

Peer Review Information

Journal: Nature Methods

Manuscript Title: Benchmarking spatial and single-cell transcriptomics integration methods for transcript distribution prediction and cell type deconvolution

Corresponding author name(s): Kun Qu

Reviewer Comments & Decisions:

Decision Letter, initial version:
--

Subject: Decision on Nature Methods submission NMETH-AS46278

Message:

29th Sep 2021

Dear Professor Qu,

Thank you very much for your patience when your Analysis entitled "Benchmarking methods for integrating spatial and single-cell transcriptomics data" is sent out for peer review. 3 reviewers agreed to review this paper but one of them hasn't sent their report despite our multiple attempts to contact them. The paper has now been seen by 2 reviewers, whose comments are attached. While they find your work of potential interest, they have raised serious concerns which in our view are sufficiently important that they preclude publication of the work in Nature Methods, at least in its present form.

As you will see, the reviewers raise concerns about different aspects of the benchmarking analysis, including its scope, method selection, implementation, findings, and other aspects. Should further extensive analyses and revisions allow you to fully address all these criticisms we would be willing to look at a extensively revised manuscript (unless, of course, something similar has by then been accepted at Nature Methods or appeared elsewhere). This includes submission or publication of a portion of this work somewhere else. We hope you understand that until we have read the revised paper in its entirety we cannot promise that it will be sent back for peer-review.

If you are interested in revising this manuscript for submission to Nature Methods in the future, please contact me to discuss your appeal before making any revisions. Otherwise, we hope that you find the reviewers' comments helpful when preparing your paper for submission elsewhere.

Sincerely,

Lin

Lin Tang, PhD
Senior Editor
Nature Methods

Reviewers' Comments:

Reviewer #1:

Remarks to the Author:

This manuscript benchmarks 7 methods for integrating single-cell RNA-sequencing and spatial transcriptomics data. The benchmark is based on evaluation of these methods on 14 matched datasets consisting of scRNA-seq and spatial transcriptomics from the same tissue. Performing such a rigorous benchmarking study over such a diverse collection of data and methods is no easy task, and the authors are commended for their extensive efforts and valuable contribution. There is substantial interest in the use and development of methods to integrate single-cell and spatial sequencing data in order to maximize the relative advantages of both technologies. Thus, this manuscript serves an important need in the community.

At the same time, the manuscript has a number of issues that decrease my enthusiasm for this work.

1) A wide diversity of technology platforms are evaluated including 3 single-cell RNA sequencing technologies (dropSeq, smartSeq, and 10X Chromium) and 9 spatial sequencing technologies. While this diversity is reflective of the fast moving state of the technology, this diversity also complicates the interpretation of the benchmarking results. Are all of these methods expected to perform similarly on these different technologies? Presumably, some methods were designed/tuned to the characteristics certain technologies. These details are not described in the text. Moreover, it's not clear that the problem of "Integrate any one of 3 scRNA-seq technologies with any one of 9 spatial sequencing technologies" is a reasonable computational problem that any of these method address. In the benchmarking, the authors run each method with a single choice of parameters, but it seems reasonable that different parameters would be required for different technologies.

2) Related to point 1, it is likely that methods would perform differently with different numbers of genes and with different spatial organization. Some discussion of these points would be helpful to guide the reader's expectations. What sequencing technologies, number of genes, number of spots is each method claimed to analyze as stated in the original publication or software documentation?

3) A whole class of recent methods that perform deconvolution of spatial transcriptomics data into single-cell resolution is missing from the comparison such as BayesSpace (Zhao et al. Nature Biotech. 2021), RCTD (Cable et al. Nature Biotech. 2021), Steroscope (Andersson et al. Commun. Biol. 2020), SPOTlight (Elosua-Bayes, et al. NAR 2021), and destVI (Lopez et al. 2021, bioRxiv). [Note that Lopez et al was posted as preprint only recently so not expected that authors would include.]. It's not clear what criteria were used to select methods for benchmarking and why not deconvolution methods were included. Presumably the evaluation criteria for deconvolution methods would be different from the criteria used in the present publication.

4) The evaluation metrics for the benchmark are based on a comparison to the matched spatial transcriptomics data. As a complement to evaluation of methods on real data, the authors might consider an evaluation on simulated data where there is a ground truth to evaluate. While simulations obviously cannot match all of the features of real data, benchmarking papers routinely use simulated data; e.g. see Saelens et al. Nature Biotechnology (2019) benchmarking of trajectory inference methods;

5) Related to the above points, it is arguably premature to perform a comprehensive benchmarking study of spatial integration methods since: the sequencing technologies are too diverse, the scope of each method is limited to a subset of technologies, the goals of the methods (cell assignment, deconvolution, etc) are too varied, both the technologies and methods are evolving too rapidly, and the standards are unclear. As a point of comparison: the Saelens et al. benchmarking of trajectory inference methods benchmarked 45 of the 70 known methods on 110 real and 229 synthetic datasets, explicitly described the trade offs of different methods according to topology of the trajectory, and included guidelines for users to choose a method based on the topology of the trajectory (cf. Fig. 5). By comparison, the present benchmarking study is relatively limited in scope and doesn't provide much practical guidance for the biologist who wants to analyze their data or the methods developer who wants to improve the state-of-the-art.

6) Given the rapid developments in spatial sequencing technologies — with 10X Visium scheduled to reach single-cell resolution next year and Slide-seq v2 approaching scRNA-seq — it's possible that interest in integration of scRNA-seq and spatial transcriptomics will be short lived. It would be helpful to hear further perspective on the future based on the benchmarking.

Minor comments:

Line 464: “10-X cross validation” —> “10-fold cross validation”

Figure 1b. Recommend adding references to this table so readers can quickly see source of datasets.

Reviewer #2:

Remarks to the Author:

In this work, the authors present a benchmarking study of methods used to integrate dissociated single cell RNA sequencing and spatial transcriptomic data. The authors evaluated 7 methods for their ability to predicted spatial expression patterns of unmeasured genes and for a subset of 4 methods their ability to predict cell locations. Based on several evaluation metrics, the authors conclude that gimVI and SpaGE have the best performance in predicting unmeasured genes and Seurat for the prediction of spatial locations.

Overall, the paper provides a much-needed independent comparison of available methods to integrate single cell and spatial data. The paper is well written and the experiments are clearly explained. The claims made by the authors are backed by the experiments. I would also like to compliment the authors on the associated GitHub repository.

My main concern is that the authors chose to compare very specific aspects of integration methods (predicting unmeasured genes and spatial localization) while other equally important tasks are ignored. I elaborate on this below in my major comments.

Major comments:

- The evaluation of integration methods focuses on two tasks (predicting unmeasured genes and spatial localization) while other tasks (e.g. cell-cell interaction deconvolution) are completely ignored. These are equally important aspects of methods integrating single cell and spatial transcriptomic data. At least the author should make this explicitly clear in the title, abstract and discussion.
- A bit related, the authors should provide a clear motivation for the inclusion and exclusion of methods. A recent review on the integration of single cell and spatial methods lists 16 methods to perform different tasks. Why did the authors ignore cell2location for the localization and stPlus for the prediction of unmeasured genes? Is it because both are preprints? Both methods are well documented and easy to use.
- The authors tested the effect of normalizing the spatial data on the results of different methods. First, I want to make sure I understand this correctly: what is tested here is integration of (normalized) scRNA-seq data with either normalized/non-normalized spatial data, correct? It is not clear to me what is the motivation here? Why not also try non-normalized scRNA-seq data? It is also not clear to me if this data

is used for evaluation (calculation of PCC and other measures) or also for visualization. I am a little concerned about outliers. In principle, if there are few cells/spots with very high values, they will drive the correlation.

- In relation to the previous comment, it seems that some methods are more sensitive to this change than others. For Supplementary Figure 3b, it will be good to plot the results per method with and without normalization. Perhaps even use a paired test to assess how different these results are per method.

- Perhaps my own lack of knowledge, but it is not clear to me why the authors restrict the prediction of cell location to spots (i.e. vizium or slide-seq like data)? Why not assess the methods ability to localize cells to locations in general? I am thinking of general experiments like in the novoSparc paper where a wide variety of spatial data are used.

- For down-sampling, the randomly set non-zero elements in the original expression matrix to zeros. This is a bit over-simplified from what happens in reality in which case, few molecules are missed (in osmFISH) or there are dropouts. Why not use a more realistic downsampling approach such as those used in the scuttle R package for single-cell data (

<http://www.bioconductor.org/packages/release/bioc/html/scuttle.html>)

- The evaluation of computational efficiency (Supplementary Figure 7, which should be a table instead) is not sufficient. Why not do this assessment for all datasets? The number of shared genes and the difficulty in aligning will have an effect (not only the number of cells/spots).

Minor comments:

- Please explicitly define what you mean by sparsity.

- There are no details of how novoSparc was run. Did the authors run it in fully de novo mode? In the novoSparc paper, this did not work well in brain tissues and using reference genes is recommended.

- The choices made while simulating the “multi-cell spot problem” in terms of the grid size and how many pixels to group are not motivated at all. Why these choices? And how do they relate to the resolution of the original data and the current vizium/slide-seq resolutions. Please elaborate.

Although we cannot publish your paper, it may be appropriate for another journal in the Nature Portfolio. If you wish to explore the journals and transfer your manuscript please use our manuscript transfer portal. If you transfer to Nature journals or the Communications journals, you will not have to re-supply manuscript metadata and files. This link can only be used once and remains active until used.

All Nature Portfolio journals are editorially independent, and the decision on your manuscript will be taken by their editors. For more information, please see our <a

http://www.nature.com/authors/author_resources/transfer_manuscripts.html?WT.mc_id=EMI_NPG_1511_AUTHORTRANSF&WT.ec_id=AUTHOR>manuscript transfer FAQ page.

Note that any decision to opt in to In Review at the original journal is not sent to the receiving journal on transfer. You can opt in to [In Review](https://www.nature.com/nature-research/for-authors/in-review) at receiving journals that support this service by choosing to modify your manuscript on transfer. In Review is available for primary research manuscript types only.

** For Nature Research Group general information and news for authors, see <http://npg.nature.com/authors>.

Decision Letter – appeal:

Subject: Decision on appeal of Nature Methods submission NMETH-AS46278A-Z
Message:

2nd Nov 2021

Dear Professor Qu,

Thank you for your letter asking us to reconsider our decision on your Analysis, "Benchmarking methods for integrating spatial and single-cell transcriptomics data". After careful consideration we have decided that we are willing to consider a revised version of your manuscript that fully addresses all the concerns raised by our reviewers with extensive new analyses and revisions.

When revising your paper:

- * include a point-by-point response to our referees and to any editorial suggestions
- * please underline/highlight any additions to the text or areas with other significant changes to facilitate review of the revised manuscript
- * address the points listed described below to conform to our open science requirements
- * ensure it complies with our general format requirements as set out in our guide to authors at www.nature.com/naturemethods
- * resubmit all the necessary files electronically by using the link below to access your home page

[REDACTED]

Note: This URL links to your confidential home page and associated information about manuscripts you may have submitted, or that you are reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage.

We hope to receive your revised paper within 8 weeks. We are very aware of the difficulties caused by the COVID-19 pandemic to the community. If you cannot send it within this time, please let us know. In this event, we will still be happy to reconsider your paper at a later date so long as nothing similar has been accepted for publication at Nature Methods or published elsewhere.

OPEN SCIENCE REQUIREMENTS

REPORTING SUMMARY AND EDITORIAL POLICY CHECKLISTS

When revising your manuscript, please submit reporting summary and editorial policy checklists.

Reporting summary: <https://www.nature.com/documents/nr-reporting-summary.zip>

Editorial policy checklist: <https://www.nature.com/documents/nr-editorial-policy-checklist.zip>

If your paper includes custom software, we also ask you to complete a supplemental reporting summary.

Software supplement: <https://www.nature.com/documents/nr-software-policy.pdf>

Please submit these with your revised manuscript. They will be available to reviewers to aid in their evaluation if the paper is re-reviewed. If you have any questions about the checklist, please see <http://www.nature.com/authors/policies/availability.html> or contact me.

Please note that these forms are dynamic ‘smart pdfs’ and must therefore be downloaded and completed in Adobe Reader. We will then flatten them for ease of use by the reviewers. If you would like to reference the guidance text as you complete the template, please access these flattened versions at <http://www.nature.com/authors/policies/availability.html>.

DATA AVAILABILITY

Please include a “Data availability” subsection in the Online Methods. This section should inform readers about the availability of the data used to support the conclusions of your study, including accession codes to public repositories, references to source data that may be published alongside the paper, unique identifiers such as URLs to data repository entries, or data set DOIs, and any other statement

about data availability. At a minimum, you should include the following statement: “The data that support the findings of this study are available from the corresponding author upon request”, describing which data is available upon request and mentioning any restrictions on availability. If DOIs are provided, please include these in the Reference list (authors, title, publisher (repository name), identifier, year). For more guidance on how to write this section please see: <http://www.nature.com/authors/policies/data/data-availability-statements-data-citations.pdf>

CODE AVAILABILITY

Please include a “Code Availability” subsection in the Online Methods which details how your custom code is made available. Only in rare cases (where code is not central to the main conclusions of the paper) is the statement “available upon request” allowed (and reasons should be specified).

We request that you deposit code in a DOI-minting repository such as Zenodo, Gigantum or Code Ocean and cite the DOI in the Reference list. We also request that you use code versioning and provide a license.

For more information on our code sharing policy and requirements, please see:

<https://www.nature.com/nature-research/editorial-policies/reporting-standards#availability-of-computer-code>

MATERIALS AVAILABILITY

As a condition of publication in Nature Methods, authors are required to make unique materials promptly available to others without undue qualifications.

Authors reporting new chemical compounds must provide chemical structure, synthesis and characterization details. Authors reporting mutant strains and cell lines are strongly encouraged to use established public repositories.

More details about our materials availability policy can be found at <https://www.nature.com/nature-portfolio/editorial-policies/reporting-standards#availability-of-materials>

ORCID

Nature Methods is committed to improving transparency in authorship. As part of our efforts in this direction, we are now requesting that all authors identified as ‘corresponding author’ on published papers create and link their Open Researcher and Contributor Identifier (ORCID) with their account on the Manuscript Tracking System (MTS), prior to acceptance. This applies to primary research papers only. ORCID helps the scientific community achieve unambiguous attribution of all scholarly contributions. You can create and link your ORCID from the home page of the MTS by clicking on

'Modify my Springer Nature account'. For more information please visit www.springernature.com/orcid.

We look forward to hearing from you soon.

Sincerely,

Lin Tang, PhD
Senior Editor
Nature Methods

Author Rebuttal to Initial comments

Responses to reviewers

We thank the reviewers for their positive assessments and thoughtful suggestions, which provided us further insights and opportunities to substantially improve the quality and scope of our study and manuscript. Following a short summary narrative, we provide our point-by-point responses below.

Briefly, we have undertaken further and extensive analyses on the scope of integration methods, on matching scRNA-seq and spatial transcriptomics datasets, and conducted additional comparisons amongst integration methods. We have revised our manuscript as follows:

1. We have now provided comparison of **9 additional integration methods** of spatial transcriptomics and scRNA-seq data that can predict the spatial distribution of undetected transcripts or perform cell type deconvolution of spots in histological sections for benchmarking, including stPlus (Chen et al. *Bioinformatics* 2021), SpatialDWLS (Dong et al. *Genome Biology* 2021), RCTD (Cable et al. *Nature Biotechnology* 2021), Stereoscope (Andersson et al. *Communication Biology* 2020), SPOTlight (Elosua-Bayes et al. *Nucleic Acids Research* 2021), DestVI (Lopez et al. 2021, *BioRxiv*), STRIDE (Sun et al. 2021, *BioRxiv*), Cell2location (Kleshchevnikov et al. 2020, *BioRxiv*) and DSTG (Song et al. *Briefings in Bioinformatics* 2021).

