OPEN ACCESS
Journal of Bioinformatics and Computational Biology
Vol. 22, No. 3 (2024) 2450007 (14 pages)

DOI: 10.1142/S0219720024500070

Gene representation bias in spatial transcriptomics

Xinling Li* and Peng Qiu†
The Wallace H. Coulter Department of Biomedical Engineering
Georgia Institute of Technology and Emory University
Atlanta, GA, United States
**xli776@gatech.edu*
†*peng.qiu@bme.gatech.edu*

Received 13 February 2024
Revised 29 March 2024
Accepted 6 April 2024
Published 20 July 2024

For sequencing-based spatial transcriptomics data, the gene-spot count matrix is highly sparse. This feature is similar to scRNA-seq. The goal of this paper is to identify whether there exist genes that are frequently under-detected in Visium compared to bulk RNA-seq, and the underlying potential mechanism of under-detection in Visium. We collected paired Visium and bulk RNA-seq data for 28 human samples and 19 mouse samples, which covered diverse tissue sources. We compared the two data types and observed that there indeed exists a collection of genes frequently under-detected in Visium compared to bulk RNA-seq. We performed a motif search to examine the last 350 bp of the frequently under-detected genes, and we observed that the poly (T) motif was significantly enriched in genes identified from both human and mouse data, which matches with our previous finding about frequently under-detected genes in scRNA-seq. We hypothesized that the poly (T) motif may be able to form a hairpin structure with the poly (A) tails of their mRNA transcripts, making it difficult for their mRNA transcripts to be captured during Visium library preparation.

Keywords: Spatial transcriptomics; bulk RNA-seq; sparsity.

1. Introduction

Spatially resolved transcriptomic methods produce transcriptomic data for individual spatial spots and locations of the spots, which enables researchers to study the spatial contexts of transcriptional profiles of cells.[1–3] There are two main categories of spatial transcriptomic technologies: Imaging-based technologies and sequencing-based technologies.[4–6] Imaging-based technologies include sequential FISH (seqFISH),[7] multiplexed error-robust fluorescence in situ hybridization (MERFISH),[8] single-molecule

† Corresponding author.

fluorescent in situ hybridization (smFISH),[9] and spatially resolved transcript amplicon readout mapping (STARmap),[10] which target 100 s of genes with single-cell and subcellular resolution. Sequencing-based technologies include Visium and Slide-seq,[11] which target the entire transcriptome with spatial resolution close to but slightly lower than single-cell resolution. Similar to scRNA-seq technologies, spatial transcriptomics have limited capture efficiency. For in-situ capturing technology, the earlier spatial transcriptomics technology's detection efficiency is as low as 6.9%, and Visium only has slightly improved efficiency.[3] It is important to achieve high capture efficiency because difficulty in capturing mRNA transcripts during Visium library preparation negatively impacts the quality of the data. In addition, spatial transcriptomics data often contain substantial amounts of zeros in their expression matrices. While some of the zeros represent true biological signals, some of them may be caused by technical issues.

Many computational tools originally developed for scRNA-seq have been extended to spatial transcriptomics. For example, Seurat[12] and Scanpy[13] can analyze both data generated by 10X Chromium and data generated by Visium. To address the sparsity in spatial transcriptomics data, multiple imputation algorithms have been developed to improve the data quality of spatial transcriptomics. Examples include FIST,[14] Tangram[15] and Sprod.[16] Meanwhile, there have been discussions suggesting that considering zero inflation is not necessary for spatial transcriptomics.[17] Therefore, there is no consensus about the best way to handle a high proportion of zeros in the gene-spot count data in spatial transcriptomics.

Among the widespread discussion of sparsity in spatial transcriptomics data, in one of the studies, it was found that the number of zeros in the gene-spot count matrix increases at higher resolution (fewer number of cells in each spot).[16] For example, the sparsity of Visium count data matrices is lower than scRNA-seq count data, as each spot of Visium sequencing technology usually contains multiple cells.[16] Slide-seq has a much higher spatial resolution (very close to single-cell resolution), and a higher level of sparsity compared to scRNA-seq, possibly due to its relatively lower per-cell sequencing depth.[16] For both Visium and Slide-seq, the percentages of zero counts increased with lower average gene expression levels.[16] In another study, it was found that MERFISH has systematically lower sparsity than scRNA-seq.[18]

In a recent study,[19] we collected paired bulk RNA-seq and scRNA-seq data for 53 samples from various biological contexts, identified frequently under-detected genes in scRNA-seq compared to bulk RNA-seq, and observed that the frequently under-detected genes in scRNA-seq have significantly enriched poly (T) motif toward the tail of the genes. We hypothesized that the poly (T) motif may be able to form a hairpin structure with the poly (A) tail of the transcripts, making them difficult to capture during scRNA-seq library preparation, which is a mechanistic conjecture of why certain genes may be consistently under-detected in scRNA-seq. Given the similarity between scRNA-seq and sequencing-based spatial transcriptomics, in this study, we collected paired bulk RNA-seq and Visium data for 28 samples from diverse human tissues and 19 samples from diverse mouse tissues and examined whether there exist genes that are consistently under-detected in Visium compared

