Tradeoff between more cells and higher read depth for single-cell RNA-seq spatial ordering analysis of the liver lobule

Morten Seirup[1,2*], Li-Fang Chu[2], Srikumar Sengupta[2], Ning Leng[2,3,#a], 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[2,4,5*], Rhonda Bacher[6,*]

[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 Cell & Regenerative Biology, University of Wisconsin School of Medicine and Public Health, Madison, Wisconsin, United States of America
[5]Department of Molecular, Cellular, & Developmental Biology, University of California Santa Barbara, Santa Barbara, California, United States of America
[6]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 single-cell experiments generate increasingly more cells at reduced sequencing depths, the value of a higher read depth may be overlooked. Using data from two contrasting single-cell RNA-seq protocols that lend themselves to having either higher read depth (Smart-seq) or many cells (MARS-seq) we evaluate the trade-offs in the context of pseudo-spatial reconstruction of the liver lobule. Overall, we find gene expression profiles after spatial-reconstruction analysis are highly reproducible between datasets. Smart-seq's higher sensitivity and read-depth allows analysis of lower expressed genes and isoforms. Our analysis emphasizes the importance of selecting a protocol based on the biological questions and features of interest. 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.

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 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 trade-off of more cells versus higher read depth in the context of pseudo-spatial reconstruction by comparing two independently produced scRNA-seq datasets on mouse liver lobule, one using Smart-seq[2]--a full-length protocol and one using MARS-seq[11]--a UMI based protocol. Although the cell number and read depth differ greatly, we find high reproducibility between protocols of gene expression profiles after spatial-reconstruction analysis. We find that the increased read depth of the Smart-seq protocol enables studies of lower expressed genes and isoforms of genes. Our results demonstrate the importance of carefully evaluating the biological question and features of interest when selecting the appropriate sequencing protocol. In applications focused on lower expressed genes or on genes with high sequence similarity, increased read depth is preferable, whereas a focus on identifying cell types based on more highly expressed genes will benefit from collecting more cells. In an ideal situation a single cell assay would result in thousands of cells that are all sequenced at a high read depth, but technical and financial restrictions make this rarely possible.

Studies comparing protocols have mainly done so with respect to performance on spike-ins or on technical variability alone[9,10]. Recently, Guo et al.[12] showed agreement of cell types and signature genes between two protocols used for single-cell RNA-seq for Fluidigm C1 and Drop-seq. However, few studies have examined comparative agreement among protocols for biological inferences beyond clustering and identifying differential gene expression, and a key question of interest with single-cell data is its ability to reflect temporal or spatial heterogeneity. For cells collected at a given time, the underlying dynamic biological process is reflected in genome-wide differences in gene expression. Computational algorithms that attempt to order cells in pseudo-time or pseudo-space based on variability in gene expression have been developed[4,13,14], and more than 45 existing algorithms were recently compared[15]. Yet, as far as we know, no comparison of single-cell protocols exists for the question of cell ordering.

Here, we chose to compare protocols on their ability to reflect the spatial patterning of the liver lobule. The main functional cells of the liver, hepatocytes, are organized spatially in a polygonal shape around a central vein (Figure 1A). From the central vein, a gradient of metabolic functions is performed extending to a portal vein at each vertex[16–20]. The gradient of differences in gene expression patterns is referred to as the zonation axis (from periportal (PP) to pericentral (PC))[21]. This coordinated spatial organization provides a particularly interesting application of single-cell techniques. For this study we obtained scRNA-seq data from 66 hepatocytes using the Fluidigm C1 system with the Smart-seq full-length protocol, and compare this dataset at the gene level to a dataset collected by Halpern et al. 2017 containing 1415 hepatocytes using the MARS-seq protocol with UMI's[22] (Figure 1A). We compare the ability of these two single-cell datasets to spatially resolve the zonation axis of the liver.

Results

By using the Fluidigm C1 coupled with the Smart-seq protocol, we were able identify on average around 38% (about 7100 genes) (Figure 1B) of all genes in the genome expressed per cell, whereas the MARS-seq dataset finds on average 12% (about 2100 genes) (Figure 1B) of all genes in the genome expressed per cell. This is in accordance with what was found by Ziegenhain et al. 2017 when they examined the methods, and underscores the increased sensitivity of the Fluidigm C1/SMART-seq protocol over MARS-seq[9]. This increased sensitivity is further illustrated in Figure 1C, which on a per gene level shows the difference in detection fraction compared to the log fold change in mean expression between the two protocols. A difference in detection fraction of zero means that the gene is detected in the same fraction of cells in both datasets and a positive value is the result of a gene detected in a larger fraction of cells in the Smart-seq protocol compared to the MARS-seq protocol, and a negative value corresponds to the opposite case where the MARS-seq protocol detects the gene in a higher fraction of cells. The difference across protocols in log2 fold-change has a linear relationship with the difference in detection fractions, which indicates a fairly constant increase in log2 expression expected as cells are sequenced with greater sensitivity. At the intercept, a difference in detection equal to zero, the log2 fold change is 3.4, indicating an experiment wide increase in sensitivity in the Smart-seq protocol of approximately 10-fold. In fact, the vast majority of genes are detected in a larger fraction of cells (positive value on the x-axis) and have a higher expression level (positive value in the y-axis) sequenced using Smart-seq protocol. Although, it is worth pointing out that around 6% of genes have higher detection using the MARS-seq protocol (negative values on x-axis) and a few of these genes also have higher expression levels (negative values on y-axis) than in the Smart-seq protocol. The subset of genes better detected in the MARS-seq dataset have higher GC content and are slightly longer (Supplementary Figure 1), which is consistent with previous reports of protocol comparisons[23,24].

