Investigating higher order interactions in single cell data with scHOT

Shila Ghazanfar[1], Yingxin Lin[2], Xianbin Su[3], David M. Lin[4], Ellis Patrick[2,5], Ze Guang Han[3], John C. Marioni[1,6*+], Jean Yee Hwa Yang[2,7*+]

[1] Cancer Research UK Cambridge Institute, University of Cambridge, Cambridge, United Kingdom

[2] School of Mathematics and Statistics, The University of Sydney, Sydney, New South Wales, 2006, Australia

[3] Key Laboratory of Systems Biomedicine (Ministry of Education) and Collaborative Innovation Center of Systems Biomedicine, Shanghai Center for Systems Biomedicine, Shanghai Jiao Tong University, Shanghai 200240, China

[4] Department of Biomedical Sciences, Cornell University, Ithaca, United States

[5] Westmead Institute for Medical Research, Sydney, New South Wales, 2145, Australia

[6] EMBL European Bioinformatics Institute, Wellcome Trust Genome Campus, Cambridge, United Kingdom

[7] Charles Perkins Centre, The University of Sydney, Sydney, New South Wales, 2006, Australia

* To whom correspondence should be addressed.; Email: marioni@ebi.ac.uk; jean.yang@sydney.edu.au

+ Equal contribution

ABSTRACT

Single-cell RNA-sequencing has transformed our ability to examine cell fate choice. For example, in the context of development and differentiation, computational ordering of cells along ‘pseudotime’ enables the expression profiles of individual genes, including key transcription factors, to be examined at fine scale temporal resolution. However, while cell fate decisions are typically marked by profound changes in expression, many such changes are observed in genes downstream of the initial cell fate decision. By contrast, the genes directly involved in the cell fate decision process are likely to interact in subtle ways, potentially resulting in observed changes in patterns of correlation and variation rather than mean expression prior to cell fate commitment. Herein, we describe a novel approach, scHOT – single cell Higher Order Testing - which provides a flexible and statistically robust framework for identifying changes in higher order interactions among genes. scHOT is general and modular in nature, can be run in multiple data contexts such as along a continuous trajectory, between discrete groups, and over spatial orientations; as well as accommodate any higher order measurement such as variability or correlation. We demonstrate the utility of scHOT by studying embryonic development of the liver, where we find coordinated changes in higher order interactions of programs related to differentiation and liver function. We also demonstrate its ability to find subtle changes in gene-gene correlation patterns across space using spatially-resolved expression data from the mouse olfactory bulb. scHOT meaningfully adds to first order effect testing, such as differential expression, and provides a framework for interrogating higher order interactions from single cell data.

INTRODUCTION

Understanding the mechanisms that underpin cell fate choices is a key challenge in developmental biology. It requires disentangling the complex interplay between cell autonomous factors, such as gene expression, and non-autonomous factors such as the signaling environment. In the former context, recent technological advances have enabled the rapid and high-throughput measurement of mRNA expression levels in individual cells. Such single-cell RNA-sequencing (scRNA-seq) datasets have facilitated the generation of atlases of cell types during development in human, mouse, zebrafish and the frog[1–5]. Using such data, cells can be computationally ordered along 'pseudotime' and changes in the expression profiles of individual genes can be subsequently determined. However, while cell fate decisions are typically associated with profound changes in expression, many such changes are downstream of the initial cell fate decision. Instead, subtle changes in patterns of variation and coexpression of genes across developmental time, sometimes not associated with substantial changes in mean expression, have been argued to play a more critical role in symmetry breaking[6,7]. Consistent with this, higher order interactions (i.e., looking beyond changes in mean expression) have proved highly informative for understanding genomics data, for example in supervised machine learning settings[8] and for estimation of unknown spatial patterning[9]. Additionally, with recent developments in high-throughput and high-resolution spatially resolved gene expression mapping (e.g., Spatial Transcriptomics[10]; seqFISH[11]; MERFISH[12]) it is now possible to explore the relationship between higher-order interactions and spatial location. For example, in the context of embryogenesis, do small numbers of spatially-localized cells display aberrantly higher variability in expression profiles prior to committing to a downstream fate?

From a computational perspective, methods for studying higher-order interactions are currently lacking. Although numerous methods have been developed for ordering cells along pseudotime, a computationally derived prediction of cell-type differentiation trajectories[13–17], methods for identifying individual genes that significantly change their expression levels across the pseudotemporal trajectory[18–20] typically focus on changes in mean expression of single genes and do not characterize subtle changes in patterns of covariation between subsets of genes across this trajectory. In those cases where higher-order interactions have been studied, a typical analysis aiming to compare correlation patterns along pseudotime first defines strict nonoverlapping sets of cells before estimation of a covariance network, either through direct thresholding on the correlation matrix or using other methods[21,22]. However, estimation of such networks is noisy[23], and ignores potentially subtle but consistent changes across a continuum, as well as requiring an often ad hoc dichotomization or classification of cells into discrete groups. As we have previously discussed[24], treating the sample ranking as a covariate and testing for an interaction effect in a linear model is restricted to identifying linear and thus monotonic interactions, which may not be present, especially in highly dynamic or complex trajectories going through multiple changes in the differentiation process. In the context of spatially resolved gene expression data, fewer methods exist, with the focus being on testing the existence of pre-defined patterns[25] (e.g., a signaling gradient); however, these require *a priori* knowledge about the spatial structures of interest.

Here we introduce single cell Higher Order Testing (scHOT), a framework for examining changes in higher order structure, such as correlation among genes across differentiation pseudotime, among discrete groups, and across spatial landscapes. scHOT builds on our previous work, DCARS (Differential Correlation Across Ranked Samples), which used bulk RNA-sequencing data to test for changes in gene-gene correlation across ranked individual samples[24]. Our approach requires one of the following types of cell-specific information (Figure 1): A) a ranking of cells, which will typically be across pseudotime, or B) spatial coordinates in either two or three dimensions. In the case where spatial coordinates are inferred[26], scHOT is also applicable using either the cell ranking along a gradient or in the inferred 2/3D space. Given this cell-specific information, as well as a scheme for determining local sample-specific weights, we calculate local higher order interaction vectors among single genes or pairs of genes, uncovering local changes in variability or covariation respectively (Figure 1). Sample-wise permutation testing is then used to assess statistical significance, while retaining the global variability or correlation structure of the original data. This framing of the significant genes and gene-pairs in terms of the set of local higher order interaction estimates allows patterns of changes across the trajectory, groups, or space to be characterized in terms of the higher order interaction, rather than simply by changes in the mean. Moreover, scHOT identifies groups of genes for which similar higher order patterns arise. For a more detailed discussion of how scHOT can be applied in practice, see the Supplementary Note. We demonstrate the utility of scHOT by studying the inferred trajectories of mouse liver hepatoblasts into structural cholangiocytes or functional liver hepatocytes, and illustrate its generalizability in a spatial setting by performing higher order testing on a Spatial Transcriptomics dataset generated to study the mouse olfactory bulb.

RESULTS

scHOT identifies multiple higher order associations during liver development

We first analyzed four single-cell RNA-sequencing datasets designed to study the early development of the mouse liver[27–30] (Methods, of which three contained hepatic cells). The integrated data encompassed 7 days of development (from embryonic (E) day 10.5 to E17.5), which covers the period where progenitor hepatoblasts transition towards more mature hepatocytes and cholangiocytes (Figure 2A). As expected, when using Monocle[17] to order the cells in pseudotime, we observed a clear bifurcation where hepatoblasts differentiated into either cholangiocytes or hepatocytes (Figure 2B). This was supported by a higher proportion of differentiated cells in the later embryonic time points as well as cell-type specific expression of known marker genes (Figure 2C).

Building on this, we first examined higher order patterns as cells transitioned from naïve hepatoblasts towards the bifurcation point where they commit to one of the two downstream lineages (Figure 2D). In total 68 genes showed a change in variability (false discovery rate (FDR) adjusted P-value < 0.1, all shown in Supplementary Figures S2C and S2D) along the trajectory independent of changes in gene expression. Of these, 58 (85%) displayed significantly increased variability along the branch (Figure 2D). These genes were enriched for involvement in processes associated with cell division, chromosomal organization and DNA replication (Figure 2E; Methods), consistent with the notion that increased plasticity can precede cell fate commitment[6,7].

We next focused on the full trajectory from naïve hepatoblast through to hepatocytes. Specifically, we investigated whether scHOT could identify changes in correlation, independent of differential expression, thus providing insight into the potentially complementary set of gene regulatory modules that are activated during the process of commitment from hepatoblast to the hepatocyte lineage. Correlation patterns identified as cells transition from naive hepatoblast to cholangiocytes can be found in Supplementary Figures S3B and S3C.

When focusing on the hepatoblast to hepatocyte lineage, we identified numerous changes in correlation between pairs of genes that did not change their individual mean expression (Methods). An example of such a gene-pair is *Cdt1* and *Top2a* (Figure 3A, FDR adjusted P-value < 0.03), which are protein-protein interacting partners[31] that have been implicated in regulation of the cell cycle in human and mouse stem cells[32]. This pair of genes changes from being strongly negatively correlated in the progenitor population to displaying no correlation in the more differentiated hepatocytes. Interestingly, when considering each gene separately, neither *Cdt1* nor *Top2a* are significantly differentially expressed along the trajectory, or significantly differentially variable (FDR adjusted P-value = 0.70 and 0.12 respectively), indicating that the association between these two genes would not be identified without using scHOT. *Top2a* encodes a DNA topoisomerase, which controls and alters the topologic states of DNA during transcription[33], while *Cdt1* is a chromatin licensing and DNA replication factor that is required for DNA replication and mitosis[34]. Our observation that these genes move from being negatively correlated to displaying no correlation suggests a trade-off between chromatin remodeling and transcription at the earlier stages of differentiation, potentially facilitating both proliferation and the global changes in gene regulatory architecture that arise when cells commit towards the hepatoblast lineage.

Across all gene-pairs tested, 224 displayed different patterns of correlation (FDR adjusted P-value < 0.2), encompassing 136 unique genes. Gene-pairs that were differentially correlated were not found to be associated with genes that were also differentially variable along the trajectory (Fisher's Exact Test P-value > 0.4) for either hepatocyte or cholangiocyte branch, suggesting an independent relationship between changes in correlation of gene-pairs and variability of the genes. The majority of local correlation patterns of these gene-pairs exhibited either a 'gain' or a 'loss' of correlation, across developmental time reflecting the prior understanding of a continuous differentiation towards the end fate (Figure 3B). Functionally annotating these clusters revealed that, in general, clusters associated with a loss of correlation (e.g., Cluster 9) contained genes linked with DNA replication and cell division. By contrast, clusters that 'gained' correlation (e.g., Cluster 7) along the trajectory were associated with hepatocyte-linked functions such as lipoprotein particle remodeling, lipoprotein metabolism, as well as mitotic cell cycle and cell division (Figure 3C, all clusters shown in Supplementary Figure S3A). A small number of gene-pairs displayed more unexpected correlation patterns, with a transient peak of coexpression at an intermediate point along the trajectory (e.g., Clusters 5 and 10) near the bifurcation point, suggesting a transient role in cell fate commitment.