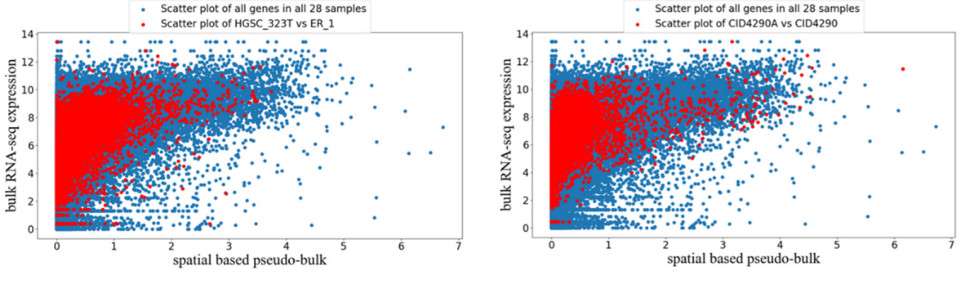
to bulk RNA-seq, as well as the potential mechanism that contributes to the under-detection.

2. Results

2.1. *Paired bulk RNA-seq and Visium data for human samples*

Through an extensive literature search, we have identified five publicly available datasets with paired bulk RNA-seq and Visium data for the same patient subject or the same tissue source. In total, these datasets provided paired bulk RNA-seq and Visium data for 28 samples. The samples originated from diverse biological contexts, such as patients with breast cancer, ovarian cancer, prostate cancer, brain metastasis, as well as human lung tissue. The bulk RNA-seq data was processed by median-of-ratios normalization and log transformation, followed by quantile normalization. For Visium data, gene counts for each spot were divided by the total counts for that spot and multiplied by 10,000, followed by natural log transformation. Finally, to compute the pseudo-bulk expression value for each gene, the mean of the log-transformed counts across all the spots was calculated.

The preprocessed bulk RNA-seq and Visium data is visualized in Fig. 1. Each dot represents the expression of one gene in one sample so there are 25,353*28 dots in one scatter plot. We visualized the paired bulk RNA-seq and Visium data for all 28 samples in blue, with one sample highlighted in red. We observed that the bulk RNA-seq and Visium-based pseudo-bulk expression were positively correlated. Across the 28 sample pairs, the Pearson correlation coefficients between the two data types had a mean and standard deviation of 0.399 ± 0.071, while the Spearman correlation coefficients between the two data types had a mean and standard deviation of 0.766 ± 0.063. This was expected, because the relationship between the two data types is not linear, as shown in Fig. 1. Our preprocessing analysis successfully aligned



(a) (b)

Fig. 1. Scatter plot visualization of paired bulk and pseudo-bulk expression data for 28 human samples. Each dot represents one gene in one sample, so there are 25,353 (genes) * 28 (samples) dots in one scatter plot. (a) A scatter plot of all genes in all 28 samples in blue, overlaid with a scatter plot for all genes in one sample "HGSC_323T (bulk RNA-seq) versus ER_1 (Visium)" in red. (b) A scatter plot of all genes in all 28 samples in blue, overlaid with all genes in another sample "CID4290A (bulk RNA-seq) versus CID4290 (Visium)" in red.

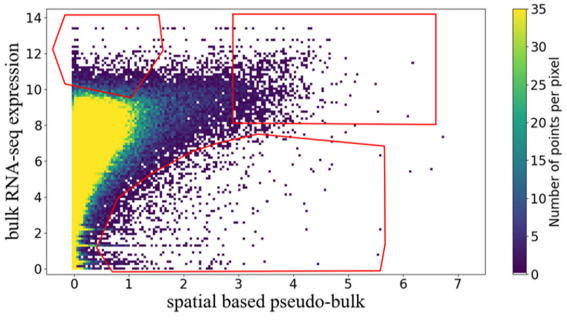
the expression data across the 28 samples, which allowed comparison across the samples to identify genes that may be consistently under-detected in Visium compare to bulk RNA-seq.

2.1.1. *Genes that tend to be under-detected in Visium*

To identify the genes that are consistently under-detected in Visium compared to bulk RNA-seq, we converted Fig. 1 into a density plot (Fig. 2(a)), and examined whether there are genes consistently appearing in the upper-left corner of the plot. We manually drew a gate in the upper-left corner of Fig. 2(a) by positioning the gate such that the high-density region was avoided, so that genes in the gate represented outlier cases whose detected expression in Visium was much lower than that detected in bulk RNA-seq. If a gene appeared in the upper-left gate multiple times, we considered it to be consistently under-detected in Visium compared to bulk RNA-seq.

