Gene count normalization in single-cell imaging-based spatially resolved transcriptomics

Authors:

Lyla Atta[1,2], Kalen Clifton[1,2], Manjari Anant[2,3], Jean Fan[1,2]

Affiliations:

1. Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD 21218, USA

2. Center for Computational Biology, Whiting School of Engineering, Johns Hopkins University, Baltimore, MD 21211, USA

3. Department of Neuroscience, Johns Hopkins University, Baltimore, MD 21218, USA

Correspondence should be addressed to:

Jean Fan (jeanfan@jhu.edu)

The authors declare that they have no competing interests.

Abstract

Recent advances in imaging-based spatially resolved transcriptomics (im-SRT) technologies now enable high-throughput profiling of targeted genes and their locations in fixed tissues. Normalization of gene expression data is often needed to account for technical factors that may confound underlying biological signals. Here, we investigate the potential impact of different gene count normalization methods with different targeted gene panels in the analysis and interpretation of im-SRT data. Using different simulated gene panels that overrepresent genes expressed in specific tissue anatomical regions or cell types, we find that normalization methods that use scaling factors derived from gene counts differentially impact normalized gene expression magnitudes in a region- or cell type-specific manner. We show that these normalization-induced effects may reduce the reliability of downstream differential gene expression and fold change analysis, introducing false positive and false negative results when compared to results obtained from gene panels that are more representative of the gene expression of the tissue's component cell types. These effects are not observed without normalization or when scaling factors are not derived from gene counts, such as with cell volume normalization. Overall, we caution that the choice of normalization method and gene panel may impact the biological interpretation of the im-SRT data.

Keywords

Normalization, scaling factor, spatial transcriptomics, differential expression

Introduction

Imaging-based spatially resolved transcriptomics (im-SRT) represent a new set of technologies that enable the high-throughput detection of targeted RNA species in fixed tissues [1]–[3]. While im-SRT data is generally acquired at a single-molecule resolution, it is often aggregated into single-cell resolution by counting RNA molecules within cell regions identified by cell segmentation [4]–[7]. This achieves single-cell resolution spatially resolved gene expression profiling to enable the interrogation of cellular function and organization within tissues to enhance our understanding of cellular spatial microenvironments in biological processes such as tissue development, disease progression, and drug response [1], [2], [8].

In order to discover biological insights, normalization of gene expression data is often needed to account for technical differences that may confound underlying biological signals [4], [9], [10]. Common normalization methods for gene expression count data include scaling by a sample's total detected gene counts (also called size factor normalization, relative counts normalization, and library size normalization). This normalization approach is commonly used in the analysis of single-cell RNA-sequencing (scRNA-seq) data to account for technical differences in RNA capture rates between cells and has also been applied to im-SRT data [11]. Alternative normalization methods such as DESeq2 and TMM consider differences in gene expression composition between samples. These normalization approaches are commonly used in analysis of bulk RNA-seq data [9], [10]. Specific to im-SRT data, volume or cell area normalization has been previously applied to account for partially imaged cell volumes [12].

Given the capacities of current im-SRT technologies, a few hundred unique gene species can typically be measured simultaneously in a given experiment. Because gene detection probes are designed based on a target gene's sequence, these gene species must be pre-selected as part of a gene panel in the experimental design [2], [4]. These gene panels may be designed with the intention of spatially profiling cell-types in a tissue of interest and may thus include genes that are consistently expressed in and/or are canonical markers of cell types of interest based on domain knowledge or informed by scRNA-seq [12]–[15]. In recent years, commercial options for im-SRT have become available, where off-the-shelf gene panels with probes for a pre-selected set of genes can be used. These gene panels may focus on specific cell-types such as immune cells or neurons, or specific pathways and processes [16]. As such, these gene panels may be skewed to enable sensitive detection and characterization of the heterogeneity of cell types of interest, potentially at the expense of sensitivity to other cell types present in the tissue.

Here, we investigate how different gene count normalization procedures may impact downstream transcriptomic analyses for im-SRT data with different gene panels. Ideally, biological conclusions drawn from im-SRT studies in the same biological system would be robust to experimental design choices such as gene panel composition. For example, if a gene is differentially expressed in a particular region or cell type in a given biological system probed with one gene panel, it should also be differentially expressed given a different gene panel, if that the gene is in both panels. However, we show using im-SRT data with simulated skewed gene panels that different normalization methods may introduce false positive and false negative results when identifying significantly differentially expressed genes as well as opposite fold change directions (Figure 1A). We further demonstrate that these concerns generalize across im-SRT technologies that have

87 different technical features, such as gene panel size and capture efficiency. Finally, using simulated
88 gene panels of different sizes, we find that these concerns can be mitigated as gene panels increase
89 in size. Overall, we caution that the choice of normalization method and gene panel may impact
90 the biological interpretation of the im-SRT data.
91
92 **Results**
93
94 To investigate the potential impact of normalizing im-SRT data with different gene panels, we
95 simulated im-SRT datasets with different gene panels. We first focused on a multiplexed error-
96 robust fluorescent *in situ* hybridization (MERFISH) im-SRT dataset of a coronal section of the
97 mouse brain that originally profiled 483 genes representative of diverse cell-types (Supp Fig 1A)
98 [5], [17].
99
100 As a proof of principle, we created different gene panels that could allow us to better characterize
101 different mouse brain regions (Supp Fig 1B-E). Many brain regions are enriched for specific cell
102 types and in turn differentially express specific genes when compared to other brain regions. To
103 achieve this, we annotated mouse brain regions by aligning and transferring region annotations
104 from the Allen Brain Atlas using STalign [18], [19]. Then, for a brain region of interest, we
105 randomly sample 100 genes that are significantly overexpressed in this region, thereby simulating
106 a 100-gene gene panel skewed to our brain region of interest. For both the original full gene panel
107 and the region-skewed panel, we compared library size, DESeq2, TMM, and cell volume
108 normalization, in addition to no normalization.
109

Skewed gene panels with different normalization methods result in region-specific effects on scaling factors and normalized gene expression magnitudes in im-SRT data

To investigate the potential impact of normalization procedures on downstream transcriptomic analyses, we first compared the scaling factors derived from the full and skewed gene panels for each normalization method (Figure 1B). Given that these scaling factors are applied to detected cell gene counts to obtain the normalized gene expression magnitudes used in downstream analysis, we expect that differences in the scaling factors between gene panels will result in differences in the normalized gene expression magnitudes, which may in turn impact downstream analyses. For normalization methods where scaling factors are derived based on gene count information (library size, DESeq2, TMM), we further expect that the scaling factors will differ depending on the gene panel used.

Under library size normalization, we found scaling factors to generally be smaller with the skewed gene panel than with the full gene panel. This is expected given that the skewed gene panel includes fewer genes than the full gene panel. However, we note that the shift in scaling factors is less pronounced for some cells than others depending on brain region. Expectedly, given a gene panel skewed towards genes enriched in the ventricle region, cells in the ventricle region have scaling factors that are more consistent in magnitude compared to scaling factors derived from the full gene panel than cells outside the ventricle region. We likewise find that these results are consistent for gene panels designed for other brain regions including the habenula, fiber tracts, and dentate gyrus (Supp Fig 2). Similar to the ventricle region, these other brain regions are enriched for specific cell types and in turn differentially express specific genes when compared to other brain

regions. As such, differences in scaling factors for regions outside of those to which a gene panel is skewed are driven by the underlying differences in cell type composition and resulting gene expression differences. We also see this region-specific effect with DESeq2 and TMM normalizations. These results indicate that in im-SRT data with region-skewed gene panels, normalization using scaling factors derived based on gene counts may introduce region-specific impacts on downstream analyses depending on the normalization procedure used. Alternatively, as cell volume is independent of the number of gene counts detected in a cell, the scaling factor is the same with the skewed and full gene panels and therefore does not result in region-specific impacts on downstream analyses.

Next, we normalized gene counts using these various scaling factors. We evaluated how consistent normalized gene expression magnitudes are when using the skewed gene panel compared to normalized gene expression magnitudes with the full gene panel. To do this, we computed the correlation in normalized gene expression magnitudes between the skewed panel and the full gene panel (Figure 1C). Normalized gene expression magnitude correlations are the lowest for TMM normalization, ranging from around 0.5 to 0.98 for different genes. Library size normalization resulted in similarly low correlations between normalized gene expression magnitudes between skewed and full gene panels. DESeq2 resulted in comparatively higher correlations, all above 0.9. Volume normalization and no normalization resulted in perfectly correlated normalized gene expression magnitudes, as expected given that scaling factors are unchanged between different gene panels. We again find that these results are consistent for gene panels skewed towards other brain regions (Supp Fig 3).

To further understand the source of differences in normalized gene expression magnitude correlations between skewed and full gene panels with different normalization methods, we looked at the normalized gene expression magnitudes for select highly expressed genes (Figure 1D). Notably, given a gene panel skewed towards genes enriched in the ventricle region, under library normalization, normalized gene expression magnitudes are similar between the full and the skewed gene panels for cells in the ventricle region, as expected given similar scaling factors. However, for other regions, normalized gene expression magnitudes are higher with the skewed gene panel compared to the full gene panel. This is consistent with the smaller scaling factors for cells in these regions with the skewed gene panel compared to the full gene panel. As with the scaling factors, region specific differences between normalized counts in skewed and full gene panels are less pronounced with the DESeq2 and TMM normalizations than with library normalization. These examples show that the region-specific impacts on scaling factors when using region-skewed gene panels with count-based normalizations can in turn differentially impact gene expression magnitudes in a region-specific manner resulting in distortions in the relationship between normalized gene expression magnitudes across cells.

Skewed gene panels with different normalization methods lead to region-specific biases in differential gene expression results

Next, to evaluate the impact of these different normalization approaches with different gene panels on downstream analysis, we applied these scaling factors and performed differential expression analysis. To do this, we compared the p-value of a Wilcoxon rank sum test of differential expression as well as the gene fold change between cells in each brain region and all other regions

179 based on normalized gene expression magnitudes achieved with the skewed gene panel versus the
180 full gene panel for each normalization method. When using library size normalization, ventricle-
181 specific genes have less significant p-values with the ventricle-skewed gene panel compared to the
182 full gene panel. In contrast, other gene p-values are similar when using the ventricle-skewed gene
183 panel and the full gene panel. This region-specific difference is less pronounced with TMM and
184 DESeq2 normalization. Under no normalization and volume normalization, the p-values are equal
185 between the skewed gene panel and the full gene panel as the normalized gene expression
186 magnitudes are the same (Figure 2A).
187
188 Likewise, with respect to gene expression fold changes, under library normalization, ventricle-
189 specific genes have lower $\log_2$ fold changes when using the skewed gene panel compared to the
190 full gene panel with some genes having negative log fold changes in the skewed gene panel and
191 positive $\log_2$ fold changes in the full gene panel. Genes not in the ventricle region specific gene
192 panel have similar fold changes between the two panels. These region-specific effects are also seen
193 with TMM normalization but are less pronounced and are not seen with DESeq2 normalization
194 (Figure 2B). As before, under no normalization and volume normalization, the fold changes are
195 equal between the skewed gene panel and the full gene panel as the normalized counts are the
196 same.
197
198 To understand the source of differences in $\log_2$ fold changes between skewed and full gene panels
199 after library size normalization, we looked at the normalized gene expression magnitudes for
200 individual highly expressed genes (Figure 2C). We observe that for the gene *Gpr75*, library size
201 normalization with the full gene panel results in normalized gene expression magnitudes that are