Finally, we explored whether the differential patterns of higher order interactions were associated with specific differentiation into the hepatocyte or cholangiocyte lineages, or a common pattern associated with maturation of cells. To do this, we used the network strength metric as a statistic to assign genes as common to either branch, or branch specific (Methods). Permuting over the topology of the gene network allowed assessment of statistical significance, revealing five genes that were associated with both branches, and ten and four genes significantly specific to the hepatocyte and cholangiocyte branches respectively (Figure 3D). In particular, the gene *Cdt1* appears associated to differentiation in general, i.e. from the hepatoblast state to either differentiated state, rather than any of the two terminal states. By contrast, genes like *Apom* and *Apoa2*, encoding apolipoprotein, were more associated with hepatocyte function. More surprisingly, we identified histone gene *H2afz* as more specific to the hepatocyte lineage, indicating a potential association with changes in global chromatin organization as cell's commit towards a hepatocyte fate.

scHOT identifies local patterns of correlation in the mouse olfactory bulb

Finally, we considered whether scHOT could also be used to identify cryptic local correlation when gene expression information is available in a spatial context. Specifically, to date, most studies of spatially resolved gene expression data have focused on clustering cells into groups or testing known patterns of correlation – we reasoned that scHOT would provide an unbiased approach for identifying local patterns of correlation that might be missed by such approaches, which rely on changes in mean expression only. To this end we considered the mouse olfactory bulb (MOB), which displays a highly stereotypical structure, with clear patterns of concentric layers corresponding to granule, internal plexiform, mitral, external plexiform, glomerular and olfactory nerve layers moving from the inside out (Figure 4A), along with distinct patterns of gene expression along this space[35]. Recently, spatial transcriptomics[10] was used to measure gene expression levels in small spatially-distinct regions of the MOB, thereby facilitating the unbiased identification of patterns of gene expression in space.

Using scHOT, we identified a set of 167 gene-pairs as significantly differentially correlated across space (FDR adjusted P-value < 0.2), with 42 non-differentially expressed highly variable genes appearing at least once among this set (Methods). Interestingly, we found that numerous pairs of genes displayed diffuse patterns of local correlation that were not apparent when visually comparing their individual expression profiles. For example, *Arrb1* (beta-arrestin-1) is widely expressed in the brain[36], consistent with its diffuse expression across the MOB in the spatial transcriptomic data. Similarly, *mTOR* (mammalian target of rapamycin), *Uchl1* (Ubiquitin carboxyl-terminal esterase L1), and *Dnm3* (Dynamin-3) are all broadly expressed[37–39]. Nevertheless, we identified all three genes as positively spatially correlated with *Arrb1* (Figure 4B and Supplementary Figure S4C). Consistent with this correlation, *Arrb1* can regulate *mTor* activation[40] and interact with the ubiquitination pathway to down regulate receptor signaling[41]. Further, both *Arrb1* and *Dnm3* function in endocytosis[42].

To explore more general patterns, we clustered all significant gene-pairs using their local spatial correlation patterns into 8 distinct groups (Figure 4C). Despite the relatively low-resolution of the data (spatial transcriptomic data is limited to a resolution of ~10 cells (approximately 100 µm)[10]), a variety of local correlation patterns were observed, often associated with distinct biologically meaningful regions of the bulb. Giving confidence in our analysis approach, clustering the cells based on their local correlation pattern largely recapitulated the symmetry of the MOB, with multiple cell groups corresponding to symmetric sets of cells, e.g. Cell groups 3, 4, and 5 (Figure 4C). Additionally, Cluster 1 contained genes that were positively correlated within the Olfactory Nerve Layer and Clusters 3 and 6 were associated with the Mitral and External plexiform layers (Figure 4D). Functional annotation of the genes belonging to these clusters revealed associations with distinct neuronal terms including signaling events such as endocytosis, phosphorylation and ubiquitination (all clusters shown in Supplementary Figure S4D). Interestingly, "myelin sheath" was highly ranked in multiple clusters (Clusters 3-7). In these clusters, the strongest patterns of correlated spatial expression occurred within more internal layers of the bulb, overlapping with the mitral and granule layers. This is consistent with the myelination of the lateral olfactory tract as it exits the bulb[43]. Clusters 1, 2, and to some extent 8, in contrast, possess spatial correlation patterns that encompass more external layers such as the olfactory nerve layer, and genes within these clusters are not highly associated with the term 'myelin sheath.' This is consistent with the fact that olfactory sensory neurons entering the bulb in the more external olfactory nerve layer are not myelinated. In sum, we have shown here that exploiting higher-order structure can reveal unexpected and spatially-coherent regions of structured heterogeneity that are independent of mean expression changes, and that approaches that focus only on the latter will fail to fully exploit the wealth of information contained within such data.

DISCUSSION

In this paper we have demonstrated the utility of higher order testing for single cell data. We examined scHOT in the context of two biological systems with distinct data characteristics – liver development and the mouse olfactory bulb. scHOT is robust due to the choice of underlying higher order metrics such as rank-based Spearman correlation; powerful as it uses a permutation framework retaining the global variability and covariance structure for inference; and extremely flexible as it can be tuned by 1) Varying the local weighting scheme in terms of shape (triangular, block, any other user defined weight) and span, 2) Choice of underlying higher order effect function (weighted Pearson correlation, weighted Spearman correlation, weighted zero-inflated Spearman or Kendall correlation, or any choice of higher order estimate when using the block weighting scheme), and 3) Choice of summarization estimate for the local higher order vector (by default the standard deviation). In general, this contrasts with other methods that estimate changes in expression across either a pseudotime trajectory or across space, which require a set of candidate hypotheses to test explicitly. In the spatial context, scHOT differs substantially to other methods such as SpatialDE[25], in that we can test either a single gene (identifying spatially variable) or two genes (identify spatially differentially correlated), and no prior suite of potential spatial structures need be provided to identify genes that are of interest.

From a biological perspective, the concept of characterizing coordinated changes over time could enable better characterization of the molecular processes underpinning cell fate choice. In particular, it will help us to better understand whether increased plasticity, as manifested in increased cell-to-cell variability, is a general feature that precedes cell commitment or whether this is restricted to specific systems such as hepatoblast differentiation. Such heterogeneity could also be a driver of differential cell fate or cell function in a spatial context: specific patterns of local correlation could indicate that a specific region of a tissue or organ is primed towards a specific fate. Intriguingly, our reanalysis of data from the mouse olfactory bulb identified patterns that were independent of changes in mean expression but associated with known spatial structure of the bulb.

In summary, scHOT is a method for inference of changes in higher order interactions, not just changes associated with the mean, and as such offers a new lens to interrogate single cell data and describe patterns of variation and covariation, offering additional and complementary insight to that obtained by examining changes in mean expression. It is enabled by a statistical framework that captures nonlinear changes in correlation structure and that uses sample ranking approaches to avoid having to discretize responses. This is especially important for continuous single cell trajectories and for studying spatial structure within ostensibly homogeneous cell types. By facilitating such analysis, scHOT will enable investigations into how highly localized patterns of variation and co-variation influence cell fate and cell function.

MATERIAL AND METHODS

Datasets

The following datasets were used to examine scHOT, and demonstrate its utility in extracting insights from diverse sources of single cell and/or spatially resolved data.

Developmental Liver Data

The 'Developmental Liver Data' is a full-transcript scRNA-Seq dataset generated using plate-based protocols from four distinct sources[27–30], across multiple mouse embryonic time points from Embryonic Day (E)10.5 to E17.5. The data were integrated using scMerge[44], taking advantage of genes that are found to be stably expressed in single cells[45]. These data comprise several cell types including hepatic cells such as hepatoblasts, cholangiocytes, and hepatocytes, among other cell types such as immune cells (Figure 2A). Monocle was used to infer a differentiation trajectory exclusively for the hepatic cells. We applied scHOT to these data, considering the following testing scenarios: changes in variability across the first branch from hepatoblasts to the cell fate decision point; and changes in correlation between pairs of highly variable genes along the entire branch from hepatoblasts to hepatocytes and the full branch from hepatoblasts to cholangiocytes.

To select genes for downstream analysis, we considered genes that were highly variable (HVGs)[46] (Supplementary Figure S2A) but that had consistent mean expression along the trajectory. To do this, we performed liberal differential gene expression testing along pseudotime by fitting, for each gene, two linear models (slope and intercept, and polynomial of degree two) and identifying a gene as differentially expressed if it was significant (F-test; unadjusted P-value < 0.05) in at least one of the tests when compared to an intercept only model. This differential expression testing was performed for the hepatoblast to hepatocyte trajectory and for the hepatoblast to cholangiocyte trajectory. The resulting sets of genes (i.e., highly variable and non-differentially expressed) were combined to form all pairwise combinations as the scaffold for higher order gene-pair testing.

Spatial data from the mouse olfactory bulb (MOB)

The ‘Mouse Olfactory Bulb’ data is a Spatial Transcriptomics dataset, where an array spotted with probes that have barcodes corresponding to defined locations was used to measure spatially-resolved gene expression levels[10]. We consider data where this technology was used to measure expression levels across a section of the mouse olfactory bulb (MOB), where each spatially resolved region contains a measure of gene expression averages across approximately 10 cells[10]. This cross section of the MOB comprises concentric layers visible with H&E staining (Figure 4A), associated with the granule, internal plexiform, mitral, external plexiform, glomerular and olfactory nerve layers moving from the inside out. The resulting expression data is derived from high throughput RNA sequencing using barcodes corresponding to the spatial locations, as well as unique molecular identifier (UMI) barcodes. We identified genes as spatially differentially expressed by performing scHOT using a first-order metric of local weighted mean expression (2,542 differentially expressed genes; unadjusted P-value < 0.05). After identifying the intersection between genes that were highly variable[46] but not differentially expressed we used scHOT to test all pairwise combinations for this set of genes.

Choice of local weighting scheme

For the trajectory-based analyses we selected a triangular weight matrix with a span of 0.25. For the spatial-based analysis we selected a two-dimensional triangular weight matrix (i.e. a cone) also with a Euclidean distance span of 0.05 (here corresponding roughly to 9 surrounding sampled points).

scHOT test statistic and inference

For single gene testing we use a local weighted variance estimate

$$s_j^2 = \sum w_{ij}^* \sum w_{ij}^* x_i^2 - \left(\sum w_{ij}^* x_i\right)^2,$$

$$w_{ij}^* = \frac{w_{ij}}{\sum_i w_{ij}},$$

where w_{ij} is the cell-specific weight for cell i and position j, and x_i is the gene expression measure for gene x and cell i. For testing pairs of genes we use a weighted Spearman correlation

$$c_j = \frac{\sum w_{ij} \sum w_{ij} r_{x_{ij}} r_{y_{ij}} - \sum w_{ij} r_{x_{ij}} \sum w_{ij} r_{y_{ij}}}{\sqrt{\left(\sum w_{ij} \sum w_{ij} r_{x_{ij}}^2 - \left(\sum w_{ij} r_{x_{ij}}\right)^2\right)\left(\sum w_{ij} \sum w_{ij} r_{y_{ij}}^2 - \left(\sum w_{ij} r_{y_{ij}}\right)^2\right)}}$$

where

$$r_{x_{ij}} = Rank\left(x_i \times I\{w_{ij} > 0\}\right),$$

and

$$r_{y_{ij}} = Rank\left(y_i \times I\{w_{ij} > 0\}\right),$$

where additionally y_i is the gene expression measure for gene y and cell i.