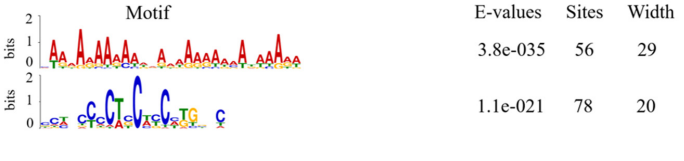
The top 20 most frequently under-detected genes in Visium are listed in Table 1, together with their frequency of occurrences among the 28 paired samples. The range of frequency is from 9 to 20, which is more than one-third of the sample size. Therefore, there seem to be genes consistently under-detected in Visium compared to bulk RNA-seq. The top 20 genes include AHNAK, MACF1, CLTC, DSP, HDLBP, EIF4G2, RNF213, RMRP, TRPS1, SNHG3, CANX, PRKDC, ITGB1, HSPA8, SRRM2, ALDOA, DDX17, XIST, MCL1, and PPP1CB. AHNAK encodes a protein involved in diverse processes such as blood-brain barrier formation, cell structure and migration, cardiac calcium channel regulation, and tumor metastasis.[20] MACF1 encodes a protein which is a member of a family of proteins that form bridges between different cytoskeleton elements.[20] EIF4G2 functions as a general repressor of translation by forming translationally inactive complexes.[20]

In Fig. 2(a), the total number of dots in the upper-left gate is 1560, and each dot represents a gene in one sample. There were 305 genes that occurred more than once in the upper-left gate. Based on the hypergeometric distribution, if 1560 points were randomly picked from the scatter plot, the expected number of genes occurring more than once by chance was 44. The fact that 305 genes occurred more than once in the upper-left gate was interesting, which was seven times the number expected by chance. To identify sequence-based motifs among the genes that occurred more than once in the upper-left gate of Fig. 2(a), we performed a motif search in the last 350 bp of the genes using the MEME suite. We observed two motifs with significant *E*-values and a large number of sites (Fig. 2(b)). For the poly (C) motif, the bit score of most positions was low and the motif was not consecutive. In contrast, for the poly (T) motif, most of the positions had a large bit score, and the motif was consecutive. Therefore, we conjectured that the poly (T) motif toward the tails of the transcripts may be associated with the under-detection of those genes in Visium as compared to bulk RNA-seq. The poly (T) motif toward the tail of the transcripts may be able to form a hairpin structure with the transcripts' poly (A) tails, making the transcripts difficult to capture during the capturing step of Visium library preparation, which is

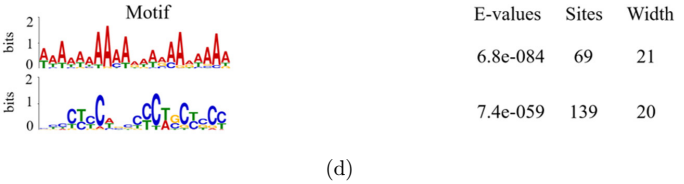


(a)

(b)



(c)



(d)

Fig. 2. Density plot of paired bulk and Visum data of 28 human samples with gates indicating three types of candidate genes. The gates were selected based on the distribution shown in the density plot. The gate in the upper-left corner represents genes that are under-detected in Visium compared to bulk RNA-seq. The gate in the upper-right corner represents genes that are highly expressed in both bulk RNA-seq and Visium data. The gate at the bottom represents genes that are over-detected in Visium compared to bulk RNA-seq (Fig. 2(a)). Significantly enriched motifs in the last 350 bp of the longest transcripts of genes that occurred more than once in each gate are shown for the upper-left gate (Fig. 2(b)), upper-right gate (Fig. 2(c)), and bottom gate (Fig. 2(d)).

a mechanistic conjecture of why those genes were consistently under-detected in Visium compare to bulk RNA-seq. This poly (T) motif associated with this mechanistic conjecture was also observed in a previous study that compared paired bulk RNA-seq and scRNA-seq data of human samples.[19]

Table 1. List of top 20 genes that are most frequently under-detected in Visium and their frequencies.

Gene name	Frequency of occurrence among 28 sample pairs
AHNAK	20
MACF1	15
CLTC	15
DSP	14
HDLBP	13
EIF4G2	13
RNF213	13
RMRP	12
TRPS1	11
SNHG3	10
CANX	10
PRKDC	10
ITGB1	10
HSPA8	10
SRRM2	10
ALDOA	10
DDX17	10
XIST	10
MCL1	9
PPP1CB	9

2.1.2. *Genes consistently highly expressed in both bulk RNA-seq and Visium*

We also manually drew a gate in the upper-right corner of Fig. 2(a). We positioned the gate such that the dense regions were avoided. There were 963 dots in total and 125 genes with more than one occurrence in the upper-right gate.

