Reproducibility across single-cell RNA-seq protocols for spatial ordering analysis

Morten Seirup[1,2*], Li-Fang Chu[2], Srikumar Sengupta[2], Ning Leng[2,3,#a], Hadley Browder[4], Kevin Kapadia[4], Christina M. Shafer[2], Bret Duffin[2], Angela L. Elwell[2,#b], Jennifer M. Bolin[2], Scott Swanson[2], Ron Stewart[2], Christina Kendziorski[3], James A. Thomson[2,5,6*], Rhonda Bacher[7,*]

[1]Molecular and Environmental Toxicology Program, University of Wisconsin Madison, Madison, Wisconsin, United States of America
[2]Morgridge Institute for Research, Madison, Wisconsin, United States of America
[3]Department of Biostatistics and Medical Informatics, University of Wisconsin Madison, Madison, Wisconsin, United States of America
[4]Department of Statistics, University of Florida, Gainesville, Florida, United States of America
[5]Department of Cell & Regenerative Biology, University of Wisconsin School of Medicine and Public Health, Madison, Wisconsin, United States of America
[6]Department of Molecular, Cellular, & Developmental Biology, University of California Santa Barbara, Santa Barbara, California, United States of America
[7]Department of Biostatistics, University of Florida, Gainesville, Florida, United States of America
[#a]Current Address: Genentech, San Francisco, California, United States of America
[#b]Current Address: Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, United States of America

***Corresponding Authors:**
Morten Seirup
E-mail: seirup@wisc.edu
James A. Thomson, V.M.D., Ph.D., Diplomate A.C.V.P.
E-mail: jthomson@morgridgeinstitute.org
Rhonda Bacher, Ph.D.
E-mail: rbacher@ufl.edu

Abstract

As newer single-cell protocols generate increasingly more cells at reduced sequencing depths, the value of a higher read depth may be overlooked. Using data from three different single-cell RNA-seq protocols that lend themselves to having either higher read depth (Smart-seq) or many cells (MARS-seq and 10X), we evaluate their ability to recapitulate biological signals in the context of pseudo-spatial reconstruction. Overall, we find gene expression profiles after spatial-reconstruction analysis are highly reproducible between datasets despite being generated by different protocols and using different computational algorithms. While UMI based protocols such as 10X and MARS-seq allow for capturing more cells, Smart-seq's higher sensitivity and read-depth allows for analysis of lower expressed genes and isoforms. Additionally, we evaluate trade-offs for each protocol by performing subsampling analyses, and find that optimizing the balance between sequencing depth and number of cells within a protocol is important for efficient use of resources. Our analysis emphasizes the importance of selecting a protocol based on the biological questions and features of interest.

Introduction

Single-cell RNA sequencing (scRNA-seq)[1–5] is a powerful tool for studying transcriptional differences between individual cells. The innovation of droplet-based techniques[6,7] and unique molecular identifiers (UMI)[8] has lowered the cost per cell and pushed the field towards obtaining data from tens of thousands of cells per experiment albeit at a reduced sequencing depth. Recent publications have compared the sensitivity, accuracy, and precision between several scRNA-seq techniques and report that the major trade-off between protocols is sensitivity, which is dependent on read depth[9,10]. With the push for sequencing an ever-increasing number of cells at the expense of read depth per cell, the value of a higher read depth might be overlooked. Here we investigate the reproducibility of biological signals across protocols that naturally lend themselves to generating data on more cells versus higher read depth.

71 Studies comparing protocols have mainly done so with respect to performance
72 on spike-ins or on technical variability alone[9,10]. Recently, Guo et al.[11] showed
73 agreement of cell types and signature genes between two platforms used for single-cell
74 RNA-seq – Fluidigm C1 and Drop-seq. However, few studies have examined
75 comparative agreement among protocols for biological inferences beyond clustering
76 and identifying differential gene expression, yet a key question of interest with single-
77 cell data is its ability to reflect temporal or spatial heterogeneity. For cells collected at a
78 given time, the underlying dynamic biological process is reflected in genome-wide
79 differences in gene expression. Computational algorithms that attempt to order cells in
80 pseudo-time or pseudo-space based on variability in gene expression have been
81 developed[4,12,13], and more than 45 existing algorithms were recently compared[14]. Yet,
82 as far as we know, no comparison of single-cell protocols exists for the question of cell
83 ordering.
84 Here, our evaluation is in the context of pseudo-spatial reconstruction in which
85 we compared three independently produced scRNA-seq datasets on the mouse liver
86 lobule. We chose to compare protocols on their ability to reflect the spatial patterning of
87 the liver lobule in which the parenchymal cells of the liver, hepatocytes, are organized
88 spatially in a polygonal shape around a central vein (Figure 1A). From the central vein,
89 a gradient of metabolic functions is performed extending to a portal triad at each
90 vertex[15–19]. The gradient of differences in gene expression patterns is referred to as the
91 zonation axis (from periportal (PP) to pericentral (PC))[20]. This coordinated spatial
92 organization provides a particularly interesting application of single-cell techniques. For
93 this study we obtained one dataset using Smart-seq—a full-length protocol, a second

94 dataset using MARS-seq[21]—a UMI and plate based protocol and the third dataset
95 generated using 10X[22]—a UMI and droplet protocol. Although the cell number and read
96 depth differ greatly across datasets, we find high reproducibility of gene expression
97 profiles after spatial-reconstruction analysis. Given the reproducibility and that each
98 protocol naturally lends itself to either producing more cells at a lower sequencing depth
99 or fewer cells at a higher depth, our results demonstrate the importance of carefully
100 evaluating the biological question and features of interest when selecting the
101 appropriate sequencing protocol. In applications focused on lower expressed genes or
102 on genes with high sequence similarity, increased read depth is preferable, whereas a
103 focus on identifying cell types based on more highly expressed genes will benefit from
104 collecting more cells. In an ideal situation a single cell assay would result in thousands
105 of cells that are all sequenced at a high read depth, but technical and financial
106 restrictions make this rarely possible.

107 **Results**

108 **Differences in detection rates**

109 By using the Fluidigm C1 coupled with the Smart-seq protocol, we were able to
110 identify on average around 38% (about 7100 genes) (Figure 1B) of all genes in the
111 genome expressed per cell, whereas the MARS-seq dataset finds on average 12%
112 (about 2200 genes) and the 10X dataset finds on average 6% (about 1100 genes)
113 (Figure 1B). This is in accordance with findings by Ziegenhain et al. 2017 when they
114 examined single-cell transcriptomic methods[9] and by Phipson et al[23] when comparing
115 biases in full-length versus UMI protocols. The increased sensitivity of the full-length
116 protocol is further illustrated in Figure 1C, which on a per gene level shows the

117 difference in detection fraction compared to the log fold change in mean expression
118 between the protocols. A difference in detection fraction of zero means that the gene is
119 detected in the same fraction of cells in both datasets. The difference across protocols
120 in log2 fold-change has a linear relationship with the difference in detection fractions,
121 which indicates a fairly constant increase in log2 expression as cells are sequenced
122 with greater sensitivity. At the intercept, a difference in detection equal to zero, the log2
123 fold change is 3.1 between Smart-seq and MARS-seq, indicating an experiment wide
124 increase in sensitivity in the Smart-seq protocol of approximately 9-fold. Between
125 Smart-seq and 10X, the increase in sensitivity is approximately 12-fold and there is a
126 similar level of sensitivity between MARS-seq and 10X. Not surprisingly, the vast
127 majority of genes are detected in a larger fraction of cells and have a higher expression
128 level in the more deeply sequenced dataset using the Smart-seq protocol. Although, it is
129 worth pointing out that around 6% of genes have higher detection using the MARS-seq
130 protocol (negative values on x-axis) and a few of these genes also have higher
131 expression levels (negative values on y-axis) than in the Smart-seq protocol. This
132 subset of genes better detected in the MARS-seq dataset have higher GC content and
133 are slightly longer (S1 Figure), which is consistent with previous reports of protocol
134 comparisons[23,24].
135
136
137
138
139