202 between cells in the ventricle region and cells in other regions. In contrast, library size
203 normalization with the ventricle skewed gene panel results in normalized gene expression
204 magnitudes in cells in the ventricle region that are lower in cells in the ventricle region compared
205 to cells in other regions in the data. This results in a negative $\log_2$ fold change with the ventricle
206 skewed gene panel that was not observed with the full gene panel. This region-specific impact is
207 consistent with the differences in library size scaling factors resulting from the skewed gene panel,
208 where cells in the ventricle region have larger scaling factors than those in other regions because
209 of their higher expression of genes included in the gene panel. Conversely, we observe that for the
210 gene *Htr2c*, library size normalization with the full gene panel results in normalized gene
211 expression magnitudes that are lower in cells in the habenula region compared to cells in other
212 regions in the data. In contrast, library size normalization with the ventricle skewed gene panel
213 results in normalized gene expression magnitudes in cells in the habenula region that are more
214 similar in magnitude to those in cells in other regions. This results in a $\log_2$ fold change that is
215 positive with the ventricle skewed gene panel compared to a negative $\log_2$ fold change with the
216 full gene panel. These examples show that the region-specific impacts on scaling factors when
217 using region-skewed gene panels with library size normalization can result in distortions in
218 normalized gene expression distributions, which in turn lead to inconsistent fold change and
219 differential gene expression results with different gene panels. Again, results are consistent for
220 gene panels designed for other brain regions (Supp Fig 4-5). These results show that the region-
221 specific effects of gene count-based normalization in im-SRT with skewed gene panels may result
222 in false negatives in differential gene expression testing, as well as fold changes with opposite
223 directions.
224

To further investigate if these observations extend to im-SRT data acquired by other technologies, we performed the same analyses with another im-SRT dataset of a coronal section of the mouse brain assayed by STARmap PLUS [20] using three gene panels skewed towards ventricles, fiber tracts, and dentate gyrus respectively and observed similar results to those seen with the MERFISH data (Supp Fig 6-8). This suggests that region-specific differences in transcriptomics analysis results using skewed gene panels also applies to im-SRT data from other imaging-based spatial transcriptomics technologies.

Normalization-induced biases can be mitigated with more representative or larger gene panels

Next, we investigated if impacts on the reliability of differential gene expression testing and fold change analysis resulting from region-skewed gene panels can be mitigated with gene panels that are more representative of the gene expression of component regions in the tissue. To do this, we simulated a non-skewed gene panel of the same size as the region-specific gene panels used in the previous simulations and compared scaling factors, and results of differential expression testing and fold change analysis between the full 483-gene gene panel and non-skewed 100-gene gene panel after normalization with the four normalization methods (Supp Fig 9). We first computed scaling factors for each of the four normalization methods and compared to those computed with the full gene panel. As expected, given the smaller gene panel size, library size scaling factors were smaller with the non-skewed gene panel than with the full gene panel. In contrast to the skewed gene panel, the decrease in library size scaling factor affected all cells similarly and did not exhibit a region-specific pattern. Similar results were observed for DESeq2 and TMM. As before, scaling factors for volume normalization are independent of gene panel and therefore the

248 same between the non-skewed and full gene panels. We then performed differential expression
249 testing and found that both p-values and fold changes were well correlated between the non-
250 skewed and the full gene panels for all genes. These results indicate that region-specific impacts
251 on differential gene expression testing and fold change analysis resulting from count-based
252 normalization can be mitigated by designing gene panels that are more representative of the gene
253 expression of component regions and cell types in the tissue.
254
255 We anticipate that as im-SRT technologies improve, more gene species will become
256 simultaneously detectable in a single experiment, reducing the need to design smaller region- or
257 cell type-specific gene panels. We therefore sought to investigate the extent to which cell type-
258 specific biases in transcriptomics analysis results will be observed with larger gene panels. To
259 evaluate the impact of the size of the skewed gene panel on the extent of cell type or region-specific
260 effects observed, we used full-transcriptome scRNA-seq data collected from sorted peripheral
261 blood mononuclear cells [21] to simulate skewed gene panels of different sizes. Specifically, we
262 simulated gene panels of 50, 100, 500, 1000, 5000 genes overexpressed in monocytes in addition
263 to a panel including all genes that passed quality filters. We also simulated gene panels of the same
264 sizes consisting of randomly selected genes. When using library size normalization, cell type-
265 specific differences are observed in the skewed gene panels of all sizes. However, the extent of
266 these differences varies depending on the size of the panel. For the monocyte specific gene panel,
267 larger differences are observed with the smaller gene panels compared to the larger panels (Figure
268 3). Again, the random gene panels that are representative of all cell-types do not show these cell
269 type-specific differences in results of transcriptomic analyses after normalization with either
270 library size, TMM, or DESeq2. Overall, these results indicate that when normalization scaling

271 factors are derived based on gene counts, region- or cell type-specific impacts on downstream
272 analysis may persist with gene panels as large as 5000 genes, though the extent of these impacts
273 will depend on the normalization method used as well as the degree of region- or cell type-specific
274 gene expression overrepresentation in the gene panel.
275
276
277 **Discussion**
278
279 Here, we investigate the impact of different normalization methods with different gene panels in
280 the analysis of im-SRT data. In particular, we simulated gene panels that overrepresent the gene
281 expression of specific component tissue anatomical regions or cell types. Given im-SRT data
282 with these skewed gene panels, we find that normalization methods that use count scaling factors
283 derived from detected gene counts such as library size, DESeq2, or TMM differentially impact
284 normalized gene expression magnitudes in component regions or cell types. We show that these
285 region or cell type-specific effects may reduce the reliability of downstream differential gene
286 expression results when compared to results obtained from gene panels that are more
287 representative of the gene expression of the tissue's component cell types.
288
289 Notably, these gene count-based normalization methods were developed for use with full-
290 transcriptome RNA sequencing data and rely on assumptions that may be violated in targeted
291 transcriptome profiling such as in im-SRT data, particularly with skewed gene panels. For
292 example, library size normalization assumes that there are no genes that are only highly
293 expressed in one subset of samples; that is, all samples have similar total gene expression [9],

294 [10]. This assumption is inherently violated by designing gene panels to include canonical cell
295 type markers for cell types of interest. Other normalization methods that attempt to account for
296 such compositional differences in gene expression, such as those used in the DESeq2 and edgeR
297 packages, rely on other assumptions. For example, DESeq2 assumes that fewer than half of
298 genes in the data are differentially expressed between samples. Similarly, by choosing one
299 sample to be a reference, TMM also assumes that most measured genes are not differentially
300 expression between samples [10]. This is often not the case in im-SRT gene panels as they are
301 often selected by identifying differentially expressed genes from prior scRNAseq data. Although
302 our simulated gene panels were intentionally skewed for demonstration purposes, the approach
303 by which these genes were selected can be used to design gene panels for im-SRT experiments,
304 which would result in unintentionally skewed gene panels.
305
306 Despite this, count-based normalization can still provide robust insights when gene panels are
307 well-designed to be representative of most of the component cell types in the tissue being
308 assayed. To this end, we recommend the use of non-spatially resolved single cell resolution, full
309 transcriptome profiling approaches such as scRNA-seq data to inform gene panel design. This
310 can be achieved by generating scRNA-seq and im-SRT datasets from the same tissue [12].
311 Furthermore, as collective atlasing efforts such as the Human BioMolecular Atlas Program
312 (HuBMAP) [22], the Human Cell Atlas [23], and others continue to generate tissue-specific gene
313 expression data, we anticipate these resources may help provide a more untargeted view of gene
314 expression across cell types in tissues of interest and enable researchers to design im-SRT gene
315 panels that are representative of most of tissues' component cell types.
316

317 In general, count normalization methods have been developed to account for variation across
318 cells due to technical factors. One of these technical factors that has motivated the development
319 of previous methods is variation in RNA capture rate in scRNA-seq. In im-SRT, we expect RNA
320 capture rate to be similar across cells in the same experiment, provided that the entire cell
321 volume is captured in the imaged region. For cells that lie at the boundaries of the imaged
322 regions, not all the cell volume will be captured resulting in a smaller gene count that is reflective
323 of technical factors rather than biological ones. If gene expression differences due to partial
324 volume capture compose a large proportion of total gene expression variation, count
325 normalization within individual im-SRT experiments may be needed to allow for reliable
326 downstream analysis. Volume or cell area normalization may be used to account for such partial
327 cell capture within imaged regions. Our results show that these normalization approaches may be
328 preferred when working with skewed gene panels as scaling factors are independent of gene
329 counts and as such gene panel differences do not impact downstream differential expression and
330 fold change analyses. However, the reliability of volume or cell area normalization is contingent
331 upon the accuracy of cell segmentation. Additional investigation may be needed to evaluate the
332 impact of cell segmentation accuracy in the analysis of normalized gene expression when
333 volume or cell area normalization is used.
334
335 Ultimately, normalization of im-SRT data should enable the removal of the effects of systematic
336 technical variation in detected gene counts. Accomplishing this goal without introducing
337 additional biases will help ensure that normalized gene expression magnitudes reflect underlying
338 biological heterogeneity to enable reliable and reproducible interpretation of im-SRT data in a
339 way that allows investigators to extract meaningful biological insights.

Methods

Datasets

Vizgen MERFISH Mouse Brain Receptor Map data was downloaded from the Vizgen and data from slice 2 replicate 2 was used [17]. Brain anatomical region annotations for each cell were obtained by transferring region annotations from the Allen Brain Atlas Common Coordinate Framework (CCF) using STalign [18], [19]. STalign was applied using the 3D reconstructed Nissl image from the Allen CCF atlas as a source and the MERFISH cell position data as a target. Anatomical regions were combined to obtain coarse anatomical annotations as specified in Supplementary Table 1. Cells in the ventricle, fiber tract, dentate gyrus, and habenula regions were used for further analysis.

STARmap PLUS data downloaded from the Broad Single Cell Portal and data from well 11 (well11raw_expression_pd.csv.gz) was used [24]. Brain anatomical region annotations were obtained from previous manually created annotations (well11_spatial.csv.gz) [25]. Anatomical regions were combined to obtain coarse anatomical annotations. Fiber_tracts_A, Fiber_tracts_C, and Fiber_tracts_D were combined to form the fiber tracts region; DG_A and DG_B formed the dentate gyrus region; VS_A and VS_B formed the ventricle region. Cells in the ventricle, fiber tract, and dentate gyrus regions were used for further analysis.

Single-cell RNA sequencing data was obtained from the 10X Genomics website [21]. Specifically data from the following datasets was downloaded and combined: human CD14+ monocytes,

363 CD19+ B-cells, CD8+ cytotoxic T-cells, CD4+ helper T-cells, CD4+/CD45RO+ memory T-cells,
364 CD4+/CD25+ Regulatory T-cells, CD8+/CD45RA+ Naive Cytotoxic T-cells,
365 CD4+/CD45RA+/CD25- naive T-cells, and CD56+ natural killer cells.
366
367 *Scaling factors and gene expression normalization*
368
369 Library size scaling factors were computed as the sum of all gene counts in a cell. DESeq2 scaling
370 factors were computed using the estimateSizeFactorsForMatrix function from the DESeq2
371 package (version 1.38.0), with method=poscounts. TMM scaling factors were computed as the
372 inverse of the output of the calcNormFactors function from the edgeR package (version 3.40.0),
373 method=TMMwsp. For the MERFISH data, cell volume scaling factors were included in the
374 downloaded data. To obtain normalized gene expression, gene counts for each gene in each cell
375 were divided by that cell's scaling factor.
376
377 *Differential gene expression testing and fold change analysis*
378
379 To test for differential gene expression, for each gene in each region, a one-sided Wilcoxon rank
380 sum test was performed between cells in that region and all other cells in the data. Significantly
381 differentially expressed genes were identified using a p-value cutoff of 0.05 after a Benjamini-
382 Hochberg multiple hypothesis correction. For each gene in each region, $\log_2$ fold change (*LFC*)
383 was computed as follows:
384

$$LFC_{A,g} = log_2 \frac{\frac{1}{N_A} \sum_i^{N_A} x_{g,i}}{\frac{1}{N_{A'}} \sum_j^{N_{A'}} x_{g,j}}$$

where $LFC_{A,g}$ is the log_2 fold change for gene g between cells in region A and all other cells in the data, $x_{g,i}$ is the normalized gene expression of gene g in cell i, N_A is the number of cells in region A, $N_{A'}$ and is the number of cells in data in regions other than in region A.