For the top 20 genes that occurred more than once, their frequency of occurrences ranged from 14 to 26, which is more than half of the sample size. Therefore, many genes are consistently highly expressed in both data types. The top 20 genes include EEF1A1, ACTB, RPL13, FTH1, TMSB10, RPS18, RPLP1, FTL, RPL10, RPS2, RPL13A, RPL37, RPL41, TPT1, RPL28, RPS27, GAPDH, B2M, RPL37A, and ACTG1. Many of them are housekeeping genes involved in diverse biological contexts. For example, EEF1A1 encodes an isoform of the alpha subunit of the elongation factor-1 complex, which is responsible for the enzymatic delivery of aminoacyl tRNAs to the ribosome.[20] This gene has been found to have multiple copies on many chromosomes, and the isoform it encodes is expressed in the brain, placenta, lung, liver, kidney, and pancreas.[20] ACTB encodes one of six different actin proteins, which is a major constituent of the contractile apparatus and one of the two non-muscle cytoskeletal actions that are ubiquitously expressed.[20] TPT1 encodes a protein that is a regulator of cellular growth and proliferation, which is involved in a variety of cellular pathways, including apoptosis, protein synthesis and cell division.[20]

For the genes that occurred more than once in the upper-right gate (Fig. 2(a)), we identified their significantly enriched motifs using the MEME suite. We found two

motifs that were significantly enriched with small E-values and a large number of sites (Fig. 2(c)). Both the poly (C) motif and the poly (A) motif were non-consecutive, and the bit score for most positions was small. The absence of a poly (T) motif for the upper-right gate further strengthened our conjecture about the formation of a hairpin structure, which may cause certain genes to be consistently under-detected in Visium.

2.1.3. *Genes that appear to be over-detected in Visium*

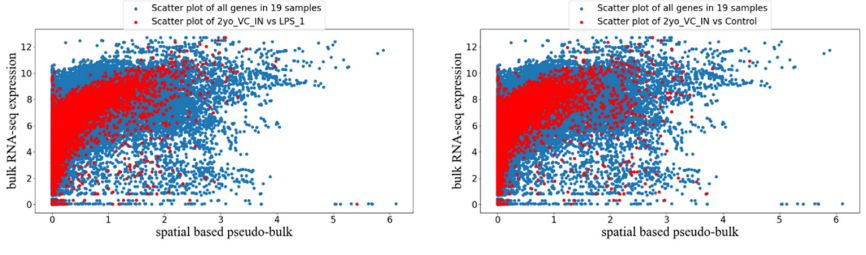
For completeness, we also tried to identify the genes that were frequently over-detected in Visium. We manually drew a third gate in the bottom region of Fig. 2(a). The total number of dots in the bottom gate (Fig. 2(a)) is 1,031, and the number of genes that occurred more than once (Fig. 2(a)) is 168.

Using the MEME suite, we observed two significantly enriched motifs for the genes that occurred more than once in the bottom gate (Fig. 2(d)), a ploy (C) motif and a poly (A) motif. These two motifs were similar to the enriched motifs in the upper-right gate that contained genes highly expressed in both bulk RNA-seq and Visium data. Once again, the absence of a poly (T) motif in the bottom gate further strengthened our conjecture that sequence-based motifs of transcripts cause certain genes to be consistently under-detected in Visium.

2.2. *Paired bulk RNA-seq and Visium data for mouse samples*

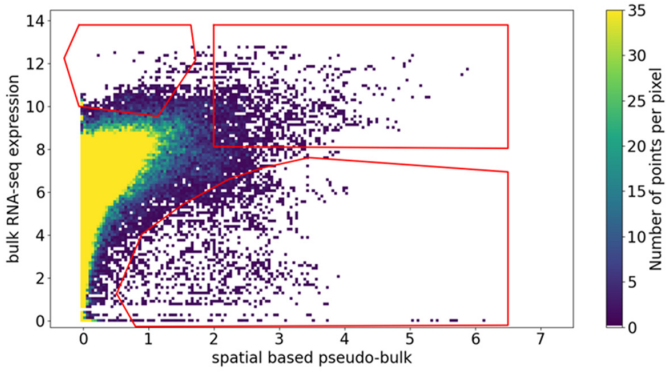
We extend this analysis to paired bulk RNA-seq and Visium data for 19 mouse samples from four studies, with the same data pre-processing steps as our analysis of the human data. These mouse datasets were from diverse sources including mouse brain, kidney, and bladder.

The paired bulk and pseudo-bulk RNA-seq data were visualized using scatter plots (Figs. 3(a) and 3(b)). Each dot represents a gene in one sample. Therefore, the

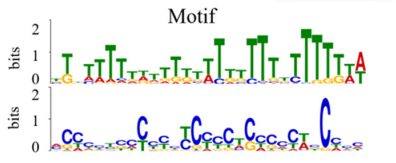


(a) (b)

Fig. 3. Scatter plot visualization of paired bulk and pseudo-bulk data for 19 mouse samples. Each dot is an expression of one gene in one sample, so there are 29,089 genes *19 dots in one scatter plot. (a) A scatter plot of all genes in all 19 samples in blue, overlaid with a scatter plot for all genes in one sample "2yo_VC_IN (bulk RNA-seq) versus LPS_1 (Visium)" in red. (b) A scatter plot of all genes in all 19 samples in blue, overlaid with all genes in another sample "2yo_VC_IN (bulk RNA-seq) versus Control (Visium)" in red.