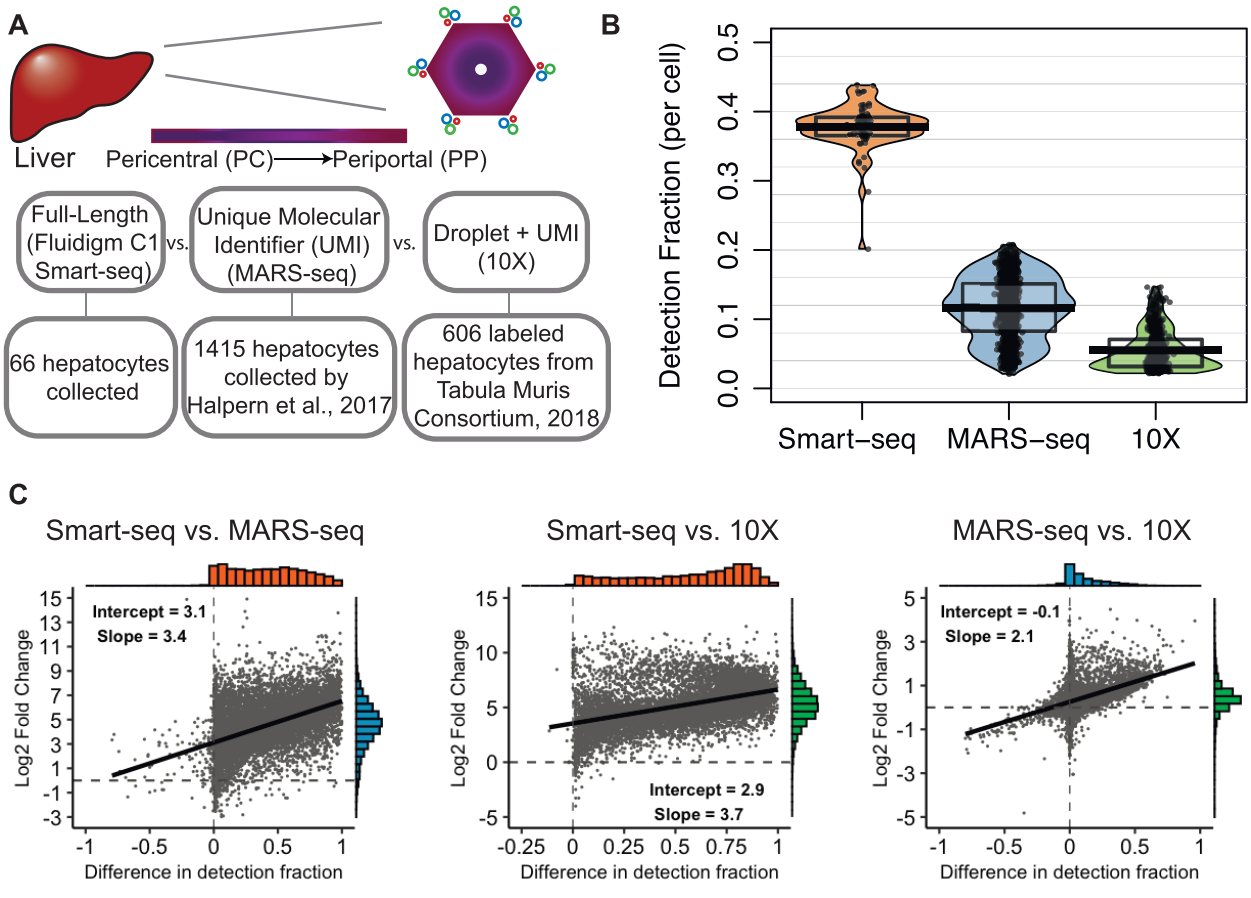


Figure 1. Illustration of the liver anatomy, and general comparison of the datasets. A) Top. Illustration of the liver lobule identifying the portal triad along the outer edges and the central vein in the middle. The color gradient represents metabolic zonation. Bottom. Highlights the main differences between the datasets compared. B) Comparison of gene detection fraction between the datasets. The detection fraction per cell (y-axis) is shown for the two datasets (x-axis). C) Left. The log2 fold-change of genes detected above an average expression level of zero in the Smart-seq dataset compared to the MARS-seq dataset (y-axis), versus the difference in gene-level detection fractions across datasets (x-axis). A linear regression line is overlaid and a histogram of the x- and y-axis are shown opposite of each axis. Middle. Similar plot shown for Smart-seq versus 10X. Right. Similar plot shown for 10X versus MARS-seq.

152

153 **Reconstructing spatial profiles of liver zonation profiles**

154 Next, to represent the spatial patterns across the liver lobule, the cells in the
155 three datasets were computationally ordered according to their expression profiles. The
156 MARS-seq dataset was spatially ordered by Halpern et al. 2017 by first performing
157 smFISH for six marker genes at various locations across the zonation axis, then single-
158 cell RNA-seq data obtained by MARS-seq assigned cells to one of nine zonation
159 locations based on each cell's expression profile of the six marker genes[21]. We ordered
160 the cells in the 10X dataset using the Monocle2 algorithm, which builds a trajectory
161 through cells based on the expression similarity among the most highly variable
162 genes[12]. For the Smart-seq protocol we used the computational algorithm Wave-Crest
163 to spatially order cells based on fifteen marker genes known in the literature to be
164 differentially expressed along the zonation axis (Figure 2A)[5]. The ordering procedure
165 uses the nearest insertion algorithm implemented in the Wave-Crest package, which
166 searches among the space of all possible orderings via a 2-opt algorithm by considering
167 insertion events and choosing orders which minimize the mean square error of a
168 polynomial regression on the marker genes expression. Of the 15 genes used, we
169 selected eight periportal expressed genes and seven pericentral expressed genes[20]. All
170 orderings assume the zonation profile and spatial organization can be represented in a
171 single dimension. A similar reconstructed order was obtained for the Smart-seq dataset
172 when applying Monocle2 (S2 Figure).

173

174

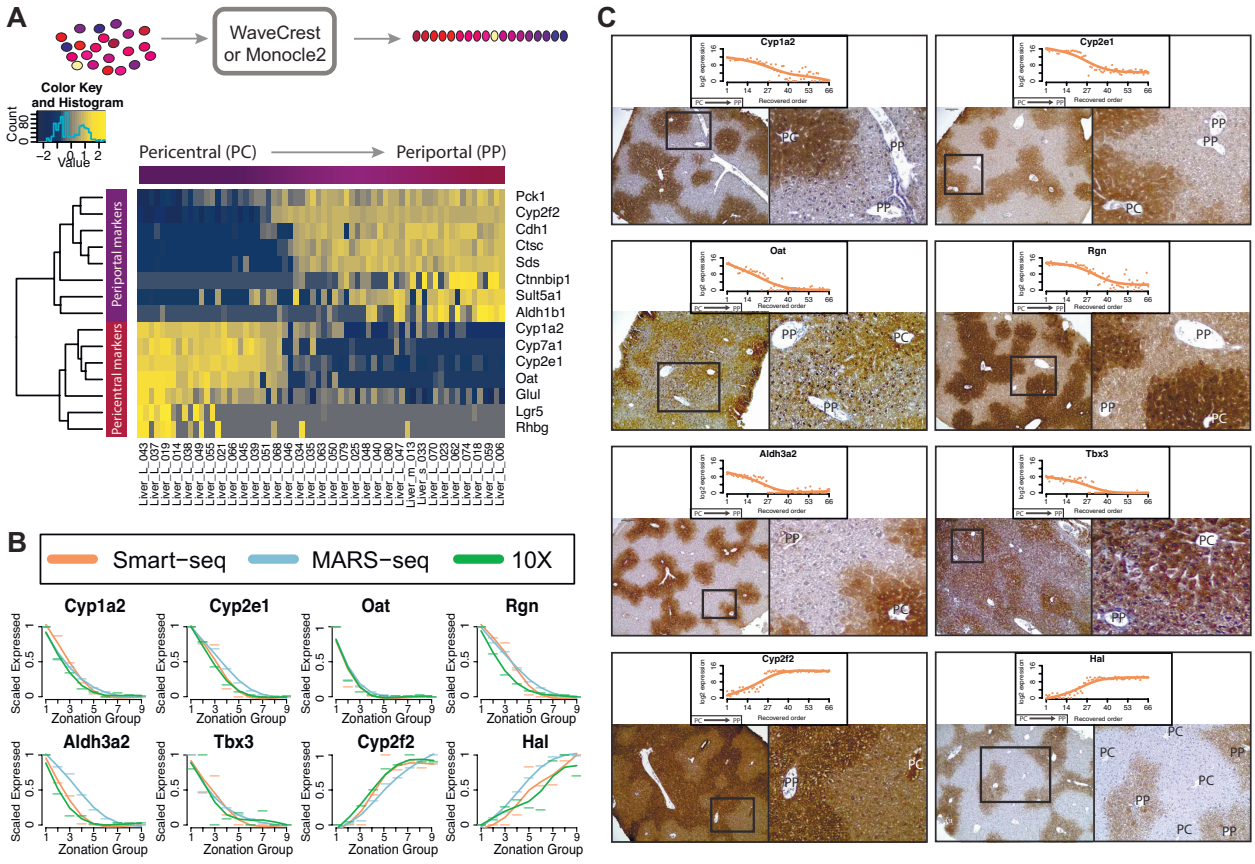


Figure 2. Pseudo-space reordering of hepatocytes, and prediction and validation of dynamically expressed genes. A) Top. Illustration of the pseudo-spatial reordering process. Bottom. Heatmap showing the pseudo-spatial reordering (x-axis) and the expression levels of the marker genes (y-axis) for the Smart-seq dataset. Pericentral cells are found on the left-hand side and Periportal cells are found on the right-hand side. B) Scaled expression profile (y-axis) of 8 dynamic genes based on the predicted pseudo-space reordering (x-axis) of the Smart-seq dataset (orange), the MARS-seq dataset (blue), and the 10X dataset (green). C) Immunohistochemistry staining of the genes highlighted in B). Above the staining is the log2 expression counts (y-axis) across the predicted pseudo-spatial order (x-axis) of the Smart-seq dataset. The left picture

186 shows the staining and the right picture is an enlarged section (black square). PP =
187 Periportal, PC = Pericentral.

