51].

#### 465 **Reads distribution in genome and junction detection**

466 To demonstrate the bias of reads distribution in genome, we calculated reads  
467 distribution over genome features, including coding sequence (CDS), 5'-  
468 untranslated region (UTR), 3'-UTR, intron, TSS\_up\_10kb (10kb upstream of  
469 transcription start site), and TES\_down\_10kb (10kb downstream of  
470 transcription end site). When genome features were overlapped, they were  
471 prioritized as follows: CDS > UTR > Intron > others.

472 We assessed sequencing depth for splicing junction detection by randomly  
473 resampling total alignments with an interval of 5%, and then detected known  
474 splice junctions from the reference gene model in GTF format.

#### 475 **Saturation analysis**

476 We resampled a series of alignment subsets (5%, 10% - 100%) and then  
477 calculated RPKM value to assess sequencing saturation, which had been  
478 described [52]. “Percent Relative Error” was used to measure how the RPKM  
479 estimated from subset of reads ( $RPKM_{est}$ ) deviates from real expression level  
480 ( $RPKM_{real}$ ). The RPKM estimated from total reads was used as approximate  
481  $RPKM_{real}$ : Percent Relative Error =  $100 * (| RPKM_{est} - RPKM_{real} |) / RPKM_{real}$ .

#### 482 **Cell clustering**

483 After filtration, a merged expression matrix of four samples was used for  
484 cell clustering by the Seurat package (version 2.3), adapting the typical  
485 pipeline [33]. In brief, gene expression was normalized by the  
486 “NormalizeData” function. Highly variable genes were calculated with the  
487 Find Variable Genes method with the default parameters. Data was scaled  
488 with mitochondrial count ratio of a cell for Smart-seq2, with total UMI number  
489 and mitochondrial count ratio of a cell for 10X. Those HVGs were used for

490 Canonical Correlation Analysis (CCA), which was used to remove batch  
491 effects of patients. Cells were clustered by the “FindClusters” method using  
492 the first 20 CCs, and UMAP was used to visualization. Subsequently, cell  
493 clusters were annotated manually, based on known markers. Hepatocyte  
494 marker genes were ALB and SERPINA1, malignant cell marker genes were  
495 STMN1, H2AFZ, CKS1B, and TUBA1B, fibroblast marker genes were RGS5  
496 and NDUFA4L2, CAF (cancer associated fibroblast) marker genes were LUM,  
497 SFRP4, DCN, FBLN1 and COL1A1, and myofibroblast marker genes were  
498 MYH11, TAGLN, and ACTA2.

#### 499 **Data visualization and statistics**

500 Microsoft R Open (version 3.5.1, <https://mran.microsoft.com/>) was used, and  
501 ggplot2 package (version 3.1.0) were used to generate data graphs. Data  
502 were presented as the mean  $\pm$  SD in figures. Results of LT (liver tumor)  
503 sample were shown in Figures, and corresponding results of other three  
504 samples were shown in supplementary files. KEGG pathway enrichment ( $P <$   
505 0.01) were performed using clusterProfiler package (version 3.9.2) [53].  
506 Differentially expressed genes were identified by the “FindMarkers” function  
507 (“logfc.threshold” = 0.25 and “min.pct” = 0.25) using the MAST method [54],  
508 and P value was adjusted using *bonferroni* correction based on the total  
509 number of gene in the dataset, with the thresholds of adjusted  $P < 0.01$ .

510

#### 511 **Authors’ contribution**

512 ZMZ supervised research. XLW and YH analyzed data. XLW and QMZ  
513 drafted the manuscript. QMZ did experiments. ZMZ and XWR revised the  
514 manuscript. All authors read and approved the final manuscript.

515

#### 516 **Competing interests**

517 The authors have declared no competing interests.

518

519 **Acknowledgements**

520 This work was supported by the National Natural Science Foundation of  
521 China (Grant No. 31530036, 81573022, and 31601063).

522

523 **Data availability**

524 Data will be accessible publicly at the time of publication.

525

526 **Ethics approval**

527 This study was approved by the Ethics Committee of Beijing Shijitan Hospital,  
528 Capital Medical University. All patients in this study provided written informed  
529 consent for sample collection and data analysis.