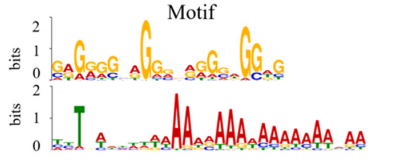


(a)



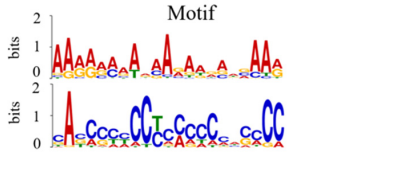
	E-values	Sites	Width
	1.6e-065	53	29
	1.6e-064	84	29

(b)



	E-values	Sites	Width
	4.8e-065	105	21
	2.2e-058	88	29

(c)



	E-values	Sites	Width
	1.8e-049	96	21
	2.6e-009	30	21

(d)

Fig. 4. Density plot of a scatter plot of 19 paired mouse samples with gates indicating candidate genes in three aspects. (Fig. 4(a)). Significantly enriched motifs in the last 350 bp of the longest transcripts of genes that occurred more than once in each gate are shown for the upper-left gate (Fig. 4(b)), upper-right gate (Fig. 4(c)), and bottom gate (Fig. 4(d)).

total number of dots in one scatter plot is 29,089 (genes) * 19 (samples). We observed a positive correlation between bulk RNA-seq and spatial-based pseudo-bulk expression, which was expected. We overlaid two individual samples (red) on all other samples (blue). Data for the other samples showed similar patterns as the ones in Figs. 3(a) and 3(b). For the 19 samples, the Pearson correlation coefficients between

the two data types is 0.496 ± 0.102, while the Spearman correlation coefficient between the two data types is 0.880 ± 0.062, similar to the correlation between bulk RNA-seq and Visium data for human samples.

Similar to the analysis of human data, we made the scatter plot and manually drew three gates on the plot (Fig. 4(a)) capturing genes that were under-detected in Visium, genes that were highly expressed in both technologies, and genes that were over-detected in Visium. The total number of dots in the upper-left, upper-right, and bottom gates (Fig. 4(a)) were 978, 1348, and 1318, respectively. The genes that occurred more than once in the upper-left, upper-right, and bottom gates (Fig. 4(a)) were 177, 231, and 139, respectively. Using the MEME suite, we identified significantly enriched motifs in the last 350bp of the genes that occurred more than once in each gate. The poly (T) motif was significantly enriched for the upper-left gate (Fig. 4(b)), while the poly (A) motif was significantly enriched in the other two gates (Figs. 4(c) and 4(d)). This result was consistent with the comparison between bulk RNA-seq and Visium data based on human data. Therefore, this result further strengthened our motif-based conjecture of genes frequently under-detected in Visium compared to bulk RNA-seq.

3. Conclusions and Discussion

In this study, we analyzed paired bulk RNA-seq and Visium data for 28 human samples and 19 mouse samples, which were generated from diverse biological contexts. We compared the bulk RNA-seq and Visium data and found a collection of genes that were consistently under-detected in Visium compared to bulk RNA-seq. The genes have significantly enriched poly (T) motif towards their 3' end. However, for the genes that are frequently over-detected in Visium and the genes that have high expression in both technologies, the poly (T) motif was absent. This result is consistent with our previous study that compared paired bulk RNA-seq and scRNA-seq samples. We hypothesize that the poly (T) motif may be able to form a hairpin structure with the poly (A) tails of the mRNA transcripts, making it difficult for the mRNA transcripts to be captured during Visium library preparation.

Among the 20 genes frequently under-detected in Visium compared to bulk RNA-seq, 8 of them are also among the top 20 most frequently under-detected genes in scRNA-seq compared to bulk RNA-seq.[19] The eight genes are AHNAK, MACF1, EIF4G2, CANX, ITGB1, SRRM2, DDX17, and XIST. It is likely that the mechanistic conjecture of under-detection in Visium also applies to scRNA-seq because these technologies share similar experimental protocols for library preparation.

The datasets analyzed in this study contain both technical and biological variabilities. The technical variabilities include choices of alignment algorithms and choices of reference genome and RNA library protocols. The biological variabilities include a potentially small number of sample pairs with mismatched sex. All these factors could impact the data and analysis results. In an ideal situation, all the samples need to be processed using the same experimental procedure, reference genome, and

alignment tool with the same version. However, this is infeasible since each dataset was obtained with slight differences in its experimental protocols. In addition, not all FASTQ files are available for the bulk RNA-seq and Visium samples in this study, and therefore we are unable to run a standardized pre-processing analysis pipeline for all the samples. Since these factors were ignored in our analysis, we were effectively embracing the variabilities in the data. Even with such variabilities in the data, we still captured a robust motif for the upper-left gate of genes under-detected in Visium. These variabilities help demonstrate the robustness of our observations.

Because there were a limited number of studies that generated both bulk and Visium data for the same biological samples, among eight out of nine datasets in this study, the bulk RNA-seq and Visium data we collected are from different research studies that examined the same tissue sources. However, even though many of the paired data in our analysis were not generated from the exact same samples, we still observed the enrichment of poly (T) motif among genes frequently under-detected in Visium for both human and mouse samples, which was encouraging.