The scHOT test statistic is a measure of variability associated with this vector of local variances or correlations. To compute this, we first calculate the sample standard deviation to estimate the variability associated with the set of local variance estimates $\{s_j^2, j = 1,2,3, \ldots\}$ or local correlation estimates $\{c_j, j = 1,2,3, \ldots\}$.

Statistical testing is then performed by randomly permuting cell labels, while keeping the overall gene expression structure constant. Thus, within each permutation round, the global correlation or global variance remains the same, while the vectors of local variability or local correlation vary. In all cases, we used sample or cell permutation and defined significance by controlling for a 0.2 Benjamini-Hochberg[47] FDR in all differential correlation tests, and at 0.1 FDR for variability tests. For correlation-based tests, we used the fact that the null distribution is based only on the two matched gene expression vectors to interpolate null distributions given the global correlation value and the number of samples (see Computational efficiency section) in order to speed up computation. In the case of discrete groupings, we use a normal approximation to the null distribution to estimate high resolution p-values.

Downstream analysis

After identifying significant gene-pairs, we took their local correlation profiles and hierarchically clustered them to identify patterns of differential correlation across either pseudotime or space, using maximum distance and Ward's minimum variance method. For the Liver Developmental Dataset correlation analysis, we smoothed local correlation vectors before hierarchical clustering using loess. For both datasets we extracted discrete clusters from the hierarchical clustering using the R function cutree with number of clusters estimated using dynamicTreeCut[48].

Comparing between trajectory branches

We implemented a statistical test for comparing the change in correlation between the two branches, by examining the normalized network strength across the tested networks per branch. We defined network strength for a given gene (node) as the sum of edge weights for significant gene-pairs associated with the gene, divided by the total edge weights across the entire network. The edge weight we selected was the -log(FDR-adjusted P-value) for each gene-pair. For each gene, we calculated the network strength of all genes per branch. We then compared these network strength values between branches using an MA-plot, i.e. comparing the sum of network strengths with the difference of network strengths. To assess significance associated with a single gene – i.e. a gene that tends to have a higher network strength than expected by chance, we repeatedly permuted the gene-pair edge weights across the network and calculated the permuted MA-plot. Individual genes were identified with a significantly nonzero network strength using the Euclidean distance from the origin as the test statistic. To identify genes with a branch-specific network strength, we considered the ratio of significance towards each branch as the test statistic.

Computational efficiency

We previously observed a relationship between the total number of samples and the null distribution of the DCARS test statistics[24]. Here we uncovered further association between the null distribution of the scHOT test statistics and the global correlation across all samples. This represents an opportunity to significantly decrease computational time as one can ‘borrow’ permutations from similarly distributed genes and gene-pairs to estimate the p-value. Our approach is to first calculate global correlations for all gene-pairs to be tested, and then take a uniform sample among the gene-pairs according to the global correlations. For this subset of gene-pairs we permute sample labels and calculate scHOT test statistics. Then for any given gene-pair of interest, we extract the desired number of permutations from this set of permuted scHOT test statistics, according to how similar their global correlation is. These are shown in Supplementary Figure S2E and S4B for the liver and MOB data respectively.

DATA AND SOFTWARE AVAILABILTY

All data analysis was performed on publicly available data. The liver developmental dataset and description is available from https://sydneybiox.github.io/scMerge/articles/Mouse_Liver_Data/Mouse_Liver_Data.html and the specific R data file downloaded from http://www.maths.usyd.edu.au/u/yingxinl/wwwnb/scMergeData/liver_scMerge.rds

The mouse olfactory bulb data was downloaded from the Spatial Research website https://www.spatialresearch.org/resources-published-datasets/doi-10-1126science-aaf2403/ and count matrix data and H&E stained brightfield image related to MOB replicate 11 was downloaded.

Instructions to access data, software, and scripts to perform the analysis presented here is available at https://github.com/MarioniLab/scHOT2019.

ACKNOWLEDGEMENT

The authors thank all their colleagues, particularly at Cancer Research UK Cambridge Institute and The University of Sydney School of Mathematics and Statistics for their support and intellectual engagement. In particular, the authors gratefully acknowledge Pengyi Yang and Michael Morgan for their helpful discussion.

AUTHOR CONTRIBUTIONS

S.G. conceived the study with input from E.P., J.Y.H.Y. and J.C.M.. S.G. developed the method and performed data analysis with input from Y.L. S.G., D.M.L and X.S. interpreted the results with input from

Z.G.H. and J.C.M.. S.G., J.C.M. and J.Y.H.Y. wrote the manuscript. All authors read and approved the final version of the manuscript.

FUNDING

The following sources of funding are gratefully acknowledged. Royal Society Newton International Fellowship (NIF\R1\181950) and funding from the Judith and Coffey Life Lab at the Charles Perkins Centre to S.G.. Australia NHMRC Career Developmental Fellowship (APP1111338) to J.Y.H.Y.. Research Training Program Tuition Fee Offset and Stipend Scholarship and Chen Family Research Scholarship to Y.L.. NIH grant (R21DC015107) to D.M.L.. SJTU-USYD Translate Medicine Fund-Systems Biomedicine (AF6260003) to J.Y.H.Y., X.S., and Z.G.H.. Core funding from EMBL and Cancer Research UK (award no. 17197) to J.C.M..

The funding source had no role in the study design; in the collection, analysis, and interpretation of data, in the writing of the manuscript, and in the decision to submit the manuscript for publication.

REFERENCES

1. Pijuan-Sala, B. *et al.* A single-cell molecular map of mouse gastrulation and early organogenesis. *Nature* (2019). doi:10.1038/s41586-019-0933-9

2. Cao, J. *et al.* The single-cell transcriptional landscape of mammalian organogenesis. *Nature* **566**, 496–502 (2019).

3. Briggs, J. A. *et al.* The dynamics of gene expression in vertebrate embryogenesis at single-cell resolution. *Science* **360**, 1–16 (2018).

4. Wagner, D. E. *et al.* Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. *Science* **360**, 981–987 (2018).

5. Polioudakis, D. *et al.* A Single-Cell Transcriptomic Atlas of Human Neocortical Development during Mid-gestation. *Neuron* (2019). doi:10.1016/j.neuron.2019.06.011

6. Mohammed, H. *et al.* Single-Cell Landscape of Transcriptional Heterogeneity and Cell Fate Decisions during Mouse Early Gastrulation. *Cell Rep.* **20**, 1215–1228 (2017).

7. Mojtahedi, M. *et al.* Cell Fate Decision as High-Dimensional Critical State Transition. *PLOS Biol.* **14**, e2000640 (2016).

8. Basu, S., Kumbier, K., Brown, J. B. & Yu, B. Iterative random forests to discover predictive and stable high-order interactions. *Proc. Natl. Acad. Sci.* **115**, 1943–1948 (2018).

9. Bageritz, J. *et al.* Gene expression atlas of a developing tissue by single cell expression correlation analysis. *Nat. Methods* **16**, 750–756 (2019).

10. Ståhl, P. L. *et al.* Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* **353**, 78–82 (2016).

11. Lubeck, E., Coskun, A. F., Zhiyentayev, T., Ahmad, M. & Cai, L. Single-cell in situ RNA profiling by sequential hybridization. *Nat. Methods* **11**, 360–361 (2014).

12. Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S. & Zhuang, X. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science* **348**, (2015).

13. Street, K. *et al.* Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics* **19**, 477 (2018).

14. Campbell, K. R. & Yau, C. A descriptive marker gene approach to single-cell pseudotime inference. *Bioinformatics* **35**, 28–35 (2018).

15. Haghverdi, L., Büttner, M., Wolf, F. A., Buettner, F. & Theis, F. J. Diffusion pseudotime robustly reconstructs lineage branching. *Nat. Methods* **13**, 845–848 (2016).

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

17. 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).

18. Berge, K. Van Den *et al.* Trajectory-based differential expression analysis for single-cell sequencing data. *bioRxiv* 623397 (2019). doi:10.1101/623397

19. Lönnberg, T. *et al.* Single-cell RNA-seq and computational analysis using temporal mixture

modeling resolves T H 1/T FH fate bifurcation in malaria. *Sci. Immunol.* **2**, eaal2192 (2017).

20. Qiu, X. *et al.* Reversed graph embedding resolves complex single-cell trajectories. *Nat. Methods* **14**, 979–982 (2017).

21. Fuller, T. F. *et al.* Weighted gene coexpression network analysis strategies applied to mouse weight. *Mamm. Genome* **18**, 463–72 (2007).

22. Treutlein, B. *et al.* Dissecting direct reprogramming from fibroblast to neuron using single-cell RNA-seq. *Nature* **534**, 391–5 (2016).

23. Chen, S. & Mar, J. C. Evaluating methods of inferring gene regulatory networks highlights their lack of performance for single cell gene expression data. *BMC Bioinformatics* **19**, 232 (2018).

24. Ghazanfar, S., Strbenac, D., Ormerod, J. T., Yang, J. Y. H. H. & Patrick, E. DCARS: Differential correlation across ranked samples. *Bioinformatics* **35**, 1–7 (2018).

25. Svensson, V., Teichmann, S. A. & Stegle, O. SpatialDE: Identification of spatially variable genes. *Nat. Methods* **15**, 343–346 (2018).

26. Nitzan, M., Karaiskos, N., Friedman, N. & Rajewsky, N. Charting a tissue from single-cell transcriptomes. *bioRxiv* 1–31 (2018). doi:dx.doi.org/10.1101/456350

27. Yang, L. *et al.* A single-cell transcriptomic analysis reveals precise pathways and regulatory mechanisms underlying hepatoblast differentiation. *Hepatology* **66**, 1387–1401 (2017).

28. Dong, J. *et al.* Single-cell RNA-seq analysis unveils a prevalent epithelial/mesenchymal hybrid state during mouse organogenesis. *Genome Biol.* **19**, 31 (2018).

29. Su, X. *et al.* Single-cell RNA-Seq analysis reveals dynamic trajectories during mouse liver development. *BMC Genomics* **18**, 946 (2017).

30. Camp, J. G. *et al.* Multilineage communication regulates human liver bud development from pluripotency. *Nature* **546**, 533–538 (2017).

31. Sugimoto, N. *et al.* Identification of novel human Cdt1-binding proteins by a proteomics approach: proteolytic regulation by APC/CCdh1. *Mol. Biol. Cell* **19**, 1007–21 (2008).

32. Zape, J. P., Lizama, C. O., Cautivo, K. M. & Zovein, A. C. Cell cycle dynamics and complement expression distinguishes mature haematopoietic subsets arising from hemogenic endothelium. *Cell Cycle* **16**, 1835–1847 (2017).

33. Thakurela, S. *et al.* Gene regulation and priming by topoisomerase IIα in embryonic stem cells. *Nat. Commun.* **4**, 2478 (2013).

34. Rialland, M., Sola, F. & Santocanale, C. Essential role of human CDT1 in DNA replication and chromatin licensing. *J. Cell Sci.* **115**, 1435–40 (2002).

35. Lin, D. M. *et al.* Spatial patterns of gene expression in the olfactory bulb. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 12718–12723 (2004).

36. Fan, X. L., Zhang, J. S., Zhang, X. Q., Yue, W. & Ma, L. Differential regulation of β-arrestin 1 and β-arrestin 2 gene expression in rat brain by morphine. *Neuroscience* **117**, 383–389 (2003).