530

531

532 **Reference**

- 533 1. Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, Wang X, Bodeau J, Tuch  
534 BB, Siddiqui A, et al: **mRNA-Seq whole-transcriptome analysis of a single cell.** *Nat*  
535 *Methods* 2009, **6**:377-382.
- 536 2. Pollen AA, Nowakowski TJ, Shuga J, Wang X, Leyrat AA, Lui JH, Li N, Szpankowski L,  
537 Fowler B, Chen P, et al: **Low-coverage single-cell mRNA sequencing reveals cellular**  
538 **heterogeneity and activated signaling pathways in developing cerebral cortex.** *Nat*  
539 *Biotechnol* 2014, **32**:1053-1058.
- 540 3. Zhang L, Yu X, Zheng L, Zhang Y, Li Y, Fang Q, Gao R, Kang B, Zhang Q, Huang JY,  
541 et al: **Lineage tracking reveals dynamic relationships of T cells in colorectal cancer.**  
542 *Nature* 2018, **564**:268-272.
- 543 4. Halpern KB, Shenhav R, Matcovitch-Natan O, Toth B, Lemze D, Golan M, Massasa  
544 EE, Baydatch S, Landen S, Moor AE, et al: **Single-cell spatial reconstruction reveals**  
545 **global division of labour in the mammalian liver.** *Nature* 2017, **542**:352-356.
- 546 5. Grover A, Sanjuan-Pla A, Thongjuea S, Carrelha J, Giustacchini A, Gambardella A,  
547 Macaulay I, Mancini E, Luis TC, Mead A, et al: **Single-cell RNA sequencing reveals**  
548 **molecular and functional platelet bias of aged haematopoietic stem cells.** *Nat*  
549 *Commun* 2016, **7**:11075.
- 550 6. Benitez JA, Cheng S, Deng Q: **Revealing allele-specific gene expression by single-cell**  
551 **transcriptomics.** *The International Journal of Biochemistry & Cell Biology* 2017,  
552 **90**:155-160.
- 553 7. Svensson V, Vento-Tormo R, Teichmann SA: **Exponential scaling of single-cell**



- 554           **RNA-seq in the past decade.** *Nat Protoc* 2018, **13**:599-604.
- 555    8.       See P, Lum J, Chen J, Ginhoux F: **A Single-Cell Sequencing Guide for Immunologists.**
- 556           *Frontiers in Immunology* 2018, **9**:2425.
- 557    9.       Picelli S, Björklund ÅK, Faridani OR, Sagasser S, Winberg G, Sandberg R:
- 558           **Smart-seq2 for sensitive full-length transcriptome profiling in single cells.** *Nature*
- 559           *Methods* 2013, **10**:1096-1098.
- 560    10.      Picelli S, Faridani OR, Bjorklund AK, Winberg G, Sagasser S, Sandberg R: **Full-length**
- 561           **RNA-seq from single cells using Smart-seq2.** *Nature Protocols* 2014, **9**:171-181.
- 562    11.      Grun D, van Oudenaarden A: **Design and Analysis of Single-Cell Sequencing**
- 563           **Experiments.** *Cell* 2015, **163**:799-810.
- 564    12.      Stegle O, Teichmann SA, Marioni JC: **Computational and analytical challenges in**
- 565           **single-cell transcriptomics.** *Nat Rev Genet* 2015, **16**:133-145.
- 566    13.      Shalek AK, Satija R, Shuga J, Trombetta JJ, Gennert D, Lu D, Chen P, Gertner RS,
- 567           Gaublomme JT, Yosef N, et al: **Single-cell RNA-seq reveals dynamic paracrine control**
- 568           **of cellular variation.** *Nature* 2014, **510**:363-369.
- 569    14.      Soneson C, Robinson MD: **Bias, robustness and scalability in single-cell differential**
- 570           **expression analysis.** *Nat Methods* 2018, **15**:255-261.
- 571    15.      Islam S, Zeisel A, Joost S, La Manno G, Zajac P, Kasper M, Lonnerberg P, Linnarsson
- 572           **S: Quantitative single-cell RNA-seq with unique molecular identifiers.** *Nat Methods*
- 573           2014, **11**:163-166.
- 574    16.      Fan J, Salathia N, Liu R, Kaeser GE, Yung YC, Herman JL, Kaper F, Fan JB, Zhang K,
- 575           Chun J, Kharchenko PV: **Characterizing transcriptional heterogeneity through**