The observed under-detection of certain genes in Visum has implications for downstream analysis and interpretation of results. If the goal of a study is to investigate a specific gene that is frequently under-detected in Visium, the expression value of the gene in Visium is less reliable compared to the expression value in bulk RNA-seq. Although imputation algorithms have been proposed to improve the quality of Visium data, such statistical approaches are often affected by systematic bias in the data, and in this case, the hypothesis of hairpin structure formation between poly (T) motif and poly (A) tail of mRNA could be a systematic bias in the data. Recognizing the possibility of such a systematic bias is beneficial for the development of both computational algorithms and experimental methods. Beyond Visium, the analysis here can be applied to study paired bulk and other sequencing-based spatial transcriptomics data to identify mechanisms that may contribute to the under-detection of certain genes in those technologies.

4. Materials and Methods

4.1. *Summary of datasets*

Paired bulk RNA-seq and Visium data for 28 human samples were obtained from five studies. Paired bulk RNA-seq and Visium data for 19 mouse samples were obtained from four studies. The paired data were generated from either the same subject or the same tissue source. A summary of the nine datasets and references is available in Table 2. Details about the accession ID of individual samples and how bulk RNA-seq and Visium data were paired can be found in Supplementary Table 1.

4.2. *Data preprocessing of bulk RNA-seq*

For each bulk RNA-seq dataset, median-of-ration normalization was performed using DESeq2. Then log transformation was performed on the normalized data. Next,

Table 2. Summary of datasets.

Bulk RNA-seq and Visium datasets	Source of samples	Number of samples
https://zenodo.org/record/4739739#.YqrNv-IWFPb (Wu *et al.*[21]), GSE176078 (Wu *et al.*[21])	Human breast cancer	6
GSE189843 (Stur *et al.*[22]) and GSE102094 (Ducie *et al.*[23])	Human high-grade serous ovarian tumor	12
GSE159697 (McCray *et al.*[24]) and TCGA	Human prostate cancer	2
GSE179572 (Sudmeier *et al.*[25]) and GSE164150 (Su *et al.*[26])	Human brain metastases	6
GSE178361 (Murthy *et al.*[27]) and GSE111892 (Clarke *et al.*[28])	Human lung tissue	2
GSE180128 (Baker *et al.*)	Mouse urinary bladder	3
GSE148612 (Hasel *et al.*[29]) and GSE99791 (Boisvert *et al.*[30])	Mouse brain	6
GSE171406 (Ferreira *et al.*[31]) and GSE141115 (Denisenko *et al.*[32])	Mouse kidney	3
GSE182127 (Buzzi *et al.*[33]) and GSE99791 (Boisvert *et al.*[30])	Mouse brain	7

overlapping genes among paired bulk RNA-seq and Visium data were identified, and a matrix representing the normalized bulk RNA-seq expression of the overlapping genes was created. Finally, quantile normalization was performed on the matrix.

4.3. *Data preprocessing of Visium*

For each Visium sample, first, gene counts for each spot were divided by the total counts for that spot and multiplied by 10,000; then, the library-size normalized data was further transformed by a natural log. Next, to compute the pseudo-bulk expression value for each gene, the mean of the log-transformed counts across all the spots was calculated. Finally, a matrix representing the normalized Visium data of the overlapping genes was created.

4.4. *Correlation analysis between bulk RNA-seq and Visium*

Pearson correlation coefficients and Spearman correlation coefficients were calculated between normalized bulk RNA-seq and Visium-based pseudo-bulk profiles for the human and mouse samples.

4.5. *Motif enrichment analysis*

MEME Suite was used to identify significantly enriched motifs of the last 350 bp of cDNA sequences of the longest transcripts of candidate genes. The longest transcripts of the genes together with their cDNA sequences were obtained from BioMart during April 2023. For motif site distribution, zero or one occurrence was selected for the analysis. MEME Suite reports *E*-value which serves as an indicator of the

statistical significance of a motif. A motif with an *E*-value smaller than 0.05 is considered to be significant.

Acknowledgments

The authors would like to thank Dr. Greg Gibson for his valuable discussions and insights during the development of this paper. This work was supported by funding from the National Institute of Health (U01CA265711) and the National Science Foundation (CCF2007029). The content is solely the responsibility of the authors and does not necessarily represent the official views of the funders.

ORCID

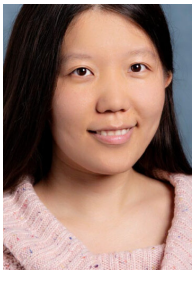
Xinling Li https://orcid.org/0000-0003-3095-6911
Peng Qiu https://orcid.org/0000-0003-3256-0734