188 Using the recreated order of the hepatocytes, we explored the dynamics of gene
189 expression across the periportal to pericentral axis. Figure 2B shows a subset of genes
190 that are predicted to be highly regulated across the axis, four of which were not in our
191 list of marker genes. Since the MARS-seq dataset placed cells into nine discrete zones
192 along the axis, we divided cells from the Smart-seq and 10X datasets into nine equally
193 sized groups in order to compare the reconstructed orderings. The zonation profiles in
194 Figure 2B have high agreement, with a median correlation of 0.95 between the three
195 datasets. Before proceeding, we also performed an additional experiment to validate
196 that our cell ordering and expression profiles reflect those of the liver lobule *in vivo*.
197 Remarkably, immunohistochemistry studies showed that selected marker gene protein
198 expression profiles also agreed with our spatial reconstructed scRNA-seq datasets: six
199 markers display a PC-high/PP-low profile and two markers display a PC-low/PP-high
200 profile in mouse liver lobule *in vivo* (Figure 2C). This confirmation in protein gradient
201 patterns corresponding to our reconstructed mRNA profiles provides us with confidence
202 for further analysis on the biological inference in comparing the three protocols in this
203 context.

204 **Comparing marker gene expression across liver zonation profiles**

205 An exciting prospect of single cell analysis is the identification of genes that have
206 non-monotonic or dynamic expression across pseudo-time or space. Several genes in
207 the bile acid synthesis pathway were shown by Halpern et al., 2017 to be non-
208 monotonically expressed in a pattern where the highest expression levels along the

lobule correspond to the functional placement of the genes in the bile acid synthesis pathway (Cyp7a1, Hsd3b7, Cyp8b1, Cyp27a1 and Baat)[21]. We find that the expression profiles for these genes are corroborated across the three datasets (Figure 3).

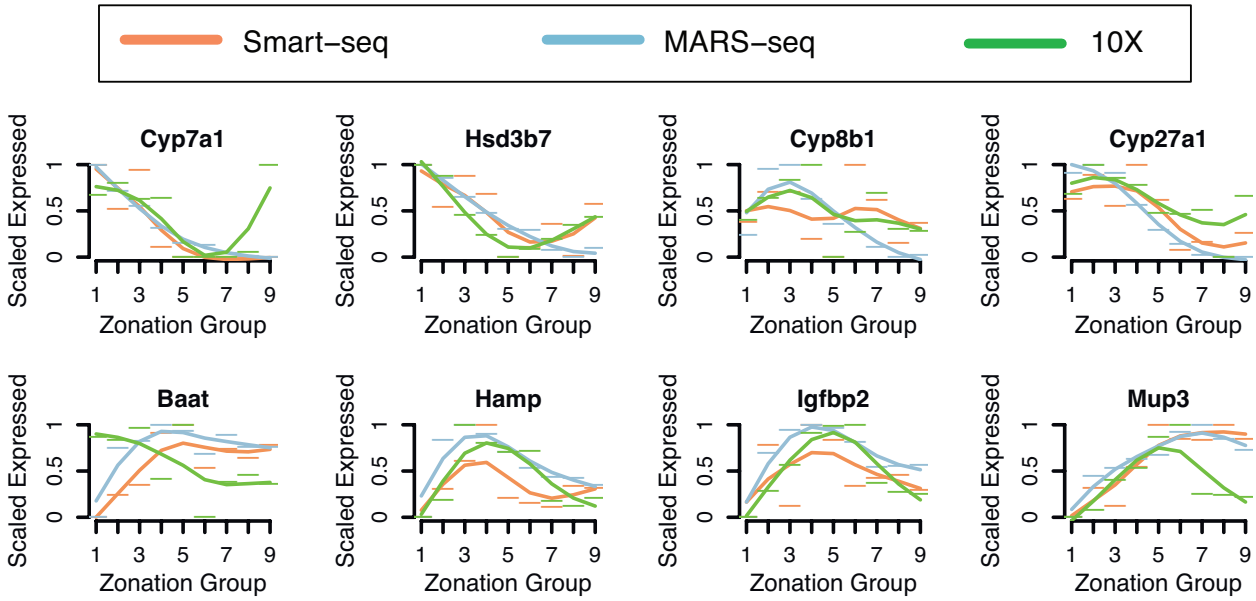


Figure 3. Comparison of zonation profiles across three datasets. Scaled expression profile (y-axis) of 8 genes non-monotonically expressed from Halpern et al. along the predicted pseudo-space reordering (x-axis) of the Smart-seq dataset (orange), the MARS-seq dataset (blue), and the 10X dataset (green).

However, in the Smart-seq dataset, Cyp8b1 is found to have largely flat expression levels along most of the lobule and lower expression toward the periportal zone and Baat appears to have an opposite trend in the 10X dataset. Other genes shown to be non-monotonically expressed such as Hamp, Igfbp2 and Mup3 in Halpern et al., 2017 display similar non-monotonic expression profiles in the Smart-seq and 10X datasets (Figure 3). The ability to identify gene expression profiles that are either high at the PP end, high at the PC end, or high in the middle of the liver lobule confirms that the sampling depth is sufficient to spatially reconstruct the liver lobule. We also investigated

225 the expression pattern of Glul in more detail as it is known to be expressed highly in a
226 one hepatocyte-wide band around the central vein[25]. Accordingly, the predicted
227 expression pattern found using all datasets demonstrated sufficient sampling of this
228 region (S3 Figure).

229 We further compared the zonation profiles between datasets by identifying genes
230 having significant differential expression along the reconstructed spatial order across
231 the periportal to pericentral axis. For genes displaying differential zonation in all
232 datasets (having adjusted p-value < .1), the Smart-seq versus MARS-seq dataset had
233 the highest median correlation (0.86), while the Smart-seq versus 10X had the lowest
234 median correlation (0.69). In Figure 4A we looked at significantly zonated genes within
235 the metabolic pathways in KEGG and found the median correlation between all datasets
236 ranged from 0.75 to 0.89. When all genes were considered the median correlation
237 ranged from 0 – 0.04.

238 Traditionally the liver lobule is divided into three zones, a periportal zone 1, a
239 pericentral zone 3, and transitioning zone 2[26,27]. The transitional nature of the liver axis
240 is reflected in the heatmap of metabolic genes that were significantly zonated in all
241 datasets (Figure 4B). Using k-means clustering, we found the Smart-seq data tended to
242 cluster into two distinct gene groups representing either the periportal or pericentral
243 zone. Examination of the two clusters by enrichment analysis of KEGG metabolic
244 pathways (Figure 4C) revealed that the predicted location along our reconstructed axis
245 of metabolic processes with known periportal or pericentral bias such as amino acid
246 metabolism (periportal), lipogenesis (pericentral), and CYP450 metabolism (pericentral)
247 corresponds to their known *in vivo* locations[27]. Despite using different reordering

algorithms and protocols, the three datasets show high agreement of expression along the recovered pericentral to periportal axis among genes that are significantly zonated in all datasets, and reliably mirror the *in vivo* patterning of the liver lobule (additional KEGG categories are shown in S4 Figure).