- 576 **pathway and gene set overdispersion analysis.** *Nature Methods* 2016, **13**:241-244.
- 577 17. Svensson V, Natarajan KN, Ly LH, Miragaia RJ, Labalette C, Macaulay IC, Cvejic A,  
578 Teichmann SA: **Power analysis of single-cell RNA-sequencing experiments.** *Nat*  
579 *Methods* 2017, **14**:381-387.
- 580 18. Baran-Gale J, Chandra T, Kirschner K: **Experimental design for single-cell RNA**  
581 **sequencing.** *Brief Funct Genomics* 2018, **17**:233-239.
- 582 19. Azizi E, Carr AJ, Plitas G, Cornish AE, Konopacki C, Prabhakaran S, Nainys J, Wu K,  
583 Kiseliovas V, Setty M, et al: **Single-Cell Map of Diverse Immune Phenotypes in the**  
584 **Breast Tumor Microenvironment.** *Cell* 2018, **0**.
- 585 20. Ilicic T, Kim JK, Kolodziejczyk AA, Bagger FO, McCarthy DJ, Marioni JC, Teichmann  
586 SA: **Classification of low quality cells from single-cell RNA-seq data.** *Genome Biol*  
587 2016, **17**:29.
- 588 21. Wagner A, Regev A, Yosef N: **Revealing the vectors of cellular identity with single-cell**  
589 **genomics.** *Nat Biotechnol* 2016, **34**:1145-1160.
- 590 22. Zheng C, Zheng L, Yoo JK, Guo H, Zhang Y, Guo X, Kang B, Hu R, Huang JY, Zhang  
591 Q, et al: **Landscape of Infiltrating T Cells in Liver Cancer Revealed by Single-Cell**  
592 **Sequencing.** *Cell* 2017, **169**:1342-1356 e1316.
- 593 23. Scialdone A, Natarajan KN, Saraiva LR, Proserpio V, Teichmann SA, Stegle O,  
594 Marioni JC, Buettner F: **Computational assignment of cell-cycle stage from single-cell**  
595 **transcriptome data.** *Methods* 2015, **85**:54-61.
- 596 24. Bacher R, Kendzierski C: **Design and computational analysis of single-cell**  
597 **RNA-sequencing experiments.** *Genome Biol* 2016, **17**:63.

- 598 25. Gladka Monika M, Molenaar B, de Ruiter H, van der Elst S, Tsui H, Versteeg D,  
599 Lacraz Grègory PA, Huibers Manon MH, van Oudenaarden A, van Rooij E: **Single-Cell**  
600 **Sequencing of the Healthy and Diseased Heart Reveals Cytoskeleton-Associated**  
601 **Protein 4 as a New Modulator of Fibroblasts Activation.** *Circulation* 2018,  
602 **138:166-180.**
- 603 26. Klein AM, Mazutis L, Akartuna I, Tallapragada N, Veres A, Li V, Peshkin L, Weitz DA,  
604 Kirschner MW: **Droplet barcoding for single-cell transcriptomics applied to embryonic**  
605 **stem cells.** *Cell* 2015, **161:1187-1201.**
- 606 27. Macosko EZ, Basu A, Satija R, Nemes J, Shekhar K, Goldman M, Tirosh I, Bialas AR,  
607 Kamitaki N, Martersteck EM, et al: **Highly Parallel Genome-wide Expression Profiling**  
608 **of Individual Cells Using Nanoliter Droplets.** *Cell* 2015, **161:1202-1214.**
- 609 28. Liu SJ, Nowakowski TJ, Pollen AA, Lui JH, Horlbeck MA, Attenello FJ, He D,  
610 Weissman JS, Kriegstein AR, Diaz AA, Lim DA: **Single-cell analysis of long**  
611 **non-coding RNAs in the developing human neocortex.** *Genome Biol* 2016, **17:67.**
- 612 29. Johnson MB, Wang PP, Atabay KD, Murphy EA, Doan RN, Hecht JL, Walsh CA:  
613 **Single-cell analysis reveals transcriptional heterogeneity of neural progenitors in**  
614 **human cortex.** *Nat Neurosci* 2015, **18:637-646.**
- 615 30. Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, Regev A, Rinn JL: **Integrative**  
616 **annotation of human large intergenic noncoding RNAs reveals global properties and**  
617 **specific subclasses.** *Genes & Development* 2011, **25:1915-1927.**
- 618 31. Hangauer MJ, Vaughn IW, Mcmanus MT: **Pervasive Transcription of the Human**  
619 **Genome Produces Thousands of Previously Unidentified Long Intergenic Noncoding**