References

1. Crosetto N, Bienko M, van Oudenaarden A, Spatially resolved transcriptomics and beyond, *Nat Rev Genet* **16**(1): 57–66, 2015.
2. Lein E, Borm LE, Linnarsson S, The promise of spatial transcriptomics for neuroscience in the era of molecular cell typing, *Science* **358**(6359): 64–69, 2017.
3. Asp M, Bergenstrahle J, Lundeberg J, Spatially resolved transcriptomes-next generation tools for tissue exploration, *Bioessays* **42**(10): e1900221, 2020.
4. Waylen LN, Nim HT, Martelotto LG, Ramialison M, From whole-mount to single-cell spatial assessment of gene expression in 3D, *Commun Biol* **3**(1): 602, 2020.
5. Moses L, Pachter L, Museum of spatial transcriptomics, *Nat Methods* **19**(5): 534–546, 2022.
6. Lewis SM, Asselin-Labat ML, Nguyen Q, Berthelet J, Tan X, Wimmer VC, Merino D, Rogers KL, Naik SH, Spatial omics and multiplexed imaging to explore cancer biology, *Nat Methods* **18**(9): 997–1012, 2021.
7. Shah S, Lubeck E, Zhou W, Cai L, In Situ transcription profiling of single cells reveals spatial organization of cells in the mouse Hippocampus, *Neuron* **92**(2): 342–357, 2016.
8. Chen KH, Boettiger AN, Moffitt JR, Wang S, Zhuang X, RNA imaging. Spatially resolved, highly multiplexed RNA profiling in single cells, *Science* **348**(6233): aaa6090, 2015.
9. Femino AM, Fay FS, Fogarty K, Singer RH, Visualization of single RNA transcripts in situ, *Science* **280**(5363): 585–590, 1998.
10. Wang X, Allen WE, Wright MA, Sylwestrak EL, Samusik N, Vesuna S, Evans K, Liu C, Ramakrishnan C, Liu J, Nolan GP, Bava FA, Deisseroth K, Three-dimensional intact-tissue sequencing of single-cell transcriptional states, *Science* **361**(6400): eaat5691, 2018.
11. Rodriques SG, Stickels RR, Goeva A, Martin CA, Murray E, Vanderburg CR, Welch J, Chen LM, Chen F, Macosko EZ, Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution, *Science* **363**(6434): 1463–1467, 2019.
12. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R, Integrating single-cell transcriptomic data across different conditions, technologies, and species, *Nat Biotechnol* **36**(5): 411–420, 2018.
13. Wolf FA, Angerer P, Theis FJ, SCANPY: Large-scale single-cell gene expression data analysis, *Genome Biol* **19**(1): 15, 2018.

14. Li Z, Song T, Yong J, Kuang R, Imputation of spatially-resolved transcriptomes by graph-regularized tensor completion, *PLoS Comput Biol* **17**(4): e1008218, 2021.
15. Biancalani T, Scalia G, Buffoni L, Avasthi R, Lu Z, Sanger A, Tokcan N, Vanderburg CR, Segerstolpe A, Zhang M, Avraham-Davidi I, Vickovic S, Nitzan M, Ma S, Subramanian A, Lipinski M, Buenrostro J, Brown NB, Fanelli D, Zhuang X, Macosko EZ, Regev A, Deep learning and alignment of spatially resolved single-cell transcriptomes with Tangram, *Nat Methods* **18**(11): 1352–1362, 2021.
16. Wang Y, Song B, Wang S, Chen M, Xie Y, Xiao G, Wang L, Wang T, Sprod for denoising spatially resolved transcriptomics data based on position and image information, *Nat Methods* **19**(8): 950–958, 2022.
17. Zhao P, Zhu J, Ma Y, Zhou X, Modeling zero inflation is not necessary for spatial transcriptomics, *Genome Biol* **23**(1): 118, 2022.
18. Liu J, Tran V, Vemuri VNP, Byrne A, Borja M, Kim YJ, Agarwal S, Wang R, Awayan K, Murti A, Taychameekiatchai A, Wang B, Emanuel G, He J, Haliburton J, Oliveira Pisco A, Neff NF, Concordance of MERFISH spatial transcriptomics with bulk and single-cell RNA sequencing, *Life Sci Alliance* **6**(1): e202201701, 2023.
19. Li X, Gibson G, Qiu P, Gene representation in scRNA-seq is correlated with common motifs at the 3' end of transcripts, *Front Bioinform* **3**: 1120290, 2023.
20. Stelzer G, Dalah I, Stein TI, Satanower Y, Rosen N, Nativ N, Oz-Levi D, Olender T, Belinky F, Bahir I, Krug H, Perco P, Mayer B, Kolker E, Safran M, Lancet D, In-silico human genomics with geneCards, *Hum Genomics* **5**(6): 709–717, 2011.
21. Wu SZ, Al-Eryani G, Roden DL, Junankar S, Harvey K, Andersson A, Thennavan A, Wang C, Torpy JR, Bartonicek N, Wang T, Larsson L, Kaczorowski D, Weisenfeld NI, Uytingco CR, Chew JG, Bent ZW, Chan CL, Gnanasambandapillai V, Dutertre CA, Gluch L, Hui MN, Beith J, Parker A, Robbins E, Segara D, Cooper C, Mak C, Chan B, Warrier S, Ginhoux F, Millar E, Powell JE, Williams SR, Liu XS, O'Toole S, Lim E, Lundeberg J, Perou CM, Swarbrick A, A single-cell and spatially resolved atlas of human breast cancers, *Nat Genet* **53**(9): 1334–1347, 2021.
22. Stur E, Corvigno S, Xu M, Chen K, Tan Y, Lee S, Liu J, Ricco E, Kraushaar D, Castro P, Zhang J, Sood AK, Spatially resolved transcriptomics of high-grade serous ovarian carcinoma, *iScience* **25**(3): 103923, 2022.
23. Ducie J, Dao F, Considine M, Olvera N, Shaw PA, Kurman RJ, Shih IM, Soslow RA, Cope L, Levine DA, Molecular analysis of high-grade serous ovarian carcinoma with and without associated serous tubal intra-epithelial carcinoma, *Nat Commun* **8**(1): 990, 2017.
24. McCray T, Pacheco JV, Loitz CC, Garcia J, Baumann B, Schlicht MJ, Valyi-Nagy K, Abern MR, Nonn L, Vitamin D sufficiency enhances differentiation of patient-derived prostate epithelial organoids, *iScience* **24**(1): 101974, 2021.
25. Sudmeier LJ, Hoang KB, Nduom EK, Wieland A, Neill SG, Schniederjan MJ, Ramalingam SS, Olson JJ, Ahmed R, Hudson WH, Distinct phenotypic states and spatial distribution of CD8(+) T cell clonotypes in human brain metastases, *Cell Rep Med* **3**(5): 100620, 2022.
26. Su J, Song Q, Qasem S, O'Neill S, Lee J, Furdui CM, Pasche B, Metheny-Barlow L, Masters AH, Lo HW, Xing F, Watabe K, Miller LD, Tatter SB, Laxton AW, Whitlow CT, Chan MD, Soike MH, Ruiz J, Multi-omics analysis of brain metastasis outcomes following craniotomy, *Front Oncol* **10**: 615472, 2020.
27. Kadur Lakshminarasimha Murthy P, Sontake V, Tata A, Kobayashi Y, Macadlo L, Okuda K, Conchola AS, Nakano S, Gregory S, Miller LA, Spence JR, Engelhardt JF, Boucher RC, Rock JR, Randell SH, Tata PR, Human distal lung maps and lineage hierarchies reveal a bipotent progenitor, *Nature* **604**(7904): 111–119, 2022.