Simulation of skewed gene panels in im-SRT data

Differential gene expression testing was performed on the normalized counts using the original, full gene panel for each normalization method as well as no normalization. After differential gene expression testing, significantly differentially expressed genes for a region were identified as those genes with a p-value <0.05 after multiple hypothesis correction and an absolute log_2 fold change >0.25 under any normalization procedure. The skewed gene panel was simulated by randomly selecting 100 genes from those identified as significantly differentially expressed. The random gene panel was simulated by randomly selecting 100 genes from all the genes in the data.

Simulation of skewed gene panels in scRNA-seq data

Differential gene expression testing was performed on the normalized counts using the original, full gene panel for each normalization method as well as no normalization. Genes were ranked according to their multiple hypothesis corrected p-value. Skewed gene panels were selected by choosing the most significantly differentially expressed genes.

Acknowledgements

Research reported in this publication was supported by the National Institute of General Medical Sciences of the National Institutes of Health under Award Number R35-GM142889. Figure 1A was created using images from BioRender.com.

References

[1] M. Asp, J. Bergenstråhle, and J. Lundeberg, "Spatially Resolved Transcriptomes—Next Generation Tools for Tissue Exploration," *BioEssays*, vol. 42, no. 10, p. 1900221, 2020, doi: 10.1002/bies.201900221.

[2] M. Zhang *et al*., "Spatial molecular profiling: platforms, applications and analysis tools," *Briefings in Bioinformatics*, vol. 22, no. 3, p. bbaa145, May 2021, doi: 10.1093/bib/bbaa145.

[3] X. Zhuang, "Spatially resolved single-cell genomics and transcriptomics by imaging," *Nat Methods*, vol. 18, no. 1, Art. no. 1, Jan. 2021, doi: 10.1038/s41592-020-01037-8.

[4] L. Heumos *et al*., "Best practices for single-cell analysis across modalities," *Nat Rev Genet*, vol. 24, no. 8, Art. no. 8, Aug. 2023, doi: 10.1038/s41576-023-00586-w.

[5] K. H. Chen, A. N. Boettiger, J. R. Moffitt, S. Wang, and X. Zhuang, "Spatially resolved, highly multiplexed RNA profiling in single cells," *Science*, vol. 348, no. 6233, Apr. 2015, doi: 10.1126/science.aaa6090.

[6] E. Lubeck, A. F. Coskun, T. Zhiyentayev, M. Ahmad, and L. Cai, "Single-cell in situ RNA profiling by sequential hybridization," *Nat Methods*, vol. 11, no. 4, Art. no. 4, Apr. 2014, doi: 10.1038/nmeth.2892.

[7] X. Wang *et al*., "Three-dimensional intact-tissue sequencing of single-cell transcriptional states," *Science*, vol. 361, no. 6400, Jul. 2018, doi: 10.1126/science.aat5691.

[8] R. Dries, J. Chen, N. del Rossi, M. M. Khan, A. Sistig, and G.-C. Yuan, "Advances in spatial transcriptomic data analysis," *Genome Res*., vol. 31, no. 10, pp. 1706–1718, Oct. 2021, doi: 10.1101/gr.275224.121.

436 [9] C. Evans, J. Hardin, and D. M. Stoebel, "Selecting between-sample RNA-Seq
437 normalization methods from the perspective of their assumptions," *Briefings in*
438 *Bioinformatics*, vol. 19, no. 5, pp. 776–792, Sep. 2018, doi: 10.1093/bib/bbx008.
439 [10] M.-A. Dillies *et al.*, "A comprehensive evaluation of normalization methods for illumina
440 high-throughput rna sequencing data analysis," *Briefings in Bioinformatics*, vol. 14, no. 6,
441 pp. 671–683, Nov. 2013, doi: 10.1093/bib/bbs046.
442 [11] C. A. Vallejos, D. Risso, A. Scialdone, S. Dudoit, and J. C. Marioni, "Normalizing single-
443 cell RNA sequencing data: challenges and opportunities," *Nat Methods*, vol. 14, no. 6, Art.
444 no. 6, Jun. 2017, doi: 10.1038/nmeth.4292.
445 [12] J. R. Moffitt *et al.*, "Molecular, spatial, and functional single-cell profiling of the
446 hypothalamic preoptic region," *Science*, vol. 362, no. 6416, Nov. 2018, doi:
447 10.1126/science.aau5324.
448 [13] L. B. Kuemmerle *et al.*, "Probe set selection for targeted spatial transcriptomics." bioRxiv,
449 p. 2022.08.16.504115, Aug. 17, 2022. doi: 10.1101/2022.08.16.504115.
450 [14] M. Zhang *et al.*, "Spatially resolved cell atlas of the mouse primary motor cortex by
451 MERFISH," *Nature*, vol. 598, no. 7879, Art. no. 7879, Oct. 2021, doi: 10.1038/s41586-
452 021-03705-x.
453 [15] R. Chen *et al.*, "Decoding molecular and cellular heterogeneity of mouse nucleus
454 accumbens," *Nat Neurosci*, vol. 24, no. 12, Art. no. 12, Dec. 2021, doi: 10.1038/s41593-
455 021-00938-x.
456 [16] "Pre-designed Xenium Gene Expression Panels - Official 10x Genomics Support," *10x*
457 *Genomics*. https://www.10xgenomics.com/support/in-situ-gene-

expression/documentation/steps/panel-design/pre-designed-xenium-gene-expression-panels (accessed Jul. 25, 2023).

[17] "Vizgen MERFISH Mouse Receptor Map." https://info.vizgen.com/mouse-brain-map (accessed Jul. 24, 2023).

[18] K. Clifton *et al*., "Alignment of spatial transcriptomics data using diffeomorphic metric mapping." bioRxiv, p. 2023.04.11.534630, Apr. 12, 2023. doi: 10.1101/2023.04.11.534630.

[19] Q. Wang *et al*., "The Allen Mouse Brain Common Coordinate Framework: A 3D Reference Atlas," *Cell*, vol. 181, no. 4, pp. 936-953.e20, May 2020, doi: 10.1016/j.cell.2020.04.007.

[20] H. Zeng *et al*., "Integrative in situ mapping of single-cell transcriptional states and tissue histopathology in a mouse model of Alzheimer's disease," *Nat Neurosci*, vol. 26, no. 3, Art. no. 3, Mar. 2023, doi: 10.1038/s41593-022-01251-x.

[21] "Datasets," *10x Genomics*. https://www.10xgenomics.com/resources/datasets (accessed Jul. 29, 2021).

[22] S. Jain *et al*., "Advances and prospects for the Human BioMolecular Atlas Program (HuBMAP)," *Nat Cell Biol*, pp. 1–12, Jul. 2023, doi: 10.1038/s41556-023-01194-w.

[23] A. Regev *et al*., "The Human Cell Atlas," *eLife*, vol. 6, p. e27041, Dec. 2017, doi: 10.7554/eLife.27041.

[24] "Spatial Atlas of Molecular Cell Types and AAV Accessibility across the Whole Mouse Brain - Single Cell Portal." https://singlecell.broadinstitute.org/single_cell/study/SCP1830/spatial-atlas-of-molecular-cell-types-and-aav-accessibility-across-the-whole-mouse-brain (accessed Jul. 26, 2023).

[25] H. Shi *et al*., “Spatial Atlas of the Mouse Central Nervous System at Molecular Resolution.” bioRxiv, p. 2022.06.20.496914, Jun. 22, 2022. doi: 10.1101/2022.06.20.496914.

Main Figures

Figure 1. Impact of region-skewed gene panel on normalization scaling factors and normalized gene expression magnitudes

A. Schematic of im-SRT normalization evaluation workflow.

(1) Im-SRT data is collected, cells are segmented to derive single cell gene counts, and region annotations are derived via manual labeling or label transfer.

(2) This yields a cell-by-gene matrix and cell region meta data. Different gene panels can be used to investigate tissue samples from the same biological systems. Ideally, biological conclusions would be robust to gene panel choice.

(3) Gene count normalization approaches include ones where cell-specific gene count scaling factors are derived based on gene counts such as in library size, DESeq2, and TMM normalization or independently from gene counts such as in cell volume or cell area normalization. Normalized gene expression is derived by dividing detected gene counts by the cell-specific scaling factors. Downstream analysis includes differential gene expression testing and fold change analysis between tissue regions to identify genes with potential biological importance and results can be impacted by the choice of gene panel and normalization method.

B. For a coronal section of the mouse brain assayed by MERFISH, scatterplots of $\log_{10}$(scaling factors) for each cell based on different normalization methods (library size normalization, DESeq2 normalization, TMM normalization, and cell volume normalization) for the full gene panel versus the ventricle-skewed gene panel. Cells within the ventricle brain region are shown in purple. Cells in the fiber tract, habenula, and dentate gyrus brain regions are in grey. Red line indicates x=y.

C. Boxplot of Pearson correlation coefficients (*r*) across genes for normalized gene expression magnitudes with the full gene panel versus the ventricle region-skewed gene panel across different normalization methods.

D. Scatterplot of normalized gene expression magnitudes when normalization was performed with the full gene panel versus the ventricle skewed gene panel for two highly expressed genes, *Htr2c* and *Gpr75*. Cells within the ventricle brain region are shown in purple. Cells in the fiber tract, habenula, and dentate gyrus regions are in grey. Red line indicates x=y.

Figure 2. Impact of region-skewed gene panel on differential expression analysis and fold change evaluation

A. Wilcoxon rank sum test p-values for one region-vs-all differential gene expression test with normalized gene expression from the full gene panel versus the ventricle-skewed gene panel. Red line indicates x=y.

B. One region-vs-all gene fold change with normalized gene expression from the full gene panel and the ventricle gene panel. Black lines indicate 0 $\log_2$ fold change and grey lines indicate 0.5 $\log_2$ fold change.

C. Library size normalized gene expression magnitudes in cells in region of interest vs cells in other regions with the full gene panel and the ventricle skewed gene panel for two representative highly expressed genes: *Htr2c* (left) and *Gpr75* (right).

In all panels, purple indicates ventricle region-specific genes as determined by a one-region-vs-all differential gene expression test using the full gene panel. Grey indicates habenula, fiber tract, or dentate gyrus region-specific genes, as determined by a one-region-vs-all differential gene expression test using the full gene panel. Red line indicates x=y.

530

531 **Figure 3. Library size normalization and downstream analysis with increasing gene panel**

532 **size**

533 **A.** Scaling factors with monocyte skewed gene panels of different sizes and full gene panel.

534 Green indicates monocyte cells and grey indicates B cells, T cells, or NK cells.

535 **B.** Wilcoxon rank sum test p-values adjusted for multiple hypothesis testing for one region-vs-all

536 differential gene expression tests with monocyte skewed gene panels of different sizes and full

537 gene panel.

538 **C.** One region-vs-all gene $\log_2$ fold change with monocyte skewed gene panels of different sizes

539 and full gene panel.

540 In B and C, green indicates monocyte specific genes in monocyte-vs-all tests and grey indicates

541 B cells, T cell-, or NK cell- specific genes in B cell-, T cell-, or NK cell-vs-all tests.

542

543

544

545 **Supplementary Figures**

546

547 **Supplementary Figure 1. Gene expression of brain anatomical regions in full and simulated**

548 **MERFISH gene panel**

549 **A.** Heatmap showing scaled mean gene expression of 481 genes in full MERFISH panel in the

550 four analyzed brain regions.