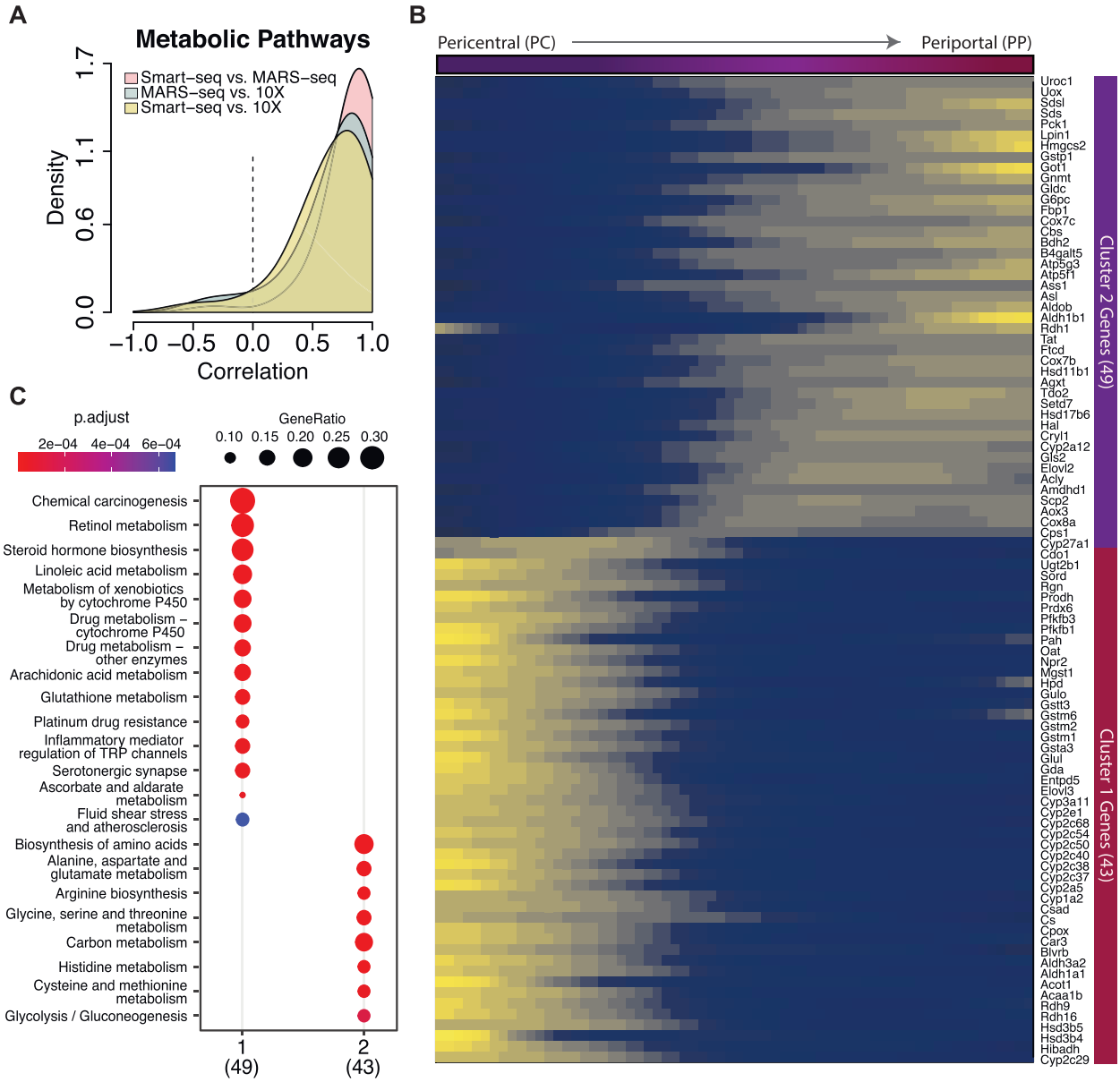


Figure 4. Correlation and Gene Ontology analysis of genes between datasets. A) Correlation analysis of significantly zonated genes annotated to the metabolic pathways in KEGG between the datasets. The pairwise correlation is shown for each

256 dataset comparison. B) Heatmap of the expression level of genes that are significantly
257 differentially zonated in all datasets and enriched in the metabolic KEGG pathway. C)
258 Breakdown of KEGG enrichment analysis of the two k-mean clusters based on the
259 genes shown in B. Dot size represents the fraction of enriched genes in each ontology,
260 and the color represents the adjusted p-value for the enrichment.
261

262 **Differences in gene profiles among lowly expressed genes and gene isoforms**

263 When we look at genes with moderate and low expression levels, we find that the
264 datasets differ to a greater degree. We identified twenty-one genes that were classified
265 as significantly zonated along the periportal to pericentral axis in the Smart-seq dataset
266 that were not detected at all in the MARS-seq dataset and thirty-five such genes not
267 detected in the 10X dataset. Compared to the Smart-seq dataset, ten genes were
268 exclusively detected in the MARS-seq dataset and no genes were exclusive to the 10X
269 dataset. Figure 5A shows the six most highly expressed genes that we were able to
270 exclusively identify in the Smart-seq dataset having significant zonation (adjusted p-
271 value < .1).

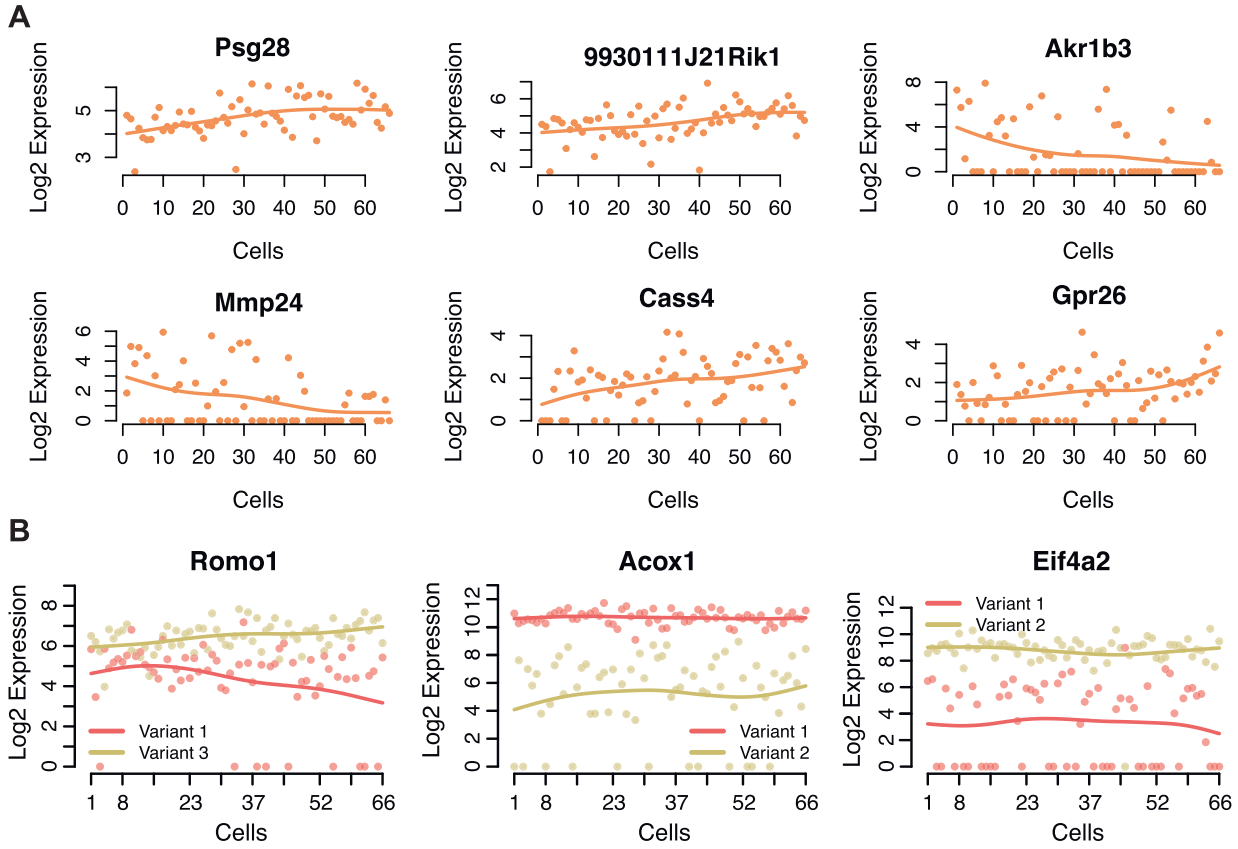


Figure 5. Genes and isoforms found in the full-length dataset and not in the UMI datasets. A) Six genes found to be zonally expressed in the Smart-seq dataset that were not detected in either the MARS-seq or 10X datasets. The log2 of expression values are represented on the y-axis and the pseudo-space ordered cells are found on the x-axis. B) Examples of genes with two transcript variants expressed differently across reordered cells from the Smart-seq dataset.