28. Clarke J, Panwar B, Madrigal A, Singh D, Gujar R, Wood O, Chee SJ, Eschweiler S, King EV, Awad AS, Hanley CJ, McCann KJ, Bhattacharyya S, Woo E, Alzetani A, Seumois G, Thomas GJ, Ganesan AP, Friedmann PS, Sanchez-Elsner T, Ay F, Ottensmeier CH, Vijayanand P, Single-cell transcriptomic analysis of tissue-resident memory T cells in human lung cancer, *J Exp Med* **216**(9): 2128–2149, 2019.
29. Hasel P, Rose IVL, Sadick JS, Kim RD, Liddelow SA, Neuroinflammatory astrocyte subtypes in the mouse brain, *Nat Neurosci* **24**(10): 1475–1487, 2021.
30. Boisvert MM, Erikson GA, Shokhirev MN, Allen NJ, The aging astrocyte transcriptome from multiple regions of the mouse brain, *Cell Rep* **22**(1): 269–285, 2018.
31. Melo Ferreira R, Sabo AR, Winfree S, Collins KS, Janosevic D, Gulbronson CJ, Cheng YH, Casbon L, Barwinska D, Ferkowicz MJ, Xuei X, Zhang C, Dunn KW, Kelly KJ, Sutton TA, Hato T, Dagher PC, El-Achkar TM, Eadon MT, Integration of spatial and single-cell transcriptomics localizes epithelial cell-immune cross-talk in kidney injury, *JCI Insight* **6**(12): e147703, 2021.
32. Denisenko E, Guo BB, Jones M, Hou R, de Kock L, Lassmann T, Poppe D, Clement O, Simmons RK, Lister R, Forrest ARR, Systematic assessment of tissue dissociation and storage biases in single-cell and single-nucleus RNA-seq workflows, *Genome Biol* **21**(1): 130, 2020.
33. Buzzi RM, Akeret K, Schwendinger N, Klohs J, Vallelian F, Hugelshofer M, Schaer DJ, Spatial transcriptome analysis defines heme as a hemopexin-targetable inflammatoxin in the brain, *Free Radic Biol Med* **179**: 277–287, 2022.



Xinling Li received her BS degree in Bioinformatics from University of California, San Diego in 2016, and MS degree in Computational Biology from Carnegie Mellon University in 2019. After which she worked at Cedars-Sinai Medical Center as a research bioinformatician for a year. She is currently a Ph.D. student in the Bioinformatics program at Georgia Institute of Technology. Her research interests include bioinformatics, mathematical modeling, and machine learning.

Peng Qiu received his Ph.D. degree in Electrical and Computer Engineering from University of Maryland in College Park. He is currently a professor in The Wallace H. Coulter Department of Biomedical Engineering, Georgia Institute of Technology and Emory University. His research interests include bioinformatics and computational biology, focusing on statistical signal processing, machine learning, control systems and optimization. He is a fellow of American Institute of Medical and Biological Engineering, an ISAC Marylou Ingram Scholar, and a Wallace H. Coulter Distinguished Faculty Fellow.