37. Macias, M. *et al.* Spatiotemporal Characterization of mTOR Kinase Activity Following Kainic Acid Induced Status Epilepticus and Analysis of Rat Brain Response to Chronic Rapamycin Treatment. *PLoS One* **8**, e64455 (2013).

38. Wilson, P. O. *et al.* The immunolocalization of protein gene product 9.5 using rabbit polyclonal and mouse monoclonal antibodies. *Br. J. Exp. Pathol.* **69**, 91–104 (1988).

39. Gray, N. W. *et al.* Dynamin 3 Is a Component of the Postsynapse, Where it Interacts with mGluR5 and Homer. *Curr. Biol.* **13**, 510–515 (2003).

40. Kendall, R. T. *et al.* Arrestin-dependent Angiotensin AT 1 Receptor Signaling Regulates Akt and mTor-mediated Protein Synthesis. *J. Biol. Chem.* **289**, 26155–26166 (2014).

41. Girnita, L. *et al.* {beta}-Arrestin is crucial for ubiquitination and down-regulation of the insulin-like growth factor-1 receptor by acting as adaptor for the MDM2 E3 ligase. *J. Biol. Chem.* **280**, 24412–9 (2005).

42. Bhatnagar, A. *et al.* The Dynamin-dependent, Arrestin-independent Internalization of 5-

Hydroxytryptamine 2A (5-HT 2A) Serotonin Receptors Reveals Differential Sorting of Arrestins and 5-HT 2A Receptors during Endocytosis. *J. Biol. Chem.* **276**, 8269–8277 (2001).

43. Jacque, C. M., Collet, A., Raoul, M., Monge, M. & Gumpel, M. Functional maturation of the oligodendrocytes and myelin basic protein expression in the olfactory bulb of the mouse. *Dev. Brain Res.* **21**, 277–282 (1985).

44. Lin, Y. *et al.* scMerge leverages factor analysis, stable expression, and pseudoreplication to merge multiple single-cell RNA-seq datasets. *Proc. Natl. Acad. Sci.* **116**, 9775–9784 (2019).

45. Lin, Y. *et al.* Evaluating stably expressed genes in single cells. *Gigascience* **8**, 229815 (2019).

46. Brennecke, P. *et al.* Accounting for technical noise in single-cell RNA-seq experiments. *Nat. Methods* **10**, 1093–5 (2013).

47. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B. Stat. Methodol.* **57**, 289–300 (1995).

48. Langfelder, P., Zhang, B. & Horvath, S. Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics* **24**, 719–720 (2008).

49. Székely, G. J., Rizzo, M. L. & Bakirov, N. K. Measuring and testing dependence by correlation of distances. *Ann. Stat.* **35**, 2769–2794 (2007).

50. Ghazanfar, S. & Yang, J. Y. H. Characterizing mutation–expression network relationships in multiple cancers. *Comput. Biol. Chem.* **63**, 73–82 (2016).

FIGURE AND TABLE LEGENDS

Figure 1. Methods workflow**. A.** Example showing a differentiation trajectory where genes are tested for changes in higher order interactions such as variability and correlation along the trajectory. A set of local higher order statistics are calculated, and significance is compared by repeatedly permuting samples (grey curves). The vector of local estimates of higher order statistics are combined using the sample standard deviation to assess how variable it is across time. **B.** Example showing that in a spatial context, scHOT calculates a field of local estimates of correlation across space, and compares the variability associated with these with permuted sample points across space.

Figure 2. A. Charts showing relative proportion of hepatic and non-hepatic cells in the Developmental Liver Data, across original dataset and embryonic stage. **B.** Monocle trajectory of hepatic cells showing a bifurcating trajectory of hepatoblasts into either hepatocytes or cholangiocytes. **C**. Panel of embryonic stage for each cell along the differentiation trajectory, and gene expression of markers of each cell type. **D.** A selection of genes significantly associated with a change in variability along the first branch of the differentiation trajectory, scatterplots show the expression of genes against the pseudotime estimates, and line plots show the local estimate of variance for the first branch (thick lines) and further for the two branches (thin lines). Examples are shown of genes that increase in variability via 'fanning' of gene expression along the trajectory (*Birc5*), by a skewed distribution arising (*H2afz*), and by changes in the modality of the expression (*Tacc3* from unimodal to bimodal). *Hmgcs2* is an example of a gene that decreases in variability. **E.** Functional enrichment analysis of genes that significantly increase in variability along the first branch.

Figure 3. A. Sequence of scatterplots showing expression of *Cdt1* and *Top2a* at equally spaced points along the entire trajectory from hepatoblast to hepatocyte. Points are colored by their position along the trajectory, and point size corresponds to the weight given to that region of the trajectory. Neither gene is significantly differentially expressed or differentially variable along the trajectory, but the gene-pair is significantly differentially correlated. **B.** Clustering of local weighted correlation of all significant gene-pairs for the hepatocyte branch, showing groups of gene-pairs that appear to gain or lose correlation across the trajectory. **C.** Functional enrichment analysis of genes belonging to the set of gene-pairs among Clusters 9 and 7 respectively for the hepatocyte branch. **D.** Comparison of hepatocyte and

cholangiocyte branches using network strength across all significant gene-pairs for either branch. Black labelled genes are significantly branch specific, while red labelled genes are significantly common across both branches (FDR-adjusted P-value < 0.05).

Figure 4. A. H&E image of mouse olfactory tissue section with labeling of known anatomical layers. **B.** Spatial expression plots of two genes, *Arrb1* and *Mtor*, which are not significantly differentially expressed across space, and are significantly differentially correlated across space. The third plot shows the local spatial correlation estimated for these two genes, recapitulating the layered pattern of the olfactory bulb. **C**. Heatmap showing all sampled points (rows) and clusters of significantly spatially differentially correlated gene-pairs (columns), with spatial maps of mean local correlation (bottom row) for each group, and highlighted positions (column on right) for the sampled points grouped into clusters. **D.** Spatial maps of local correlation for selected clusters of gene-pairs and barplots showing functional enrichment analysis.

SUPPLEMENTARY FIGURE LEGENDS

Supplementary Figure S1. A. Example showing testing for correlation differences in three distinct groups. A set of local higher order statistics are calculated, and significance is compared by repeatedly permuting samples (grey boxplots). The set of local estimates of higher order statistics are combined using the sample standard deviation to assess how variable they are between groups. **B.** Possible schemes for the testing scaffold using gene networks, including: i) a gene-gene network; ii) a gene set scaffold where all pairwise combinations within a gene set are included; and iii) selected genes of interest versus all others. **C.** Examples of weighted higher order functions including weighted Pearson correlation, weighted Spearman correlation, weighted variance. Note that any user defined function can be used.

Supplementary Figure S2. A. HVG selection for Developmental Liver Data. **B.** Barplot for Gene Ontology testing for genes showing loss of variability along hepatoblast branch. **C.** Scatter and ribbon plot of genes showing loss of variability along hepatoblast branch. **D.** Scatter and ribbon plot of genes showing gain of variability along hepatoblast branch. **E.** Global correlation and null scHOT correlation test statistics for both hepatocyte and cholangiocyte branches.

Supplementary Figure S3. A. Gene ontology barplots for all scHOT hepatocyte clusters. **B.** Line plots of clustered significant scHOT gene-pairs for full cholangiocyte branch. **C.** Gene ontology barplots for all scHOT cholangiocyte clusters.

Supplementary Figure S4. A. HVG selection for Spatial Transcriptomics analysis. **B.** Global correlation and null scHOT correlation test statistics for spatial MOB data. **C.** Spatial expression plots of two gene-pairs *Arrb1* and *Uchl1* as well as *Arrb1* and *Dnm3* which are not significantly differentially expressed across space, but are significantly differentially correlated across space. The third plot shows the local spatial correlation estimated for these two genes, recapitulating the layered pattern of the olfactory bulb. **D.** Spatial maps of mean local correlation and gene ontology functional enrichment barplots for all MOB scHOT clusters.

SUPPLEMENTARY TEXT

scHOT extensions and considerations

scHOT is a flexible framework within which multiple aspects can be modified to facilitate bespoke analysis. For example, higher order patterns can be studied along trajectories, across space, or among discrete groups (Supplementary Figure S1A). Moreover, distinct sets of genes or gene-pairs can be interrogated depending on the biological question of interest (Supplementary Figure S1B). Of particular interest, the local weighting scheme and concordance function can also be adapted, depending upon the biological context (some examples are given in Supplementary Figure S1C). For example, if one were interested in identifying changes in higher order interactions along a circular trajectory, e.g. the cell cycle, one could define a local weighting scheme that was also circular – by ensuring that the two ends match given any starting point. Another example is for discrete groups that are either completely distinct, or ordered in some way – e.g. over discrete time points along a time course experiment, one may define the weight matrix to incorporate the discrete grouping, while also accounting for the flanking groups. More generally, one may wish to place a higher local weight over a particular local region and a smaller weight over a different region. We note, however, that these changes to the weighting scheme may affect the generalizability of the null distributions across multiple genes or gene-pairs, so the user should take care in ensuring that the null distributions appear similar when employing computational speed-up steps.

Any concordance metric can be 'plugged in' if using a block type of weight matrix. That is, one is able to use the fast implementation of distance correlation between two distributions[49], mutual information, partial correlation, or any concordance metric suited especially to other data types such as ordinal or binarized single cell data[50], without needing to explicitly derive weighted formulations of these concordance metrics. Additionally, any concordance metric that doesn't necessarily have a 'weighted' formulation and/or implementation can be utilized using block weighting scheme. This makes scHOT versatile, by enabling user-defined metrics. For summarizing the vector of local higher order statistics, users may wish to substitute the sample standard deviation with any other choice of variability or change estimate – e.g. if the goal is to examine how monotonic a change in higher order interaction is, a measure of monotonicity such as mutual information or Spearman correlation with the weighting scheme index could be used.