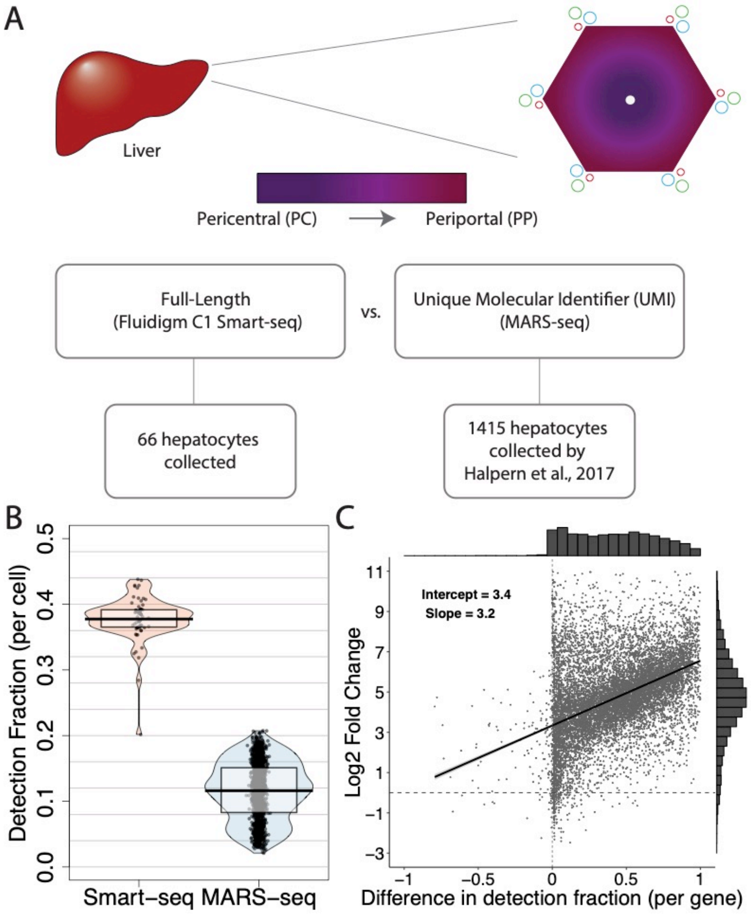

Figure 1. Illustration of the liver anatomy, and general comparison of the datasets. A) Top. Illustration of the liver lobule identifying the portal triad along the outer edges and the central vein in the middle. The color gradient represents metabolic zonation. A) Bottom. Highlights the main differences 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) 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.

Next, to represent the spatial patterns across the liver lobule, the cells in the two datasets were computationally ordered according to their expression profiles. The MARS-seq dataset was spatially ordered by Halpern et al. 2017 by first performing smFISH for six marker genes at various locations across the zonation axis, then single-cell RNA-seq data obtained by MARS-seq were assigned into one of nine zonation locations based on each cell's expression profile of the six marker genes[22]. For the Smart-seq protocol we used a computational algorithm called Wave-Crest to spatially order the 66 cells along the zonation axis (Figure 2A)[5]. The ordering is based on fifteen marker genes known in the literature to be differentially expressed along the zonation axis. Cells were ordered using the nearest insertion algorithm implemented in the Wave-Crest package. The algorithm searches among the space of all possible orderings via a 2-opt algorithm by considering insertion events and choosing orders which minimize the mean square error of a polynomial regression on the marker genes expression. Of the 15 genes used, we selected eight periportal expressed genes and seven pericentral expressed genes[21]. Both orderings assume the zonation profile and spatial organization can be represented in a single dimension.