- 620            **RNAs.** *Plos Genetics* 2013, **9**:e1003569.
- 621    32.    Wu AR, Neff NF, Kalisky T, Dalerba P, Treutlein B, Rothenberg ME, Mburu FM,  
622            Mantalas GL, Sim S, Clarke MF, Quake SR: **Quantitative assessment of single-cell**  
623            **RNA-sequencing methods.** *Nature methods* 2014, **11**:41-46.
- 624    33.    Satija R, Farrell JA, Gennert D, Schier AF, Regev A: **Spatial reconstruction of**  
625            **single-cell gene expression data.** *Nat Biotechnol* 2015, **33**:495-502.
- 626    34.    Lambrechts D, Wauters E, Boeckx B, Aibar S, Nittner D, Burton O, Bassez A,  
627            Decaluwe H, Pircher A, Van den Eynde K, et al: **Phenotype molding of stromal cells in**  
628            **the lung tumor microenvironment.** *Nat Med* 2018, **24**:1277-1289.
- 629    35.    Puram SV, Tirosh I, Parikh AS, Patel AP, Yizhak K, Gillespie S, Rodman C, Luo CL,  
630            Mroz EA, Emerick KS, et al: **Single-Cell Transcriptomic Analysis of Primary and**  
631            **Metastatic Tumor Ecosystems in Head and Neck Cancer.** *Cell* 2017, **171**:1611-1624  
632            e1624.
- 633    36.    Hashimshony T, Senderovich N, Avital G, Klochendler A, de Leeuw Y, Anavy L,  
634            Gennert D, Li S, Livak KJ, Rozenblatt-Rosen O, et al: **CEL-Seq2: sensitive**  
635            **highly-multiplexed single-cell RNA-Seq.** *Genome Biology* 2016, **17**:77.
- 636    37.    Brennecke P, Anders S, Kim JK, Kolodziejczyk AA, Zhang X, Proserpio V, Baying B,  
637            Benes V, Teichmann SA, Marioni JC, Heisler MG: **Accounting for technical noise in**  
638            **single-cell RNA-seq experiments.** *Nat Methods* 2013, **10**:1093-1095.
- 639    38.    Bourgon R, Gentleman R, Huber W: **Independent filtering increases detection power**  
640            **for high-throughput experiments.** *Proc Natl Acad Sci U S A* 2010, **107**:9546-9551.
- 641    39.    Deng Q, Ramskold D, Reinius B, Sandberg R: **Single-cell RNA-seq reveals dynamic,**

- 642            **random monoallelic gene expression in mammalian cells.** *Science* 2014, **343**:193-196.
- 643    40.    Reinius B, Mold JE, Ramskold D, Deng Q, Johnsson P, Michaelsson J, Frisen J,  
644            Sandberg R: **Analysis of allelic expression patterns in clonal somatic cells by**  
645            **single-cell RNA-seq.** *Nat Genet* 2016, **48**:1430-1435.
- 646    41.    La Manno G, Soldatov R, Zeisel A, Braun E, Hochgerner H, Petukhov V, Lidschreiber  
647            K, Kastri ME, Lonnerberg P, Furlan A, et al: **RNA velocity of single cells.** *Nature* 2018,  
648            **560**:494-498.
- 649    42.    Ziegenhain C, Vieth B, Parekh S, Reinius B, Guillaumet-Adkins A, Smets M,  
650            Leonhardt H, Heyn H, Hellmann I, Enard W: **Comparative Analysis of Single-Cell RNA**  
651            **Sequencing Methods.** *Molecular Cell* 2017, **65**:631-643.e634.
- 652    43.    Chen W, Li Y, Easton J, Finkelstein D, Wu G, Chen X: **UMI-count modeling and**  
653            **differential expression analysis for single-cell RNA sequencing.** *Genome Biol* 2018,  
654            **19**:70.
- 655    44.    Giladi A, Amit I: **Single-Cell Genomics: A Stepping Stone for Future Immunology**  
656            **Discoveries.** *Cell* 2018, **172**:14-21.
- 657    45.    Smillie CS, Biton M, Ordovas-Montanes J, Sullivan KM, Burgin G, Graham DB, Herbst  
658            RH, Rogel N, Slyper M, Waldman J, et al: **Rewiring of the cellular and inter-cellular**  
659            **landscape of the human colon during ulcerative colitis.** *bioRxiv* 2018:455451.
- 660    46.    Oetjen KA, Lindblad KE, Goswami M, Gui G, Dagur PK, Lai C, Dillon LW, McCoy JP,  
661            Hourigan CS: **Human bone marrow assessment by single-cell RNA sequencing, mass**  
662            **cytometry, and flow cytometry.** *JCI Insight* 2018, **3**:e124928.
- 663    47.    Tanay A, Regev A: **Scaling single-cell genomics from phenomenology to mechanism.**