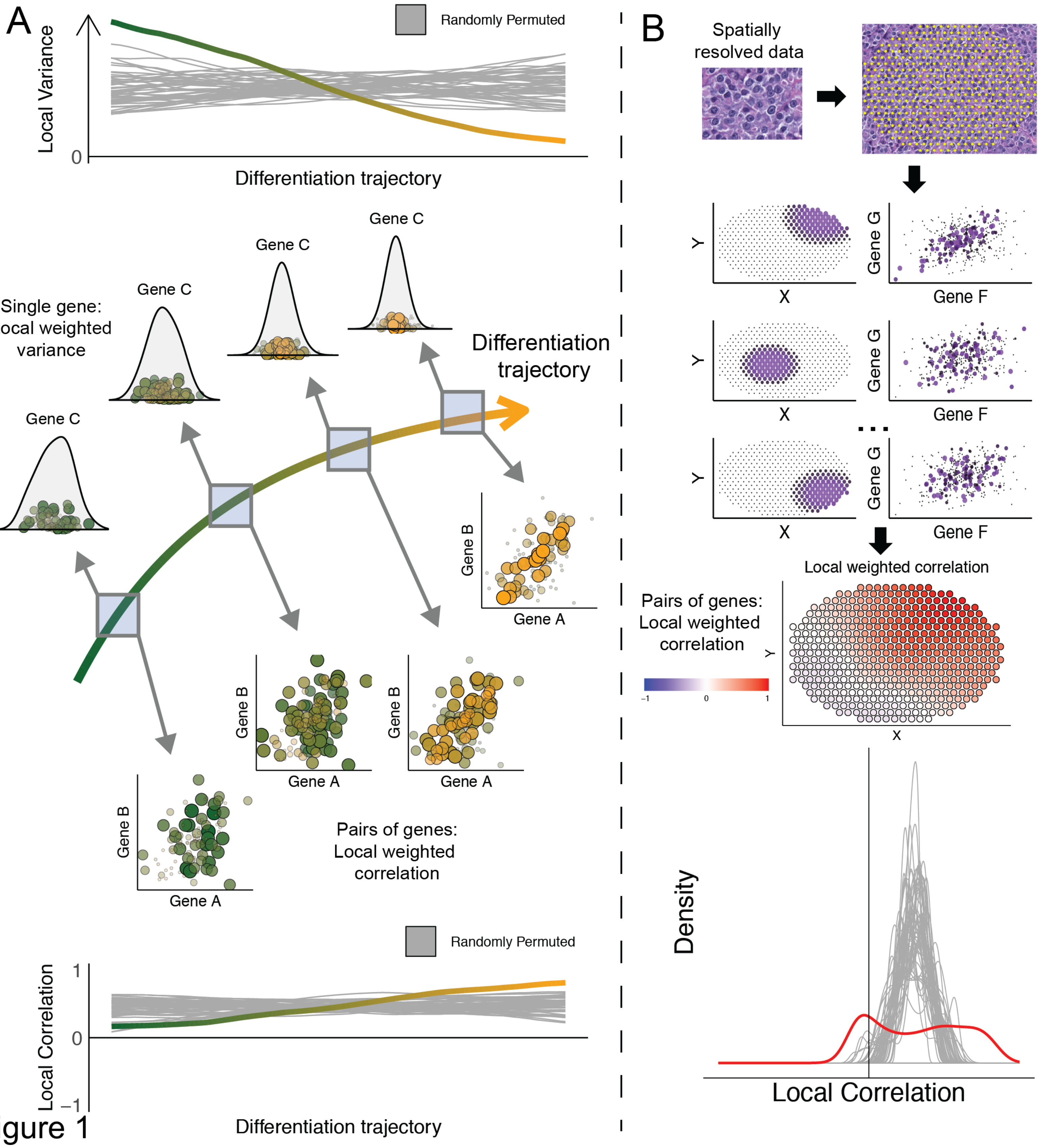

A
Randomly Permuted
Local Variance
0
Differentiation trajectory
Gene C
Gene C
Gene C
Gene C
Single gene:
Local weighted
variance
Differentiation
trajectory
Gene B
Gene A
Gene B
Gene A
Gene B
Gene A
Gene B
Gene A
Pairs of genes:
Local weighted
correlation
Randomly Permuted
1
0
−1
Local Correlation
Differentiation trajectory
B
Spatially
resolved data
Y
X
Gene G
Gene F
Y
X
Gene G
Gene F
Y
X
Gene G
Gene F
Local weighted correlation
Pairs of genes:
Local weighted
correlation
−1
0
1
Y
X
Density
Local Correlation

Figure 1

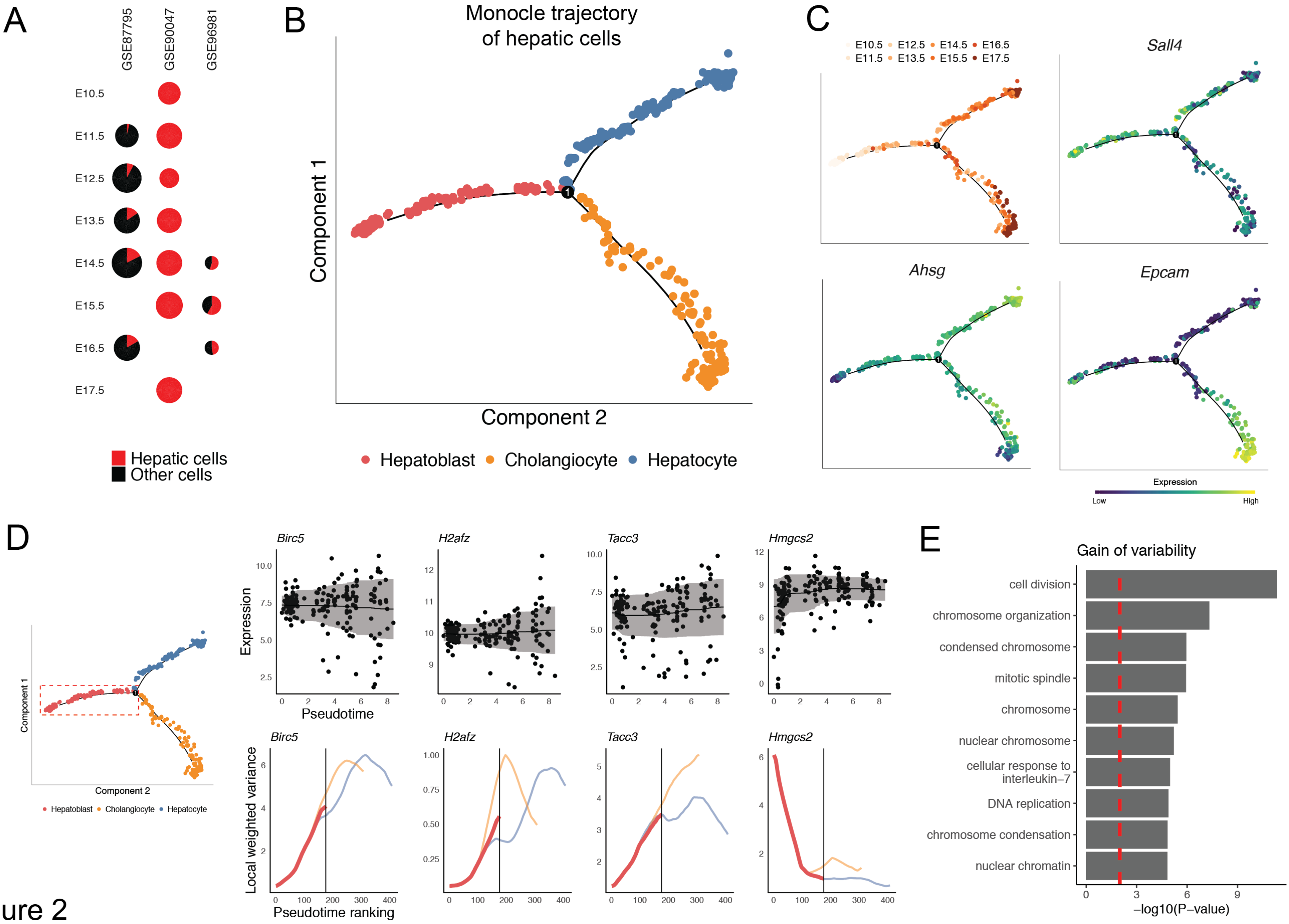

A
GSE87795
GSE90047
GSE96981
E10.5
E11.5
E12.5
E13.5
E14.5
E15.5
E16.5
E17.5
🟥 Hepatic cells
⬛ Other cells
B
Monocle trajectory
of hepatic cells
Component 1
Component 2
● Hepatoblast ● Cholangiocyte ● Hepatocyte
C
E10.5 E11.5 E12.5 E13.5 E14.5 E15.5 E16.5 E17.5
Sall4
Ahsg
Epcam
Expression
Low
High
D
Component 1
Component 2
Hepatoblast Cholangiocyte Hepatocyte
Birc5
H2afz
Tacc3
Hmgcs2
Expression
Pseudotime
Local weighted variance
Pseudotime ranking
E
Gain of variability
cell division
chromosome organization
condensed chromosome
mitotic spindle
chromosome
nuclear chromosome
cellular response to interleukin−7
DNA replication
chromosome condensation
nuclear chromatin
0
3
6
9
−log10(P−value)