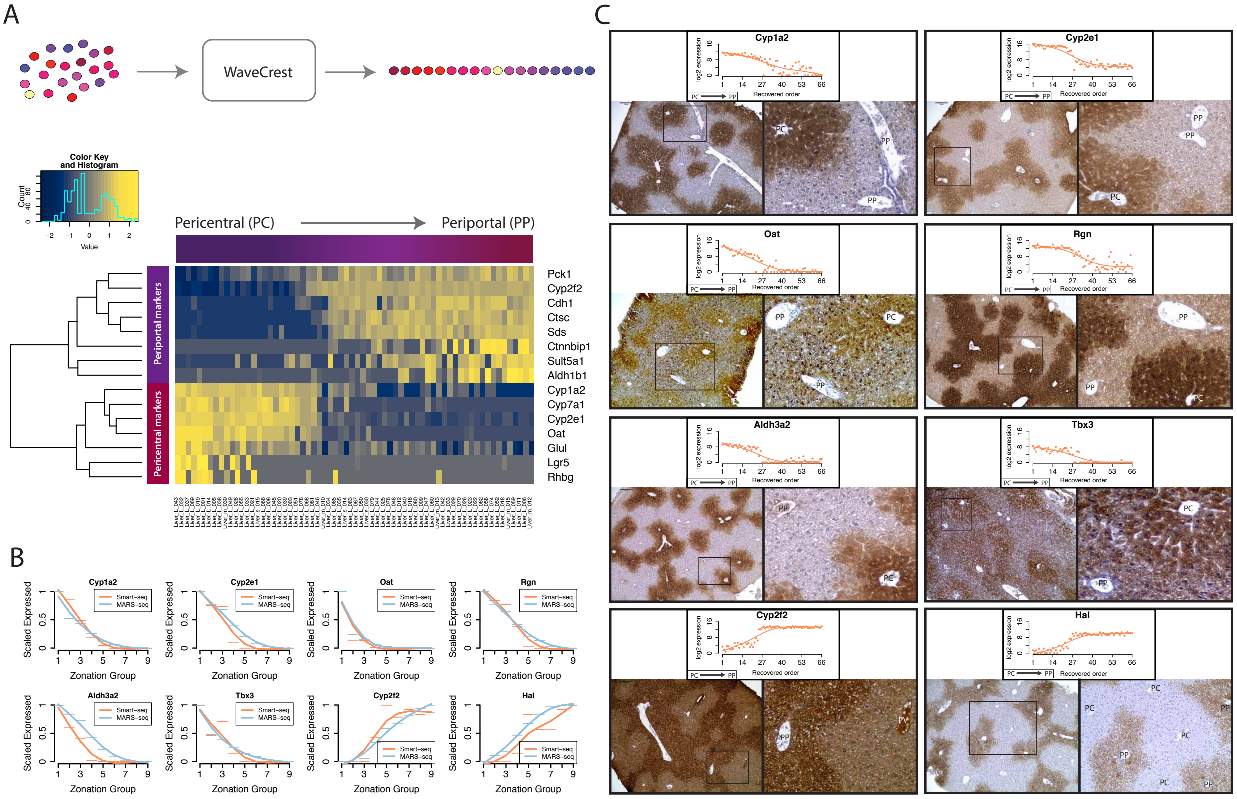

Figure 2. Pseudo-space reordering of hepatocytes, and prediction and validation of dynamically expressed genes. A) Top. Illustration of the pseudo-spatial reordering of the Smart-seq experiment. Bottom. Heatmap showing the pseudo-spatial reordering (x-axis) and the expression levels of the marker genes (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), and the MARS-seq dataset (blue). C) Immunohistochemistry staining of the genes highlighted in B). Above the staining is the predicted log2 expression levels (y-axis) across the pseudo-spatial order (x-axis). The left picture shows the staining and the right picture is an enlarged section (black square). PP = Periportal, PC = Pericentral.

Using the recreated order of the hepatocytes we explored dynamic gene expression across the periportal to pericentral axis. Figure 2B shows a subset of genes that are predicted to be highly regulated across the axis, four of which were not in our list of marker genes. We first compared their expression across the zonation axis in the Smart-seq dataset to that from the MARS-seq dataset. Since the MARS-seq dataset placed cells into nine discrete zones along the axis, we divided cells from the Smart-seq dataset into nine equally sized groups. The zonation profiles in Figure 2B have high agreement, with a median Spearman correlation of 0.93. Before proceeding, we also performed an additional experiment to validate that our cell ordering and expression profiles reflect those of the liver lobule *in vivo*. Immunohistochemistry was performed on sections of paraffin embedded livers with antibodies against select genes from either category (Figure 2C). A complete list of dynamic genes across the zonation axis from the Smart-seq dataset is provided in Additional File 2, and scatter plots are in Additional File 3.

An exciting prospect of single cell analysis is the identification of genes that have non-monotonic or dynamic expression along the liver lobule. Several genes in the bile acid synthesis pathway was shown by Halpern et al. to be non-monotonically expressed in a pattern where the highest expression levels along the lobule corresponds to the functional placement of the genes in the bile acid synthesis pathway (Cyp7a1, Hsd3b7, Cyp8b1, Cyp27a1 and Baat). We find that the expression profiles for these genes, besides Cyp8b1, found in the Smart-seq dataset match the patterns found in the MARS-seq dataset (Supplementary Figure 3A). 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. Other genes shown to be non-monotonically expressed such as Hamp, Igfbp2 and Mup3 in Halpern et al. were also identified to be non-monotonically expressed in the Smart-seq dataset (Supplementary Figure 3B). 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 the expression pattern of Glul in more detail as it is known to be expressed highly in a one hepatocyte wide band around the central vein[25]. Accordingly, the predicted expression pattern found using the Smart-seq dataset demonstrated sufficient sampling of this region (Supplementary Figure 3C).

We further compared the zonation profiles between datasets and found a high correlation of gene expression and spatial location of transcripts across the periportal to pericentral axis. For genes significantly zonated in both datasets (having adjusted p-value < .1) the median Spearman correlation is 0.73. In Figure 3A we looked at zonated genes within the metabolic pathways in KEGG, and found the median correlation between datasets (highlighted in dark pink) is 0.82. Among all genes in that pathway (light pink) the correlation is moderate with a median of 0.18, and no correlation is found when all genes are considered (grey).

Traditionally the liver lobule is divided into three zones, a periportal zone 1, a pericentral zone 3 and transitioning zone 2[26,27]. The transitional nature of the liver axis is reflected in the heatmap of metabolic genes that were significantly zonated in both datasets (Figure 3B). Using k-means clustering, we found the Smart-seq data tended to cluster into two distinct gene groups representing either the periportal or pericentral zone. Examination of the two clusters by enrichment analysis of KEGG metabolic pathways (Figure 3C) revealed that the predicted location along our reconstructed axis of metabolic processes with known periportal or pericentral bias such as amino acid metabolism (periportal), lipogenesis (pericentral) and CYP450 metabolism (pericentral) corresponds to their known *in vivo* locations[27]. Despite using different reordering algorithms and protocols, the two datasets show high agreement of expression along the recovered pericentral to periportal axis among genes that are detectable in both datasets, and both reliably mirror the *in vivo* patterning of the liver lobule (additional KEGG categories are shown in Supplementary Figure 2).