2. We have now collected a total of **45 publicly available paired datasets** comprising both spatial transcriptomics and scRNA-seq data, and we have integrated them into our benchmarking study to evaluate the performance of different integration methods. Note that we also now provide an evaluation of **32 simulated datasets** (*i.e.*, where the ground truth is known).

3. We have now expanded the dimension and scope of our comparison amongst the distinct integration methods, which now includes performance in deconvoluting cell-types. We have also tried to expand our content in comparing integration methods for cell-cell interaction analyses in considerable detail, but found that the vast discrepancies in the results provided from different methods precludes the possibility of informative conclusion. We have illustrated our findings in this response letter for the reviewers' reference.

4. We have added more description and explanation for the applicability of the integration methods in our revised manuscript, and provide optimal parameters for the performance of the integration methods in this response letter; the corresponding scripts can be found at <https://github.com/QuKunLab/SpatialBenchmarking>.

5. Note that an updated version of Tangram was released recently (Oct 28, 2021). Accordingly, we have re-performed all of the benchmarking analysis related to Tangram.

NMETH-AS46278B
Point-by-point-response

Li et al., (Qu) p.2

6. We also followed the reviewer's suggestions to i) evaluate the impact of the normalized/non-normalized scRNA-seq expression matrices on the performance of different integration methods, ii) use the recommended Scuttle R package and the additional Splatter R package for the down-sampling content, and iii) provide a comprehensive assessment of computational efficiency for the integration methods in integrating 45 available paired spatial transcriptomics and scRNA-seq datasets.

7. We have revised the main text, methods, figure legends and overall language with the assistance of professional English editors to improve the manuscript's cogency and organization.

Reviewer #1:

Remarks to the Author:

This manuscript benchmarks 7 methods for integrating single-cell RNA-sequencing and spatial transcriptomics data. The benchmark is based on evaluation of these methods on 14 matched datasets consisting of scRNA-seq and spatial transcriptomics from the same tissue. Performing such a rigorous benchmarking study over such a diverse collection of data and methods is no easy task, and the authors are commended for their extensive efforts and valuable contribution. There is substantial interest in the use and development of methods to integrate single-cell and spatial sequencing data in order to maximize the relative advantages of both technologies. Thus, this manuscript serves an important need in the community.

At the same time, the manuscript has a number of issues that decrease my enthusiasm for this work.

1) A wide diversity of technology platforms are evaluated including 3 single-cell RNA sequencing technologies (dropSeq, smartSeq, and 10X Chromium) and 9 spatial sequencing technologies. While this diversity is reflective of the fast moving state of the technology, this diversity also complicates the interpretation of the benchmarking results. Are all of these methods expected to perform similarly on these different technologies? Presumably, some methods were designed/tuned to the characteristics certain technologies. These details are not described in the text. Moreover, it's not clear that the problem of "Integrate any one of 3 scRNA-seq technologies with any one of 9 spatial sequencing technologies" is a reasonable computational problem that any of these method address. In the benchmarking, the authors run each method with a single choice of parameters, but it seems reasonable that different parameters would be required for different technologies.

Response: We appreciate the supportive assessment of our study as filling a need in the community, and we thank the reviewer for focusing our attention on this important concern. We agree with the reviewer's comment that the diversity of sequencing technologies complicates the benchmarking study. To demonstrate the applicability of the integration methods to different sequencing technologies, we have employed the following strategies:

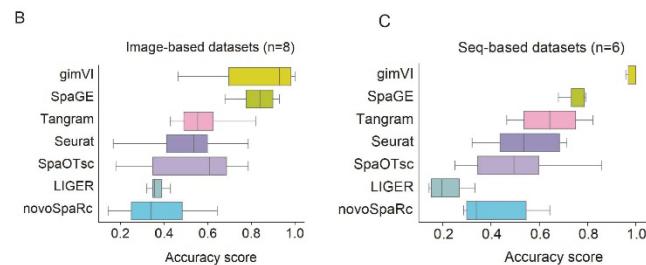
(1) We divided the spatial transcriptome datasets into two categories (seq-based technologies and image-based technologies) in our previous submission. We found that the rankings of the accuracy scores of these integration methods for processing these two categories of datasets are the same (**Supplementary Figure 2b, c in the previous submission**). We have now included 45 paired datasets and one extra integration method (stPlus) to predict the spatial distribution of undetected transcripts in the revised manuscript. We also found that the rankings of the integration methods are the same for the seq-based and image-based datasets (**revised Supplementary Figure 3a, b**).

(2) As 10X Visium, seqFISH, MERFISH, and Slide-seq have released more than 3 datasets, in the revised manuscript we added the comparison of the integration methods on the four types of spatial transcriptomics technologies (**revised Supplementary Figure 3d-g**). We found that Tangram, gimVI, and SpaGE outperformed other integration methods for datasets generated from 10X Visium, seqFISH, and MERFISH platforms, and Tangram and gimVI are top ranked methods in processing Slide-seq datasets. This result implies that based on the current collection of datasets, the diversity of the four spatial transcriptomics technologies did not affect the performance rankings of the integration methods.

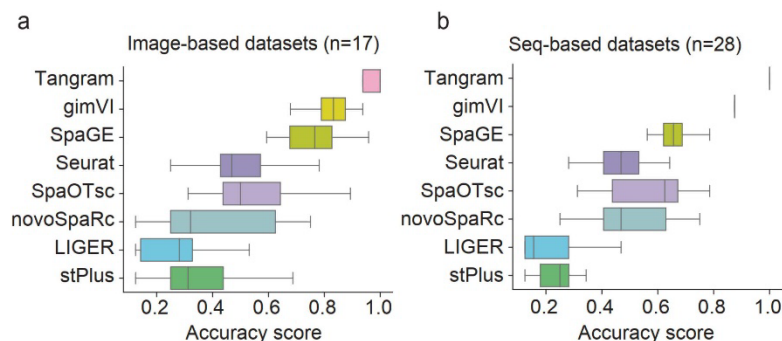
(3) In addition, we used several intrinsic parameters to denote the characteristics of the datasets generated by different spatial transcriptomics technologies, including the sparsity of expression matrix, number of measured genes, number of spots, size of spots, etc. We found that only the sparsity of the spatial datasets is linearly correlated with the Jensen-Shannon divergence (JS) values of the integration methods ($R^2 \geq 0.50$, P-values $< 10^{-6}$) (**revised Supplementary Figure 6**), but other intrinsic parameters have no significant correlation with the four-accuracy metrics of these methods (*i.e.* PCC, SSIM, RMSE, and JS). This trend still stands after we added more paired datasets and integration methods in the revised manuscript (**revised Supplementary Figure 6**).

(4) To find out whether the performance of the integration methods depends on parameters when processing different datasets, we searched the optimal parameters of each integration method when processing the 12 datasets used in our previous submission (**Table R1**), and then ranked the eight integration methods by their performance when using the optimal parameters (**Graph 1&2**). We found that regardless of whether the default parameters or the optimal parameters were used, the performance of Tangram (new version), gimVI, and SpaGE invariably outperform the other integration methods to predict the spatial distribution of undetected transcripts, and the rankings of other integration methods are also the same.

(5) In the original papers of three integration methods (gimVI, Seurat, and novoSpaRc), the authors simply stated that these methods can be applied on datasets from spatial transcriptomics technologies in general; however, their preference on different sequencing technologies was not clarified. Since the functions of these integration methods can be easily extended to process data obtained by different technologies, and aiming to make our benchmarking fair and practical, we evaluated the performance of these integration methods by applying them to all of the collected datasets. Now we also provide the original application scenarios of the integration methods and provide **Table R2** to describe the types of the spatial transcriptomics datasets used in the original paper for each method, as well as the gene numbers and spot numbers in these datasets.



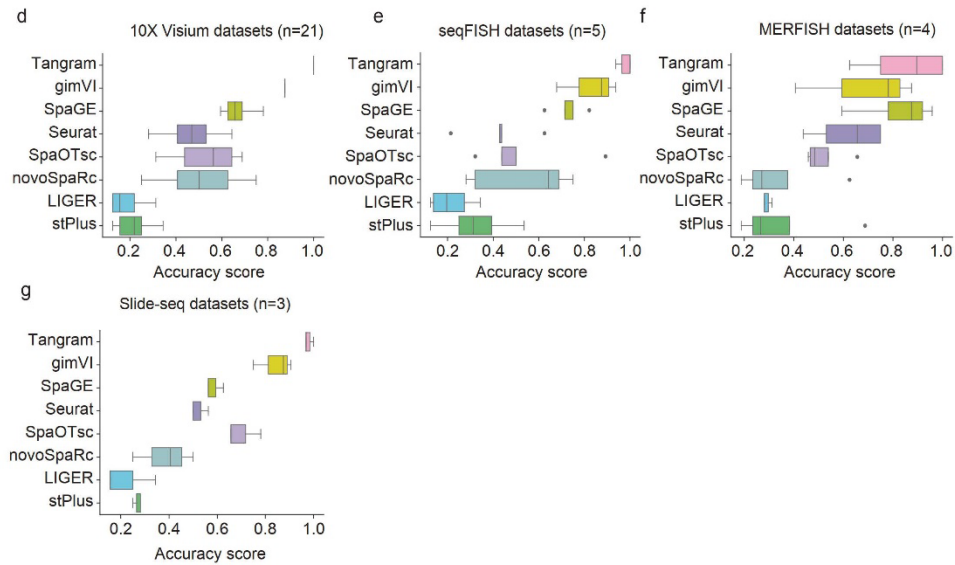
Supplementary Figure 2b, c in the previous submission. Boxplot of AS (accuracy score, aggregated from PCC, SSIM, RMSE, and JS) of the integration methods for transcripts in the Image-based datasets (**b**) and the Seq-based datasets (**c**). Center line, median; box limits, upper and lower quartiles; whiskers, 1.5× interquartile range.



Revised Supplementary Figure 3a, b. Boxplots of AS (accuracy score, aggregated from PCC, SSIM, RMSE, and JS; see Methods) of the integration methods for transcripts in the 28 image-based datasets (**a**) and 28 seq-based datasets (**b**). Center line, median; box limits, upper and lower quartiles; whiskers, 1.5× interquartile range.

NMETH-AS46278B
Point-by-point-response

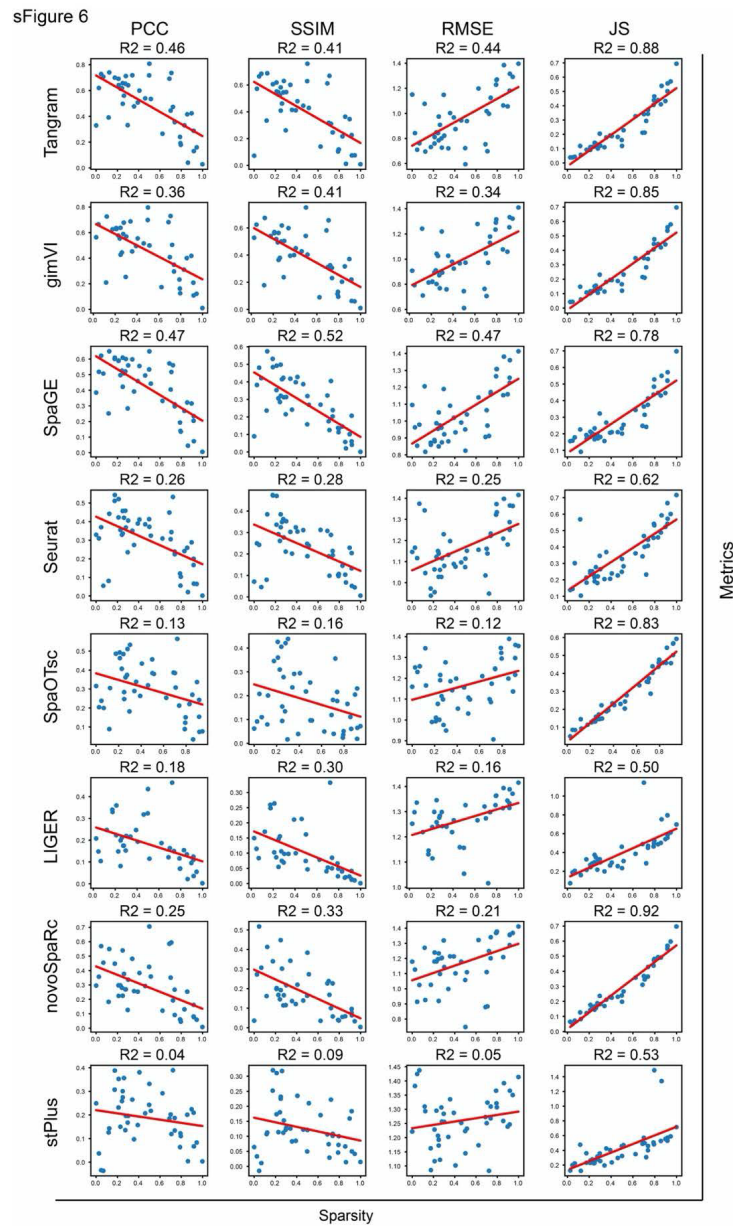
Li et al., (Qu) p.5



Revised **Supplementary Figure 3d-g**, Boxplots of AS (accuracy score, aggregated from PCC, SSIM, RMSE, and JS) of the integration methods for transcripts in the 21 10X visium datasets (d), 5 seqFISH datasets (e), 4 MERFISH datasets (f), 3 Slide-seq datasets (g). Center line, median; box limits, upper and lower quartiles; whiskers, $1.5 \times$ interquartile range.

NMETH-AS46278B
Point-by-point-response

Li et al., (Qu) p.6

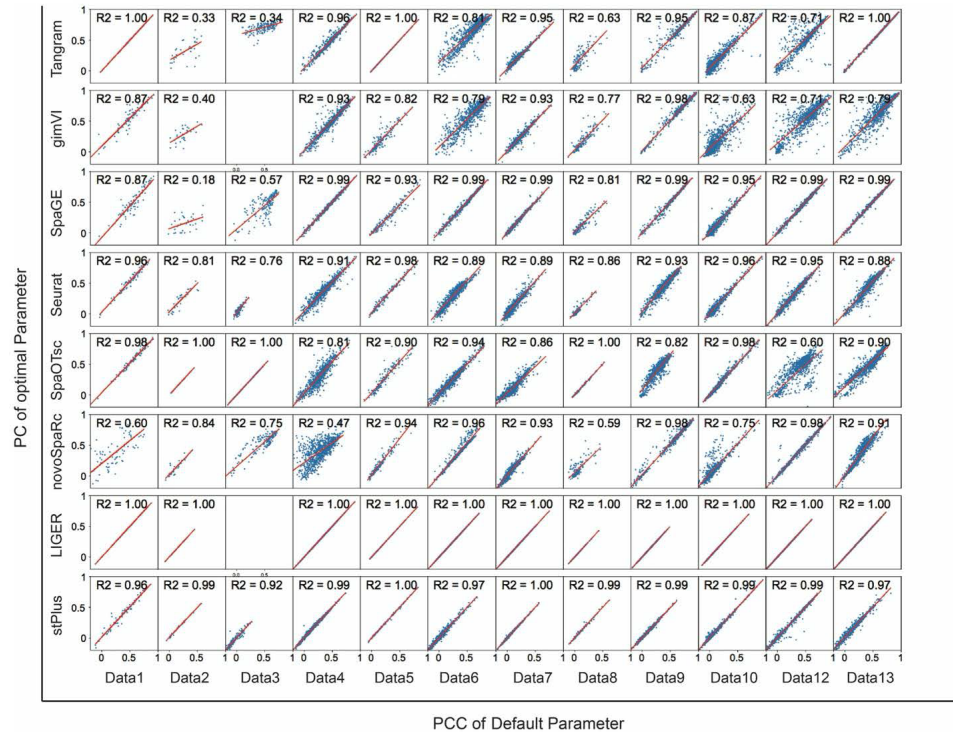


Revised Supplementary Figure 6. Correlation between the four metrics (PCC, SSIM, RMSE, and JS) and the sparsity of each examined spatial expression matrix. For all the eight integration methods that

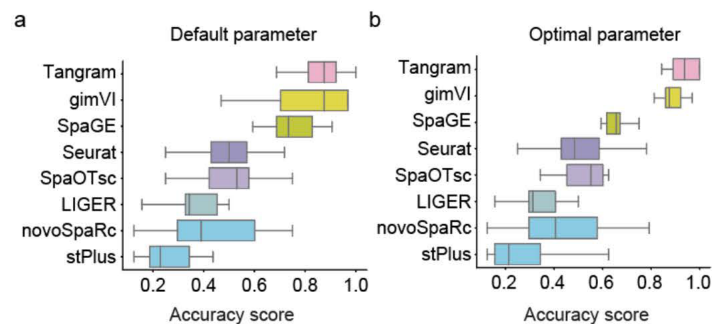
NMETH-AS46278B
Point-by-point-response

Li et al., (Qu) p.7

can predict the spatial distribution of transcripts, the JS values are linearly positively correlated with the sparsity of expression matrices of the spatial transcriptomics data ($R^2 \geq 0.50$).