551 **B.** Heatmap showing scaled mean gene expression of the 100 genes in 100-gene gene panel

552 skewed towards ventricle brain region for the four analyzed brain regions.

553 **C.** Heatmap showing scaled mean gene expression of the 100 genes in 100-gene gene panel

554 skewed towards habenula brain region for the four analyzed brain regions.

555 **D.** Heatmap showing scaled mean gene expression of the 100 genes in 100-gene gene panel

556 skewed towards fiber tract brain regions for the four analyzed brain regions.

557 **E.** Heatmap showing scaled mean gene expression of the 100 genes in 100-gene gene panel

558 skewed towards dentate gyrus brain regions for the four analyzed brain regions.

559 **F.** Heatmap showing scaled mean gene expression of the 100 genes in 1100-gene non-skewed

560 gene panel for the four analyzed brain regions.

561 In all panels, purple, green, blue, and red indicate the ventricles, habenula, fiber tracts, and

562 dentate gyrus regions, respectively.

563

564 **Supplementary Figure 2: Scaling factors with MERFISH gene panels skewed towards the**

565 **habenula, fiber tract, and dentate gyrus regions**

566 **A.** For a coronal section of the mouse brain assayed by MERFISH, scatterplots of $\log_{10}$(scaling

567 factors) for each cell based on different normalization methods (library size normalization,

568 DESeq2 normalization, TMM normalization, and cell volume normalization) for the full gene
569 panel versus the habenula region-skewed gene panel. Cells within the habenula brain region are
570 shown in green. Cells in the fiber tract, ventricle, and dentate gyrus brain regions are shown in
571 grey. Red line indicates x=y.
572 **B.** Scatterplots of $\log_{10}$(scaling factors) for each cell based on different normalization methods
573 (library size normalization, DESeq2 normalization, TMM normalization, and cell volume
574 normalization) for the full gene panel versus the fiber tract region-skewed gene panel. Cells
575 within the fiber tract brain regions are shown in blue. Cells in the habenula, ventricle, and dentate
576 gyrus brain regions are shown in grey. Red line indicates x=y.
577 **C.** Scatterplots of $\log_{10}$(scaling factors) for each cell based on different normalization methods
578 (library size normalization, DESeq2 normalization, TMM normalization, and cell volume
579 normalization) for the full gene panel versus the dentate gyrus region-skewed gene panel. Cells
580 within the dentate gyrus brain region are shown in red. Cells in the habenula, ventricle, and fiber
581 tract brain regions are shown in grey. Red line indicates x=y.
582
583 **Supplementary Figure 3: Normalized gene expression correlations between full and region-**
584 **skewed MERFISH gene panels**
585 **A.** Boxplot of Pearson correlation coefficients (*r*) across genes for normalized gene expression
586 magnitudes with the full gene panel versus the habenula region-skewed gene panel across
587 different normalization methods evaluated.
588 **B.** Boxplot of Pearson correlation coefficients (*r*) across genes for normalized gene expression
589 magnitudes with the full gene panel versus the fiber tract region-skewed gene panel across
590 different normalization methods evaluated.

591 **C.** Boxplot of Pearson correlation coefficients (*r*) across genes for normalized gene expression
592 magnitudes with the full gene panel versus the dentate gyrus region-skewed gene panel across
593 different normalization methods evaluated.
594
595 **Supplementary Figure 4. Differential gene expression p-values with full and region-skewed**
596 **MERFISH gene panels**
597 **A.** Wilcoxon rank sum test p-values adjusted for multiple hypothesis testing for one region-vs-all
598 differential gene expression test with normalized gene expression from the full gene panel versus
599 the habenula region-skewed gene panel. Green indicates habenula region-specific genes, as
600 determined by a one-region-vs-all differential gene expression test using the full gene panel.
601 Grey indicates ventricle, fiber tract, or dentate gyrus region-specific genes, as determined by a
602 one-region-vs-all differential gene expression test using the full gene panel. Red line indicates
603 x=y.
604 **B.** Wilcoxon rank sum test p-values adjusted for multiple hypothesis testing for one region-vs-all
605 differential gene expression test with normalized gene expression from the full gene panel versus
606 the fiber tract region-skewed gene panel. Blue indicates fiber tract region-specific genes, as
607 determined by a one-region-vs-all differential gene expression test using the full gene panel.
608 Grey indicates ventricle, habenula, or dentate gyrus region-specific genes, as determined by a
609 one-region-vs-all differential gene expression test using the full gene panel. Red line indicates
610 x=y.
611 **C.** Wilcoxon rank sum test p-values adjusted for multiple hypothesis testing for one region-vs-all
612 differential gene expression test with normalized gene expression from the full gene panel versus
613 the dentate gyrus region-skewed gene panel. Red indicates dentate gyrus region-specific genes,

614 as determined by a one-region-vs-all differential gene expression test using the full gene panel.
615 Grey indicates ventricle, fiber tract, or habenula region-specific genes, as determined by a one-
616 region-vs-all differential gene expression test using the full gene panel. Red line indicates x=y.
617 In all panels, green, blue, and red indicate habenula, fiber tract, and dentate gyrus-specific genes,
618 respectively.
619
620 **Supplementary Figure 5. Gene $\log_2$ fold change with full and region-skewed MERFISH**
621 **gene panels**
622 **A.** One region-vs-all gene fold change with normalized gene expression from the full gene panel
623 versus the habenula region-skewed gene panel. Green indicates habenula region-specific genes,
624 as determined by a one-region-vs-all differential gene expression test using the full gene panel.
625 Grey indicates ventricle, fiber tract, or dentate gyrus region-specific genes, as determined by a
626 one-region-vs-all differential gene expression test using the full gene panel.
627 **B.** One region-vs-all gene fold change with normalized gene expression from the full gene panel
628 versus the fiber tract region-skewed gene panel. Blue indicates fiber tract region-specific genes,
629 as determined by a one-region-vs-all differential gene expression test using the full gene panel.
630 Grey indicates ventricle, habenula, or dentate gyrus region-specific genes, as determined by a
631 one-region-vs-all differential gene expression test using the full gene panel.
632 **C.** One region-vs-all gene fold change with normalized gene expression from the full gene panel
633 versus the dentate gyrus region-skewed gene panel. Red indicates dentate gyrus region-specific
634 genes, as determined by a one-region-vs-all differential gene expression test using the full gene
635 panel. Grey indicates ventricle, fiber tract, or habenula region-specific genes, as determined by a
636 one-region-vs-all differential gene expression test using the full gene panel.

637 In all panels, black lines indicate 0 $\log_2$ fold change, grey lines indicate 0.5 $\log_2$ fold change, red
638 line indicates x=y.
639
640 **Supplementary Figure 6. Scaling factors with full and region-skewed STARmap gene**
641 **panels**
642 **A.** For a coronal section of the mouse brain assayed by STARmap, scatterplots of $\log_{10}$(scaling
643 factors) for each cell based on different normalization methods (library size normalization,
644 DESeq2 normalization, and TMM normalization) for the full gene panel versus the ventricle
645 region-skewed gene panel. Cells within the ventricle brain region are shown in purple. Cells in
646 the fiber tract and dentate gyrus brain regions are shown in grey.
647 **B.** Scatterplots of $\log_{10}$(scaling factors) for each cell based on different normalization methods
648 (library size normalization, DESeq2 normalization, and TMM normalization) for the full gene
649 panel versus the fiber tract region-skewed gene panel. Cells within the fiber tract brain regions
650 are shown in blue. Cells in the ventricle and dentate gyrus brain regions are shown in grey.
651 **C.** Scatterplots of $\log_{10}$(scaling factors) for each cell based on different normalization methods
652 (library size normalization, DESeq2 normalization, and TMM normalization) for the full gene
653 panel versus the dentate gyrus region-skewed gene panel. Cells within the dentate gyrus brain
654 region are shown in red. Cells in the ventricle and fiber tract brain regions are shown in grey.
655 In all panels, red line indicates x=y.
656
657 **Supplementary Figure 7. Differential gene expression p-values with full and region-skewed**
658 **STARmap gene panels**

659 **A.** Wilcoxon rank sum test p-values adjusted for multiple hypothesis testing for one region-vs-all
660 differential gene expression test with normalized gene expression from the full gene panel versus
661 the ventricle region-skewed gene panel. Purple indicates ventricle region-specific genes, as
662 determined by a one-region-vs-all differential gene expression test using the full gene panel.
663 Grey indicates fiber tract or dentate gyrus region-specific genes, as determined by a one-region-
664 vs-all differential gene expression test using the full gene panel.
665 **B.** Wilcoxon rank sum test p-values adjusted for multiple hypothesis testing for one region-vs-all
666 differential gene expression test with normalized gene expression from the full gene panel versus
667 the fiber tract region-skewed gene panel. Blue indicates fiber tract region-specific genes, as
668 determined by a one-region-vs-all differential gene expression test using the full gene panel.
669 Grey indicates ventricle or dentate gyrus region-specific genes, as determined by a one-region-
670 vs-all differential gene expression test using the full gene panel.
671 **C.** Wilcoxon rank sum test p-values adjusted for multiple hypothesis testing for one region-vs-all
672 differential gene expression test with normalized gene expression from the full gene panel versus
673 the dentate gyrus region-skewed gene panel. Red indicates dentate gyrus region-specific genes,
674 as determined by a one-region-vs-all differential gene expression test using the full gene panel.
675 Grey indicates ventricle or fiber tract region-specific genes, as determined by a one-region-vs-all
676 differential gene expression test using the full gene panel.
677 In all panels, red line indicates x=y.
678
679 **Supplementary Figure 8. Gene $\log_2$ fold change with full and region-skewed STARmap**
680 **gene panels**

681 **A.** One region-vs-all gene $\log_2$ fold change with normalized gene expression from the full gene
682 panel and the ventricle region-skewed gene panel. Purple indicates ventricle region-specific
683 genes, as determined by a one-region-vs-all differential gene expression test using the full gene
684 panel. Grey indicates fiber tract or dentate gyrus region-specific genes, as determined by a one-
685 region-vs-all differential gene expression test using the full gene panel.
686 **B.** One region-vs-all gene $\log_2$ fold change with normalized gene expression from the full gene
687 panel versus the habenula region-skewed gene panel. Blue indicates fiber tract region-specific
688 genes, as determined by a one-region-vs-all differential gene expression test using the full gene
689 panel. Grey indicates ventricle or dentate gyrus region-specific genes, as determined by a one-
690 region-vs-all differential gene expression test using the full gene panel.
691 **C.** One region-vs-all gene $\log_2$ fold change with normalized gene expression from the full gene
692 panel versus the dentate gyrus region-skewed gene panel. Red indicates dentate gyrus region-
693 specific genes, as determined by a one-region-vs-all differential gene expression test using the
694 full gene panel. Grey indicates ventricle or fiber tract region-specific genes, as determined by a
695 one-region-vs-all differential gene expression test using the full gene panel.
696 In all panels, black lines indicate 0 $\log_2$ fold change, grey lines indicate 0.5 $\log_2$ fold change, red
697 line indicates x=y.
698
699 **Supplementary Figure 9. Non-skewed MERFISH gene panel**
700 **A.** For a coronal section of the mouse brain assayed by MERFISH, scatterplots of $\log_{10}$(scaling
701 factors) for each cell based on different normalization methods (library size normalization,
702 DESeq2 normalization, TMM normalization, and cell volume normalization) for the full gene
703 panel versus the non-skewed 100-gene gene panel. Cells within the ventricle, habenula, fiber

704 tract, and dentate gyrus brain regions are shown in purple, green, blue, and red, respectively. Red
705 line is x=y.
706 **B.** Pearson correlation coefficient (*r*) distribution across genes for normalized gene expression
707 magnitudes from the full gene panel versus the non-skewed gene panel across different
708 normalization methods evaluated.
709 **C.** Wilcoxon rank sum test p-values adjusted for multiple hypothesis testing for one region-vs-all
710 differential gene expression test with normalized gene expression from the full gene panel versus
711 the non-skewed 100-gene gene panel. Purple, green, blue, and red indicate ventricle, habenula,
712 fiber tract, and dentate gyrus region-specific genes, respectively, as determined by a one-region-
713 vs-all differential gene expression test using the full gene panel. Red line indicates x=y.
714 **D.** One region-vs-all gene fold change with normalized gene expression from the full gene panel
715 versus the non-skewed 100-gene gene panel. Purple, green, blue, and red indicate ventricle,
716 habenula, fiber tract, and dentate gyrus region-specific genes, respectively, as determined by a
717 one-region-vs-all differential gene expression test using the full gene panel. Black lines indicate
718 0 $\log_2$ fold change, grey lines indicate 0.5 $\log_2$ fold change, red line indicates x=y.