- 664 *Nature* 2017, **541**:331-338.
- 665 48. Buettner F, Natarajan KN, Casale FP, Proserpio V, Scialdone A, Theis FJ, Teichmann  
666 SA, Marioni JC, Stegle O: **Computational analysis of cell-to-cell heterogeneity in**  
667 **single-cell RNA-sequencing data reveals hidden subpopulations of cells.** *Nat*  
668 *Biotechnol* 2015, **33**:155-160.
- 669 49. Skinner SO, Xu H, Nagarkar-Jaiswal S, Freire PR, Zwaka TP, Golding I: **Single-cell**  
670 **analysis of transcription kinetics across the cell cycle.** *Elife* 2016, **5**:e12175.
- 671 50. Marinov GK, Williams BA, McCue K, Schroth GP, Gertz J, Myers RM, Wold BJ: **From**  
672 **single-cell to cell-pool transcriptomes: stochasticity in gene expression and RNA**  
673 **splicing.** *Genome Res* 2014, **24**:496-510.
- 674 51. Lun AT, McCarthy DJ, Marioni JC: **A step-by-step workflow for low-level analysis of**  
675 **single-cell RNA-seq data with Bioconductor.** *F1000Res* 2016, **5**:2122.
- 676 52. Wang L, Wang S, Li W: **RSeQC: quality control of RNA-seq experiments.**  
677 *Bioinformatics* 2012, **28**:2184-2185.
- 678 53. Yu GC, Wang LG, Han YY, He QY: **clusterProfiler: an R Package for Comparing**  
679 **Biological Themes Among Gene Clusters.** *Omics-a Journal Of Integrative Biology*  
680 2012, **16**:284-287.
- 681 54. Finak G, McDavid A, Yajima M, Deng J, Gersuk V, Shalek AK, Slichter CK, Miller HW,  
682 McElrath MJ, Prlic M, et al: **MAST: a flexible statistical framework for assessing**  
683 **transcriptional changes and characterizing heterogeneity in single-cell RNA**  
684 **sequencing data.** *Genome Biol* 2015, **16**:278.

685

686

## 687 **Figure Legends**

### 688 **Figure 1 Cell evaluation**

689 **A.** The total reads number of each cell. The proportion of reads of genes in the  
690 GO:0005576 “extracellular region” term (**B**), GO:0016020 “membrane” term  
691 (**C**), and GO:0005737 “cytoplasm” term (**D**). **E.** The ratio of cells in the G1,  
692 G2M, and S phases. The proportion of reads of mitochondrial gene (**F**) and  
693 genes in GO:0005840 “ribosome” term (**G**).

### 694 **Figure 2 Comparison of lncRNA**

695 The ratio of reads of protein coding (PC) genes (**A**), lncRNA (**B**),  
696 house-keeping (HK) genes (**C**), transcription factors (TFs) (**D**). Overlap of  
697 highly variable genes (HVGs) identified from 10X and Smart-seq2 (**E**). Types  
698 of HVGs (**F**).

### 699 **Figure 3 Comparison of detected genes and their expression**

700 **A.** The number of detected genes in every cell. **B.** Overlap of all detected  
701 genes between 10X and Smart-seq2. **C.** Distribution of detected genes based  
702 on their expression levels. **D.** Saturation analysis by resampling a series of  
703 subsets of total reads. **E.** The ratio of reads of the top10 high expressed genes.  
704 **F.** Overlap of the top25% high expressed genes among 10X, Smart-seq2, and  
705 bulk RNA-seq. **G.** Correlation of expression of common detected genes among  
706 10X, Smart-seq2, and bulk RNA-seq.

### 707 **Figure 4 Results of cells clustering and differentially expressed genes** 708 **(DEGs)**

709 Cell clustering results for 10X (**A**) and Smart-seq2 (**B**). **C.** Overlap of DEGs of  
710 LT (liver tumor) sample with other three samples identified by 10X and  
711 Smart-seq2. Comparison of KEGG enrichment results of LT sample (**D**) and  
712 fibroblasts (**F**). **E.** Overlap of DEGs of each cell type compared with remaining  
713 types between 10X and Smart-seq2.