Graph 1. Correlation between the prediction accuracy (PCC value) of using the default parameter and that of using the optimal parameter, including the eight integration methods that can predict the spatial distribution of undetected transcripts, and the 12 datasets used in our previous submission.



NMETH-AS46278B

Li et al., (Qu) p.8

Point-by-point-response

Graph 2. a, b. Boxplots of AS (accuracy score, aggregated from PCC, SSIM, RMSE, and JS; see Methods) of the integration methods in predicting the spatial distribution of undetected transcripts, using default parameters (**a**) and/or optimal parameters (**b**).

Table R1. Search range of the optimal parameters of each integration method.

Methods	Paramant	The range of tested parameters	The number of tested parameters
SpaGE	n_pv	Ratio : 0.1~1	10
	n_pv	10~100	10
LIGER	knn_k = 20	10, 20, 30	3
	weight = FALSE	FALSE, TRUE	2
	k	20, 30	2
	lambda = 5	5, 20	2
	knn_k = 20	10, 20, 30	3
	eps = 0.9	0.5, 0.7, 0.9	3
Seurat	dims = 1:30	20, 30, 40	3
	k.anchor = 5	5, 10, 15	3
	k.filter = 200	150, 200, 250, NA	4
	k.score = 30	20, 30, 40	3
	k.weight = 50	40, 50, 60	3
Tangram	cluster_label = cell	cell, clusters	2
	density_prior = None	uniform, ma_count_based	2
stPlus	neighbors = 50	40, 50, 60	3
	tmin = 5	1, 5, 9	3
gimVI	generative_distributions	['zinb', 'nb'], ['zinb', 'poisson'], ['nb', 'nb'], ['nb', 'poisson']	4
	model_library_size	[True, False], [True, True]	2
	n_latent	10, 20, 40	3
novoSpaRc	num_neighbors_source	5, 10, 20	3
	num_neighbors_target		
	locations_metric_p	1, 2	2
	expression_metric_p		
SpaOTsc	alpha	0.1, 0.3, 0.5, 0.7	4
	alpha	0	1
	epsilon	0.1, 0.4, 1	3

Table R2. Each integration method was designed for what type of spatial transcriptomics data, and the complexity of the datasets used in the original paper.

	Methods	Spatial transcriptome data used in the original paper	Gene #	Spot #	Original paper
Transcript spatial distribution	gimVI	smFISH, STARmap	33,166	4462, 3704	<i>arXiv:1905.02269</i> (2019)
	SpaGE	smFISH, MERFISH, seqFISH+, STARmap	33, 155, 10000, 1020	3405, 64373, 524, 1549	<i>Nucleic Acids Res.</i> 48(18), e107 (2020)
	LIGER	STARmap	1020	2500	<i>Cell</i> 177(7), 1873-1887.e17 (2019)
	stPlus	smFISH, MERFISH, STARmap	33, 155, 1020	3405, 64373, 1549	<i>Bioinformatics</i> 37, i289-i307 (2021)
Transcript spatial distribution & Cell type deconvolution	Tangram	MERFISH, STARmap, smFISH, Visium	254, 1020, 22, 31053	4951, 1550, 4840, (162, 161, 134)	<i>Nat. Methods</i> 18(11), 1352-1362 (2021)
	Seurat	STARmap	(1020, 1020)	(1539, 890)	<i>Cell</i> 177(7), 1888-1902.e21 (2019)
	SpaOTsc	STARmap, Slide-seq, seqFISH+, FISII	1020, 18838, 10000, (47, 84)	1549, 26316, 2050, (64, 3039)	<i>Nat. Commun.</i> 11(1), 2084 (2020)
	novoSpaRc	FISH, Slide-seq	84, NA	3639, 46376	<i>Nature</i> 576(7785), 132-137 (2019)
Cell type deconvolution	BC2D	Slide-seq	NA	NA	<i>Nat. Biotechnol.</i> doi: 10.1038/s41587-021-00830-w (2021)
	SpatialDWLS	seqFISH+, Visium, ST	10000, 6018, 1700	523, 2698, 3115	<i>Genome Biology</i> 22(1):145 (2021)
	Stereoscope	Visium, ST, Slide-seq	39739, NA, (19869, 19130)	1515, NA, (38666, 29275)	<i>Commun. Biol.</i> 3(1), 565 (2020)
	SPOTlight	ST	(19738, 25753, 32285, 32285, 32285, 32285)	(321, 956, 2695, 2825, 3355, 3289)	<i>Nucleic Acids Res.</i> 49(9), e50 (2021)
	DeaVI	Visium	(32285, 32285, 32285, 32285)	(2199, 2113, 841, 698)	<i>BioRxiv</i> doi:10.1101/2021.05.10.445517
	cell2location	Visium, slide-seq	33538, NA	4639, NA	<i>BioRxiv</i> doi:10.1101/2020.11.15.378125
	STRIDE	Slide-seqV2, ST	NA, (17344, 39739)	NA, (1933, 3111)	<i>BioRxiv</i> doi:10.1101/2021.09.08.459458
	DSTG	Pseudo-ST	500-5000	500-4000	<i>Brief. Bioinform.</i> 22(5), bbab14 (2021)

NMETH-AS46278B
Point-by-point-response

Li et al., (Qu) p.9

2) Related to point 1, it is likely that methods would perform differently with different numbers of genes and with different spatial organization. Some discussion of these points would be helpful to guide the reader's expectations. What sequencing technologies, number of genes, number of spots is each method claimed to analyze as stated in the original publication or software documentation?

Response: We thank the reviewer for helping us to re-consider the applicability of the integration methods in light of this important concern. We provide **Table R2** to describe the applicability (including sequencing technologies, number of genes, number of spots) of the integration methods, which can be seen in our response to comment 1. We also added the following content in the **Discussion** section to guide the reader's expectations in the revised manuscript (Line 374-389, highlighted in yellow), which reads as follows:

"The wide diversity and fast-moving state of scRNA-seq technologies (DropSeq, SmartSeq, and 10X Chromium) and spatial sequencing technologies (FISH, osmFISH, seqFISH, MERFISH, STARmap, ISS, EXseq, BaristaSeq, ST, 10X Visium, Slide-seq, Seq-scope, and HDST) complicates the task of this benchmarking analysis. We adopted three distinct perspectives to overcome the diversity of sequencing technologies: (1) We divided the spatial transcriptome datasets into two categories (seq-based technologies and image-based technologies); (2) we added an independent comparison of the integration methods for the set of four spatial transcriptomics technologies (MERFISH, seqFISH, Slide-seq, and 10X Visium) that have to date released > three datasets; (3) we used several intrinsic parameters (e.g., the number of captured genes, the number of captured spots, and the sparsity of the expression matrix) to depict the characteristics of the datasets generated by different spatial transcriptomics technologies. There are still aspects (i.e., different numbers of genes and different spatial organization) that may affect the performance of the integration methods and the user's expectations. Nevertheless, based on the current collection of datasets, we found that the performance rankings of these integration methods are barely affected by the spatial transcriptome technologies that generated these data."

3) A whole class of recent methods that perform deconvolution of spatial transcriptomics data into single-cell resolution is missing from the comparison such as BayesSpace (Zhao et al. Nature Biotech. 2021), RCTD (Cable et al. Nature Biotech. 2021), Steroscope (Andersson et al. Commun. Biol. 2020), SPOTlight (Elosua-Bayes, et al. NAR 2021), and destVI (Lopez et al. 2021, bioRxiv). [Note that Lopez et al was posted as preprint only recently so not expected

NMETH-AS46278B
Point-by-point-response

Li et al., (Qu) p.10

that authors would include.]. It's not clear what criteria were used to select methods for benchmarking and why not deconvolution methods were included. Presumably the evaluation criteria for deconvolution methods would be different from the criteria used in the present publication.

Response: We thank the reviewer for highlighting this important concern. In our previous submission, we selected the integration methods that can predict the spatial distribution of undetected transcripts or assign cells from scRNA-seq data to spatial locations. In the revised manuscript, we selected the integration methods that can predict the spatial distribution of undetected transcripts or perform the cell-type deconvolution of spots of the transcriptome data; accordingly, more integration methods are now included. These criteria are clarified in our revised manuscript (Line 93-97, highlighted in yellow), which reads as follows:

"Here, we used multiple metrics to systematically benchmark the performance of sixteen integration methods that can predict the spatial distribution of undetected transcripts, or the cell type composition of spots in histological sections (Fig. 1a), based on processing of 45 paired datasets containing both spatial transcriptomics data and scRNA-seq data and 32 simulated datasets (Fig. 1b)."

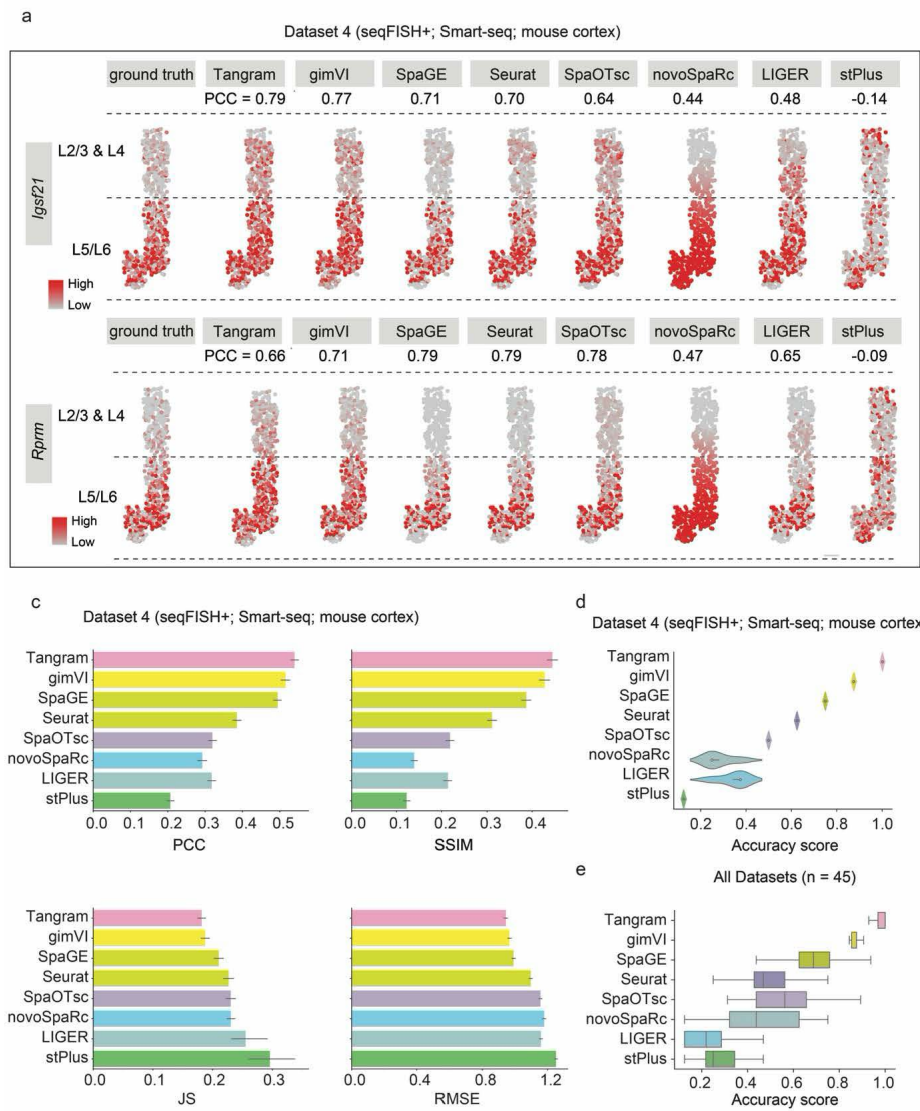
Guided by the reviewers, we have now included comparisons of eight additional methods that can perform cell type deconvolution, including SpatialDWLS (Dong et al. *Genome Biology* 2021), RCTD (Cable et al. *Nature Biotechnology* 2021), Stereoscope (Andersson et al. *Communication Biology* 2020), SPOTlight (Elosua-Bayes et al. *Nucleic Acids Research* 2021), DestVI (Lopez et al. 2021, *BioRxiv*), Cell2location (Kleshchevnikov et al. 2021, *BioRxiv*), DSTG (Song et al. *Briefings in Bioinformatics* 2021), and STRIDE (Sun et al. 2021, *BioRxiv*) in our benchmark analyses, although it bears mention that the last three of these methods have not yet been published in peer-reviewed journals. We also added stPlus (Chen et al. *Bioinformatics* 2021) in the revised manuscript to benchmark its performance for predicting the spatial distribution of undetected transcripts. Note that BayesSpace (Zhao et al. *Nature Biotechnology* 2021) is designed for spatial resolution enhancement and cell clustering, but not for cell type deconvolution (nor the prediction of undetected transcripts), so it was not included in this benchmarking study.

We then examined the performance of these integration methods on a total of 45 publicly available paired datasets (comprising both spatial transcriptomics and scRNA-seq data) and 32 simulated datasets. We found that Tangram (new version), gimVI, and SpaGE outperformed other integration methods for predicting the spatial distribution of gene expression (revised **Figure 2&3**), while Cell2location, SpatialDWLS and RCTD outperformed other methods for cell type deconvolution of spots in histological sections (revised **Figure 4**).