Further, an exciting field of study that benefits from an enhanced resolution of scRNA-seq is isoform analysis[28–30]. Many genes in the genome have two or more isoforms that are distinctly expressed and can change properties such as structure, function, and localization of the resulting protein[31]. Due to the increased sensitivity of full-length cDNA libraries generated by Smart-seq protocol, we were able to examine

genes with known isoforms and identify cases where the transcript variants for each isoform has distinct expression from each other across the periportal to pericentral axis, which is not possible with less sensitive protocols. In Figure 5B the transcript variants of Romo1 are seen to display opposite trends in expression across the zonation axis, where the Romo1 variant 3 is increasing in expression from the pericentral end towards the periportal end and the Romo1 variant 1 is decreasing in expression along the same axis. We also highlight genes Acox1 and Eif4a2 whose variants both show constant expression across the zonation axis but at different levels. Both of these genes are known to have isoform-specific expression in the liver lobule[32,33]. (For Ensembl and ENTEREZ IDs for transcript variants see S6 Table).

Due to UMI based protocols capturing only one end of the transcript compared to full-length cDNA procedures, there is an inability to resolve not just isoforms but also many genes that are closely related. For instance, there were 242 concatenated genes in the MARS-seq set that correspond to 539 unique genes. An example of this is seen in S5 Figure where we highlight a concatenate of Ugt1a enzymes. Eight genes are concatenated (annotated together) and when combined, the average expression level is shown to be high at the pericentral end of the lobule and low at the periportal end. Again, it is clear that not all the members of this concatenated group follow this trend as Ugt1a6a can be seen to have consistent expression levels across the pericentral to periportal axis.

Evaluating the trade-off within each protocol *in silico*

To further study the trade-offs between higher depth versus more cells, we performed a subsampling experiment. For each dataset, we held either the number of

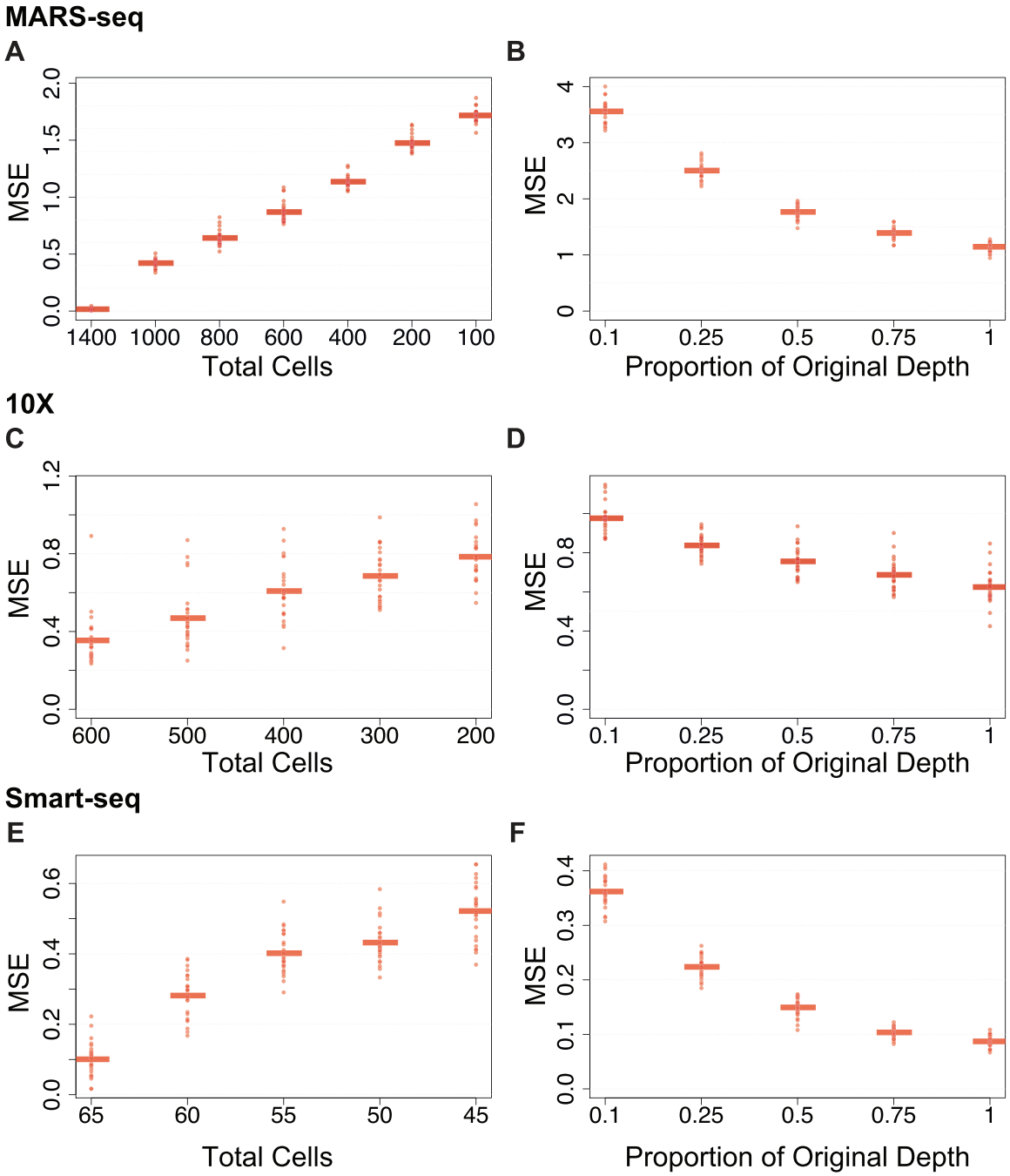
307 cells or the sequencing depth constant while varying the other. For the Smart-seq and
308 10X datasets, we evaluated the effect on the cell ordering as well as the gene-specific
309 zonation profiles. For the MARS-seq dataset, the assignment of each cell to a zonation
310 group depended on external data and was independent of the other cells profiled, and
311 thus we only evaluated the effect on zonation profiles. We estimate the MSE (mean
312 squared error) as the difference in zonation profiles in the subsampled dataset versus
313 the original dataset. In Figure 6A, the MARS-seq dataset displayed an approximately
314 linear tradeoff in zonation profile error for fewer cells at the original read depth.
315 However, at reduced read depths using the original 1,415 cells, the error increased
316 exponentially (Fig.6B). Within a dataset, we can compare the MSE between the two
317 trade-off scenarios and we find that for the MARS-seq dataset resequencing at the
318 same depth results in error that is equivalent to the reduction observed in MSE by going
319 from 600 to 1400 total cells. For the 10X dataset, we also find an approximately linear
320 tradeoff in zonation profile error for fewer cells at the original read depth (Fig. 6C).
321 However, at reduced read depths using the original 606 cells, we observe a gradual
322 increase in error as total depth decreases (Fig.6D). Similarly, by comparing the MSE
323 trade-off, it appears that resequencing at the same depth results in error that is
324 equivalent to reducing the total cells from 600 to around 400. Thus, in scenarios with
325 very low sequencing depth (average of 3-12k total UMIs per cell), sequencing deeper
326 may be more beneficial than adding more cells. For the Smart-seq dataset, we found
327 the spatial ordering to be quite robust to reduced sequencing depth, even as low as
328 50% fewer reads only marginal increased the average MSE as shown in Figure 6F. The
329 average sequencing depth for the Smart-seq cells was 3.5 million counts per cell, well

330 beyond the suggested sequencing saturation for single-cell data that occurs close to

331 one million total reads[34]. We do see more dramatic increases in error related to zonation

332 profiles when profiling fewer cells (Figure 6E). For Smart-seq data, sequencing to even

333 half of the current depth and increasing the number of cells would be beneficial.



334

335 Figure 6. Subsampling total numbers of cells and sequencing depth. A) For 25
336 subsamplings at various total numbers of cells in the MARS-seq dataset, the mean
337 squared error (MSE) of the zonation profile over 500 randomly selected genes is shown.
338 B) Similar to A, but for 25 subsamplings at various total read depths. C-D) Similar to A-
339 B, but for the Smart-seq dataset. E-F) Similar to A-B, but for the 10X dataset.
340

341 **Discussion**

342 In summary, we compared three scRNA-seq datasets of mouse hepatocytes
343 where two, MARS-seq and 10X, are wide but shallow and the other, Smart-seq, is
344 narrow but deeply sequenced. We find that the three different protocols present highly
345 reproducible liver zonation profiles in single cells, and for the vast majority of genes that
346 are highly expressed we observe highly comparable results. Our results were not
347 dependent on any one computational method or pre-processing pipeline. We do
348 however find that when we look at medium to low expressed genes, the increased
349 sensitivity of the C1/Smart-seq protocol is able to identify several genes exclusive to this
350 dataset. This increased sensitivity also allowed us to identify several genes with
351 isoforms that behaved differently across the periportal to pericentral axis. Though in
352 general, there are still limitations of short reads in regard to isoform analysis and if more
353 accuracy is needed, the newly developed technique ScISOr-seq[35] might be better
354 suited. We do however believe that this full-length data allows for more reliable
355 preliminary isoform analysis compared to either UMI method. However, the main
356 weakness of using fewer cells is that it is less likely that rare cell types will be sampled.
357 In cases where such rare cells are of high interest, protocols that produce a large

358 number of cells are preferable. In an ideal case, one would sample many cells and
359 sequence all of them deeply; unfortunately, this is not always possible in practice and
360 the decision of whether to sample many cells shallowly or fewer cells deeply comes
361 down to whether rare cell types are of interest or if higher resolution of the individual
362 cells is preferred.