Figure 2

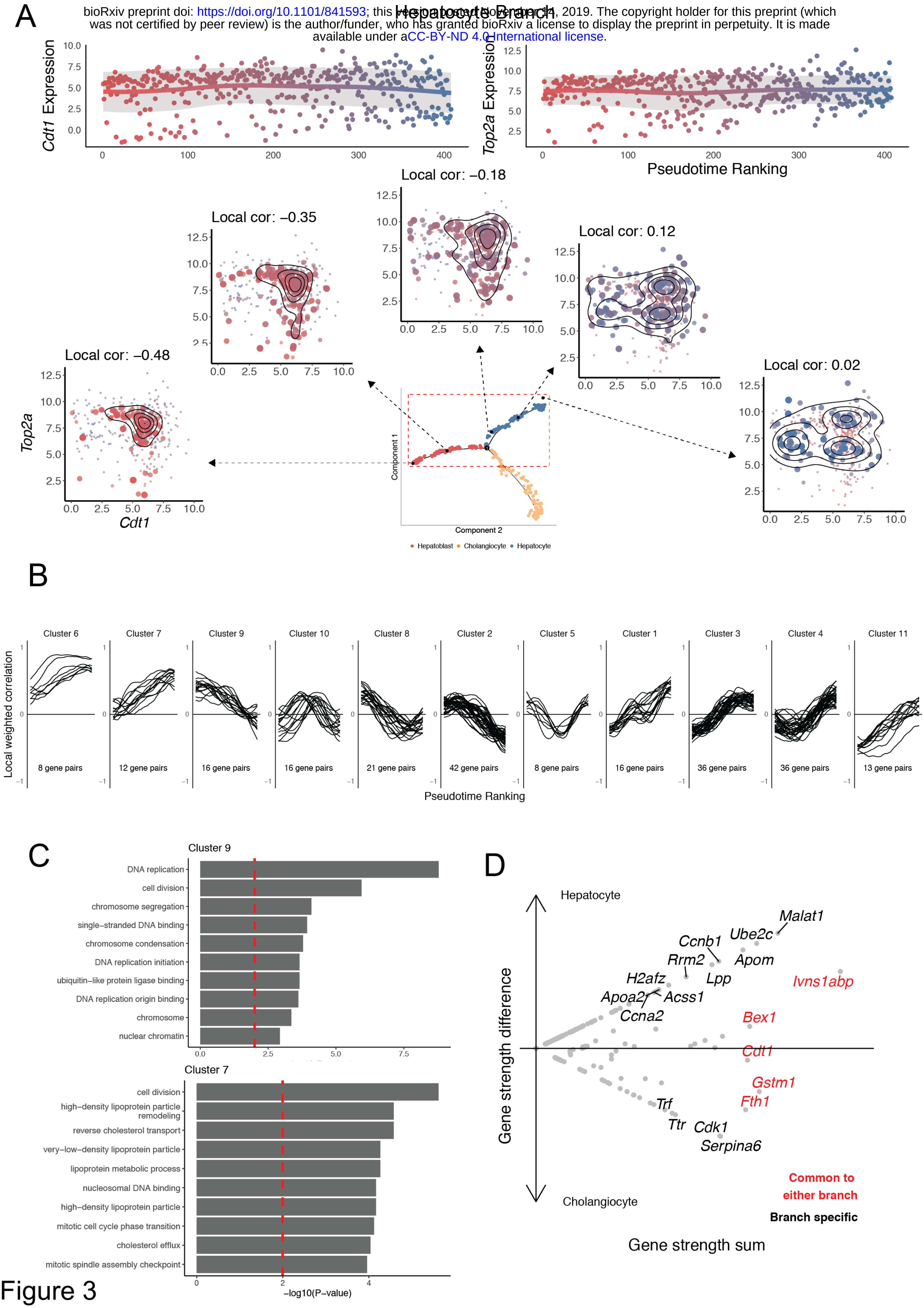

A
Hepatocyte Branch
Cdt1 Expression
Top2a Expression
Pseudotime Ranking
0
100
200
300
400
Local cor: −0.48
Local cor: −0.35
Local cor: −0.18
Local cor: 0.12
Local cor: 0.02
Top2a
Cdt1
Component 1
Component 2
Hepatoblast
Cholangiocyte
Hepatocyte
B
Cluster 6
Cluster 7
Cluster 9
Cluster 10
Cluster 8
Cluster 2
Cluster 5
Cluster 1
Cluster 3
Cluster 4
Cluster 11
8 gene pairs
12 gene pairs
16 gene pairs
16 gene pairs
21 gene pairs
42 gene pairs
8 gene pairs
16 gene pairs
36 gene pairs
36 gene pairs
13 gene pairs
Local weighted correlation
Pseudotime Ranking
C
Cluster 9
DNA replication
cell division
chromosome segregation
single−stranded DNA binding
chromosome condensation
DNA replication initiation
ubiquitin−like protein ligase binding
DNA replication origin binding
chromosome
nuclear chromatin
Cluster 7
cell division
high−density lipoprotein particle remodeling
reverse cholesterol transport
very−low−density lipoprotein particle
lipoprotein metabolic process
nucleosomal DNA binding
high−density lipoprotein particle
mitotic cell cycle phase transition
cholesterol efflux
mitotic spindle assembly checkpoint
−log10(P−value)
D
Hepatocyte
Cholangiocyte
Gene strength difference
Gene strength sum
Malat1
Ube2c
Ccnb1
Rrm2
Apom
H2afz
Lpp
Ivns1abp
Apoa2
Acss1
Ccna2
Bex1
Cdt1
Gstm1
Fth1
Trf
Ttr
Cdk1
Serpina6
Common to either branch
Branch specific

Figure 3

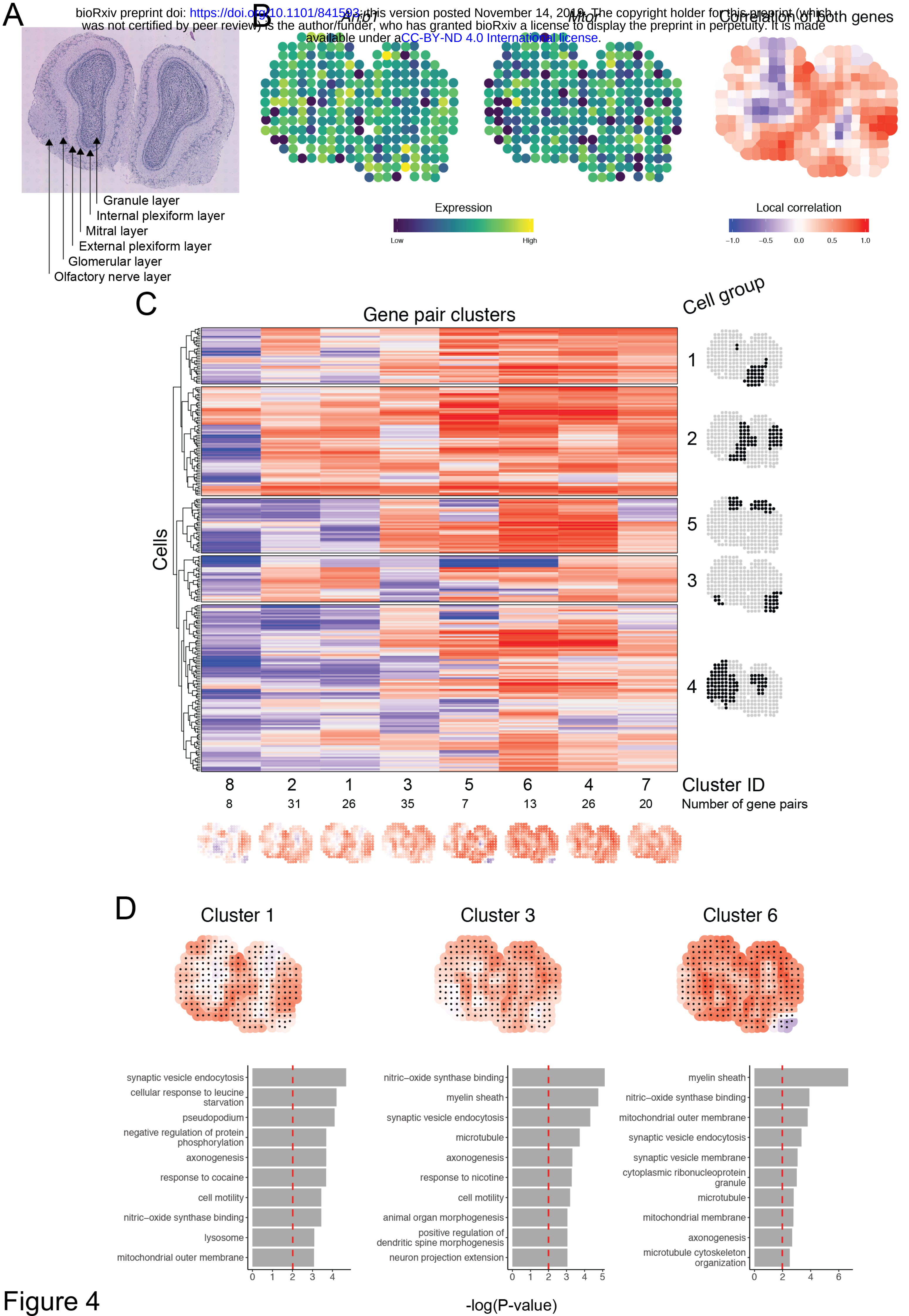
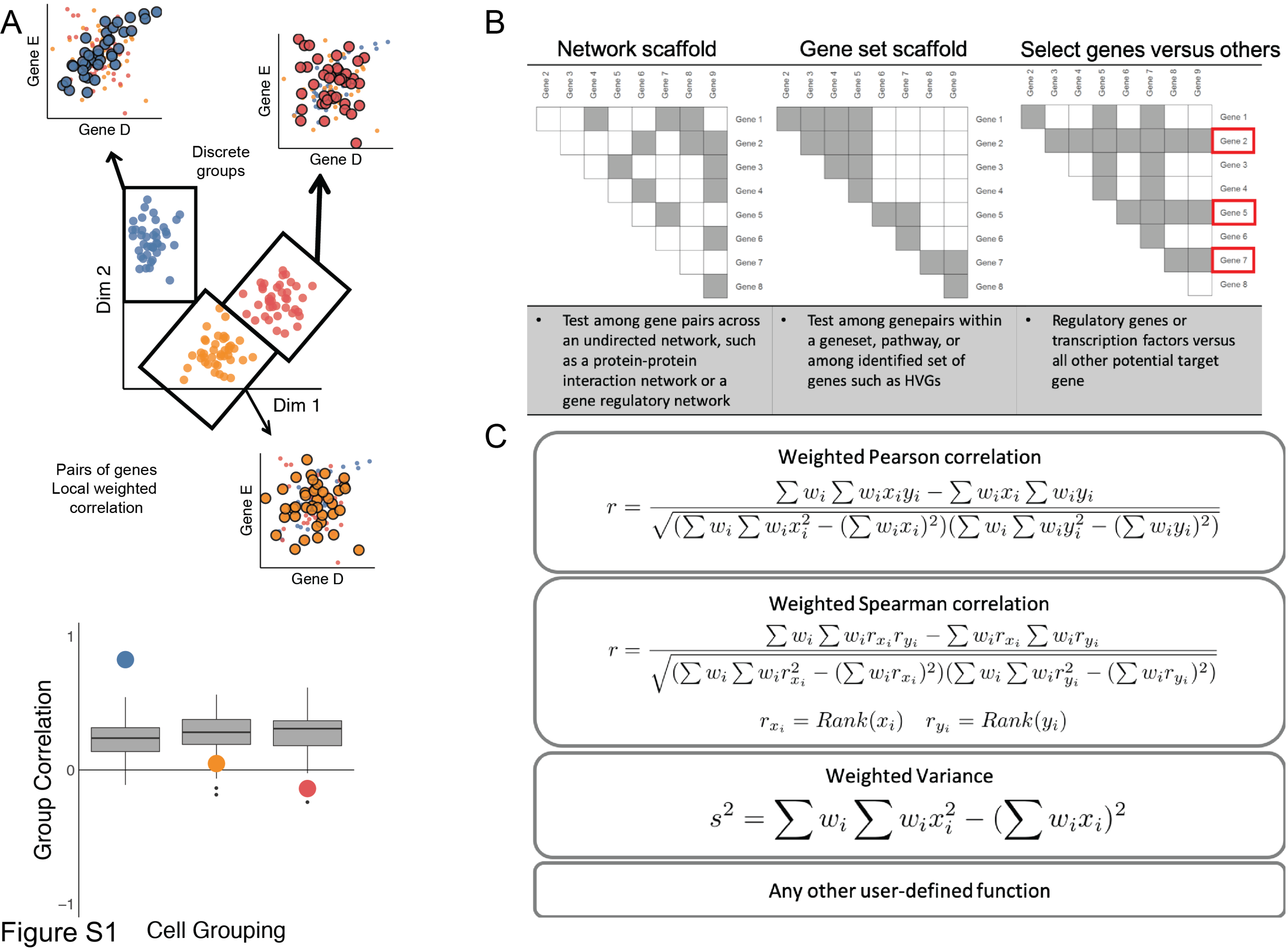

A
Granule layer
Internal plexiform layer
Mitral layer
External plexiform layer
Glomerular layer
Olfactory nerve layer
B
Arrb1
Mtor
Correlation of both genes
Expression
Low
High
Local correlation
−1.0
−0.5
0.0
0.5
1.0
C
Gene pair clusters
Cell group
Cells
1
2
5
3
4
8
2
1
3
5
6
4
7
Cluster ID
8
31
26
35
7
13
26
20
Number of gene pairs
D
Cluster 1
Cluster 3
Cluster 6
synaptic vesicle endocytosis
cellular response to leucine starvation
pseudopodium
negative regulation of protein phosphorylation
axonogenesis
response to cocaine
cell motility
nitric−oxide synthase binding
lysosome
mitochondrial outer membrane
0
1
2
3
4
nitric−oxide synthase binding
myelin sheath
synaptic vesicle endocytosis
microtubule
axonogenesis
response to nicotine
cell motility
animal organ morphogenesis
positive regulation of dendritic spine morphogenesis
neuron projection extension
0
1
2
3
4
5
myelin sheath
nitric−oxide synthase binding
mitochondrial outer membrane
synaptic vesicle endocytosis
synaptic vesicle membrane
cytoplasmic ribonucleoprotein granule
microtubule
mitochondrial membrane
axonogenesis
microtubule cytoskeleton organization
0
2
4
6
-log(P-value)

Figure 4



- Test among gene pairs across an undirected network, such as a protein-protein interaction network or a gene regulatory network
- Test among genepairs within a geneset, pathway, or among identified set of genes such as HVGs
- Regulatory genes or transcription factors versus all other potential target gene