Figure 1

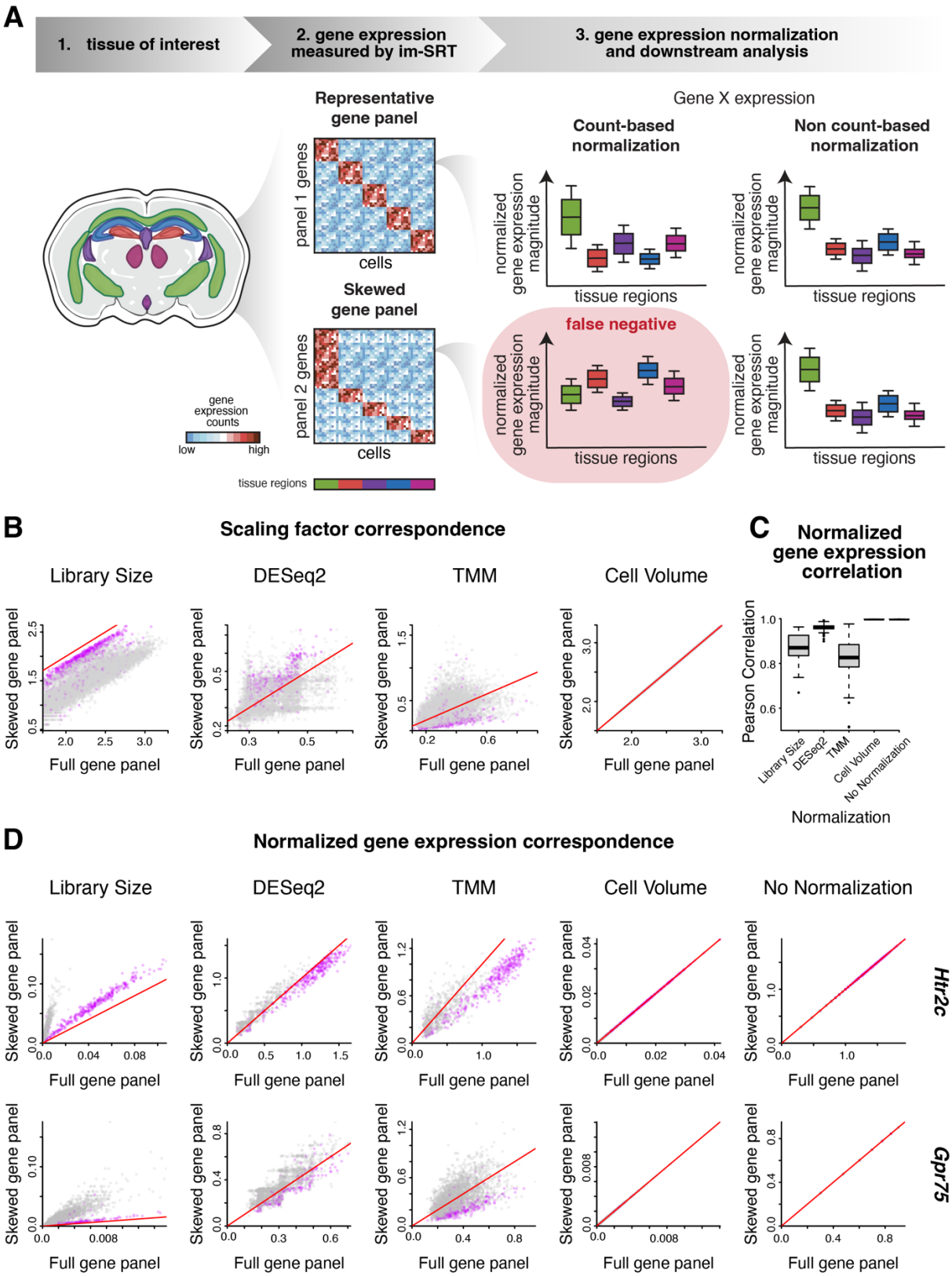

A
1. tissue of interest
2. gene expression measured by im-SRT
3. gene expression normalization and downstream analysis
Representative gene panel
panel 1 genes
cells
Skewed gene panel
panel 2 genes
cells
gene expression counts
low
high
tissue regions
Gene X expression
Count-based normalization
Non count-based normalization
normalized gene expression magnitude
tissue regions
false negative
B
Scaling factor correspondence
Library Size
DESeq2
TMM
Cell Volume
Skewed gene panel
Full gene panel
C
Normalized gene expression correlation
Pearson Correlation
Library Size
DESeq2
TMM
Cell Volume
No Normalization
Normalization
D
Normalized gene expression correspondence
Library Size
DESeq2
TMM
Cell Volume
No Normalization
Htr2c
Gpr75
Skewed gene panel
Full gene panel

Figure 2

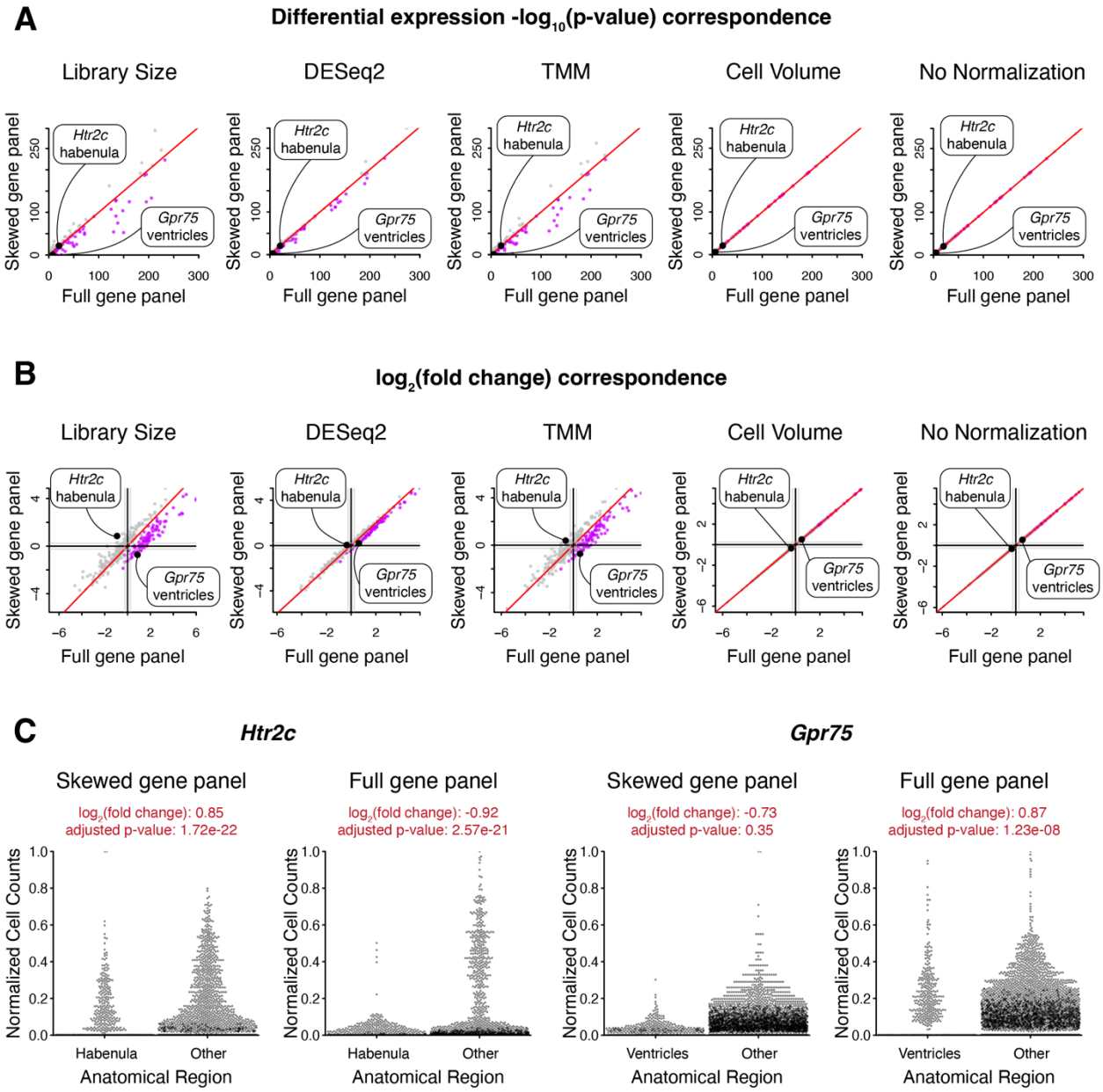
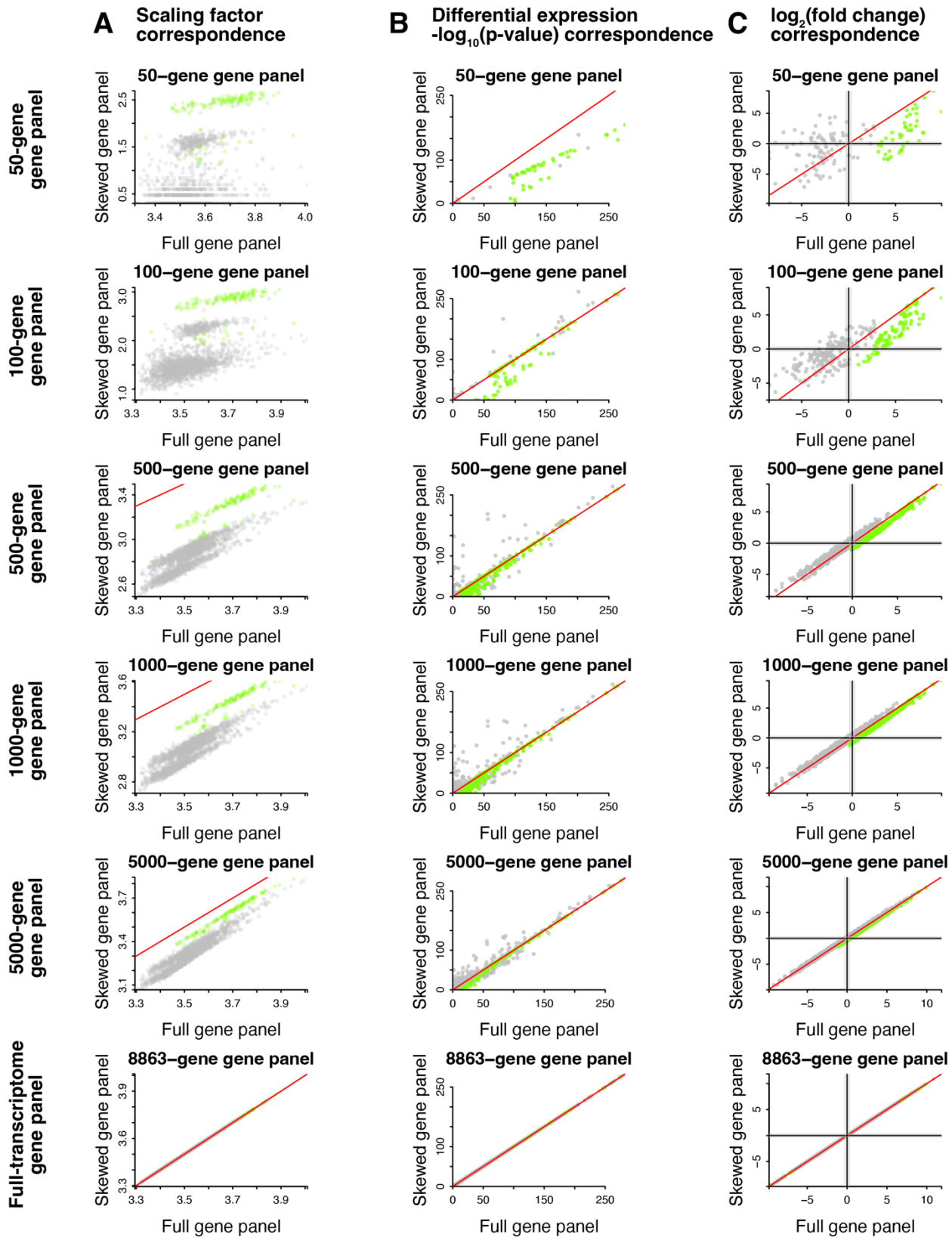
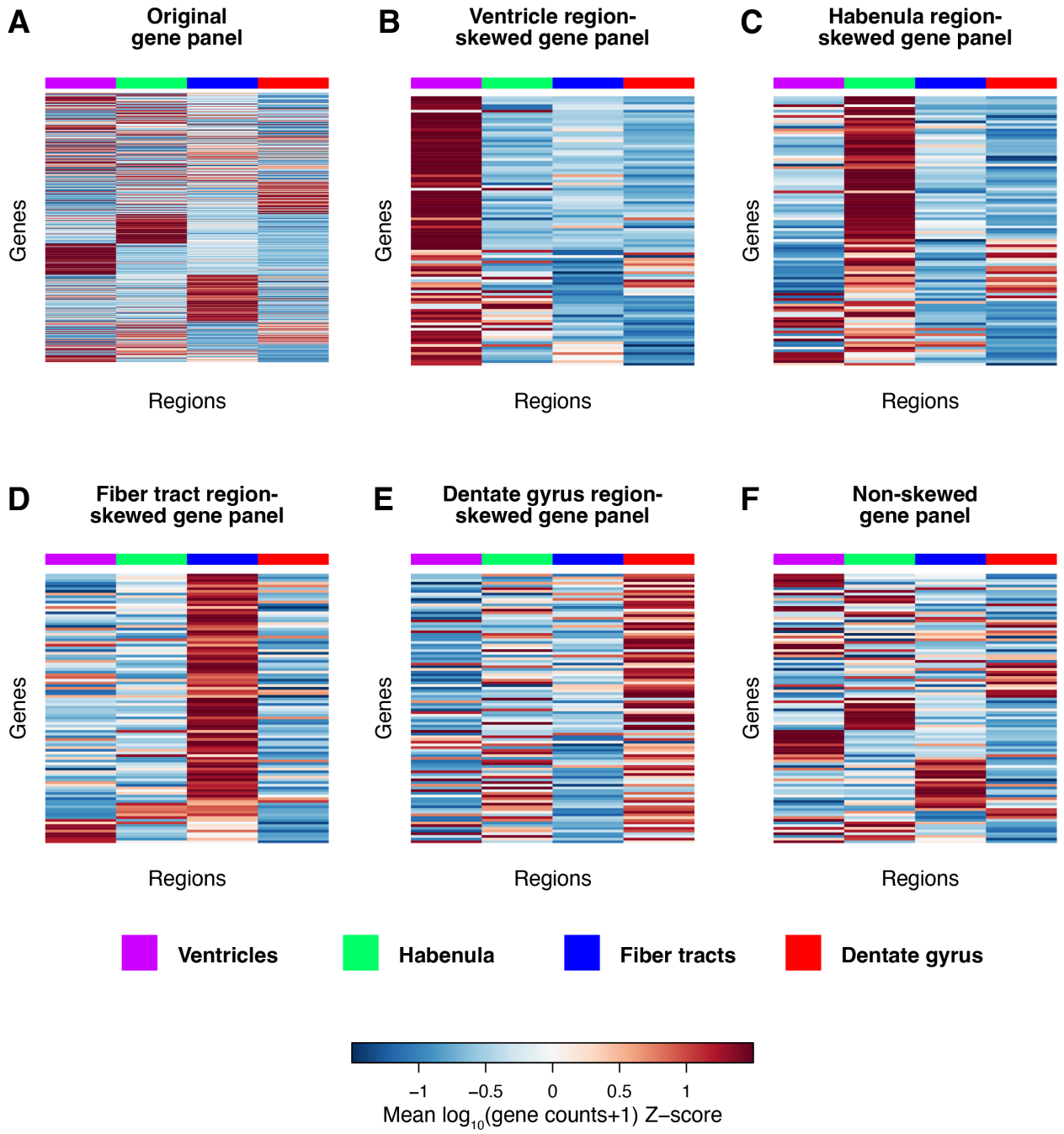