### 714 **Figure 5 Dropout assessment**

715 **A.** Comparison of dropout ratios between 10X and Smart-seq2. **B.** Two

716 examples of house-keeping genes to show dropout events. **C.** The relationship  
717 of dropout ratios and the average expression for each gene. **D.** Number of  
718 expressing cells against the average expression of each gene. **E.** CV  
719 (coefficient of variation) distribution of each detected gene. **F.** The relationship  
720 between CV and gene expression levels. **G.** Dropout ratios of gene with CV  
721 more than 800.

## 722 **Figure 6 Comparison of gene structural information**

723 **A.** The reads coverage over gene body. **B.** Reads distribution in genome. **C.**  
724 Detection of known splice junctions. **D.** Gene length was divided into  
725 consecutive 100 bins, we counted the number of detected genes in each bin,  
726 PCCs (Pearson correlation coefficients) of gene number between Smart-seq2  
727 and 10X were calculated.

728

## 729 **Supplementary material**

730 **Table S1 Cell number of each sample**

731 **Table S2 List of the most highly expressed genes (Top10)**

732 **Table S3 KEGG enrichment results of 10X-specific, bulk-specific, and**  
733 **Smart-seq2-specific genes in the top25% list**

734 **Table S4 DEGs among samples with the change trends conflicting**

735 **Table S5 DEGs among cell types with the change trends conflicting**

736 **Table S6 List of genes with zero dropout ratio in a sample**

737

## 738 **Figure S1 Cell evaluation of other three samples**

739 **A.** The total reads of each cell. The proportion of reads of genes in the  
740 GO:0005576 “extracellular region” term (**B**), GO:0016020 “membrane” term  
741 (**C**), and GO:0005737 “cytoplasm” term (**D**). **E.** The ratios of cells in the G1,  
742 G2M, and S phases. The proportion of reads of mitochondrial gene (**F**) and  
743 genes in GO:0005840 “ribosome” term (**G**). **H.** Reads proportion of rDNA.

## 744 **Figure S2 Assessment of each cell**



745 **A.** The unique mapping reads of each sample. **B.** Cell cycle phase scores of  
746 each cell.

### 747 **Figure S3 Comparison of certain classes of genes**

748 The expression proportion of protein coding (PC) genes (**A**), lncRNA (**B**),  
749 house-keeping (HK) genes (**C**), transcription factors (TFs) (**D**). **E.** KEGG  
750 enrichment results of 10X-specific, Smart-seq2-specific, and shared highly  
751 variable genes (HVGs).

### 752 **Figure S4 Comparison of expression profiles**

753 **A.** The number of detected genes in every cell. **B.** Overlap of all the detected  
754 genes between two platforms. **C.** Distribution of detected genes based on their  
755 expression levels. **D.** Saturation analysis. Y axis is “Percent Relative Error”  
756 which is used to measure how the RPKM estimated from subset of reads  
757 deviates from real expression level. **E.** Percentage of total counts assigned to  
758 the top 10 most highly-abundant genes. **F.** Overlap of the top25% high  
759 expressed genes among 10X, Smart-seq2, and bulk RNA-seq. **G.** Correlation  
760 of common detected genes expression among 10X, Smart-seq2, and bulk  
761 RNA-seq.

### 762 **Figure S5 Results of differentially expressed genes (DEGs)**

763 **A.** Overlap of DEGs of remaining samples between Smart-seq2 and 10X.  
764 Overlap of Up-regulated and down-regulated DEGs for each sample (**B**) and  
765 each cell type (**C**) between Smart-seq2 and 10X.

### 766 **Figure S6 Dropout events assessment of other three samples**

767 **A.** Comparison of dropout ratios between 10X and Smart-seq2. **B.** Two  
768 examples of house-keeping genes. **C.** The relationships of dropout ratios and  
769 the average gene expression levels. **D.** Number of expressing cells against the  
770 average expression for each gene. **E.** CV (coefficient of variation) distribution  
771 of each detected gene. **F.** The relationship between CV and gene expression  
772 levels. Dropout ratios of genes with CV less than 800 (**G**) and genes with CV  
773 more than 800 (**H**).

### 774 **Figure S7 Comparison of 3'-end VS full-length capture**

775 **A.** Reads coverage over gene body. **B.** Reads distribution in genome. **C.**  
776 Detection of known splice junctions. **D.** PCC (Pearson correlation coefficient)  
777 of gene number in consecutive 100 bins divided by gene lengths between  
778 Smart-seq2 and 10X.  
779