NMETH-AS46278B
Point-by-point-response

Li et al., (Qu) p.11

Figure 2

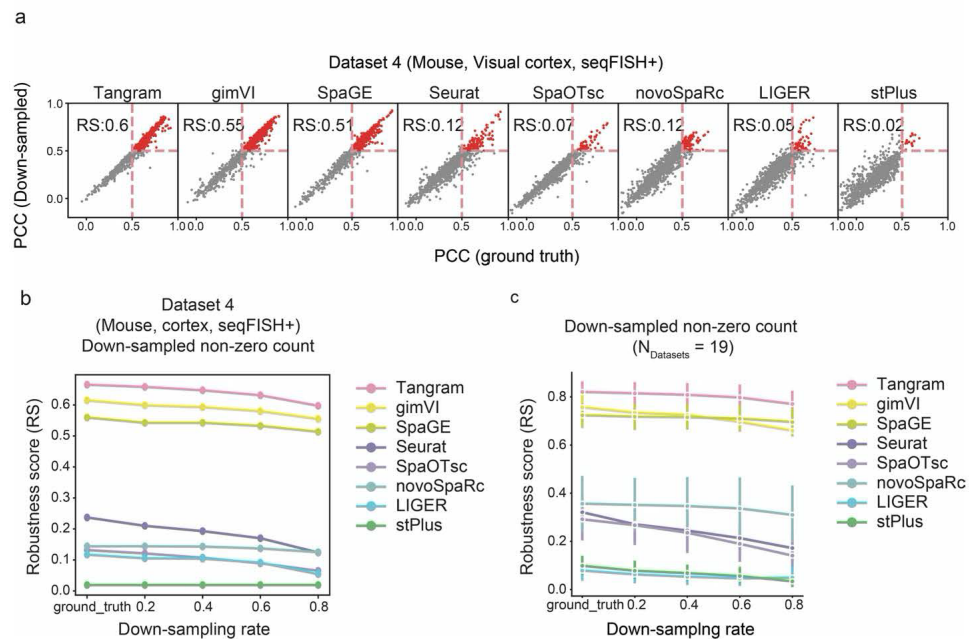


NMETH-AS46278B
Point-by-point-response

Li et al., (Qu) p.12

Revised Figure 2. Comparing the accuracy of eight integration methods capable of predicting the spatial distribution of RNA transcripts. **a,b**, The spatial distribution of *Igsf21* (a) and *Rprm* (b) in dataset 4 (seqFISH+; Smart-seq; mouse cortex), including the ground truth and the predicted result from each of the integration methods. PCC: Pearson Correlation Coefficient between the expression vector of a transcript in the ground truth and that of the predicted result. **c**, The bar plots of PCC, SSIM, RMSE, and JS of each integration method in predicting the spatial distribution of transcripts in dataset 4. SSIM: Structural Similarity Index; RMSE: Root Mean Square Error; JS: Jensen-Shannon divergence. The confidence interval represented by the error bar is 95%. **d**, Violin plots of AS (accuracy score, which is aggregated from the PCC, SSIM, RMSE, and JS values; see Methods) of the eight integration methods for transcripts in dataset 4. Center line, median; box limits, upper and lower quartiles; whiskers, $1.5 \times$ interquartile range. **e**, Boxplots of AS of the eight integration methods for all 45 paired datasets. Center line, median; box limits, upper and lower quartiles; whiskers, $1.5 \times$ interquartile range.

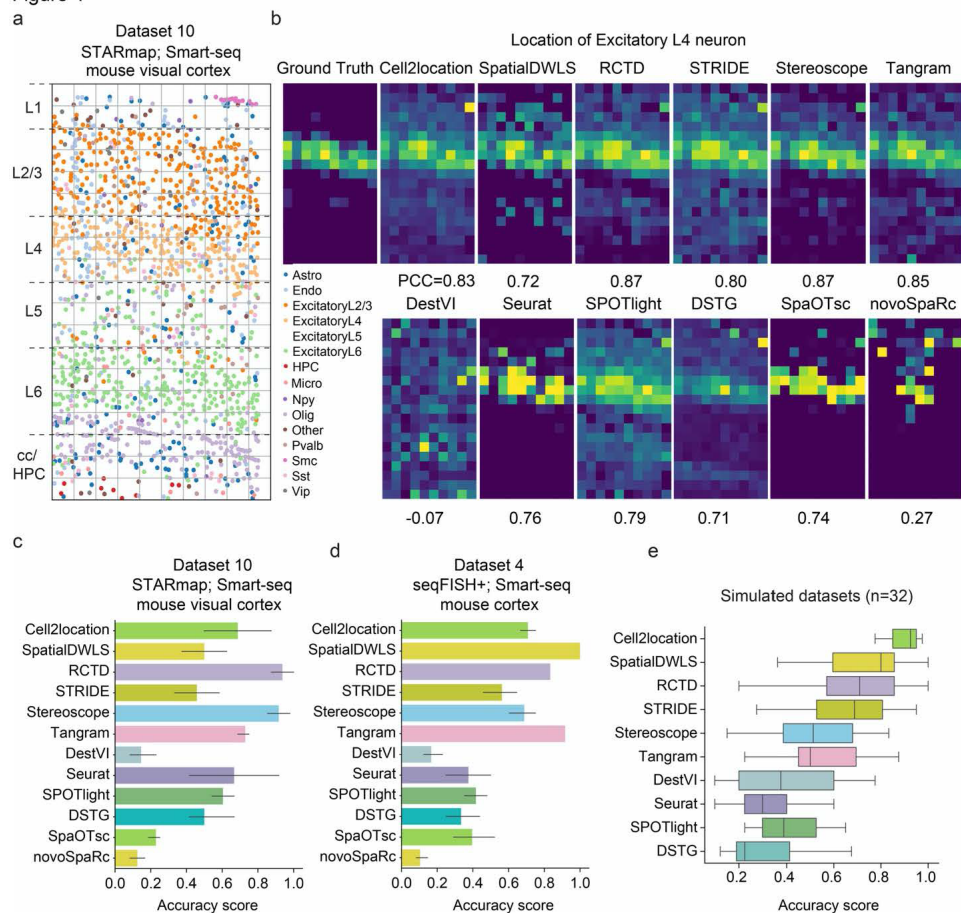
Figure 3



Revised Figure 3. Comparing the accuracy of the eight integration methods for sparse spatial expression matrices down-sampled from the original datasets using Splatter. **a**, PCC of the spatial distribution of transcripts predicted from the original data and down-sampled data from dataset 4. The PCC values of the red-colored transcripts are greater than 0.5 for both the original and the down-sampled data.

The proportion of the red-colored transcripts in all transcripts was defined as the “robustness score” (RS). **b**, RS values of the eight integration methods when processing sparse expression matrices down-sampled from dataset 4 at different down-sampling rates. **c**, RS values of the eight integration methods when processing the sparse expression matrices of the down-sampled datasets. The original datasets (used to generate the down-sampled datasets) capture >1000 genes from >100 spots, and the sparsity of the expression matrices is <0.7. The confidence interval represented by the error bar is 90%.

Figure 4



Revised Figure 4. Comparing the performance of the twelve integration method capable of deconvoluting cell types of each histological spot. a, A STARmap slide of dataset 10 (STARmap; Smart-seq; mouse visual cortex), with cells annotated by cell type. Each grid represents a simulated spot containing multiple cells. **b**, The proportion of L4 excitatory neurons in the spots simulated from dataset 10, including the ground truth and the predicted results of twelve integration methods. **c,d**, Barplots of AS (accuracy score, aggregated from PCC, SSIM, RMSE, and JS; see Methods) of the cell type composition of the

NMETH-AS46278B
Point-by-point-response

Li et al., (Qu) p.14

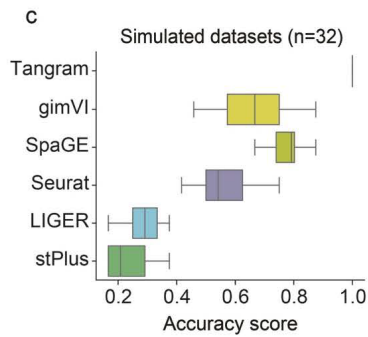
histological spots simulated from dataset 10 (c) and dataset 4 (d), predicted by twelve integration methods. The confidence interval represented by the error bar is 95%. e, Boxplots of AS of the ten integration methods for all the 32 simulated datasets. SpaOTsc and novoSpaRc are excluded, as they require spatial location information for each spot, which is not available in the simulated datasets. Center line, median; box limits, upper and lower quartiles; whiskers, $1.5 \times$ interquartile range.

4) The evaluation metrics for the benchmark are based on a comparison to the matched spatial transcriptomics data. As a complement to evaluation of methods on real data, the authors might consider an evaluation on simulated data where there is a ground truth to evaluate. While simulations obviously cannot match all of the features of real data, benchmarking papers routinely use simulated data; e.g. see Saelens et al. *Nature Biotechnology* (2019) benchmarking of trajectory inference methods;

Response: We thank the reviewer for this highly constructive suggestion; we have now provided an evaluation on **32** simulated datasets where there is a ground truth. We used the approach introduced by RCTD (Cable et al. *Nature Biotechnology* 2021) and Stereoscope (Andersson et al. *Communication Biology* 2020) to construct the simulation datasets, and the details has been added to the **Methods** section of the revised manuscript (Line 872-886, highlighted in yellow), which reads as follows:

"For the 32 simulated datasets, we devised the generation procedure of spatial transcriptomics data simulation by referring to the algorithm introduced by RCTD and Stereoscope. For each simulated spot, we first sampled cell numbers (N_c) in a uniform distribution in the range 5-15, and sampled the number of cell types (N_t) in a uniform distribution in the range 2-6. Then we assumed that these cell types have equal distribution possibility $P=1/N_t$ in the spot, and randomly assigned cells from each cell types of the scRNA-seq data to the spot. To obtain the gene expression values at each spatial location, we summed the gene expression values of all cells in one spot. Referring to the method for constructing simulated datasets used in RCTD, we used Scuttle (<http://bioconductor.org/packages/release/bioc/html/scuttle.html>) to down-sample the number of counts per spot to 10% of the original value. We can obtain the percentage of a cell type at each spot by counting the number of cells corresponding to the cell type. For the large simulated dataset used for assessing the efficiency of each integration method, we used the same algorithm as above, but set the number of spots to 20000, the number of cells to 10000, and the number of cell types to 56."

We quantified the prediction performance of each integration method on simulated datasets by calculating the accuracy score (AS), which was defined by aggregating the four metrics (PCC, SSIM, RMSE, and JS) to simplify the evaluation of the accuracy of each integration method. Here we excluded novoSpaRe and SpaOTsc, as the two methods required location information of spots, which is not available for the simulated datasets. We found that the new version of Tangram, gimVI, SpaGE outperformed the other integration methods for predicting the spatial distribution of unmeasured gene. Cell2location, SpatialDWLS, and RCTD outperformed other integration methods for the cell-type deconvolution of spots (**revised Supplementary Figure 3c & revised Figure 4**).

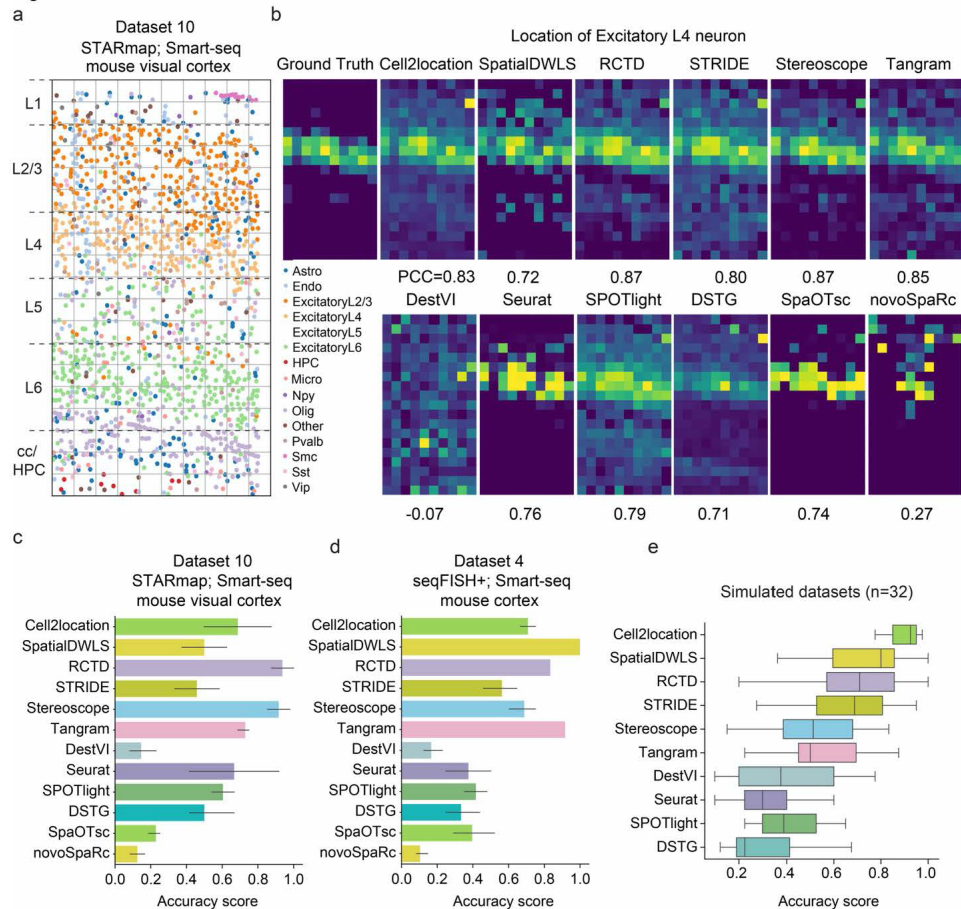


Revised Supplementary Figure 3c, Boxplots of AS (accuracy score, aggregated from PCC, SSIM, RMSE, and JS) of the integration methods for transcripts in the 32 simulated datasets. Center line, median; box limits, upper and lower quartiles; whiskers, 1.5× interquartile range.

NMETH-AS46278B
Point-by-point-response

Li et al., (Qu) p.16

Figure 4



Revised Figure 4. Comparing the performance of the twelve integration method capable of deconvoluting cell types of each histological spot. **a**, A STARmap slide of dataset 10 (STARmap; Smart-seq; mouse visual cortex), with cells annotated by cell type. Each grid represents a simulated spot containing multiple cells. **b**, The proportion of L4 excitatory neurons in the spots simulated from dataset 10, including the ground truth and the predicted results of twelve integration methods. **c,d**, Barplots of AS (accuracy score, aggregated from PCC, SSIM, RMSE, and JS; see Methods) of the cell type composition of the histological spots simulated from dataset 10 (**c**) and dataset 4 (**d**), predicted by twelve integration methods. The confidence interval represented by the error bar is 95%. **e**, Boxplots of AS of the ten integration methods for all the 32 simulated datasets. SpaOTsc and novoSpaRc are excluded, as they require spatial location information for each spot, which is not available in the simulated datasets. Center line, median; box limits, upper and lower quartiles; whiskers, $1.5 \times$ interquartile range.

5) Related to the above points, it is arguably premature to perform a comprehensive benchmarking study of spatial integration methods since: the sequencing technologies are too diverse, the scope of each method is limited to a subset of technologies, the goals of the methods (cell assignment, deconvolution, etc) are too varied, both the technologies and methods are evolving too rapidly, and the standards are unclear. As a point of comparison: the Saelens et al. benchmarking of trajectory inference methods benchmarked 45 of the 70 known methods on 110 real and 229 synthetic datasets, explicitly described the trade offs of different methods according to topology of the trajectory, and included guidelines for users to choose a method based on the topology of the trajectory (cf. Fig. 5). By comparison, the present benchmarking study is relatively limited in scope and doesn't provide much practical guidance for the biologist who wants to analyze their data or the methods developer who wants to improve the state-of-the-art.

Response: We thank the reviewer for focusing our attention on this important concern, and agree that both the technologies and methods in spatial transcriptomics analysis are evolving very rapidly. For instance, we were able to gather ~3 times more datasets and 9 more analytical tools for benchmarking since the first submission of our manuscript. Nevertheless, and indeed because of this context, we believe there is urgent need to systematically benchmark the currently available tools “for the biologist who wants to analyze their data or the methods developer who wants to improve the state-of-the-art”.

Inspired by the reviewer's comment and Saelens et al.'s benchmarking study on trajectory inference methods, we have now included 45 paired scRNA-seq and spatial transcriptomics datasets and 32 simulated datasets for the benchmarking of the integration methods in the revised manuscript. We have also included the comparisons of eight additional integration methods that can perform cell type deconvolution of spatial transcriptomics data, including SpatialDWLS (Dong et al. *Genome Biology* 2021), RCTD (Cable et al. *Nature Biotechnology* 2021), Stereoscope (Andersson et al. *Communication Biology* 2020), SPOTlight (Elosua-Bayes et al. *Nucleic Acids Research* 2021), DestVI (Lopez et al. 2021, *BioRxiv*), Cell2location (Kleshchevnikov et al. 2021, *BioRxiv*), DSTG (Song et al. *Briefings in Bioinformatics* 2021) and STRIDE (Sun et al. 2021, *BioRxiv*), although the last three have not been published on peer-reviewed journals yet, for deconvoluting cell-types of each spot, and one additional integration method stPlus (Chen et al. *Bioinformatics* 2021) for predicting the spatial distribution of undetected transcripts.