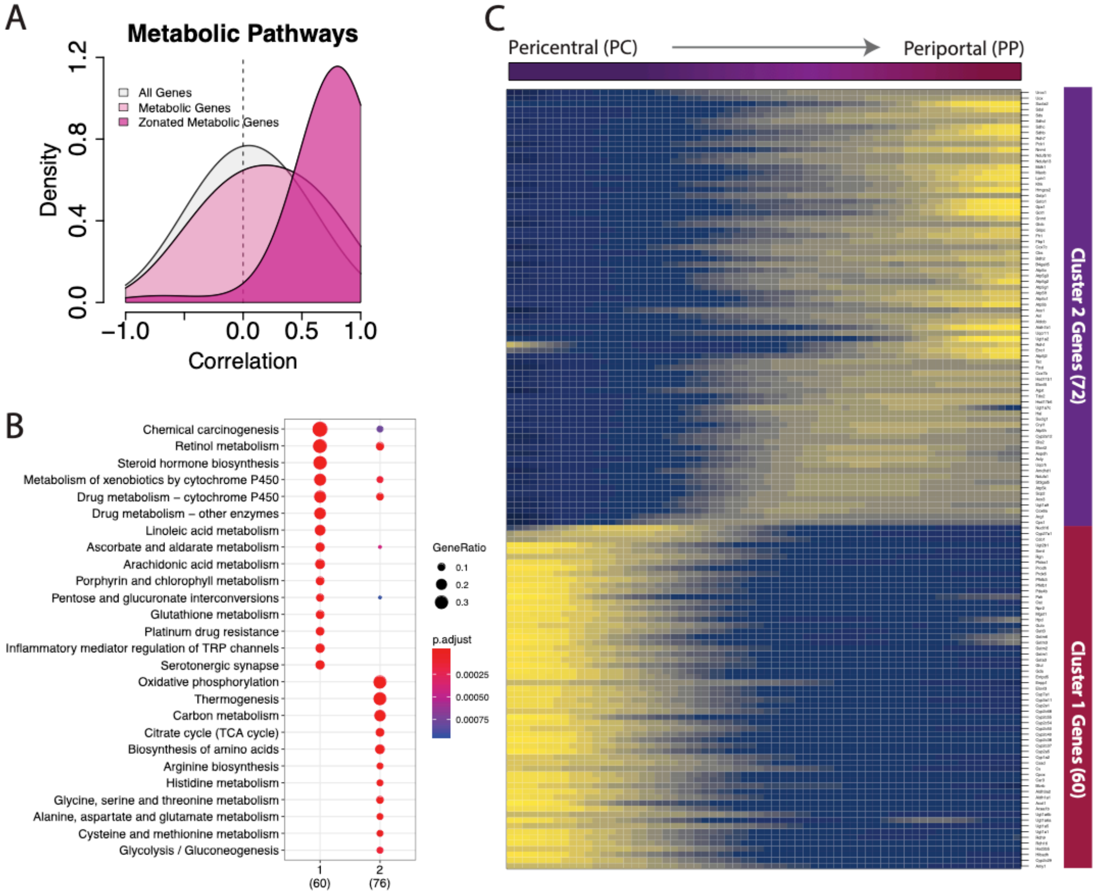

Figure 3. Correlation and Gene Ontology analysis of genes between datasets. A) Correlation analysis of genes annotated to the metabolic pathways in KEGG between the datasets. The dark pink density is the correlation of genes from the

metabolic pathways with significant zonation profiles in both datasets. The light pink density displays the correlation of all genes in the metabolic pathway and the grey density displays the correlation of all genes. B) Heatmap of the expression level of genes that are significantly differentially zonated in both datasets and enriched in the metabolic KEGG pathway. C) Breakdown of KEGG enrichment analysis of the two k-mean clusters based on the genes shown in B. Dot size represents the fraction of enriched genes in each ontology, and the color represents the adjusted p-value for the enrichment.

When we look at genes with moderate and low expression levels, we find that the two datasets differ to a greater degree. We identified twenty genes that were classified as significantly zonated along the periportal to pericentral axis in the Smart-seq dataset that were not detected at all in the MARS-seq dataset, whereas only three such genes were exclusive to the MARS-seq dataset. Figure 4A shows six most highly expressed genes that we were able to exclusively identify in the Smart-seq dataset having significant zonation (adjusted p-value < 0.10). This is not a surprising result due to the well-known sensitivity advantage the C1/Smart-seq technique holds over the MARS-seq technique.

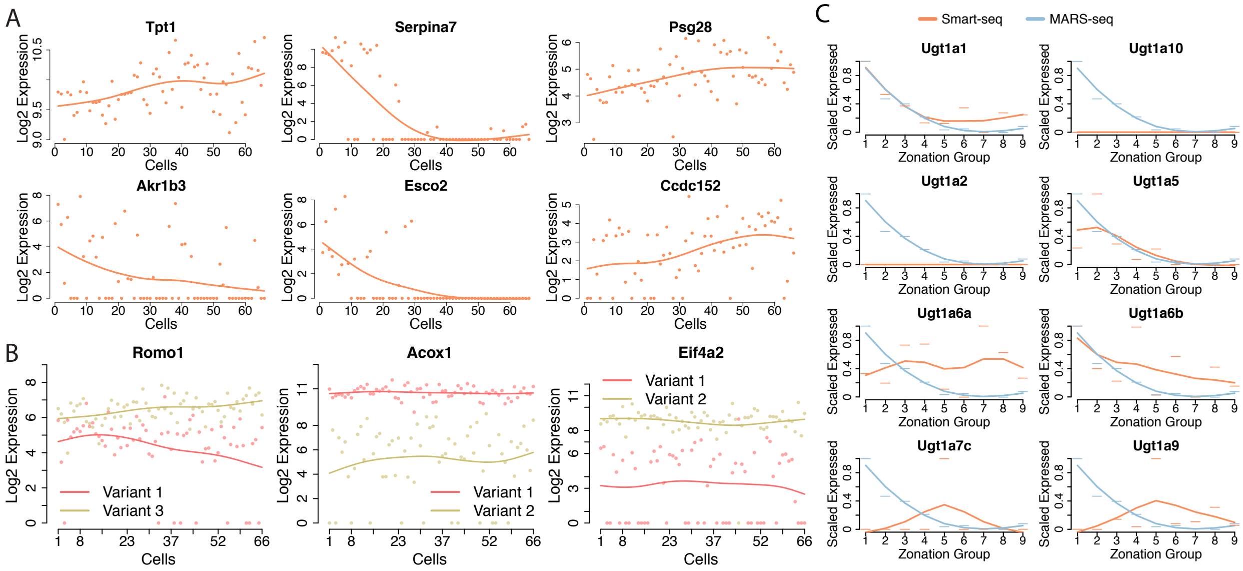

Figure 4. Genes and isoforms found in the full-length dataset and not in the UMI dataset. A) Six genes found to be zonally expressed in the Smart-seq dataset that were not detected in the MARS-seq dataset. The log2 of expression values are represented on the y-axis and the pseudo-space ordered cells are found on the x-axis. B) Examples of genes with two transcript variants expressed differently across reordered cells from the Smart-seq dataset. C) Eight Ugt1a genes that were concatenated in the MARS-seq dataset (blue on all graphs), but can be resolved in the Smart-seq dataset (orange line).