363 Given the distinct advantages of the protocols, we emphasize that the biological
364 question should be the driving factor when deciding on protocol. Within a chosen
365 protocol, achieving balance between the sequencing depth and the number of cells is
366 still an important consideration for optimal use of resources. Based on our simulations
367 of datasets at opposite ends of the sequencing depth versus number of cells trade-off,
368 there is eventually a detriment to sacrificing reads for additional cells or sequencing
369 beyond the attainable sensitivity level on too few cells. We expect that the extent of the
370 cells versus depth trade-off will vary for other cell types or tissues and it will largely
371 depend on the heterogeneity of the biological system under study.

372

373

374 **Methods**

375 **Animals and handling.**

376 All animals were kept under standard husbandry conditions. A wildtype 8-week-old male
377 C57BL/6 (Jackson laboratories) was used in this experiment. Using isoflurane, the
378 mouse was anesthetized before euthanizing by cervical dislocation. Animal experiments
379 and procedures were approved by the University of Wisconsin Medical School's Animal

380 Care and Use Committee and conducted in accordance with the Animal Welfare Act
381 and Health Research Extension Act.
382
383 **Cell isolation.**
384 The euthanized mouse was pinned to a Styrofoam plate using 20 ga needles to aid in
385 dissection. The abdominal cavity was opened, and the portal vein exposed. A piece of
386 4-0 suture thread (Ethicon vicryl coated) was threaded under the portal vein and used to
387 secure a 26 ga catheter inserted into the portal vein (Butler Schein animal health 26 G
388 IV Catheter, Fisher Scientific). Hepatocytes were isolated using a 2-step perfusion
389 protocol. First, Liver Perfusion Medium (Gibco) warmed to 37°C was pumped through
390 the catheter for 10 minutes using a peristaltic pump at 7 ml/min flowrate. Then, Liver
391 Digest Medium (Gibco) warmed to 37°C was pumped through the liver at the same
392 settings for 10 minutes. After perfusion, the liver was excised and transferred to a 10 cm
393 dish containing 20 ml liver digest medium. The liver was dissected, allowing the cells to
394 spill into the media. The cells were then filtered through a 40 µm cell strainer into a 50
395 ml tube and 30 ml media (Williams E media + 2 µg/ml human insulin + 1x glutamax +
396 10% FBS) were added and placed on ice. The hepatocytes were purified by
397 centrifugation at 50 x G, 4 times for 3 minutes each, each time discarding the
398 supernatant and adding media.
399
400 **Single cell RNA sequencing- Full-length dataset**
401 Single-cell RNA sequencing was performed as previously described[4,5] with the following
402 modifications. In this study, we used small (5-10 µm), medium (10-17 µm), and large

403 (17-25 μm) plate sizes. ERCC RNA Spike-In (ThermoFisher Cat. No. 4456740) was
404 diluted in the lysis mix following the manufacturer's user guide and previous studies[36].
405 Single end reads of 51 bp were sequenced on an Illumina HiSeq 2500 system.
406 Sequencer outputs were processed using Illumina's CASAVA-1.8.2. The demultiplexed
407 reads were trimmed and filtered to eliminate adapter sequence and low-quality
408 basecalls. The reads were mapped to an mm10 mRNA transcript reference (extended
409 with ERCC transcripts) using bowtie-0.12.9[37]; expression estimates were generated
410 using RSEM v.1.2.3[38]. Using the Fluidigm C1 system to capture and synthesize cDNA
411 from single cells in the liver, we generated transcriptomes for 149 cells. To exclude low
412 quality transcriptomes, we removed cells in which the fraction of ERCC spike-in made
413 up 20% or more of the total assigned reads. This left 66 high quality cells that were
414 used in the downstream analysis. Finally, the data was normalized using SCnorm (R
415 package v 1.5.7)[39].
416
417 **Pseudo-spatial reordering- Full-length dataset**
418 For the full-length data, the cells were computationally ordered using the Wave-Crest
419 method as described in Chu et al. 2016[5]. For the reordering step, gene expression
420 values were rescaled to mean 0 and variance 1 to ensure the values across different
421 genes are comparable. The Wave-Crest algorithm implements an extended nearest
422 insertion algorithm that iteratively adds cells to the order and selects the insertion
423 location as the location producing the smallest mean squared error in a linear
424 regression of the proposed order versus gene expression. A 2-opt algorithm is then
425 used to find an optimal cell order by considering adjacent cell exchanges. The cell

426 ordering step uses the expression profiles of pre-selected known marker genes of liver
427 zonation. Thus, the resulting linear profile of ordered cells represents the periportal to
428 pericentral axis. The known marker genes used to construct the periportal to pericentral
429 axis in Wave-Crest include the following pericentral markers: cytochrome P450 7a1
430 (Cyp7a1), cytochrome P450 2e1 (Cyp2e1), ornithine aminotransferase (Oat),
431 cytochrome P450 1a2 (Cyp1a2), rh family, B glycoprotein (Rhbg), leucine-rich repeat-
432 containing G-protein coupled receptor 5 (Lgr5), glutamate-ammonia ligase (Glul); and
433 the following periportal markers: phosphoenolpyruvate carboxykinase 1 (Pck1), catenin
434 beta interacting protein 1 (Ctnnbip1), aldehyde dehydrogenase 1 family member B1
435 (Aldh1b1), sulfotransferase family 5A, member 1 (Sult5a1), cytochrome P450 2f2
436 (Cyp2f2), cathepsin C (Ctsc), serine dehydratase (Sds), and E-cadherin (Cdh1). All
437 markers were selected based on their expression ratio as reported by Braeuning et al.
438 2006[20].
439
440 A detection step was done to identify additional genes that follow the one-dimensional
441 periportal to pericentral axis by fitting a linear regression to the relationship between
442 each gene's expression and the Wave-Crest cell order. To determine if a gene is
443 significantly dynamic (differentially zonated) along the recovered axis, we tested
444 whether the regression slope is different from zero. We reported the Benjamini-
445 Hochberg adjusted p-values to control the false discovery rate. For genes having an
446 adjusted p-value < .01, the direction of the expression profile was assigned based on
447 the sign of the regression slope (periportal: positive slope, pericentral: negative slope).
448 We also calculated the linear fitting mean squared error (MSE) for each gene. Genes

449 with a smoother trend over the recovered cell order are expected to have a smaller
450 MSE. We report the full list of genes, sorted by their MSE, in S7 Table; scatter plots for
451 genes having adjusted p-value < .01 are shown in S8 File.
452
453 **Pseudo-spatial reordering- 10X dataset**
454 The 10X dataset was downloaded from the Tabula Muris compendium public resource
455 via Figshare[22] . The 10X data was originally processed using the CellRanger version
456 2.0.1. Within the liver cells, the authors originally identified 975 hepatocytes. For our
457 analysis, we performed a second quality control step to identify cells with low RNA
458 content, possible doublets, or dead/damaged cells, where we filtered cells based on the
459 total number of genes expressed per cell. Using the Seurat R package v3.1.5,
460 hepatocytes were further filtered to those having between 200 and 3000 genes detected
461 per cell (only one cell had more than 5000 genes detected per cell). Next, we clustered
462 the cells using Seurat, where a k-nearest neighbors (KNN) graph used was constructed
463 based on the first 20 principle components to create a shared nearest neighbors graph
464 based on the Jaccard index between each cell and its 20 nearest neighbors, as
465 implemented in the FindNeighbors function. Clusters were then identified by partitioning
466 this graph using the Louvain community detection algorithm with a resolution of .5, as
467 implemented in the FindClusters function. The cells clustered into three distinct larger
468 groups and we retained only the largest grouping of cells that clustered together,
469 resulting in 606 total cells. The data was then normalized using scran v1.12.1. Next, we
470 used Monocle v2.12.0 to order the cells, basing the ordering on the top 200 highly
471 variable genes estimated using the mean variance relationship via the