During our revision, after including 9 additional integration methods, 45 publicly available paired datasets, and 32 simulated datasets, we re-performed all the comparison analysis among the integration methods and have accordingly updated both the figures and the manuscript text. Note that the most recently released version of Tangram has become one of the best performing methods

for predicting the spatial distribution of undetected transcripts (**revised Figure 2&3**). Cell2location, SpatialDWLS, and RCTD outperformed other integration methods for the cell type deconvolution of spots (**revised Figure 4**). Thus, the benchmarking study provided practical guidance for the biologists who want to choose an appropriate method to integrate the spatial transcriptomics data and the scRNA-seq data, and to predict the spatial distribution of undetected transcripts or the cell type composition of spots in histological sections.

Further, and as mentioned in our response to the Reviewer's comment 1, we have now adopted three distinct perspectives to overcome the diversity of sequencing technologies:

(1) We divided the spatial transcriptome datasets into two categories (seq-based technologies and image-based technologies) in our previous submission. We found that the rankings of the accuracy scores of these integration methods for processing these two categories of datasets are the same (**Supplementary Figure 2b, c** in the previous submission). We have now included 45 paired datasets and one extra integration method (stPlus) to predict the spatial distribution of undetected transcripts in the revised manuscript. We also found that the rankings of the integration methods are the same for the seq-based and image-based datasets (**revised Supplementary Figure 3a, b**).

(2) Since 10X Visium, seqFISH, MERFISH, and Slide-seq have released more than 3 datasets, in the revised manuscript we added the comparison of the integration methods on the four types of spatial transcriptomics technologies (**revised Supplementary Figure 3d-g**). We found that Tangram, gimVI, and SpaGE outperformed other integration methods for datasets generated from 10X Visium, seqFISH, and MERFISH platforms and Tangram and gimVI are top ranked methods in processing Slide-seq datasets. This result implies that based on the current collection of datasets, the diversity of the four spatial transcriptomics technologies did not affect the performance rankings of the integration methods.

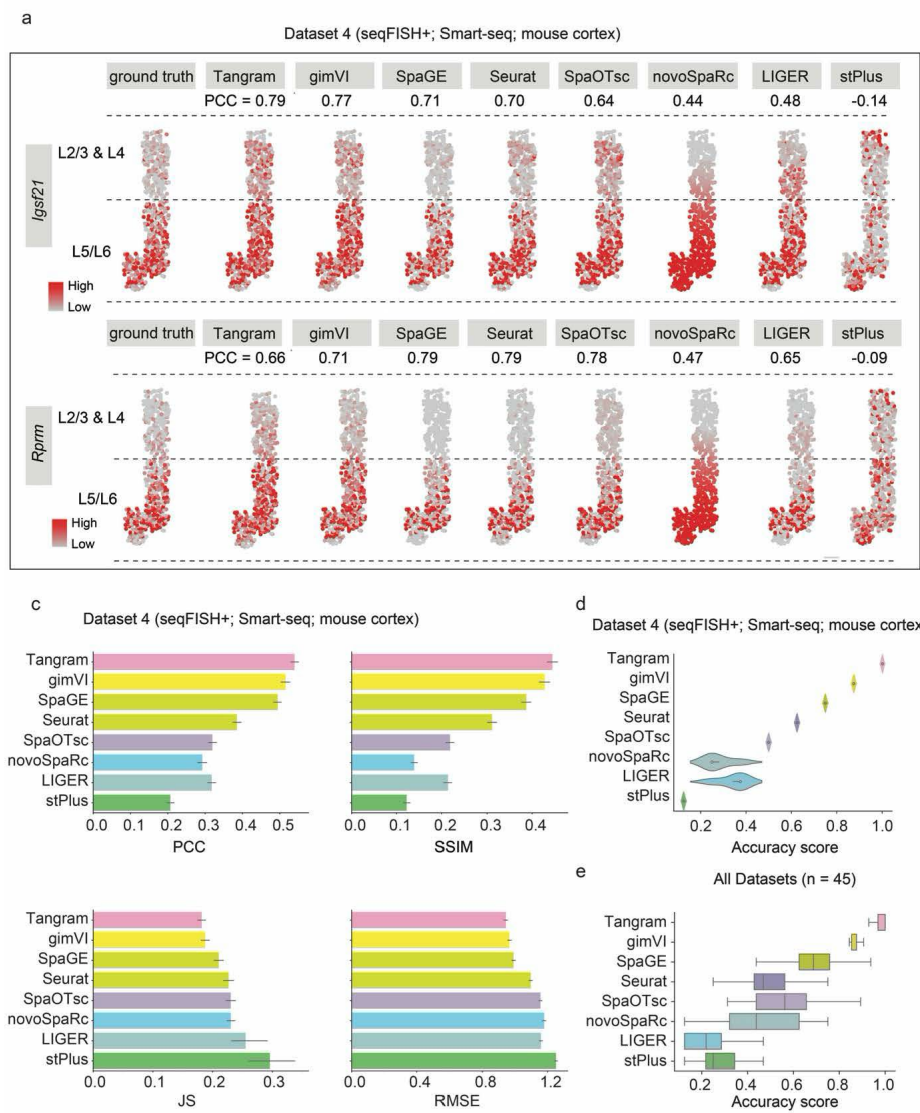
(3) In addition, we used several intrinsic parameters to denote the characteristics of the datasets generated by different spatial transcriptomics technologies, including the sparsity of expression matrix, number of measured genes, number of spots, size of spots, etc. We found that only the sparsity of the spatial datasets is linearly correlated with the Jensen-Shannon divergence (JS) values of the integration methods ($R^2 \geq 0.50$, P-values $< 10^{-6}$) (**revised Supplementary Figure 6**), but other intrinsic parameters have no significant correlation with the four-accuracy metrics of these methods (*i.e.* PCC, SSIM, RMSE, and JS). This trend still stands after we added more paired datasets and integration methods in the revised manuscript (**revised Supplementary Figure 6**).

In summary, based on the current collection of datasets, the performance rankings of these integration methods are barely affected by the spatial transcriptome technologies that generated these datasets.

NMETH-AS46278B
Point-by-point-response

Li et al., (Qu) p.19

Figure 2

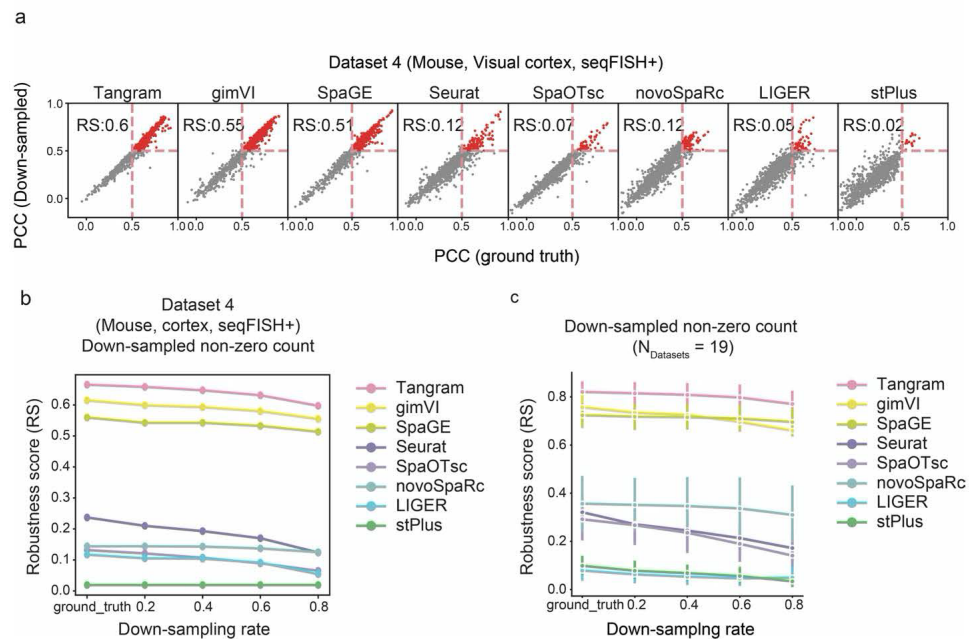


NMETH-AS46278B
Point-by-point-response

Li et al., (Qu) p.20

Revised Figure 2. Comparing the accuracy of eight integration methods capable of predicting the spatial distribution of RNA transcripts. **a,b**, The spatial distribution of *Igsf21* (a) and *Rprm* (b) in dataset 4 (seqFISH+; Smart-seq; mouse cortex), including the ground truth and the predicted result from each of the integration methods. PCC: Pearson Correlation Coefficient between the expression vector of a transcript in the ground truth and that of the predicted result. **c**, The bar plots of PCC, SSIM, RMSE, and JS of each integration method in predicting the spatial distribution of transcripts in dataset 4. SSIM: Structural Similarity Index; RMSE: Root Mean Square Error; JS: Jensen-Shannon divergence. The confidence interval represented by the error bar is 95%. **d**, Violin plots of AS (accuracy score, which is aggregated from the PCC, SSIM, RMSE, and JS values; see Methods) of the eight integration methods for transcripts in dataset 4. Center line, median; box limits, upper and lower quartiles; whiskers, $1.5 \times$ interquartile range. **e**, Boxplots of AS of the eight integration methods for all 45 paired datasets. Center line, median; box limits, upper and lower quartiles; whiskers, $1.5 \times$ interquartile range.

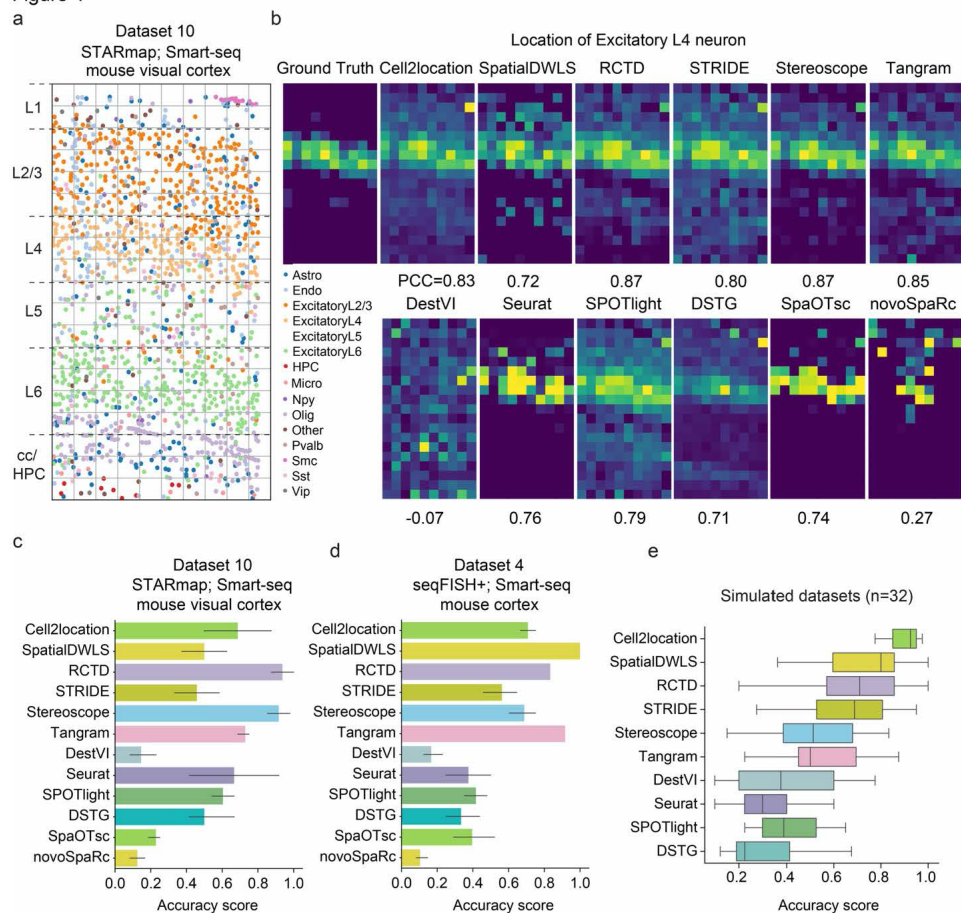
Figure 3



Revised Figure 3. Comparing the accuracy of the eight integration methods for sparse spatial expression matrices down-sampled from the original datasets using Splatter. **a**, PCC of the spatial distribution of transcripts predicted from the original data and down-sampled data from dataset 4. The PCC values of the red-colored transcripts are greater than 0.5 for both the original and the down-sampled data.

The proportion of the red-colored transcripts in all transcripts was defined as the “robustness score” (RS). **b**, RS values of the eight integration methods when processing sparse expression matrices down-sampled from dataset 4 at different down-sampling rates. **c**, RS values of the eight integration methods when processing the sparse expression matrices of the down-sampled datasets. The original datasets (used to generate the down-sampled datasets) capture >1000 genes from >100 spots, and the sparsity of the expression matrices is <0.7. The confidence interval represented by the error bar is 90%.

Figure 4



Revised Figure 4. Comparing the performance of the twelve integration method capable of deconvoluting cell types of each histological spot. a, A STARmap slide of dataset 10 (STARmap; Smart-seq; mouse visual cortex), with cells annotated by cell type. Each grid represents a simulated spot containing multiple cells. **b**, The proportion of L4 excitatory neurons in the spots simulated from dataset 10, including the ground truth and the predicted results of twelve integration methods. **c**, **d**, Barplots of AS (accuracy score, aggregated from PCC, SSIM, RMSE, and JS; see Methods) of the cell type composition of the

NMETH-AS46278B
Point-by-point-response

Li et al., (Qu) p.22

histological spots simulated from dataset 10 (c) and dataset 4 (d), predicted by twelve integration methods. The confidence interval represented by the error bar is 95%. **e**, Boxplots of AS of the ten integration methods for all the 32 simulated datasets. SpaOTsc and novoSpaRc are excluded, as they require spatial location information for each spot, which is not available in the simulated datasets. Center line, median; box limits, upper and lower quartiles; whiskers, $1.5 \times$ interquartile range.

6) Given the rapid developments in spatial sequencing technologies — with 10X Visium scheduled to reach single-cell resolution next year and Slide-seq v2 approaching scRNA-seq — it's possible that interest in integration of scRNA-seq and spatial transcriptomics will be short lived. It would be helpful to hear further perspective on the future based on the benchmarking.

Response: We thank the reviewer for this food for thought. Based on the technical details of 10X Visium scheduled in 2022 (spot, $\sim 4\mu\text{m}$), Slide-seq v2 (spot, $\sim 10\mu\text{m}$), and BGI Stereo-seq (spot, $\sim 200\text{nm}$), each spot contains cell fragments from one to multiple cells. As each spot will not correspond exactly to a single cell based on these anticipated technologies in the near future, we tend to believe that the integration of scRNA-seq data and spatial transcriptomics data will remain necessary and will still be challenging.

In addition, consider that there are a very large number of academic studies using spatial transcriptomics technologies that are now underway yet have not been reported in the literature; some of these are near-certain to represent seminal resources that enable step-change advances in their respective research fields. Moreover, despite the rapid developments in spatial sequencing technologies, the publicly available spatial transcriptomics datasets maintain high value to a plethora of research communities. Thus, we contend there will still be a need and strong interest for the integration of spatial transcriptomics and scRNA-seq data to determine the spatial distribution of cells or undetected transcripts.