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 the C1/Smart-seq protocol compared to MARS-seq 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 4B 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 Supplementary Table 1).

We also note that due to the nature of the MARS-seq protocol there is also an inability to resolve not just isoforms but many genes that are closely related. There were 242 concatenated genes in the MARS-seq set corresponding to 539 unique genes. An example of this is seen in Figure 4C where we highlight a concatenate of Ugt1a enzymes as another example of this. Eight genes are concatenated 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 and Ugt1a6a can be seen to have consistent expression levels across the pericentral to periportal axis.

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 cells or the sequencing depth constant while varying the other. For the Smart-seq dataset, we evaluated the effect on the cell ordering as well as the gene-specific zonation profiles. For the MARS-seq dataset, the assignment of each cell to a zonation group depended on external data and was independent of the other cells profiled, thus we only evaluated the effect on zonation profiles. In Supplementary Figure 4A&B, the MARS-seq dataset displayed an approximately linear tradeoff in zonation profile error for fewer cells at the original read depth. While, at reduced read depth using the original 1,415 cells, a linear increase in error only existed up to 70% of the total depth, and at lower levels the error increased exponentially. The average mean squared error we observed in zonation profiles through subsampling in the MARS-seq dataset indicates that resequencing at the same depth results in error that is equivalent to reducing the total cells by about 400. Thus, in scenarios with such low sequencing depth (average of 11.7k total UMIs per cell), sequencing deeper would be more beneficial than adding more cells. For the Smart-seq dataset, we found the spatial ordering to be quite robust to reduced sequencing depth, even as low as 50% fewer reads and only marginal increases in gene-specific zonation error as shown in Supplementary Figure 4C&D. The average sequencing depth for the Smart-seq cells was 3.5 million counts per cell, well beyond the commonly suggested sequencing saturation for single-cell data that occurs close to one million total reads[34]. We do see more significant increases in error related to zonation profiles when profiling fewer cells in Supplementary Figure 4E. Here the tradeoff of sequencing to even half of the current depth and increasing the number of cells would be beneficial.

Discussion

In summary, we compared two scRNA-seq datasets of mouse hepatocytes where one, MARS-seq, is wide but shallow (1500 cells and about 3000 genes per cell) and the other, C1/Smart-seq is narrow but deep (66 cells and 8000 genes per cell). We find that the two different protocols present highly reproducible liver zonation profiles in single cells, and for the vast majority of genes that are highly expressed we observe highly comparable results. We do however find that when we look at medium to low expressed genes the increased sensitivity of the C1/Smart-seq protocol is able to identify several genes exclusive to this dataset. This increased sensitivity also allowed us to identify several genes with isoforms that behaved differently across the periportal to pericentral axis. We are aware of the limitation of short reads in regard to isoform analysis and if more accuracy is needed, the newly developed technique ScISOr-seq[35] might be better suited. We do however believe that this data allows for preliminary isoform analysis. We were able to resolve and identify individual genes with differing spatial patterns that lower sensitivity techniques are unable to distinguish. The main weakness of using fewer cells is that it is less likely that rare cell types will be sampled. In cases where such rare cells are of high interest, protocols that produce a large number of cells are preferable. In an ideal case, one would sample many cells and sequence all of them deeply, unfortunately, this is not always possible in practice and the decision of whether to sample many cells shallowly or fewer cells deeply comes down to whether rare cell types are of interest or if higher resolution of the individual cells is preferred. Given the distinct advantages, we emphasize that the biological question should be the driving factor when deciding on protocol. Within a chosen protocol, achieving balance between the sequencing depth and the number of cells is still an important consideration for optimal use of resources. Based on our simulations of two datasets at opposite ends of the sequencing depth versus number of cells trade-off, there is eventually a detriment to sacrificing reads for additional cells or sequencing beyond the attainable sensitivity level on too few cells. We expect that the extent of the cells versus depth trade-off will vary for other cell types or tissues and it will largely depend on the heterogeneity of the biological system under study.

Author contributions.

Morten Seirup and James A. Thomson designed the experiments. Morten Seirup, Li-Fang Chu and Srikumar Sengupta performed the experiments. Angela L. Elwell, and Jennifer M. Bolin prepared sequencing libraries. Bret Duffin provided animal husbandry. Christina M. Shafer and Scott Swanson developed the sequencing and alignments pipeline. Rhonda Bacher performed statistical analyses (with input from Ron Stewart and Christina Kendziorski and Ning Leng). Morten Seirup, James A. Thomson and Rhonda Bacher supervised the project. Morten Seirup, James A. Thomson, and Rhonda Bacher wrote the paper. All authors read and approved the final manuscript.

Competing interests.

The authors declare that they have no competing interests

References.