472 FindVariableFeatures function in Seurat. To determine if a gene is significantly dynamic
473 (differentially zonated) along the recovered axis, the Monocle2 function
474 differentialGeneTest was used to fit a spline on gene expression versus the estimated
475 pseudo-time.
476
477 **Comparative Analysis**
478 Smoothed densities (bean plots) with overlaid raw data, the mean, and a box
479 representing the interquartile range of the cellular detection fractions were created using
480 the pirateplot function in the yarrr R package (v0.1.5). The cellular detection fraction
481 was calculated per cell as the proportion of genes having expression greater than zero.
482 The fold-change for each gene between the two datasets (A versus B) was calculated
483 as the log2 fold-change of the dataset A over dataset B, where each gene mean was
484 calculated as the average expression among non-zero counts across all cells in the
485 datasets. The heatmap in Figure 2 of marker gene expression on the normalized Smart-
486 seq data was generated by setting values above the 95th percentile or below the 5nd
487 percentile to the 95th percentile or 5nd percentile value, respectively.
488
489 Due to the datasets having different dynamic ranges, we used scaled expression plots
490 to compare expression profiles, where the ordered cells in the full-length dataset and
491 10X were each divided into nine equally sized groups to correspond to the nine layers in
492 the UMI dataset. For the full-length and 10X dataset, for a given gene, the median (full-
493 length) or mean (10X) expression in each group was calculated, then the nine values
494 were scaled between zero and one. Smoothed fits were overlaid using the

495 smooth.spline function in R with the degrees of freedom parameter df=4. Expression
496 correlations along the zonation axis between datasets were calculated using Pearson
497 correlation. Enrichment of genes in KEGG pathways or GO was done using the R
498 package clusterProfiler (v. 3.10.1)[40]. For the enrichment analysis, since different
499 statistical methods were used to assess zonation profiles, genes were considered
500 significantly zonated if they had an adjusted p-value < .1 in all datasets. The heatmap in
501 Figure 3 is a smoothed heatmap, where a smoothing spline was first fit to the log
502 expression (pseudo-count of one added) of each gene using the smooth.spline function
503 in R with the smoothing parameter df=4 which provided profiles that were not over- or
504 underfit in either dataset. Then the smoothed expression was scaled and outliers above
505 the 98th percentile or below the 2nd percentile were set to the 98th percentile or 2nd
506 percentile value, respectively. Additional KEGG categories from this analysis can be
507 interactively viewed on Github
508 https://github.com/rhondabacher/scSpatialReconstructCompare-Paper.
509

510 **Subsampling Analysis**

511 In all subsamplings described below, each scenario was repeated a total of 25 times
512 and the zonation group means were scaled to be between zero and one.
513

514 For the MARS-seq dataset, zonation group means were recalculated on a subsampled
515 set of cells using the posterior probability matrix and original UMI counts from Halpern
516 et al. 2017. In each sampling, the mean squared error (MSE) was calculated based on a
517 random sample of 500 genes as $\sum_{i=1}^{500}\sum_{j=1}^{9}(Z_{i,j} - \hat{Z}_{i,j})^2 \,/500$, where $Z_{i,j}$ represents the

518 mean expression of gene i in zonation group j in the original dataset and $\hat{Z}_{i,j}$ is the
519 corresponding value for the subsampled dataset. For subsampling at lower read depths,
520 we fixed the number of cells at the original total of 1415 cells and simulated each cell's
521 gene counts individually using a multinomial distribution. For each cell, the subsampled
522 total counts were set to X% of the original total read counts for that cell (for X =
523 (10,20,30,40,50,60,70,80,90,100)) and each gene's cell-specific probability was
524 calculated as its original count divided by the original total counts for that cell. The MSE
525 was calculated for each subsampled set as described above.
526
527 For the Smart-seq dataset, we reran Wave-Crest when subsampling the total number of
528 cells using the original parameter settings and marker genes. Then, as before, the
529 ordered cells were assigned zonation groups by dividing cells into nine equally sized
530 groups. The zonation profile error was estimated using MSE and calculated as
531 described above with the exception that since Wave-Crest orders can be flipped, we
532 calculated the MSE on the returned order and its reverse, and kept the minimum MSE
533 of the two. To evaluate the zonation profile error with lower read depths, we used a
534 similar approach as described above for the MARS-seq dataset, fixing the number of
535 cells to be the same as the original total of 66 and, since the order correlation was
536 shown to be consistently high, we used the original Wave-Crest order for every scenario
537 when evaluating zonation profile error. For the 10X dataset, the subsampling was
538 performed similarly as for the Smart-seq dataset, however Monocle2's ordering was
539 more variable as it was not based on marker genes and thus we did not fix the order
540 when evaluating the zonation profile error. Trade-offs in MSE are directly comparable

541 within a dataset but due to intrinsic differences in the original processing and in
542 subsampling, the MSE should not be compared across the datasets.
543
544 **Immunohistochemistry**
545 An 8-week-old male C57BL/6 mouse was anesthetized using isoflurane before
546 euthanizing by cervical dislocation. The liver was excised, sliced as thinly as possible
547 with a razor blade, and fixed in formaldehyde overnight. The liver slices were paraffin
548 embedded and sectioned. Sections were stained following the protocol published by
549 Abcam (http://www.abcam.com/ps/pdf/protocols/ihc_p.pdf). In short, the slices are
550 deparaffinized by dipping into sequential solutions of 100% xylene, 50-50% xylene-
551 ethanol, 100% ethanol, 95% ethanol, 70% ethanol, 50% ethanol, and tap water. The
552 antigens were then retrieved by placing the slides in Tris-EDTA buffer (10 mM Tris
553 Base, 1 mM EDTA Solution, 0.05% Tween 20, pH 9.0) and incubating them in a
554 decloaking chamber (Biocare Medical Decloaking Chamber #DC2008US) with the
555 following settings: delayed start 30 sec.; preheat 80°C, 2 min.; heat 101°C, 3 min. 30
556 sec.; and fan on. The slides were washed 2 x 5 min in TBS + 0.025% Triton X-100
557 before they were blocked for two hours at room temperature in 10% normal serum in
558 1% BSA. The appropriate primary antibody was then diluted in the same 10% normal
559 serum in 1% BSA, added to the slides, and incubated at 4ºC overnight in an incubation
560 chamber. The next day the slides were washed 2 x 5 min in TBS + 0.025% Triton X-100
561 followed by 15 min incubation in 0.3% H_2O_2 at room temperature. Next, the appropriate
562 secondary antibody was diluted into 10% normal serum in 1% BSA before it was added
563 to the slides and incubated for 1 hour at room temperature. The slides were then

564 washed 3 x 5 min in TBS before DAB (#ab103723) staining mixed according to
565 manufacturer instruction was applied and incubated under a microscope to stop the
566 reaction after sufficient staining. The slides were rinsed in tap water for 5 min before
567 being counterstained with Mayer's hematoxylin (#MHS1-100ML) for 30 sec. The stain
568 was developed in running tap water for 5 min. The slides were then dehydrated by
569 sequentially dipping in 50% ethanol, 70% ethanol, 95% ethanol, 100% ethanol, 50-50%
570 xylene-ethanol, and 100% xylene before Poly-Mount (#08381-120) was added and a
571 coverslip placed on top. The following primary antibodies were added: Aldh3a4 1:250
572 (AB184171), Cyp2e1 1:50 (AB28146), Cyp1a2 1:50 (R31007), Rgn 1:100 (NBP1-
573 80849), Oat 1:50 (AB137679), Cyp2f2 1:100 (SC-67283), Hal 1:50 (AV45694), and
574 Tbx3 1:50 (SC-31657). The following secondary antibodies were used: goat-anti-rabbit
575 HRP conjugated (ab97051) and donkey-anti-goat HRP conjugated (ab97110) at a
576 concentration of 1:500.

577 **Acknowledgements**

578 Not applicable.

579 **References**
580

581 1. Tang, F. *et al.* mRNA-Seq whole-transcriptome analysis of a single cell. *Nat Methods* **6**,
582 377–382 (2009).
583 2. Ramsköld, D. *et al.* Full-length mRNA-Seq from single-cell levels of RNA and individual
584 circulating tumor cells. *Nat Biotechnol* **30**, 777–782 (2012).
585 3. Islam, S. *et al.* Characterization of the single-cell transcriptional landscape by highly
586 multiplex RNA-seq. *Genome Research* **21**, 1160–1167 (2011).