We also added the following content in the revised **Discussion** section (Line 390-400, highlighted in yellow):

"Our benchmarking study evaluated sixteen integration methods on 45 paired datasets consisting of scRNA-seq and spatial transcriptomics from the same tissue and 32 simulated datasets where the ground truth is known. Despite the rapid developments in spatial sequencing technologies, it bears emphasis that many currently available and soon-to-be-published public spatial transcriptomics datasets will almost certainly be strongly influential in the research agendas of many scientific communities. In summary, this study presents a much-needed independent comparison of available integration methods of spatial transcriptomics and scRNA-seq data to determine the cell type deconvolution or the spatial distribution of undetected genes,

NMETH-AS46278B
Point-by-point-response

Li et al., (Qu) p.23

and also provides a need and strong interest for the biologist who wants to analyze their data or the methods developer who wants to improve the state-of-the-art."

Minor comments:

- Line 464: "10-X cross validation"—> "10-fold cross validation"

Response: We apologize for this mistake. We have changed "10-X cross validation" to "10-fold cross validation" in the revised submission.

- Figure 1b. Recommend adding references to this table so readers can quickly see source of datasets.

Response: We thank the reviewer for this helpful recommendation. We have added references to the table in revised **Supplementary Table 1**.

We would like to take this opportunity to express our gratitude to the reviewer for the very helpful guidance about how to improve our study. Many thanks.

Reviewer #2:

Remarks to the Author:

In this work, the authors present a benchmarking study of methods used to integrate dissociated single cell RNA sequencing and spatial transcriptomic data. The authors evaluated 7 methods for their ability to predicted spatial expression patterns of undetected transcripts and for a subset of 4 methods their ability to predict cell locations. Based on several evaluation metrics, the authors conclude that gimVI and SpaGE have the best performance in predicting undetected transcripts and Seurat for the prediction of spatial locations.

Overall, the paper provides a much-needed independent comparison of available methods to integrate single cell and spatial data. The paper is well written and the experiments are clearly explained. The claims made by the authors are backed by the experiments. I would also like to compliment the authors on the associated GitHub repository.

NMETH-AS46278B
Point-by-point-response

Li et al., (Qu) p.24

My main concern is that the authors chose to compare very specific aspects of integration methods (predicting undetected transcripts and spatial localization) while other equally important tasks are ignored. I elaborate on this below in my major comments.

Major comments:

1) The evaluation of integration methods focuses on two tasks (predicting undetected transcripts and spatial localization) while other tasks (e.g. cell-cell interaction deconvolution) are completely ignored. These are equally important aspects of methods integrating single cell and spatial transcriptomic data. At least the author should make this explicitly clear in the title, abstract and discussion.

Response: We agree with this point about the additional tasks of integration methods, and appreciate the reviewer's bringing this omission to our attention. Accordingly, we have now properly presented the scope of integration method tasks in our revised manuscript, and have modified our title, abstract, and discussion to reflect the specific content that we investigated in our study (detailed below).

(1) We have changed the title of the manuscript to

“Benchmarking spatial and single-cell transcriptomics integration methods for transcript distribution prediction and cell type deconvolution”

(2) We have revised the **Abstract** in the revised manuscript (Line 27-41, highlighted in yellow), which now reads as follows:

“Spatial transcriptomics approaches have substantially advanced our capacity to detect the spatial distribution of RNA transcripts in tissues, yet it remains challenging to characterize data at whole-transcriptome-level data for single cells in space. Addressing this need, researchers have developed integration methods to combine spatial transcriptomic data with single-cell RNA-seq data to i) predict the spatial distribution of undetected transcripts and/or ii) perform cell type deconvolution of spots in histological sections. However, to date no studies have comparatively analyzed these integration methods to benchmark their performance. Here, we examined 45 paired datasets (comprising both spatial transcriptomics and scRNA-seq data) and 32 simulated datasets and found that Tangram, gimVI, and SpaGE outperformed other integration methods for predicting the spatial distribution of RNA transcripts, while Cell2location, SpatialDWLS, and RCTD are the top-performing methods for the cell type deconvolution of spots in histological sections. Based on these comparative results from data obtained using highly diverse spatial transcriptomics approaches, we provide example guidance (<https://github.com/QuKunLab/SpatialBenchmarking>) to help researchers select optimal integration methods for working with their datasets.”

(3) We have added a clear statement on the selection criteria of integration methods in the revised manuscript (Line 93-97, highlighted in yellow), which reads as follows:

"Here, we used multiple metrics to systematically benchmark the performance of sixteen integration methods that can predict the spatial distribution of undetected transcripts, or the cell type composition of spots in histological sections (Fig. 1a), based on processing of 45 paired datasets containing both spatial transcriptomics data and scRNA-seq data and 32 simulated datasets (Fig. 1b)."

Addressing the potential to perform benchmarking for cell-cell interaction deconvolution, we have thought about this in considerable detail, and present our current thinking (and preliminary findings) as follows:

We have now assessed the performance of six methods (including SpaOTsc [Cang, Z. et al. *Nature Communications* 2020], Giotto [Dries, R. et al. *Genome Biology* 2021], CellChat [Jin, S. et al. *Nature Communications* 2021], NicheNet [Browaeys, R. et al. *Nature Methods* 2020], ICELLNET [Noël, F. et al. *Nature Communications* 2021], SingleCellSignalR [Simon Cabello-Aguilar et al. *Nucleic Acids Research* 2021]) in predicting cell-cell interactions. Note that SpaOTsc and Giotto required spatial transcriptomics data. The other four methods only used scRNA-seq data to generate predictions of intercellular interactions.

First, we used the six methods to predict cell-cell interactions between the L2/3 neurons and L4 neurons in dataset 4. We used the expression metrics of the spatial transcriptomics data as the inputs for all the six methods; we found that these methods generated 0-68 interacted ligand-receptor pairs between the two types of cells, although none of these interactions were commonly present in the output results for all six methods, or even in the output for any five of the six methods (**Graph 3a**).

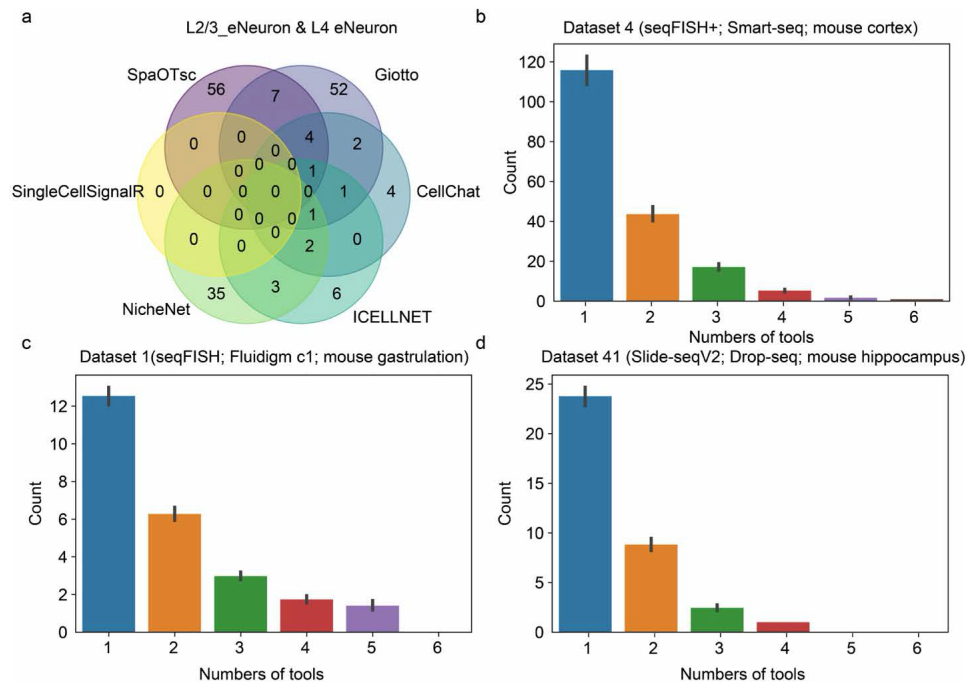
In addition, we used the six methods to predict cell-cell interactions among all cell types in dataset 1, 4, and 41, as the types of the cells in these 3 datasets are available in the data source papers (Palla, G. et al. *BioRxiv* 2021; Dries, R. et al. *Genome Biology* 2021). We found that 1) only 8 out of 184 predicted ligand-receptor pairs are common in the results for > three methods for dataset 4 (**Graph 3b**); 2) only 3 out of 25 predicted ligand-receptor pairs are common in the results for > three methods for dataset 1 (**Graph 3c**); 3) only 1 out of 36 predicted ligand-receptor pairs is common to the results for > three methods for dataset 41 (**Graph 3d**).

As we trust the reviewer will agree, these findings illustrate the extreme diversity of the cell-cell interactions predicted by these methods; at present, we take the view that we cannot assess the

NMETH-AS46278B
Point-by-point-response

Li et al., (Qu) p.26

performance of these methods in predicting cell-cell interactions based on the above results alone. Thus, even if we wanted to benchmark these methods in predicting cell-cell interactions, the vast discrepancies in the results provided from different methods precludes the possibility of informative comparison. We see no way to informatively compare the performance of these methods in predicting cell-cell interactions, and any such comparison would be severely biased (to the extent of logical incoherence). Although (as noted) we have now considered the cell-cell interactions in detail within dataset 1, 4, and 41, please note that we have excluded any direct comparison of the prediction of cell-cell interactions from our revised manuscript.



Graph 3. Comparison of ligand-receptor interactions predicted by six methods capable of predicting cell-cell interactions (SpaOTsc, Giotto, CellChat, ICELLNET, NicheNet, and SingleCellSignalR). a, The number of overlapped ligand-receptor interactions of these methods between L2/3 eNeuron and L4 eNeuron in dataset 4. **b-d,** The numbers of ligand-receptor interactions found by different methods for each cell-type pair in dataset 1 (seqFISH; Fluidigm c1; mouse gastrulation) (b), dataset 4 (seqFISH+; Smart-seq; mouse cortex) (c), and dataset 41 (Slide-seqV2; Drop-seq; mouse hippocampus) (d).

2) A bit related, the authors should provide a clear motivation for the inclusion and exclusion of methods. A recent review on the integration of single cell and spatial methods lists 16

NMETH-AS46278B
Point-by-point-response

Li et al., (Qu) p.27

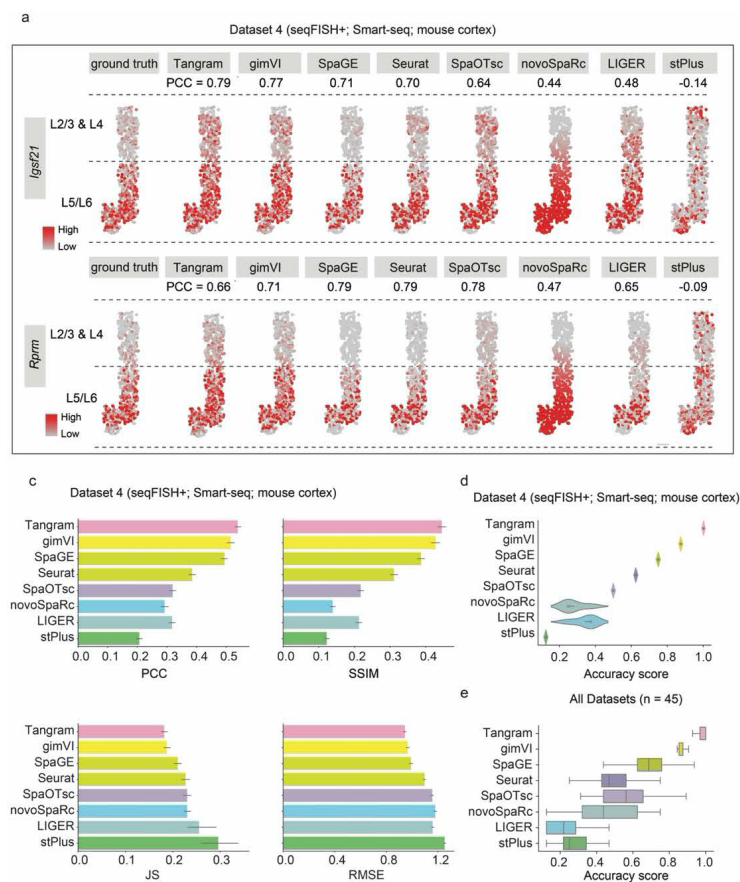
methods to perform different tasks. Why did the authors ignore cell2location for the localization and stPlus for the prediction of undetected transcripts? Is it because both are preprints? Both methods are well documented and easy to use.

Response: We thank the reviewer for focusing our attention on this point. There are two fundamental limitations or challenges of the current state-of-the-art spatial transcriptome approaches, which integration methods can help to overcome: (1) approaches based on in-situ hybridization and fluorescence microscopy (image-based) can detect the spatial distribution of transcripts with high resolution and accuracy, but are limited in the total number of RNA transcripts they can detect; (2) approaches based on next generation sequencing (seq-based) can capture RNA expression at a whole-transcriptome scale from spots in space, but are limited to detect the spatial information of these transcripts at single cell resolution, since each spot may contain multiple cells. Therefore, we selected those methods that can perform the prediction of undetected transcripts or cell-type deconvolution by integrating spatial transcriptome data and scRNA-seq data for benchmarking. This motivation is now clarified in our revised manuscript (Line 89-97, highlighted in yellow), which reads as follows:

"The emergence of these integration methods has undoubtedly helped to deepen our understanding of spatial transcriptomics data and related biological and pathological processes. However, to the best of our knowledge, no study has compared the performance of these integration methods for the prediction of the spatial distribution of transcripts or for the cell type deconvolution of spots in histological sections. Here, we used multiple metrics to systematically benchmark the performance of sixteen integration methods that can predict the spatial distribution of undetected transcripts, or the cell type composition of spots in histological sections (Fig. 1a), based on processing of 45 paired datasets containing both spatial transcriptomics data and scRNA-seq data and 32 simulated datasets (Fig. 1b)."

During our revision, we have included comparisons of 9 additional integration methods that perform cell type deconvolution of spatial transcriptomics data, including SpatialDWLS (Dong et al. *Genome Biology* 2021), RCTD (Cable et al. *Nature Biotechnology* 2021), Stereoscope (Andersson et al. *Communication Biology* 2020), SPOTlight (Elosua-Bayes et al. *Nucleic Acids Research* 2021), DestVI (Lopez et al. 2021, *BioRxiv*), Cell2location (Kleshchevnikov et al. 2021, *BioRxiv*), DSTG (Song et al. *Briefings in Bioinformatics* 2021) and STRIDE (Sun et al. 2021, *BioRxiv*), although the last three have not been published on peer-reviewed journals; and one additional integration method stPlus (Chen et al. *Bioinformatics* 2021) for predicting the spatial distribution of undetected transcripts.

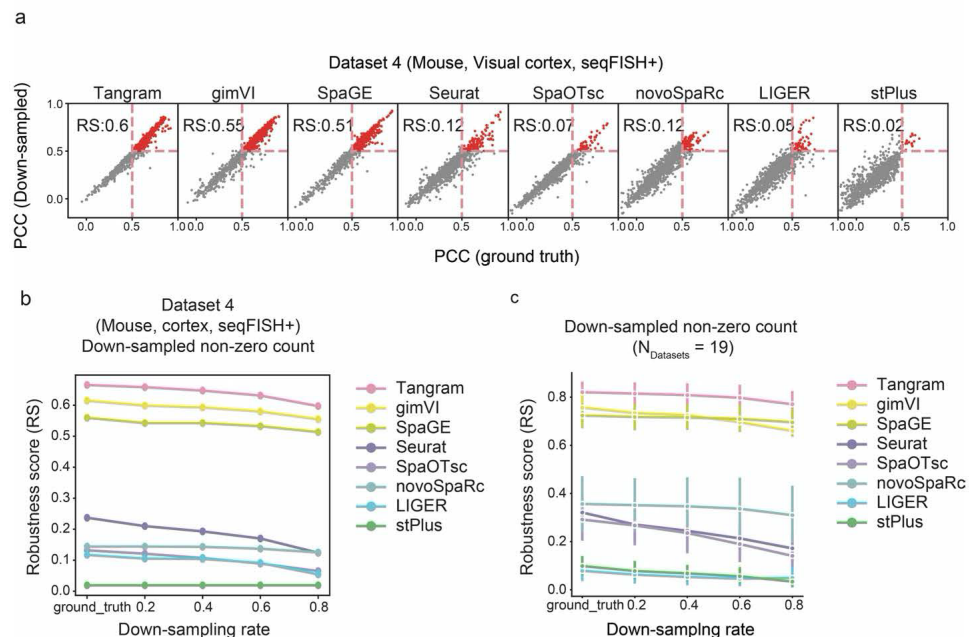
We then re-performed all the comparison analysis among the integration methods on a total of 45 publicly available paired datasets (comprising both spatial transcriptomics and scRNA-seq data) and have accordingly updated both the figures and the manuscript text. We found that Tangram (new version), gimVI, and SpaGE outperformed other integration methods for predicting the spatial distribution of undetected transcripts, while Cell2location, SpatialDWLS, and RCTD outperformed other methods on a total of 32 simulated datasets for deconvoluting the cell-types of each spot (**revised Figure 2-4**).