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

4. Leng, N. *et al.* Oscope identifies oscillatory genes in unsynchronized single-cell RNA-seq experiments. *Nat Methods* **12**, 947–950 (2015).
5. Chu, L.-F. *et al.* Single-cell RNA-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm. *Genome Biol* **17**, 173 (2016).
6. Klein, A. M. *et al.* Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells. *Cell* **161**, 1187–1201 (2015).
7. Macosko, E. Z. *et al.* Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161**, 1202–1214 (2015).
8. Islam, S. *et al.* Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat Methods* **11**, 163–166 (2014).
9. Ziegenhain, C. *et al.* Comparative Analysis of Single-Cell RNA Sequencing Methods. *Molecular Cell* **65**, 631-643.e4 (2017).
10. Svensson, V. *et al.* Power analysis of single-cell RNA-sequencing experiments. *Nat Methods* **14**, 381–387 (2017).
11. Jaitin, D. A. *et al.* Massively Parallel Single-Cell RNA-Seq for Marker-Free Decomposition of Tissues into Cell Types. *Science* **343**, 776–779 (2014).
12. Guo, M. *et al.* Single cell RNA analysis identifies cellular heterogeneity and adaptive responses of the lung at birth. *Nat Commun* **10**, 37 (2019).
13. Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol* **32**, 381–386 (2014).
14. Ji, Z. & Ji, H. TSCAN: Pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis. *Nucleic Acids Res* **44**, e117–e117 (2016).

15. Saelens, W., Cannoodt, R., Todorov, H. & Saeys, Y. A comparison of single-cell trajectory inference methods. *Nat Biotechnol* **37**, 547–554 (2019).

16. Burger, H.-J., Gebhardt, R., Mayer, C. & Mecke, D. Different capacities for amino acid transport in periportal and perivenous hepatocytes isolated by digitonin/collagenase perfusion. *Hepatology* **9**, 22–28 (1989).

17. Pösö, A. R., Penttilä, K. E., Suolinna, E. M. & Lindros, K. O. Urea synthesis in freshly isolated and in cultured periportal and perivenous hepatocytes. *Biochem. J.* **239**, 263–267 (1986).

18. Tosh, D., Alberti, G. M. M. & Agius, L. Glucagon regulation of gluconeogenesis and ketogenesis in periportal and perivenous rat hepatocytes. Heterogeneity of hormone action and of the mitochondrial redox state. *Biochem. J.* **256**, 197–204 (1988).

19. Guzmán, M. & Castro, J. Zonation of fatty acid metabolism in rat liver. *Biochem. J.* **264**, 107–113 (1989).

20. Anundi, I., Lähteenmäki, T., Rundgren, M., Moldeus, P. & Lindros, K. O. Zonation of acetaminophen metabolism and cytochrome P450 2E1-mediated toxicity studied in isolated periportal and perivenous hepatocytes. *Biochemical Pharmacology* **45**, 1251–1259 (1993).

21. Braeuning, A. *et al.* Differential gene expression in periportal and perivenous mouse hepatocytes. *FEBS Journal* **273**, 5051–5061 (2006).

22. Halpern, K. B. *et al.* Single-cell spatial reconstruction reveals global division of labour in the mammalian liver. *Nature* **542**, 352–356 (2017).

23. Parekh, S., Ziegenhain, C., Vieth, B., Enard, W. & Hellmann, I. The impact of amplification on differential expression analyses by RNA-seq. *Sci Rep* **6**, 25533 (2016).

24. Phipson, B., Zappia, L. & Oshlack, A. Gene length and detection bias in single cell RNA sequencing protocols. *F1000Res* **6**, 595 (2017).

25. Gebhardt, R. & Mecke, D. Heterogeneous distribution of glutamine synthetase among rat liver parenchymal cells in situ and in primary culture. *The EMBO Journal* **2**, 567–570 (1983).

26. Bhatia, S. N. *et al.* Zonal liver cell heterogeneity: effects of oxygen on metabolic functions of hepatocytes. *Cell Eng* **1**, 125–135 (1996).

27. Kietzmann, T. Metabolic zonation of the liver: The oxygen gradient revisited. *Redox Biology* **11**, 622–630 (2017).

28. Song, Y. *et al.* Single-Cell Alternative Splicing Analysis with Expedition Reveals Splicing Dynamics during Neuron Differentiation. *Molecular Cell* **67**, 148-161.e5 (2017).

29. Karlsson, K., Lönnerberg, P. & Linnarsson, S. Alternative TSSs are co-regulated in single cells in the mouse brain. *Mol Syst Biol* **13**, 930 (2017).

30. Arzalluz-Luque, Á. & Conesa, A. Single-cell RNAseq for the study of isoforms—how is that possible? *Genome Biol* **19**, 110 (2018).

31. Wang, E. T. *et al.* Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**, 470–476 (2008).

32. Gruppuso, P. A., Tsai, S.-W., Boylan, J. M. & Sanders, J. A. Hepatic translation control in the late-gestation fetal rat. *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology* **295**, R558–R567 (2008).

33. Oaxaca-Castillo, D. *et al.* Biochemical characterization of two functional human liver acyl-CoA oxidase isoforms 1a and 1b encoded by a single gene. *Biochemical and Biophysical Research Communications* **360**, 314–319 (2007).

34. Haque, A., Engel, J., Teichmann, S. A. & Lönnberg, T. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Med* **9**, 75 (2017).

35. Gupta, I. *et al.* Single-cell isoform RNA sequencing characterizes isoforms in thousands of cerebellar cells. *Nat Biotechnol* **36**, 1197–1202 (2018).

36. Brennecke, P. *et al.* Accounting for technical noise in single-cell RNA-seq experiments. *Nat Methods* **10**, 1093–1095 (2013).

37. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**, R25 (2009).

38. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).

39. Bacher, R. *et al.* SCnorm: robust normalization of single-cell RNA-seq data. *Nat Methods* **14**, 584–586 (2017).

40. Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. *OMICS: A Journal of Integrative Biology* **16**, 284–287 (2012).

Materials and Methods

Animals and handling.

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

Care and Use Committee and conducted in accordance with the Animal Welfare Act and Health Research Extension Act.

Cell isolation.