Weighted Pearson correlation

$$r = \frac{\sum w_i \sum w_i x_i y_i - \sum w_i x_i \sum w_i y_i}{\sqrt{(\sum w_i \sum w_i x_i^2 - (\sum w_i x_i)^2)(\sum w_i \sum w_i y_i^2 - (\sum w_i y_i)^2)}}$$

Weighted Spearman correlation

$$r = \frac{\sum w_i \sum w_i r_{x_i} r_{y_i} - \sum w_i r_{x_i} \sum w_i r_{y_i}}{\sqrt{(\sum w_i \sum w_i r_{x_i}^2 - (\sum w_i r_{x_i})^2)(\sum w_i \sum w_i r_{y_i}^2 - (\sum w_i r_{y_i})^2)}}$$

$$r_{x_i} = Rank(x_i) \quad r_{y_i} = Rank(y_i)$$

Weighted Variance

$$s^2 = \sum w_i \sum w_i x_i^2 - (\sum w_i x_i)^2$$

Any other user-defined function

Figure S1

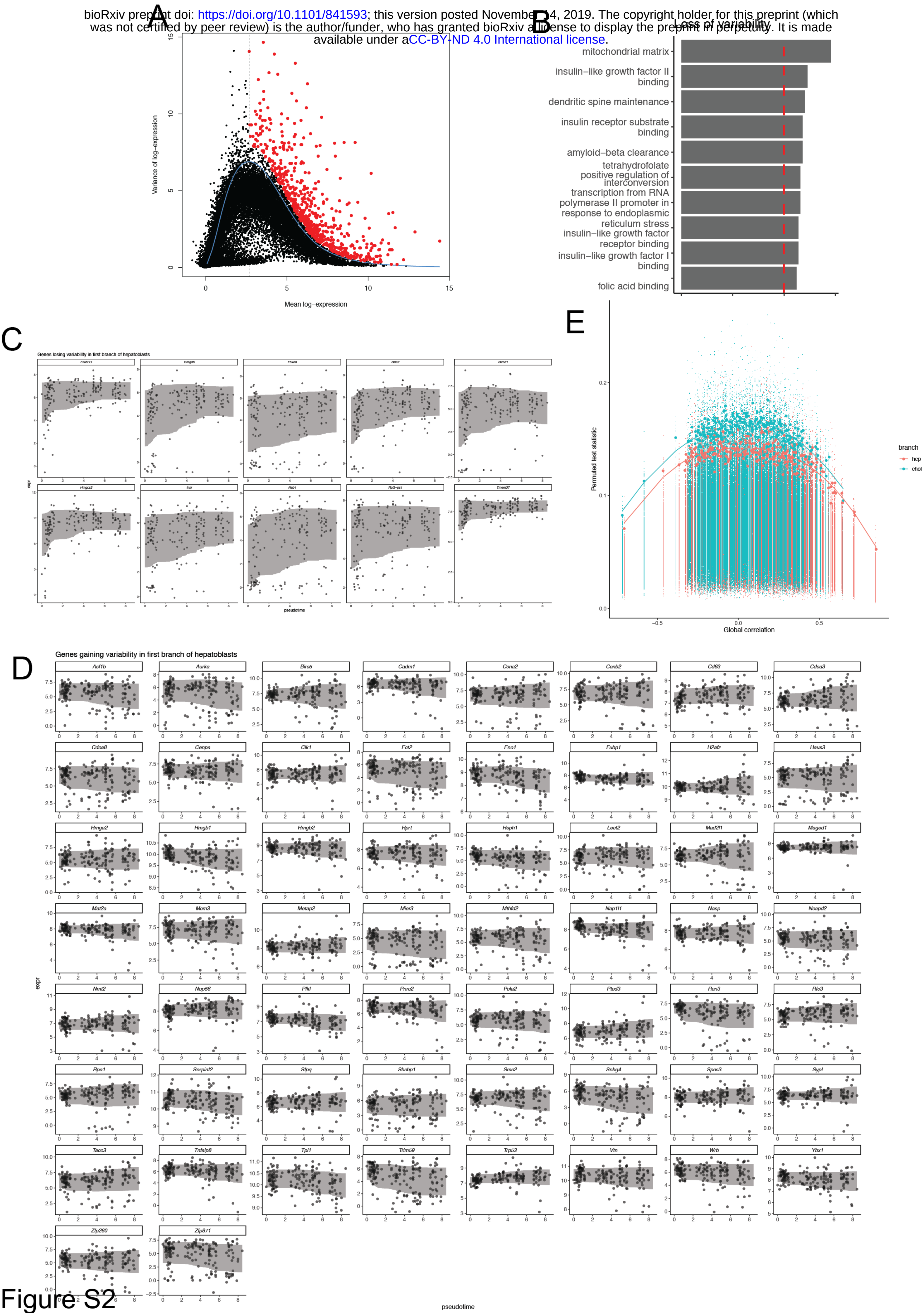

A
Variance of log−expression
Mean log−expression
B
Loss of variability
mitochondrial matrix
insulin−like growth factor II binding
dendritic spine maintenance
insulin receptor substrate binding
amyloid−beta clearance
positive regulation of transcription from RNA polymerase II promoter in response to endoplasmic reticulum stress
tetrahydrofolate interconversion
insulin−like growth factor receptor binding
insulin−like growth factor I binding
folic acid binding
C
Genes losing variability in first branch of hepatoblasts
Creb3l3
Dmgdh
Fbxo8
Gls2
Gmd1
Hmgcs2
Insr
Nab1
Rpl3-ps1
Tmem37
expr
pseudotime
E
Permuted test statistic
Global correlation
branch
hep
chol
D
Genes gaining variability in first branch of hepatoblasts
Asf1b
Aurka
Birc5
Cadm1
Ccna2
Ccnb2
Cd63
Cdca3
Cdca8
Cenpa
Clk1
Ect2
Eno1
Fubp1
H2afz
Haus3
Hmga2
Hmgb1
Hmgb2
Hprt
Hsph1
Lect2
Mad2l1
Maged1
Mat2a
Mcm3
Metap2
Mier3
Mthfd2
Nap1l1
Nasp
Ncapd2
Nmt2
Nop56
Pfkl
Pnrc2
Pola2
Ptod3
Rcn3
Rfc3
Rpa1
Serpinf2
Sfpq
Shcbp1
Smc2
Snhg4
Spcs3
Sypl
Tacc3
Tnfaip8
Tpi1
Trim59
Trp53
Vtn
Wrb
Ybx1
Zfp260
Zfp871
expr
pseudotime