Revised Figure 2. Comparing the accuracy of eight integration methods capable of predicting the spatial distribution of RNA transcripts. **a,b**, The spatial distribution of *Igsf21* (a) and *Rprm* (b) in dataset 4 (seqFISH+; Smart-seq; mouse cortex), including the ground truth and the predicted result from each of the integration methods. PCC: Pearson Correlation Coefficient between the expression vector of a transcript

in the ground truth and that of the predicted result. **c**, The bar plots of PCC, SSIM, RMSE, and JS of each integration method in predicting the spatial distribution of transcripts in dataset 4. SSIM: Structural Similarity Index; RMSE: Root Mean Square Error; JS: Jensen-Shannon divergence. The confidence interval represented by the error bar is 95%. **d**, Violin plots of AS (accuracy score, which is aggregated from the PCC, SSIM, RMSE, and JS values; see Methods) of the eight integration methods for transcripts in dataset 4. Center line, median; box limits, upper and lower quartiles; whiskers, $1.5 \times$ interquartile range. **e**, Boxplots of AS of the eight integration methods for all 45 paired datasets. Center line, median; box limits, upper and lower quartiles; whiskers, $1.5 \times$ interquartile range.

Figure 3

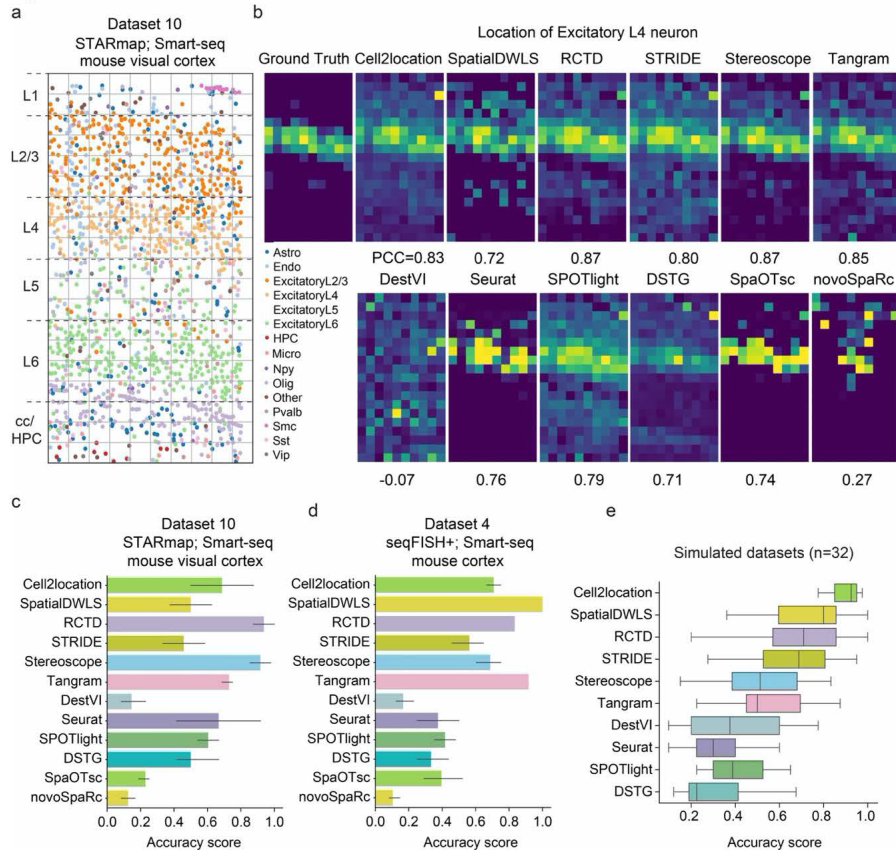


Revised Figure 3. Comparing the accuracy of the eight integration methods for sparse spatial expression matrices down-sampled from the original datasets using Splatter. **a**, PCC of the spatial distribution of transcripts predicted from the original data and down-sampled data from dataset 4. The PCC values of the red-colored transcripts are greater than 0.5 for both the original and the down-sampled data. The proportion of the red-colored transcripts in all transcripts was defined as the “robustness score” (RS). **b**, RS values of the eight integration methods when processing sparse expression matrices down-sampled from dataset 4 at different down-sampling rates. **c**, RS values of the eight integration methods when processing the sparse expression matrices of the down-sampled datasets. The original datasets (used to generate the down-sampled datasets) capture >1000 genes from >100 spots, and the sparsity of the expression matrices is <0.7. The confidence interval represented by the error bar is 90%.

NMETH-AS46278B
Point-by-point-response

Li et al., (Qu) p.30

Figure 4



Revised Figure 4. Comparing the performance of the twelve integration method capable of deconvoluting cell types of each histological spot. **a**, A STARmap slide of dataset 10 (STARmap; Smart-seq; mouse visual cortex), with cells annotated by cell type. Each grid represents a simulated spot containing multiple cells. **b**, The proportion of L4 excitatory neurons in the spots simulated from dataset 10, including the ground truth and the predicted results of twelve integration methods. **c,d**, Barplots of AS (accuracy score, aggregated from PCC, SSIM, RMSE, and JS; see Methods) of the cell type composition of the histological spots simulated from dataset 10 (**c**) and dataset 4 (**d**), predicted by twelve integration methods. The confidence interval represented by the error bar is 95%. **e**, Boxplots of AS of the ten integration methods for all the 32 simulated datasets. SpaOTsc and novoSpaRc are excluded, as they require spatial location information for each spot, which is not available in the simulated datasets. Center line, median; box limits, upper and lower quartiles; whiskers, $1.5 \times$ interquartile range.

NMETH-AS46278B
Point-by-point-response

Li et al., (Qu) p.31

3) The authors tested the effect of normalizing the spatial data on the results of different methods. First, I want to make sure I understand this correctly: what is tested here is integration of (normalized) scRNA-seq data with either normalized/non-normalized spatial data, correct? It is not clear to me what is the motivation here? Why not also try non-normalized scRNA-seq data? It is also not clear to me if this data is used for evaluation (calculation of PCC and other measures) or also for visualization. I am a little concerned about outliers. In principle, if there are few cells/spots with very high values, they will drive the correlation.

Response: Yes, we did test the effect of the normalization of the spatial data on the results of the integration methods in our previous submission, since different strategies for the input data were used by the benchmarked methods. For example: 1) Seurat, LIGER, SpaGE, and stPlus normalized the total RNA count in each spot of the spatial data to a fixed number; 2) gimVI used the raw count matrices of spatial transcriptomics data as the input for integration; 3) Tangram, SpaOTsc, and novoSpaRc did not clearly denote whether they were customized to process the normalized or non-normalized count matrices of spatial transcriptomics data. To benchmark these methods in an unbiased way, we tested all of the integration methods with normalized and non-normalized spatial data in previous submission. We implemented the normalization for the spatial transcriptomics data following the same strategy reported for SpaGE and stPlus.

Inspired by the reviewer, we have added a new section about evaluating the normalization effect for both the spatial transcriptomics data and the scRNA-seq data on the benchmarked methods in the revised manuscript. Briefly, we adopted the Seurat and LIGER strategy for the normalization of the scRNA-seq data (Stuart, T. et al. *Cell*, 2019; Welch, JD. et al. *Cell*, 2019). We tested four schemes of input expression matrices: raw expression matrix of spatial data and raw expression matrix of scRNA-seq data (R-R); normalized expression matrix of spatial data and raw expression matrix of scRNA-seq data (N-R); raw expression matrix of spatial data and normalized expression matrix of scRNA-seq data (R-N); normalized expression matrix of spatial data and normalized expression matrix of scRNA-seq data (N-N). We found that:

(1) In terms of the seq-based datasets, Tangram (new version), gimVI, SpaGE, Seurat, SpaOTsc, and LIGER performed significantly better when using the raw expression matrices than when using normalized expression matrices of the spatial data (P-value<0.01, paired t-test) (revised **Supplementary Figure 5a**). SpaGE, Seurat, SpaOTsc, and novoSpaRc performed better when using the R-R schemes for the input matrices than when using the other three schemes (P-value<0.01).

(2) In terms of the image-based datasets, Tangram (new version), gimVI, SpaGE, and Seurat performed significantly better when using the raw expression matrices than when using the normalized expression matrices of the spatial data (P-value < 0.05, paired t-test) (revised

Supplementary Figure 5b). Moreover, SpaGE, Seurat, and LIGER performed significantly better when using the R-R schemes for the input matrices than when using the other three schemes (P-value < 0.05).

(3) Seeking to prevent bias from a few spots with very high values potentially driving a correlation, we have now performed the same analysis after filtering out outliers; specifically, this analysis only examined values in the 10%~90% confidence interval. Similar conclusions were obtained (revised **Supplementary Figure 5**).

(4) Note that Tangram and SpaGE outperformed the other integration methods in predicting the spatial distribution of transcripts when combining spatial transcriptomics datasets and scRNA-seq datasets (with or without normalization) (revised **Supplementary Figure 5c-f**).

In the revised manuscript, we have modified two paragraphs to consider the impact of the normalization of the spatial data and scRNA-seq data (Line 212-236, highlighted in yellow), which read as follows:

“Several integration methods (e.g., Seurat, LIGER, SpaGE, and stPlus) normalized the spatial transcriptomics data by default prior to integration. Here, we tested four schemes of input expression matrices: (1) raw expression matrix of spatial data and raw expression matrix of scRNA-seq data (R-R); (2) normalized expression matrix of spatial data and raw expression matrix of scRNA-seq data (N-R); (3) raw expression matrix of spatial data and normalized expression matrix of scRNA-seq data (R-N); (4) normalized expression matrix of spatial data and normalized expression matrix of scRNA-seq data (N-N).

Interestingly, for 28 paired seq-based datasets, the transcript spatial distributions generated by Tangram, gimVI, SpaGE, Seurat, SpaOTsc, and LIGER have significantly higher PCC-values when using R-R and R-N input schemes than when using an N-R or N-N input scheme, and this trend was observed for 16 out of the 28 paired datasets (P-values < 0.01, paired t-test) (Supplementary Fig. 4&5a); for SpaGE, Seurat, SpaOTsc, and novoSpaRc, the PCC values of the results with the R-R input scheme were higher than those with the other input schemes in 19 of the 28 paired datasets (P-values < 0.01); and stPlus generated results with higher PCC values when using an R-R input scheme than an N-N input scheme in 18 of the 28 paired datasets (P-value < 0.05). For 15 paired image-based datasets (Supplementary Fig. 4&5b), the transcript spatial distributions generated by Tangram, gimVI, SpaGE, and Seurat have higher PCC-values when

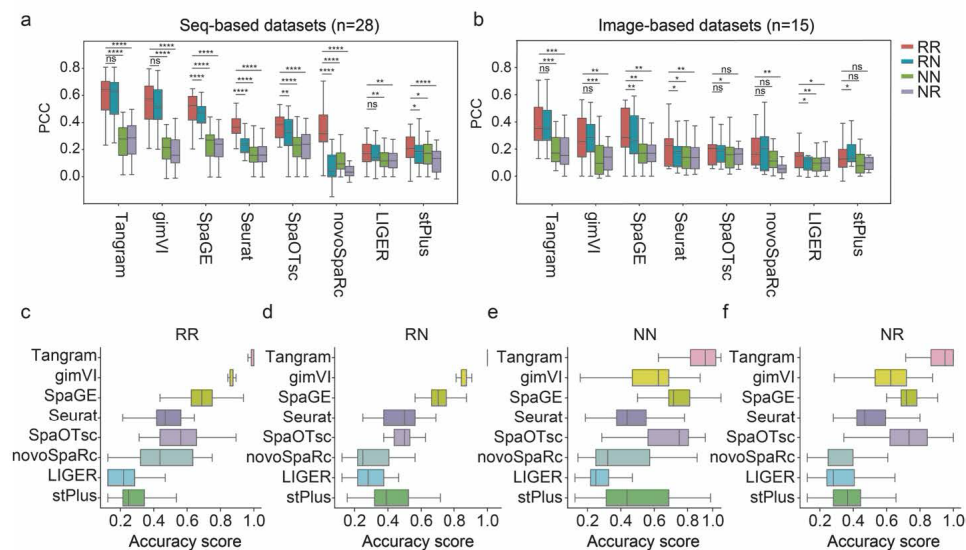
NMETH-AS46278B
Point-by-point-response

Li et al., (Qu) p.33

using R-R and R-N input schemes than when using N-R or N-N input schemes (11 out of 15 datasets, P -value < 0.05); SpaGE, Seurat, and LIGER have higher PCC-values when using R-R input schemes than when using the other input schemes (11 out of 15 datasets, P -value < 0.05); SpaOTsc have higher PCC-values when using R-R input schemes than when using NN input schemes (12 out of 15 datasets, P -value < 0.05). Nevertheless, it should be emphasized that regardless of what input scheme was used, Tangram invariably outperformed the other integration methods (Supplementary Fig. 5c-f)."

In short, based on the examined collection of datasets, those integration methods with relatively higher prediction accuracy (e.g. Tangram, gimVI, and SpaGE) performed better when using the raw expression matrices of the spatial transcriptome data.

sFigure 5



Revised Supplementary Figure 5. Impact of normalization on the accuracy of eight integration methods that can predict the spatial distribution of RNA transcripts. **a, b**, Boxplots of the PCC values of the eight integration methods for seq-based (a) or image-based (b) datasets when using the four schemes of input expression matrices (i.e. R-R, R-N, N-R, and N-N). For the genes predicted by each method, we removed outliers using 10%-90% confidence interval. Statistical significance was analyzed with paired t-test, * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$, and **** $P < 0.0001$. Center line, median; box limits, upper and lower quartiles; whiskers, $1.5 \times$ interquartile range. **c-f**, Boxplots of the AS values of the eight integration methods for all the 45 paired datasets when using the four schemes of input expression matrices. For the

NMETH-AS46278B
Point-by-point-response

Li et al., (Qu) p.34

genes predicted by each method, we removed outliers using 10%-90% confidence interval. Center line, median; box limits, upper and lower quartiles; whiskers, $1.5 \times$ interquartile range.

4) In relation to the previous comment, it seems that some methods are more sensitive to this change than others. For Supplementary Figure 3b, it will be good to plot the results per method with and without normalization. Perhaps even use a paired test to assess how different these results are per method.

Response: We thank the reviewer for this constructive suggestion. As suggested, in the revised **Supplementary Fig. 4 & Supplementary Fig. 5** we have now marked the normalized/non-normalized results per method and used a paired test to assess the differences between normalized and non-normalized results for each method. We found that:

(1) In terms of the seq-based datasets, Tangram (new version), gimVI, SpaGE, Seurat, SpaOTsc, and LIGER performed significantly better when using the raw expression matrices than when using normalized expression matrices of the spatial data (P-value<0.01, paired t-test) (revised **Supplementary Figure 5a**). SpaGE, Seurat, SpaOTsc, and novoSpaRc performed better when using the R-R schemes for the input matrices than when using the other three schemes (P-value<0.01).