587 4. Leng, N. *et al.* Oscope identifies oscillatory genes in unsynchronized single-cell RNA-seq
588 experiments. *Nat Methods* **12**, 947–950 (2015).
589 5. Chu, L.-F. *et al.* Single-cell RNA-seq reveals novel regulators of human embryonic stem cell
590 differentiation to definitive endoderm. *Genome Biol* **17**, 173 (2016).
591 6. Klein, A. M. *et al.* Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic
592 Stem Cells. *Cell* **161**, 1187–1201 (2015).
593 7. Macosko, E. Z. *et al.* Highly Parallel Genome-wide Expression Profiling of Individual Cells
594 Using Nanoliter Droplets. *Cell* **161**, 1202–1214 (2015).
595 8. Islam, S. *et al.* Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat*
596 *Methods* **11**, 163–166 (2014).
597 9. Ziegenhain, C. *et al.* Comparative Analysis of Single-Cell RNA Sequencing Methods.
598 *Molecular Cell* **65**, 631-643.e4 (2017).
599 10. Svensson, V. *et al.* Power analysis of single-cell RNA-sequencing experiments. *Nat Methods*
600 **14**, 381–387 (2017).
601 11. Guo, M. *et al.* Single cell RNA analysis identifies cellular heterogeneity and adaptive
602 responses of the lung at birth. *Nat Commun* **10**, 37 (2019).
603 12. Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by
604 pseudotemporal ordering of single cells. *Nat Biotechnol* **32**, 381–386 (2014).
605 13. Ji, Z. & Ji, H. TSCAN: Pseudo-time reconstruction and evaluation in single-cell RNA-seq
606 analysis. *Nucleic Acids Res* **44**, e117–e117 (2016).
607 14. Saelens, W., Cannoodt, R., Todorov, H. & Saeys, Y. A comparison of single-cell trajectory
608 inference methods. *Nat Biotechnol* **37**, 547–554 (2019).

609 15. Burger, H.-J., Gebhardt, R., Mayer, C. & Mecke, D. Different capacities for amino acid
610 transport in periportal and perivenous hepatocytes isolated by digitonin/collagenase
611 perfusion. *Hepatology* **9**, 22–28 (1989).
612 16. Pösö, A. R., Penttilä, K. E., Suolinna, E. M. & Lindros, K. O. Urea synthesis in freshly
613 isolated and in cultured periportal and perivenous hepatocytes. *Biochem. J.* **239**, 263–267
614 (1986).
615 17. Tosh, D., Alberti, G. M. M. & Agius, L. Glucagon regulation of gluconeogenesis and
616 ketogenesis in periportal and perivenous rat hepatocytes. Heterogeneity of hormone action
617 and of the mitochondrial redox state. *Biochem. J.* **256**, 197–204 (1988).
618 18. Guzmán, M. & Castro, J. Zonation of fatty acid metabolism in rat liver. *Biochem. J.* **264**,
619 107–113 (1989).
620 19. Anundi, I., Lähteenmäki, T., Rundgren, M., Moldeus, P. & Lindros, K. O. Zonation of
621 acetaminophen metabolism and cytochrome P450 2E1-mediated toxicity studied in isolated
622 periportal and perivenous hepatocytes. *Biochemical Pharmacology* **45**, 1251–1259 (1993).
623 20. Braeuning, A. *et al.* Differential gene expression in periportal and perivenous mouse
624 hepatocytes. *FEBS Journal* **273**, 5051–5061 (2006).
625 21. Halpern, K. B. *et al.* Single-cell spatial reconstruction reveals global division of labour in the
626 mammalian liver. *Nature* **542**, 352–356 (2017).
627 22. The Tabula Muris Consortium *et al.* Single-cell transcriptomics of 20 mouse organs creates a
628 Tabula Muris. *Nature* **562**, 367–372 (2018).
629 23. Phipson, B., Zappia, L. & Oshlack, A. Gene length and detection bias in single cell RNA
630 sequencing protocols. *F1000Res* **6**, 595 (2017).

631 24. Parekh, S., Ziegenhain, C., Vieth, B., Enard, W. & Hellmann, I. The impact of amplification
632 on differential expression analyses by RNA-seq. *Sci Rep* **6**, 25533 (2016).
633 25. Gebhardt, R. & Mecke, D. Heterogeneous distribution of glutamine synthetase among rat
634 liver parenchymal cells in situ and in primary culture. *The EMBO Journal* **2**, 567–570
635 (1983).
636 26. Bhatia, S. N. *et al.* Zonal liver cell heterogeneity: effects of oxygen on metabolic functions of
637 hepatocytes. *Cell Eng* **1**, 125–135 (1996).
638 27. Kietzmann, T. Metabolic zonation of the liver: The oxygen gradient revisited. *Redox Biology*
639 **11**, 622–630 (2017).
640 28. Song, Y. *et al.* Single-Cell Alternative Splicing Analysis with Expedition Reveals Splicing
641 Dynamics during Neuron Differentiation. *Molecular Cell* **67**, 148-161.e5 (2017).
642 29. Karlsson, K., Lönnerberg, P. & Linnarsson, S. Alternative TSSs are co-regulated in single
643 cells in the mouse brain. *Mol Syst Biol* **13**, 930 (2017).
644 30. Arzalluz-Luque, Á. & Conesa, A. Single-cell RNAseq for the study of isoforms—how is that
645 possible? *Genome Biol* **19**, 110 (2018).
646 31. Wang, E. T. *et al.* Alternative isoform regulation in human tissue transcriptomes. *Nature*
647 **456**, 470–476 (2008).
648 32. Gruppuso, P. A., Tsai, S.-W., Boylan, J. M. & Sanders, J. A. Hepatic translation control in
649 the late-gestation fetal rat. *American Journal of Physiology-Regulatory, Integrative and*
650 *Comparative Physiology* **295**, R558–R567 (2008).
651 33. Oaxaca-Castillo, D. *et al.* Biochemical characterization of two functional human liver acyl-
652 CoA oxidase isoforms 1a and 1b encoded by a single gene. *Biochemical and Biophysical*
653 *Research Communications* **360**, 314–319 (2007).

654 34. Haque, A., Engel, J., Teichmann, S. A. & Lönnberg, T. A practical guide to single-cell RNA-
655 sequencing for biomedical research and clinical applications. *Genome Med* **9**, 75 (2017).
656 35. Gupta, I. *et al.* Single-cell isoform RNA sequencing characterizes isoforms in thousands of
657 cerebellar cells. *Nat Biotechnol* **36**, 1197–1202 (2018).
658 36. Brennecke, P. *et al.* Accounting for technical noise in single-cell RNA-seq experiments. *Nat*
659 *Methods* **10**, 1093–1095 (2013).
660 37. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient
661 alignment of short DNA sequences to the human genome. *Genome Biol* **10**, R25 (2009).
662 38. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or
663 without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).
664 39. Bacher, R. *et al.* SCnorm: robust normalization of single-cell RNA-seq data. *Nat Methods*
665 **14**, 584–586 (2017).
666 40. Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterProfiler: an R Package for Comparing
667 Biological Themes Among Gene Clusters. *OMICS: A Journal of Integrative Biology* **16**,
668 284–287 (2012).
669
670
671
672
673
674 **Supporting information captions**
675 S1 Figure – Examining GC content and gene length in genes with a higher detection
676 fraction in either dataset. Top) The GC content (left) and gene length (right) are shown

677 for genes having a higher detection fraction in either the Smart-seq dataset (gray) or the
678 MARS-seq dataset (blue). A dotted line is shown for genes having a larger mean in
679 either dataset. The two lines closely correspond since the genes having a high detection
680 fraction typically have a higher mean. Bottom) Similar to the top for comparing the
681 Smart-seq and 10X datasets.
682 S2 Figure – Correlation between WaveCrest and Monocle methods for ordering cells in
683 the Smart-seq dataset.
684 S3 Figure – Expression of Glul. Scaled expression plots of Glul showing high correlation
685 among all three datasets.
686 S4 Figure – Correlation analysis of more KEGG pathways. A) Top left: Correlation
687 analysis for genes in the KEGG pathway “Complement and coagulation cascade”. The
688 pairwise correlation is shown for each dataset comparison. Following are plots for the
689 eight highest correlated genes between the any two datasets in that pathway. On the
690 right is a smoothed heatmap of the Smart-seq expression data for the gene expression
691 of all significantly zonated genes enriched in that KEGG pathway. B) Similar to (A) but
692 for the “Drug metabolism – cytochrome P450” pathway. C) Similar to (A) but for the
693 “Biosynthesis of amino acids” pathway.
694 S5 Figure – Additional genes in Smart-seq dataset but not in the MARS-seq dataset.
695 Eight Ugt1a genes that were concatenated in the MARS-seq dataset (blue on all
696 graphs), but can be resolved in the Smart-seq dataset (orange line).
697 S6 Table – Ensembl and RefSeq ID’s for genes with transcript variants.
698 S7 Table – Summary of genes with dynamic expression across the zonation axis
699 identified using Wave-Crest.

700 S8 File – Scatter plots of dynamic genes listed in S6 Table.

701 S9 Dataset – Normalized Smart-Seq single-cell data with cells in the Wave-Crest order.

702