The euthanized mouse was pinned to a Styrofoam plate using 20 ga needles to aid in dissection. The abdominal cavity was opened, and the portal vein exposed. A piece of 4-0 suture thread (Ethicon vicryl coated) was threaded under the portal vein and used to secure a 26 ga catheter inserted into the portal vein (Butler Schein animal health 26 G IV Catheter, Fisher Scientific). Hepatocytes were isolated using a 2-step perfusion protocol. First, Liver Perfusion Medium (Gibco) warmed to 37°C was pumped through the catheter for 10 minutes using a peristaltic pump at 7 ml/min flowrate. Then, Liver Digest Medium (Gibco) warmed to 37°C was pumped through the liver at the same settings for 10 minutes. After perfusion, the liver was excised and transferred to a 10 cm dish containing 20 ml liver digest medium. The liver was dissected allowing the cells to spill into the media. The cells were then filtered through a 40 μm cell strainer into a 50 ml tube and 30 ml media (Williams E media + 2 μg/ml human insulin + 1x glutamax + 10% FBS) were added and placed on ice. The hepatocytes were purified by centrifugation at 50 x G, 4 times for 3 minutes each, each time discarding the supernatant and adding media.

Single cell RNA sequencing.

Single-cell RNA sequencing was performed as previously described[4,5] with the following modifications. In this study, we used small (5-10 μm), medium (10-17 μm), and large

(17-25 μm) plate sizes. ERCC RNA Spike-In (ThermoFisher Cat. No. 4456740) was diluted in the lysis mix following the manufacturer's user guide and previous studies[36]. Single end reads of 51 bp were sequenced on an Illumina HiSeq 2500 system. Sequencer outputs were processed using Illumina's CASAVA-1.8.2. The demultiplexed reads were trimmed and filtered to eliminate adapter sequence and low-quality basecalls. The reads were mapped to an mm10 mRNA transcript reference (extended with ERCC transcripts) using bowtie-0.12.9[37]; expression estimates were generated using RSEM v.1.2.3[38].

Using the Fluidigm C1 system to capture and synthesize cDNA from single cells in the liver, we generated transcriptomes for 149 cells. To exclude low quality transcriptomes, we removed cells in which the fraction of ERCC spike-in made op 20% or more of the total assigned reads. This left 66 high quality cells, that were used in the downstream analysis. Finally, the data was normalized using SCnorm (R package v 1.5.7)[39].

Data availability.

scRNA-sequencing data that support the findings of this study have been deposited in NCBI's Gene Expression Omnibus with the GEO Series accession code "GSE116140" https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE116140. The normalized and ordered expression data is provided as Additional File 4.

All code used in the analysis and figures is available on Github at https://github.com/rhondabacher/LiverSpatialCompare.

Pseudo-spatial reordering.

For the full-length data, the cells were computationally ordered using the Wave-Crest method as described by Chu et al. 2016[5]. Prior to reordering, gene expression values were rescaled to mean 0 and variance 1 to ensure the values across different genes are comparable. The Wave-Crest algorithm implements an extended nearest insertion algorithm that iteratively adds cells to the order and selects the insertion location as the location producing the smallest mean squared error in a linear regression of the proposed order versus gene expression. A 2-opt algorithm is then used to find an optimal cell order by considering adjacent cell exchanges. The cell ordering step uses the expression profiles of pre-selected known marker genes of liver zonation. Thus, the resulting linear profile of ordered cells represents the periportal to pericentral axis. The known marker genes used to construct the periportal to pericentral axis in Wave-Crest include the following pericentral markers: cytochrome P450 7a1 (Cyp7a1), cytochrome P450 2e1 (Cyp2e1), ornithine aminotransferase (Oat), cytochrome P450 1a2 (Cyp1a2), rh family, B glycoprotein (Rhbg), leucine-rich repeat-containing G-protein coupled receptor 5 (Lgr5), glutamate-ammonia ligase (Glul); and the following periportal markers: phosphoenolpyruvate carboxykinase 1 (Pck1), catenin beta interacting protein 1 (Ctnnbip1), aldehyde dehydrogenase 1 family member B1 (Aldh1b1), sulfotransferase family 5A, member 1 (Sult5a1), cytochrome P450 2f2 (Cyp2f2), cathepsin C (Ctsc), serine dehydratase (Sds), and E-cadherin (Cdh1). All markers were selected based on their expression ratio as reported by Braeuning et al. 2006[21].

A detection step was done to identify additional genes that follow the one-dimensional periportal to pericentral axis by fitting a linear regression to the relationship between each gene's expression and the Wave-Crest cell order. To determine if a gene is significantly dynamic (zonated) along the recovered axis, we tested whether the regression slope is different from zero. We reported the Benjamini-Hochberg adjusted p-values to control the false discovery rate. For genes having an adjusted p-value < .01, the direction of the expression profile was assigned based on the sign of the regression slope (periportal: positive slope, pericentral: negative slope). We also calculated the linear fitting mean squared error (MSE) for each significant gene. Genes with a smoother trend over the recovered cell order are expected to have a smaller MSE. We reported the full list of significant genes, sorted by their MSE, in Additional File 2; scatter plots are shown in Additional File 3.

Comparative Analysis

Smoothed densities (bean plots) with overlaid raw data, the mean, and a box representing the interquartile range of the cellular detection fractions were created using the pirateplot function in the yarrr R package (v0.1.5). The cellular detection fraction was calculated per cell as the proportion of genes having expression greater than zero. The fold-change for each gene between the two datasets was calculated as the log2 fold-change of the full-length gene mean over the UMI gene mean, where each gene mean was calculated as the average expression among non-zero counts across all cells in the datasets. The heatmap in Figure 2 of marker gene expression on the normalized Smart-seq data was generated by setting values above the 95th percentile or below the 5nd percentile to the 95th percentile or 5nd percentile value, respectively