Figure S2

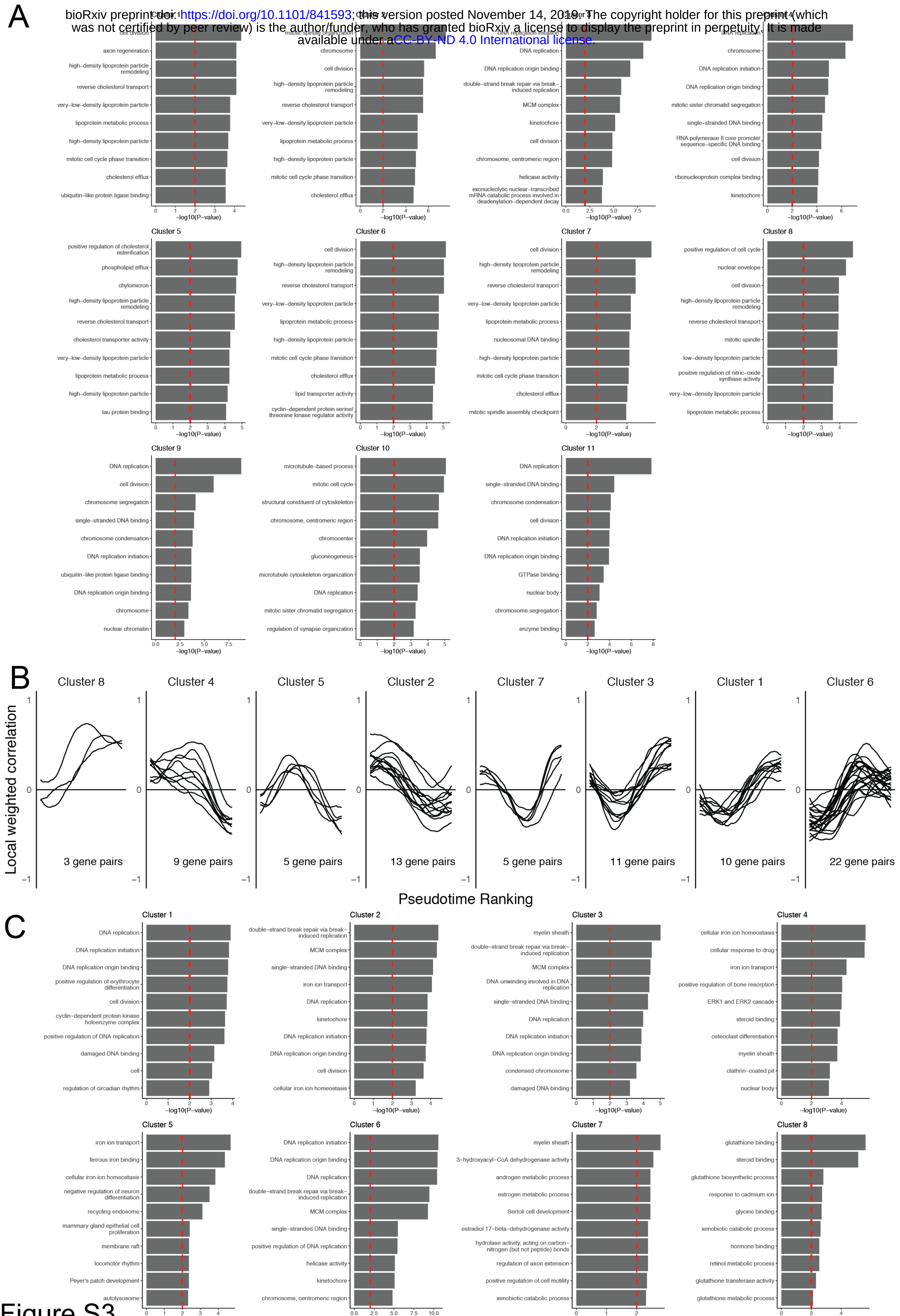

A
Cluster 1
cell division
axon regeneration
high−density lipoprotein particle remodeling
reverse cholesterol transport
very−low−density lipoprotein particle
lipoprotein metabolic process
high−density lipoprotein particle
mitotic cell cycle phase transition
cholesterol efflux
ubiquitin−like protein ligase binding
−log10(P−value)
Cluster 2
mitotic spindle organization
chromosome
cell division
high−density lipoprotein particle remodeling
reverse cholesterol transport
very−low−density lipoprotein particle
lipoprotein metabolic process
high−density lipoprotein particle
mitotic cell cycle phase transition
cholesterol efflux
Cluster 3
DNA replication initiation
DNA replication
DNA replication origin binding
double−strand break repair via break−induced replication
MCM complex
kinetochore
cell division
chromosome, centromeric region
helicase activity
exonucleolytic nuclear−transcribed mRNA catabolic process involved in deadenylation−dependent decay
Cluster 4
DNA replication
chromosome
DNA replication initiation
DNA replication origin binding
mitotic sister chromatid segregation
single−stranded DNA binding
RNA polymerase II core promoter sequence−specific DNA binding
cell division
ribonucleoprotein complex binding
kinetochore
Cluster 5
positive regulation of cholesterol esterification
phospholipid efflux
chylomicron
high−density lipoprotein particle remodeling
reverse cholesterol transport
cholesterol transporter activity
very−low−density lipoprotein particle
lipoprotein metabolic process
high−density lipoprotein particle
tau protein binding
Cluster 6
cell division
high−density lipoprotein particle remodeling
reverse cholesterol transport
very−low−density lipoprotein particle
lipoprotein metabolic process
high−density lipoprotein particle
mitotic cell cycle phase transition
cholesterol efflux
lipid transporter activity
cyclin−dependent protein serine/threonine kinase regulator activity
Cluster 7
cell division
high−density lipoprotein particle remodeling
reverse cholesterol transport
very−low−density lipoprotein particle
lipoprotein metabolic process
nucleosomal DNA binding
high−density lipoprotein particle
mitotic cell cycle phase transition
cholesterol efflux
mitotic spindle assembly checkpoint
Cluster 8
positive regulation of cell cycle
nuclear envelope
cell division
high−density lipoprotein particle remodeling
reverse cholesterol transport
mitotic spindle
low−density lipoprotein particle
positive regulation of nitric−oxide synthase activity
very−low−density lipoprotein particle
lipoprotein metabolic process
Cluster 9
DNA replication
cell division
chromosome segregation
single−stranded DNA binding
chromosome condensation
DNA replication initiation
ubiquitin−like protein ligase binding
DNA replication origin binding
chromosome
nuclear chromatin
Cluster 10
microtubule−based process
mitotic cell cycle
structural constituent of cytoskeleton
chromosome, centromeric region
chromocenter
gluconeogenesis
microtubule cytoskeleton organization
DNA replication
mitotic sister chromatid segregation
regulation of synapse organization
Cluster 11
DNA replication
single−stranded DNA binding
chromosome condensation
cell division
DNA replication initiation
DNA replication origin binding
GTPase binding
nuclear body
chromosome segregation
enzyme binding
B
Local weighted correlation
Cluster 8
3 gene pairs
Cluster 4
9 gene pairs
Cluster 5
5 gene pairs
Cluster 2
13 gene pairs
Cluster 7
5 gene pairs
Cluster 3
11 gene pairs
Cluster 1
10 gene pairs
Cluster 6
22 gene pairs
1
0
−1
Pseudotime Ranking
C
Cluster 1
DNA replication
DNA replication initiation
DNA replication origin binding
positive regulation of erythrocyte differentiation
cell division
cyclin−dependent protein kinase holoenzyme complex
positive regulation of DNA replication
damaged DNA binding
cell
regulation of circadian rhythm
Cluster 2
double−strand break repair via break−induced replication
MCM complex
single−stranded DNA binding
iron ion transport
DNA replication
kinetochore
DNA replication initiation
DNA replication origin binding
cell division
cellular iron ion homeostasis
Cluster 3
myelin sheath
double−strand break repair via break−induced replication
MCM complex
DNA unwinding involved in DNA replication
single−stranded DNA binding
DNA replication
DNA replication initiation
DNA replication origin binding
condensed chromosome
damaged DNA binding
Cluster 4
cellular iron ion homeostasis
cellular response to drug
iron ion transport
positive regulation of bone resorption
ERK1 and ERK2 cascade
steroid binding
osteoclast differentiation
myelin sheath
clathrin−coated pit
nuclear body
Cluster 5
iron ion transport
ferrous iron binding
cellular iron ion homeostasis
negative regulation of neuron differentiation
recycling endosome
mammary gland epithelial cell proliferation
membrane raft
locomotor rhythm
Peyer's patch development
autolysosome
Cluster 6
DNA replication initiation
DNA replication origin binding
DNA replication
double−strand break repair via break−induced replication
MCM complex
single−stranded DNA binding
positive regulation of DNA replication
helicase activity
kinetochore
chromosome, centromeric region
Cluster 7
myelin sheath
3−hydroxyacyl−CoA dehydrogenase activity
androgen metabolic process
estrogen metabolic process
Sertoli cell development
estradiol 17−beta−dehydrogenase activity
hydrolase activity, acting on carbon−nitrogen (but not peptide) bonds
regulation of axon extension
positive regulation of cell motility
xenobiotic catabolic process
Cluster 8
glutathione binding
steroid binding
glutathione biosynthetic process
response to cadmium ion
glycine binding
xenobiotic catabolic process
hormone binding
retinol metabolic process
glutathione transferase activity
glutathione metabolic process
−log10(P−value)

Figure S3

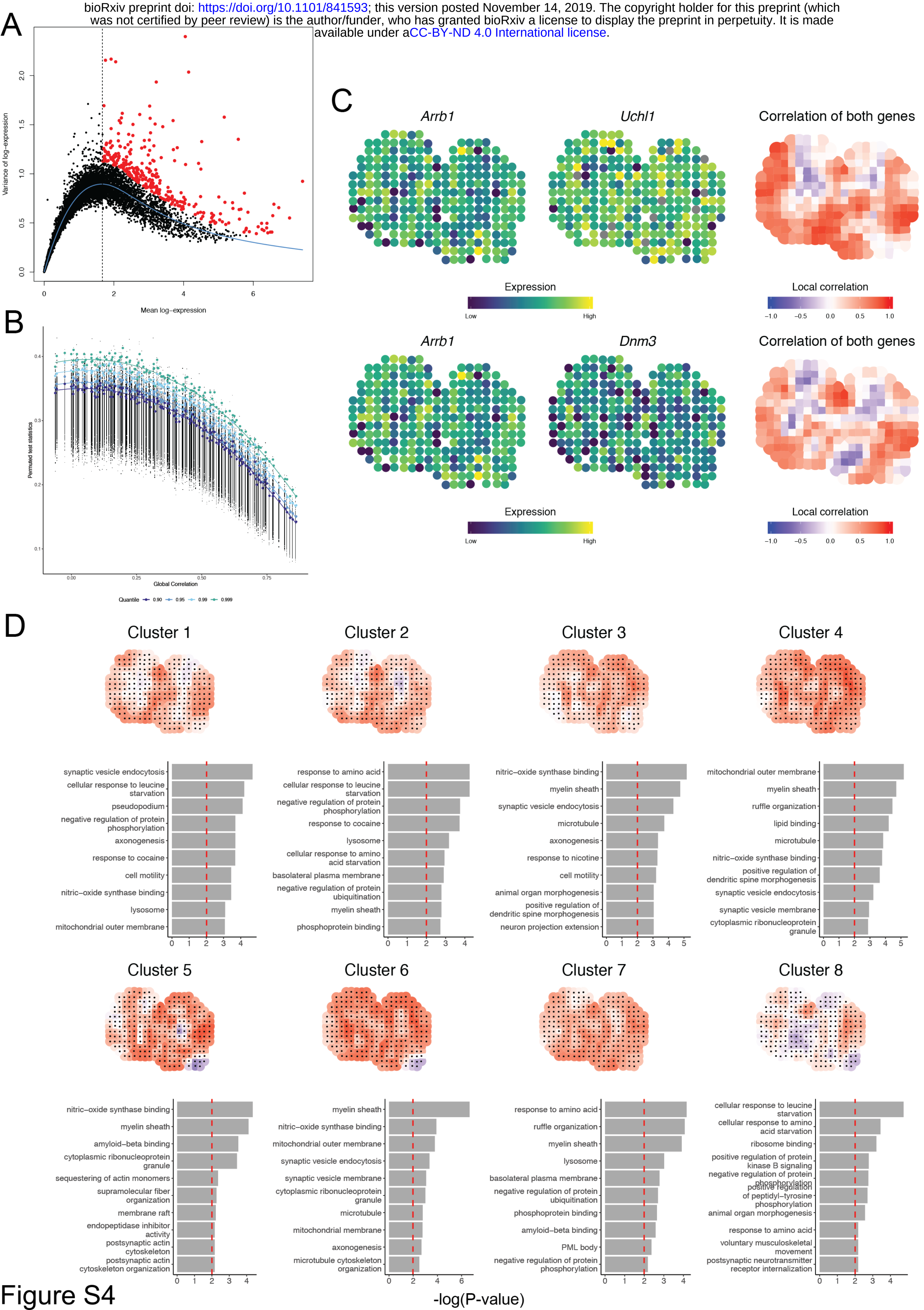

A
Variance of log−expression
Mean log−expression
0.0
0.5
1.0
1.5
2.0
0
2
4
6
B
Permuted test statistics
Global Correlation
0.1
0.2
0.3
0.4
0.00
0.25
0.50
0.75
Quantile 0.90 0.95 0.99 0.999
C
Arrb1
Uchl1
Correlation of both genes
Expression
Low
High
Local correlation
−1.0
−0.5
0.0
0.5
1.0
Arrb1
Dnm3
Correlation of both genes
Expression
Low
High
Local correlation
−1.0
−0.5
0.0
0.5
1.0
D
Cluster 1
synaptic vesicle endocytosis
cellular response to leucine starvation
pseudopodium
negative regulation of protein phosphorylation
axonogenesis
response to cocaine
cell motility
nitric−oxide synthase binding
lysosome
mitochondrial outer membrane
Cluster 2
response to amino acid
cellular response to leucine starvation
negative regulation of protein phosphorylation
response to cocaine
lysosome
cellular response to amino acid starvation
basolateral plasma membrane
negative regulation of protein ubiquitination
myelin sheath
phosphoprotein binding
Cluster 3
nitric−oxide synthase binding
myelin sheath
synaptic vesicle endocytosis
microtubule
axonogenesis
response to nicotine
cell motility
animal organ morphogenesis
positive regulation of dendritic spine morphogenesis
neuron projection extension
Cluster 4
mitochondrial outer membrane
myelin sheath
ruffle organization
lipid binding
microtubule
nitric−oxide synthase binding
positive regulation of dendritic spine morphogenesis
synaptic vesicle endocytosis
synaptic vesicle membrane
cytoplasmic ribonucleoprotein granule
Cluster 5
nitric−oxide synthase binding
myelin sheath
amyloid−beta binding
cytoplasmic ribonucleoprotein granule
sequestering of actin monomers
supramolecular fiber organization
membrane raft
endopeptidase inhibitor activity
postsynaptic actin cytoskeleton
postsynaptic actin cytoskeleton organization
Cluster 6
myelin sheath
nitric−oxide synthase binding
mitochondrial outer membrane
synaptic vesicle endocytosis
synaptic vesicle membrane
cytoplasmic ribonucleoprotein granule
microtubule
mitochondrial membrane
axonogenesis
microtubule cytoskeleton organization
Cluster 7
response to amino acid
ruffle organization
myelin sheath
lysosome
basolateral plasma membrane
negative regulation of protein ubiquitination
phosphoprotein binding
amyloid−beta binding
PML body
negative regulation of protein phosphorylation
Cluster 8
cellular response to leucine starvation
cellular response to amino acid starvation
ribosome binding
positive regulation of protein kinase B signaling
negative regulation of protein phosphorylation
positive regulation of peptidyl−tyrosine phosphorylation
animal organ morphogenesis
response to amino acid
voluntary musculoskeletal movement
postsynaptic neurotransmitter receptor internalization
-log(P-value)

Figure S4