A
Differential expression -log10(p-value) correspondence
Library Size
DESeq2
TMM
Cell Volume
No Normalization
Htr2c habenula
Gpr75 ventricles
Skewed gene panel
Full gene panel
B
log2(fold change) correspondence
C
Htr2c
Gpr75
Skewed gene panel
Full gene panel
log2(fold change): 0.85
adjusted p-value: 1.72e-22
log2(fold change): -0.92
adjusted p-value: 2.57e-21
log2(fold change): -0.73
adjusted p-value: 0.35
log2(fold change): 0.87
adjusted p-value: 1.23e-08
Normalized Cell Counts
Habenula
Other
Ventricles
Anatomical Region

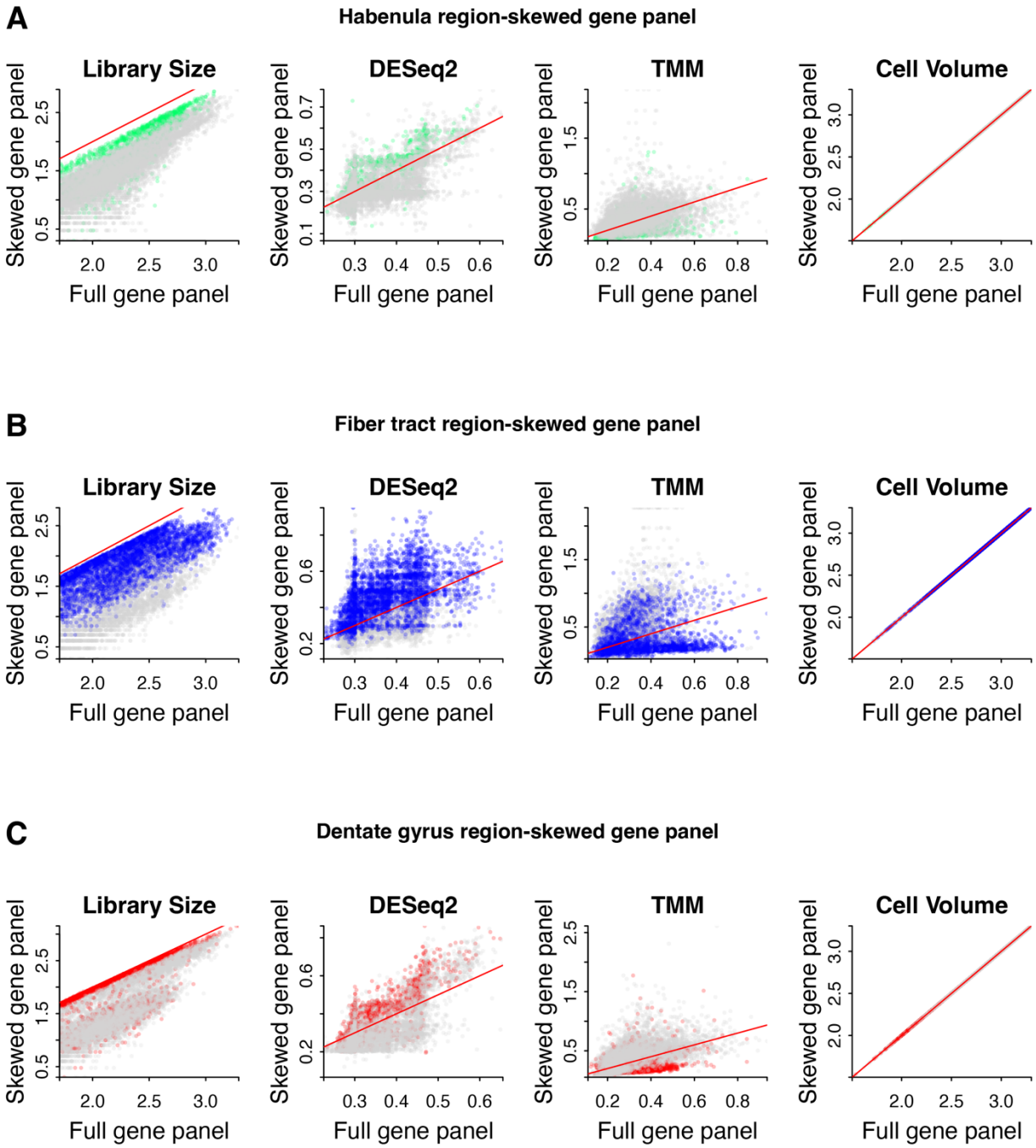
Figure 3


A
Scaling factor correspondence
B
Differential expression -log10(p-value) correspondence
C
log2(fold change) correspondence
50-gene gene panel
100-gene gene panel
500-gene gene panel
1000-gene gene panel
5000-gene gene panel
Full-transcriptome gene panel
50–gene gene panel
100–gene gene panel
500–gene gene panel
1000–gene gene panel
5000–gene gene panel
8863–gene gene panel
Skewed gene panel
Full gene panel