When comparing the two datasets having different dynamic ranges, we used scaled expression plots, where the ordered cells in the full-length dataset were divided into nine equally sized groups to correspond to the nine layers in the UMI dataset. For the full-length dataset, for a given gene, the median expression in each group was calculated, then the nine means were scaled between zero and one. Smoothed fits were overlaid using the smooth.spline function in R with the degrees of freedom parameter df=4. Expression correlations along the zonation axis between datasets were calculated using Spearman correlation. Enrichment of genes in KEGG pathways or GO was done using the R package clusterProfiler (v. 3.10.1)[40]. For the enrichment analysis, since different statistical methods were used to assess zonation profiles, genes were considered significantly zonated if they had an adjusted p-value < .1 in both datasets and more than 10 non-zero expression values. The heatmap in Figure 3 is a smoothed heatmap, where a smoothing spline was first fit to the log expression (pseudo-count of one added) of each gene using the smooth.spline function in R with the smoothing parameter df=4 which provided profiles that were not over- or underfit in either dataset. Then the smoothed expression was scaled and outliers above the 98th percentile or below the 2nd percentile were set to the 98th percentile or 2nd percentile value, respectively. Additional KEGG categories from this analysis can be interactively viewed on Github https://github.com/rhondabacher/LiverSpatialCompare.

Subsampling Analysis

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

For the MARS-seq dataset, zonation group means were recalculated on a subsampled set of cells using the posterior probability matrix and original UMI counts from Halpern et al. 2017. In each sampling, the mean squared error (MSE) was calculated based on a 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 mean expression of gene i in zonation group j in the original dataset and $\hat{Z}_{i,j}$ is the corresponding value for the subsampled dataset. For subsampling at lower read depths, we fixed the number of cells at the original total of 1415 cells and simulated each cell's gene counts individually using a multinomial distribution. For each cell, the subsampled total counts were set to X% of the original total read counts for that cell (for X = (10,20,30,40,50,60,70,80,90,100)) and each gene's cell-specific probability was calculated as its original count divided by the original total counts for that cell. The MSE was calculated for each subsampled set as described above.

For the Smart-seq dataset, we reran Wave-Crest when subsampling the total number of cells using the original parameter settings and marker genes. Then, as before, the ordered cells were assigned zonation groups by dividing cells into nine equally sized groups. The zonation profile error was estimated using MSE and calculated as described above with the exception that since Wave-Crest orders can be flipped, we calculated the MSE on the returned order and its reverse, and kept the minimum MSE of the two. We also computed the MSE similarly on random permuted orders of the full 66 cells to assess the maximal MSE distribution. For evaluating lower read depths, we first determined the effect of lower read depth on the ordering accuracy by re-running

Wave-Crest on lower read-depth subsampled datasets and calculating the correlation of the original order to the cell order obtained on the subsampled data. To evaluate the zonation profile error with lower read depths, we used a similar approach as described above for the MARS-seq dataset, fixing the number of cells to be the same as the original total of 66 and, since the order correlation was shown to be consistently high, we used the original Wave-Crest order for every scenario when evaluating zonation profile error.

Immunohistochemistry.

An 8-week-old male C57BL/6 mouse was anesthetized using isoflurane before euthanizing by cervical dislocation. The liver was excised, sliced as thinly as possible with a razor blade, and fixed in formaldehyde overnight. The liver slices were paraffin embedded and sectioned. Sections were stained following the protocol published by Abcam (http://www.abcam.com/ps/pdf/protocols/ihc_p.pdf). In short, the slices are deparaffinized by dipping into sequential solutions of 100% xylene, 50-50% xylene-ethanol, 100% ethanol, 95% ethanol, 70% ethanol, 50% ethanol, and tap water. The antigens were then retrieved by placing the slides in Tris-EDTA buffer (10 mM Tris Base, 1 mM EDTA Solution, 0.05% Tween 20, pH 9.0) and incubating them in a decloaking chamber (Biocare Medical Decloaking Chamber #DC2008US) with the following settings: delayed start 30 sec.; preheat 80°C, 2 min.; heat 101°C, 3 min. 30 sec.; and fan on. The slides were washed 2 x 5 min in TBS + 0.025% Triton X-100 before they were blocked for two hours at room temperature in 10% normal serum in 1% BSA. The appropriate primary antibody was then diluted in the same 10% normal serum in 1% BSA, added to the slides, and incubated at 4ºC overnight in an incubation chamber. The next day the slides were washed 2 x 5 min in TBS + 0.025% Triton X-100 followed by 15 min incubation in 0.3% H_2O_2 at room temperature. Next, the appropriate secondary antibody was diluted into 10% normal serum in 1% BSA before it was added to the slides and incubated for 1 hour at room temperature. The slides were then washed 3 x 5 min in TBS before DAB (#ab103723) staining mixed according to manufacturer instruction was applied and incubated under a microscope to stop the reaction after sufficient staining. The slides were rinsed in tap water for 5 min before being counterstained with Mayer's hematoxylin (#MHS1-100ML) for 30 sec. The stain was developed in running tap water for 5 min. The slides were then dehydrated by sequentially dipping in 50% ethanol, 70% ethanol, 95% ethanol, 100% ethanol, 50-50% xylene-ethanol, and 100% xylene before Poly-Mount (#08381-120) was added and a coverslip placed on top. The following primary antibodies were added: Aldh3a4 1:250 (AB184171), Cyp2e1 1:50 (AB28146), Cyp1a2 1:50 (R31007), Rgn 1:100 (NBP1-80849), Oat 1:50 (AB137679), Cyp2f2 1:100 (SC-67283), Hal 1:50 (AV45694), and Tbx3 1:50 (SC-31657). The following secondary antibodies were used: goat-anti-rabbit HRP conjugated (ab97051) and donkey-anti-goat HRP conjugated (ab97110) at a concentration of 1:500.

Additional Files

Additional File 1 – Supplementary Figures and Tables.

Additional File 2 – Summary of genes with dynamic expression across the zonation axis identified using Wave-Crest.

Additional File 3 – Scatter plots of dynamic genes listed in Additional File 2.

Additional File 4 – Normalized Smart-Seq single-cell data with cells in the Wave-Crest order.