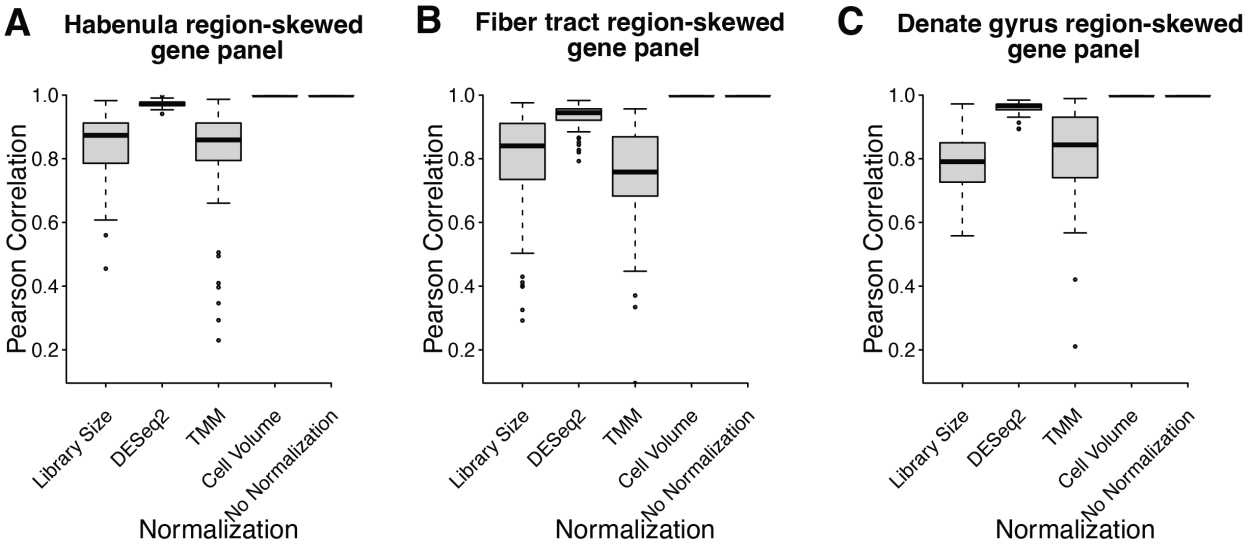
Supplementary Figure 1



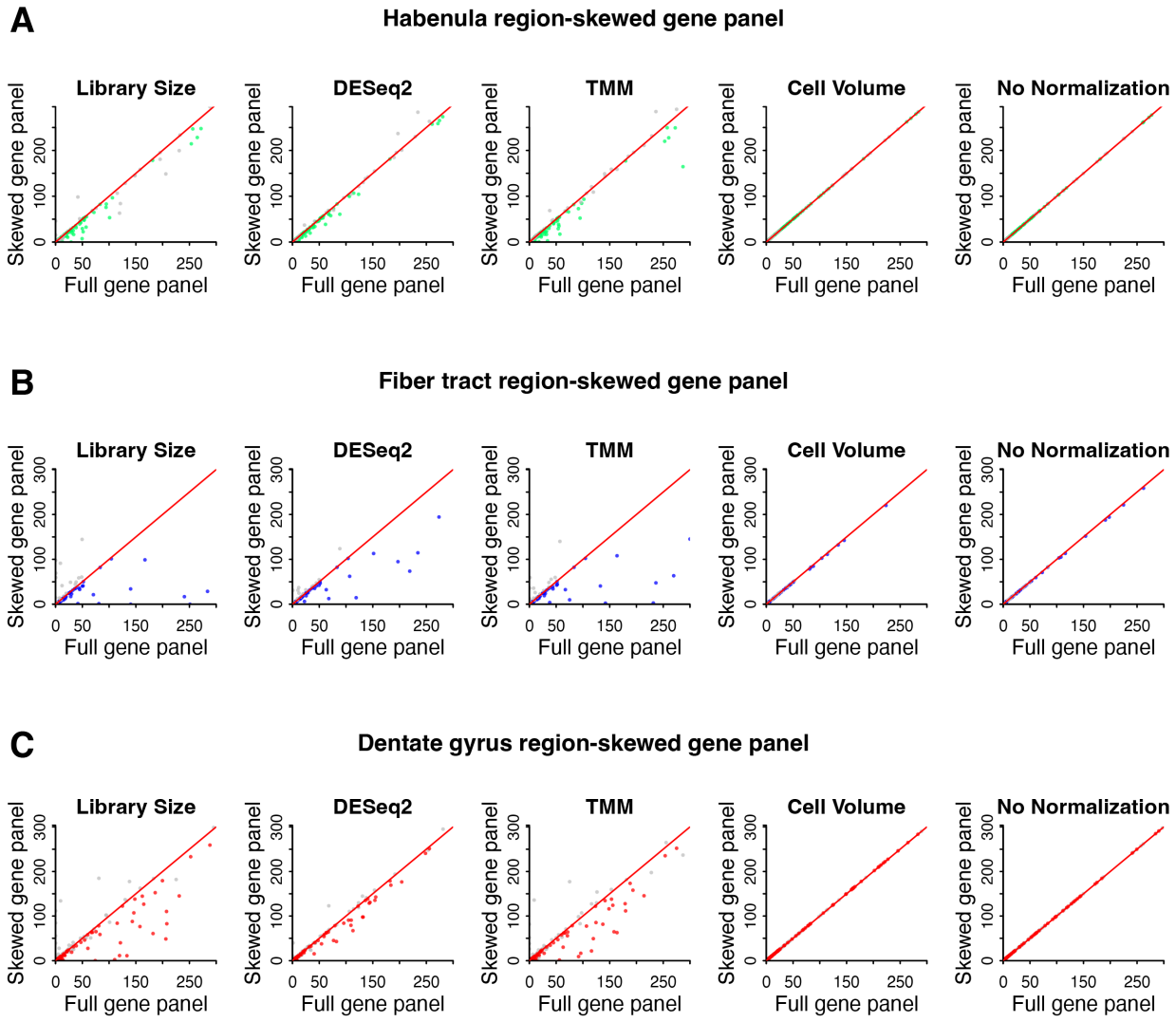
Supplementary Figure 2



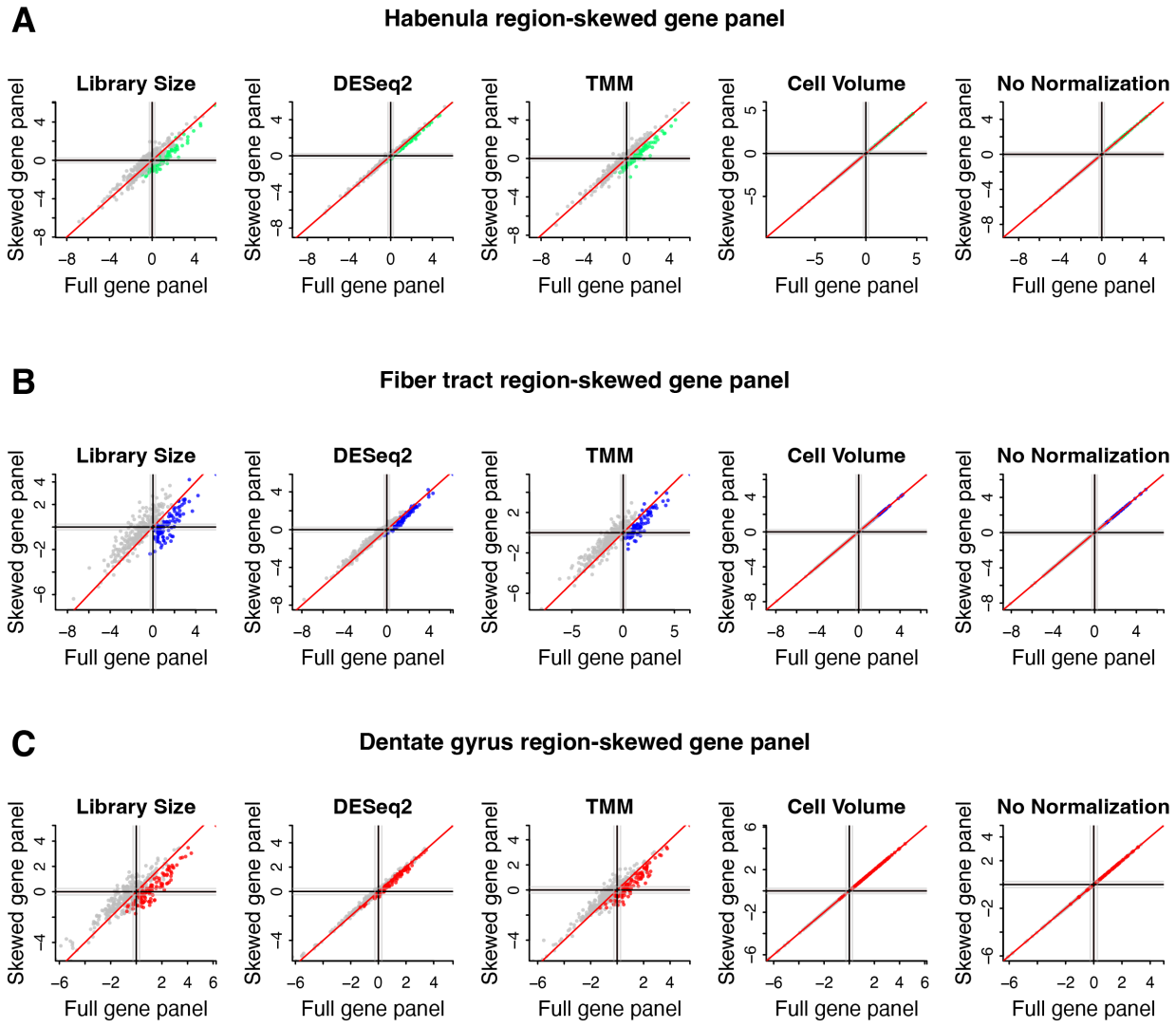
Supplementary Figure 3



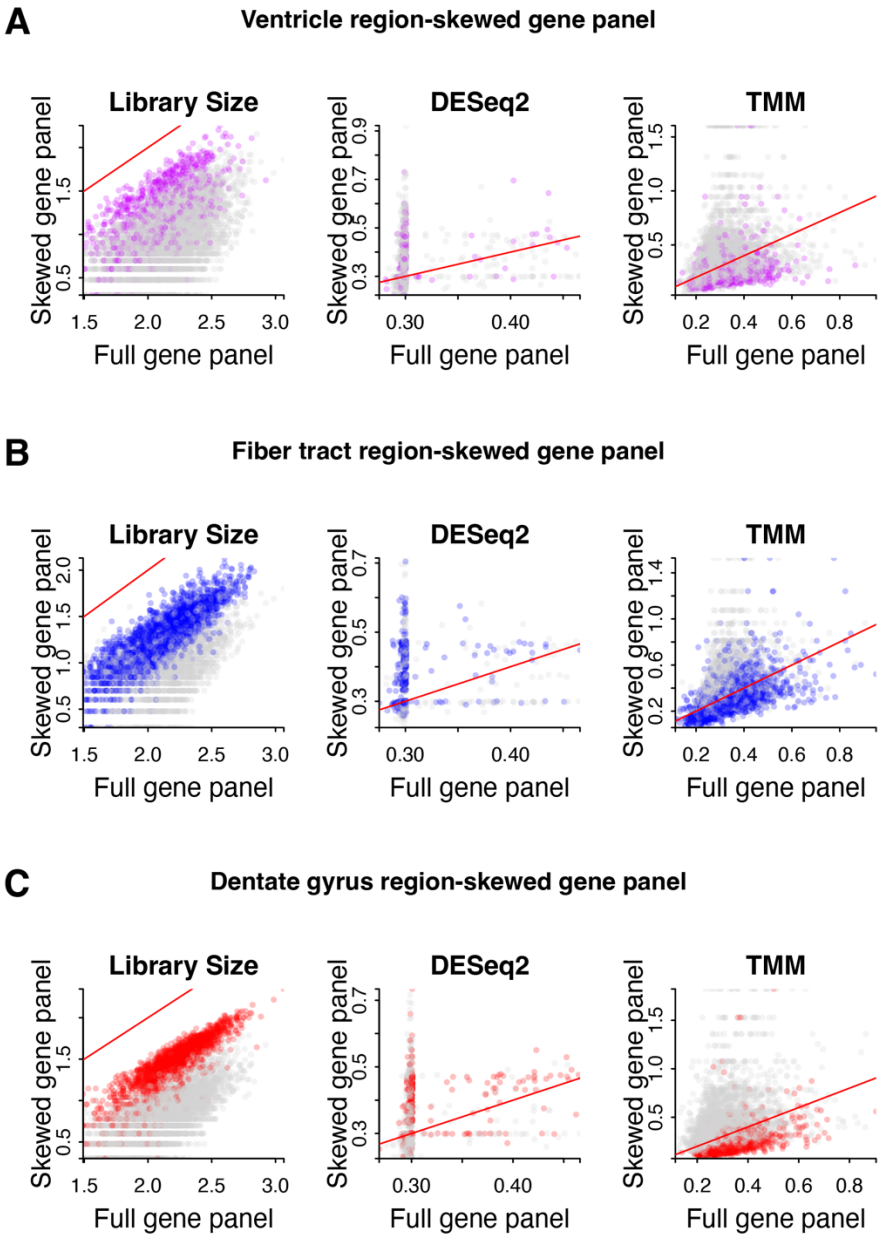
Supplementary Figure 4



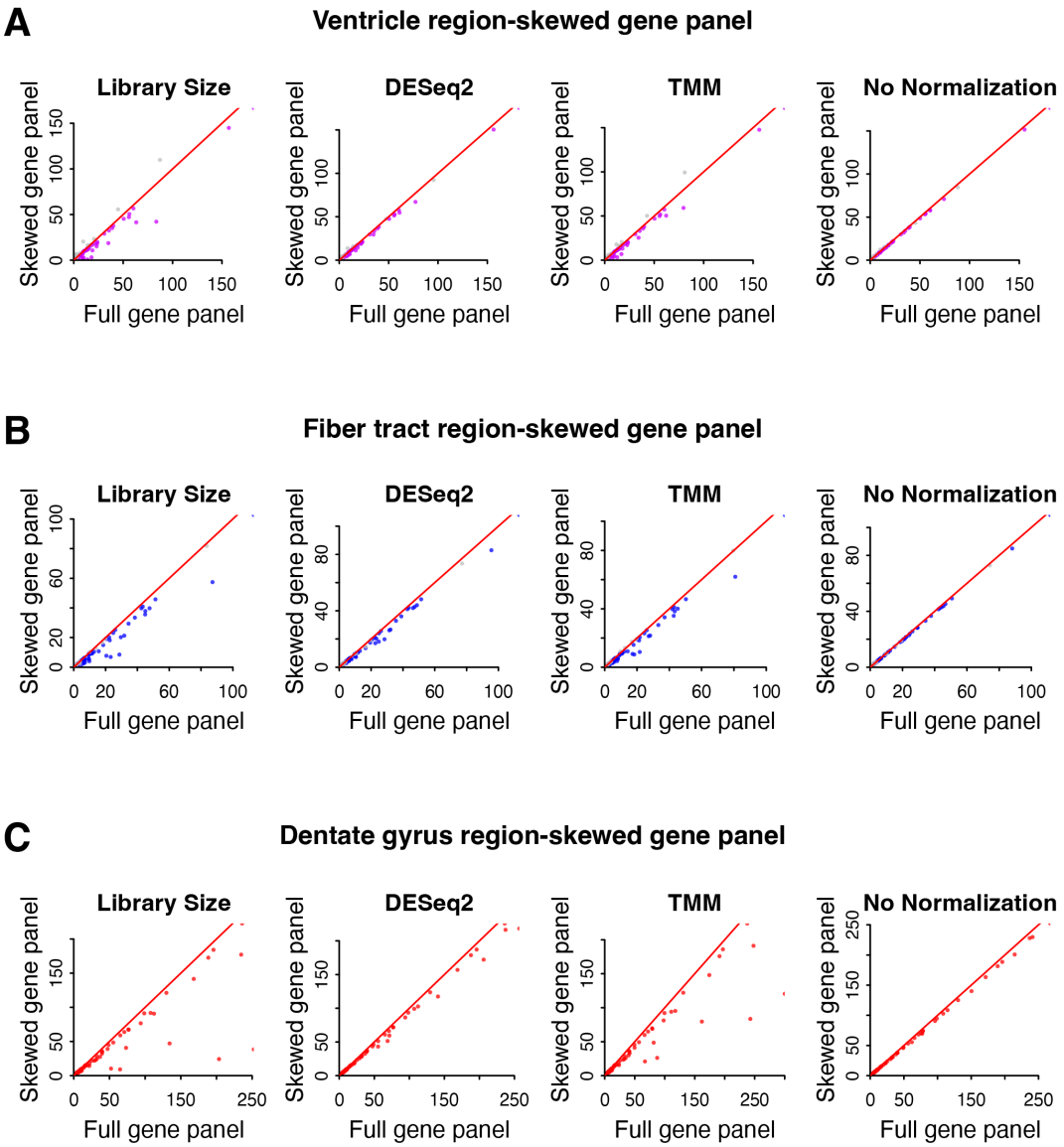
Supplementary Figure 5



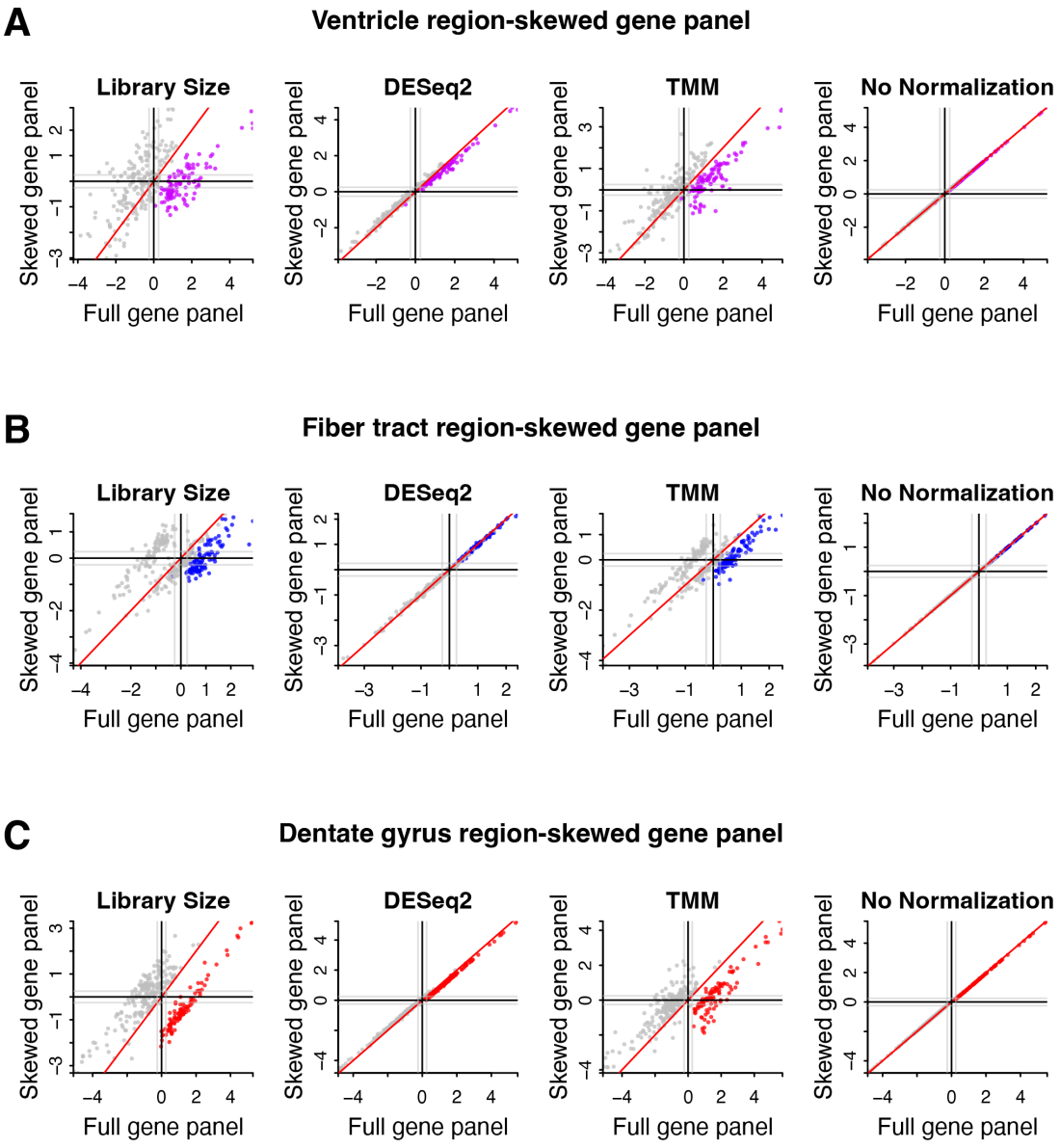
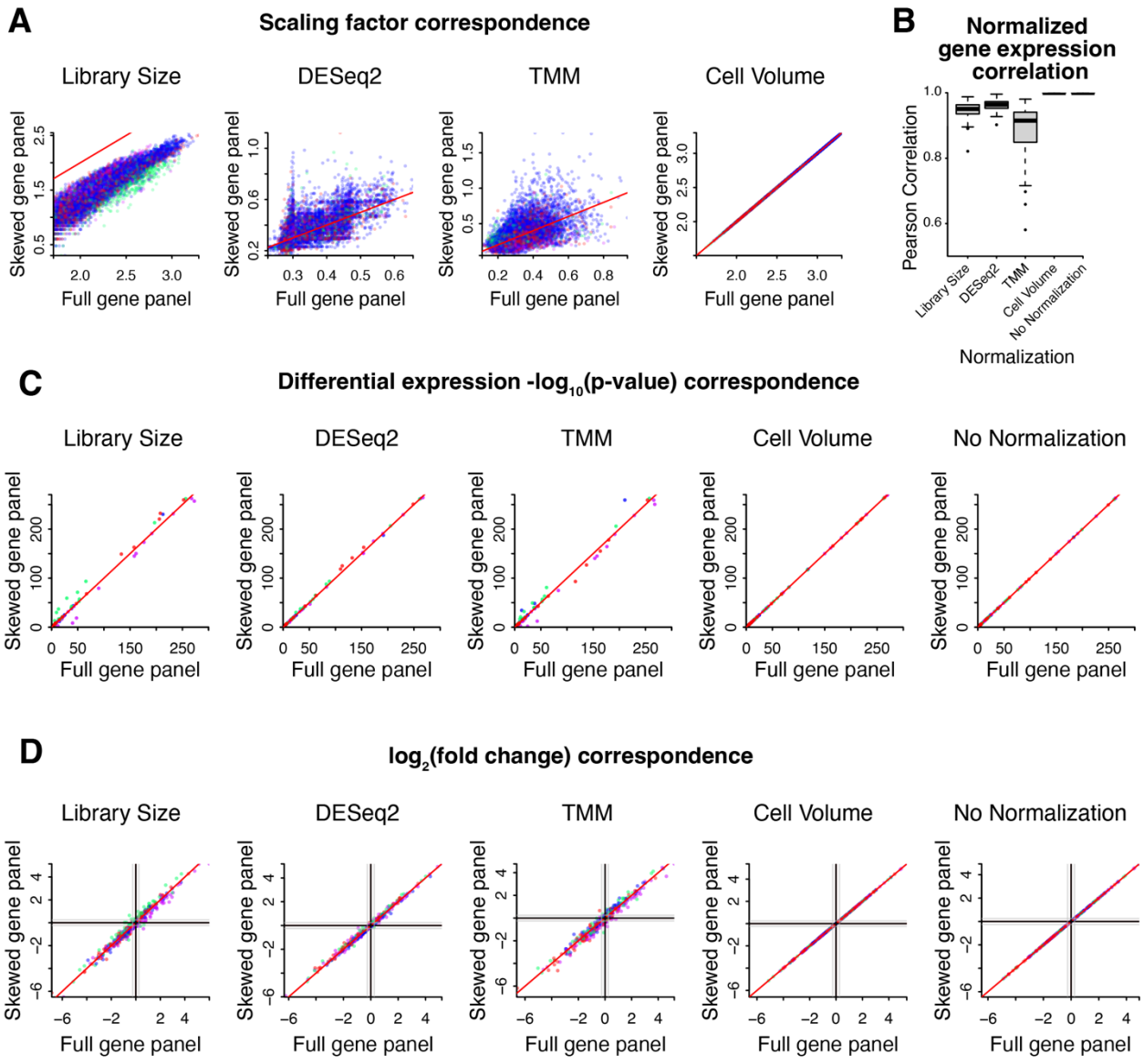
Supplementary Figure 6



Supplementary Figure 7



Supplementary Figure 8



Supplementary Figure 9