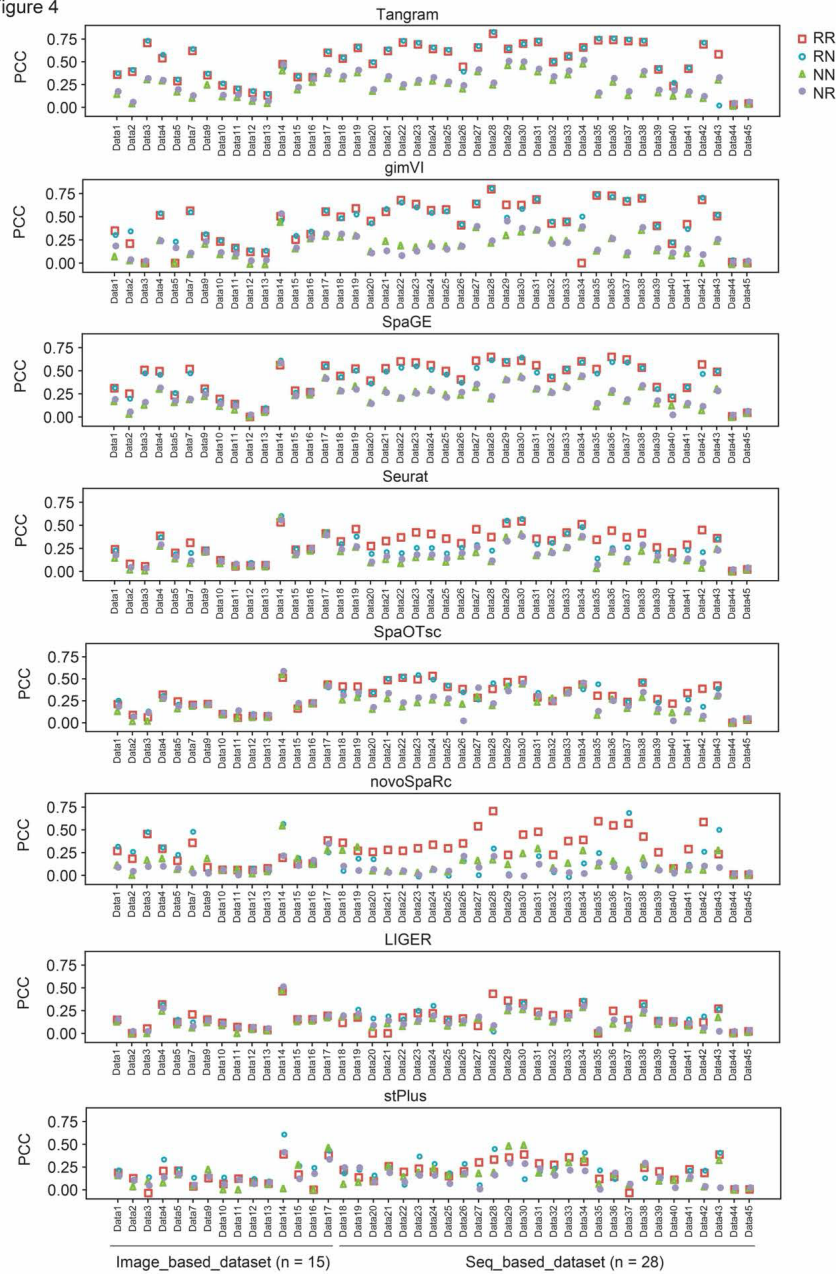
(2) In terms of the image-based datasets, Tangram (new version), gimVI, SpaGE, and Seurat performed significantly better when using the raw expression matrices than when using the normalized expression matrices of the spatial data (P-value < 0.05, paired t-test) (revised **Supplementary Figure 5b**). Moreover, SpaGE, Seurat, and LIGER performed significantly better when using the R-R schemes for the input matrices than when using the other three schemes (P-value < 0.05).

In short, based on the current collection of datasets, integration methods especially those with higher prediction accuracy (*i.e.* Tangram, gimVI, and SpaGE) performed better when using non-normalized than the normalized expression matrix of the spatial transcriptome.

NMETH-AS46278B
Point-by-point-response

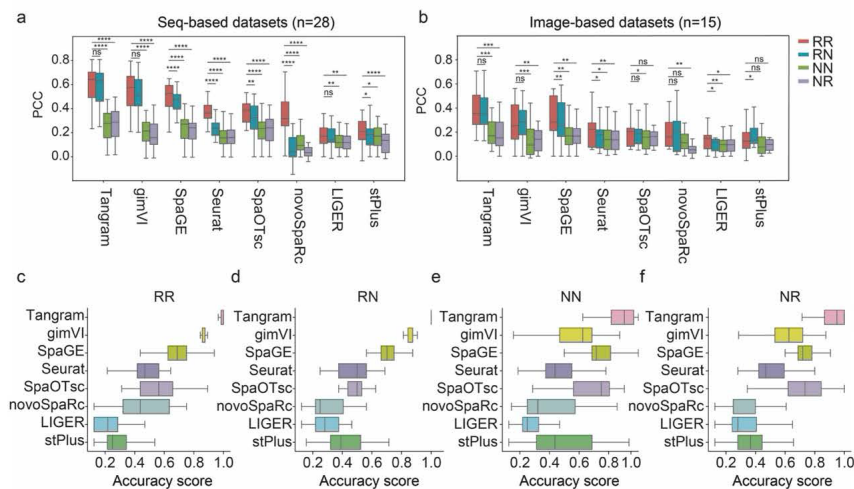
Li et al., (Qu) p.35

sFigure 4



Revised Supplementary Figure 4. The PCC values of each integration method when processing the raw expression matrices and the normalized expression matrices. R-R: raw expression matrix of spatial data and raw expression matrix of scRNA-seq data; N-R: normalized expression matrix of spatial data and raw expression matrix of scRNA-seq data; R-N: raw expression matrix of spatial data and normalized expression matrix of scRNA-seq data; N-N: normalized expression matrix of spatial data and normalized expression matrix of scRNA-seq data.

sFigure 5



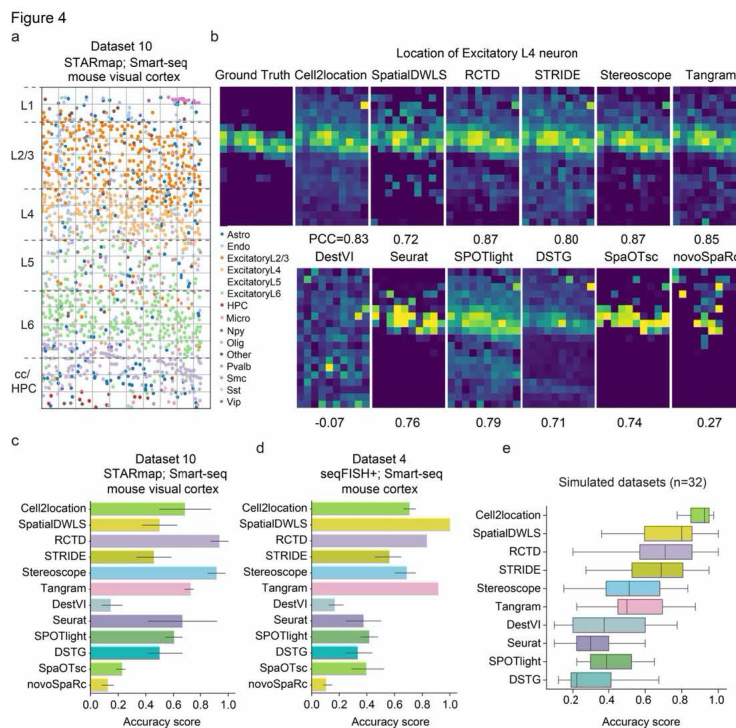
Revised Supplementary Figure 5. Impact of normalization on the accuracy of eight integration methods that can predict the spatial distribution of RNA transcripts. **a, b,** Boxplots of the PCC values of the eight integration methods for seq-based (a) or image-based (b) datasets when using the four schemes of input expression matrices (i.e. R-R, R-N, N-R, and N-N). For the genes predicted by each method, we removed outliers using 10%-90% confidence interval. Statistical significance was analyzed with paired t-test, * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$, and **** $P < 0.0001$. Center line, median; box limits, upper and lower quartiles; whiskers, $1.5 \times$ interquartile range. **c-f,** Boxplots of the AS values of the eight integration methods for all the 45 paired datasets when using the four schemes of input expression matrices. For the genes predicted by each method, we removed outliers using 10%-90% confidence interval. Center line, median; box limits, upper and lower quartiles; whiskers, $1.5 \times$ interquartile range.

5) Perhaps my own lack of knowledge, but it is not clear to me why the authors restrict the prediction of cell location to spots (i.e. vizium or slide-seq like data)? Why not assess the methods ability to localize cells to locations in general? I am thinking of general experiments like in the novoSpaRc paper where a wide variety of spatial data are used.

Response: We thank the reviewer for this comment. In addition to predicting cell locations for spots, in the revised manuscript we have also assessed the performance of the integration methods

in predicting physical localization of cell types in tissues, and thereby included more integration methods for cell type deconvolution into the benchmarking, including SpatialDWLS (Dong et al. *Genome Biology* 2021), RCTD (Cable et al. *Nature Biotechnology* 2021), Stereoscope (Andersson et al. *Communication Biology* 2020), SPOTlight (Elosua-Bayes et al. *Nucleic Acids Research* 2021), and DSTG (Song et al. *Briefings in Bioinformatics* 2021). We have also added Cell2location (Kleshchevnikov et al. 2021, *BioRxiv*), DestVI (Lopez et al. 2021, *BioRxiv*) and STRIDE (Sun et al. 2021, *BioRxiv*), although they have not been published on peer-reviewed journals, for the cell type deconvolution of each spot.

In our new study, we adopted the algorithms proposed by RCTD and Stereoscope to generate 32 simulated datasets (revised **Supplementary Table 2**) from 16 paired scRNA-seq datasets (revised **Supplementary Table 3**). In these simulated datasets, the location of each cell type is known and can therefore be used as ground truth for informative comparison of the integration methods that can deconvolute the cell types of the histological spots. Our examination of the integration methods performance using the 32 simulated datasets showed that Cell2location, SpatialDWLS, and RCTD outperformed other methods for the cell type deconvolution of each spot (revised **Figure 4**).



Revised Figure 4. Comparing the performance of the twelve integration method capable of deconvoluting cell types of each histological spot. **a**, A STARmap slide of dataset 10 (STARmap; Smart-seq; mouse visual cortex), with cells annotated by cell type. Each grid represents a simulated spot containing multiple cells. **b**, The proportion of L4 excitatory neurons in the spots simulated from dataset 10, including the ground truth and the predicted results of twelve integration methods. **c,d**, Barplots of AS (accuracy score, aggregated from PCC, SSIM, RMSE, and JS; see Methods) of the cell type composition of the histological spots simulated from dataset 10 (**c**) and dataset 4 (**d**), predicted by twelve integration methods. The confidence interval represented by the error bar is 95%. **e**, Boxplots of AS of the ten integration methods for all the 32 simulated datasets. SpaOTsc and novoSpaRc are excluded, as they require spatial location information for each spot, which is not available in the simulated datasets. Center line, median; box limits, upper and lower quartiles; whiskers, $1.5 \times$ interquartile range.

6) For down-sampling, the randomly set non-zero elements in the original expression matrix to zeros. This is a bit over-simplified from what happens in reality in which case, few molecules are missed (in osmFISH) or there are dropouts. Why not use a more realistic downsampling approach such as those used in the scuttle R package for single-cell data (<http://www.bioconductor.org/packages/release/bioc/html/scuttle.html>)

Response: We thank the reviewer for focusing our attention on this point. Guided by the reviewer, we applied the Scuttle package (McCarthy, D. J. et al., *Bioinformatics*, 2017) (<http://www.bioconductor.org/packages/release/bioc/html/scuttle.html>) for the down-sampling of the expression matrices of the datasets, and we also used the Splatter package (Zappia, L. et al., *Genome Biology*, 2017) to down-sample the datasets, as both of these methods have been widely used in robustness evaluation of single-cell analysis methods. We found that Tangram (new version), gimVI, and SpaGE outperformed other integration methods in predicting the spatial distribution of transcripts for highly sparse datasets (revised **Figure 3**). Note that using Scuttle and Splatter, our main conclusions about the performance in predicting the spatial distribution of transcripts for highly sparse datasets still stand.

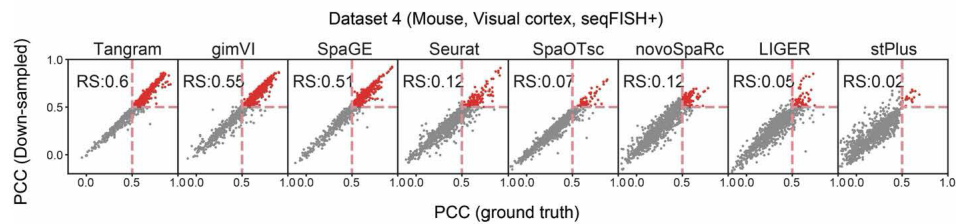
We have added a revised **Supplementary Figure 7** to support assessment of matrix sparsity on the prediction of the spatial distribution of transcripts. We also present a description of revised **Supplementary Figure 7** in the revised main text (Line 263-275, **highlighted in yellow**), which reads as follows:

“We then assessed the performance of each integration method by counting the proportion of transcripts in a dataset exceeding a PCC threshold of 0.5 for both the original and down-sampled data, which we deemed the “robustness score” (RS). For dataset 4, Tangram had the highest RS value (0.60), followed by gimVI (0.55) and then SpaGE (0.51) (Fig. 3a). Moreover, we noted that i) the RS values decreased as the down-sampling rate increased and ii) the RS values for Tangram,

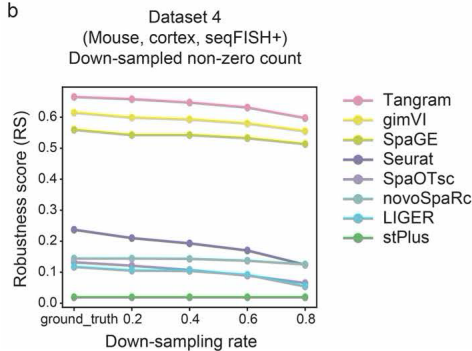
gimVI, and *SpaGE* were consistently higher than those for *Seurat*, *SpaOTsc*, *LIGER*, *novoSpaRc*, or *stPlus* (Fig. 3b). A combined analysis which included the down-sampled data from 19 datasets again highlighted the strong performance of *Tangram*, *gimVI*, and *SpaGE*: even when the down-sampling rate reached 0.8, the average RS values of these three methods remained ≥ 0.50 (Fig. 3c, down-sampled by *Splatter*; Supplementary Fig. 7b, c, down-sampled by *Scuttle*). In summary, *Tangram*, *gimVI*, and *SpaGE* outperformed other integration methods in predicting the spatial distribution of transcripts for highly sparse datasets.”

Figure 3

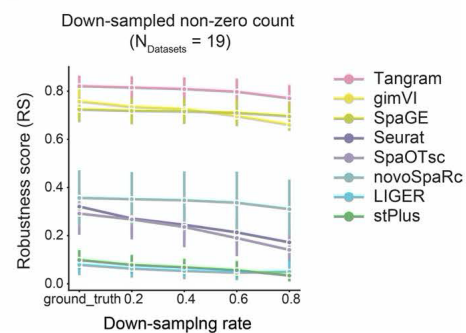
a



b



c

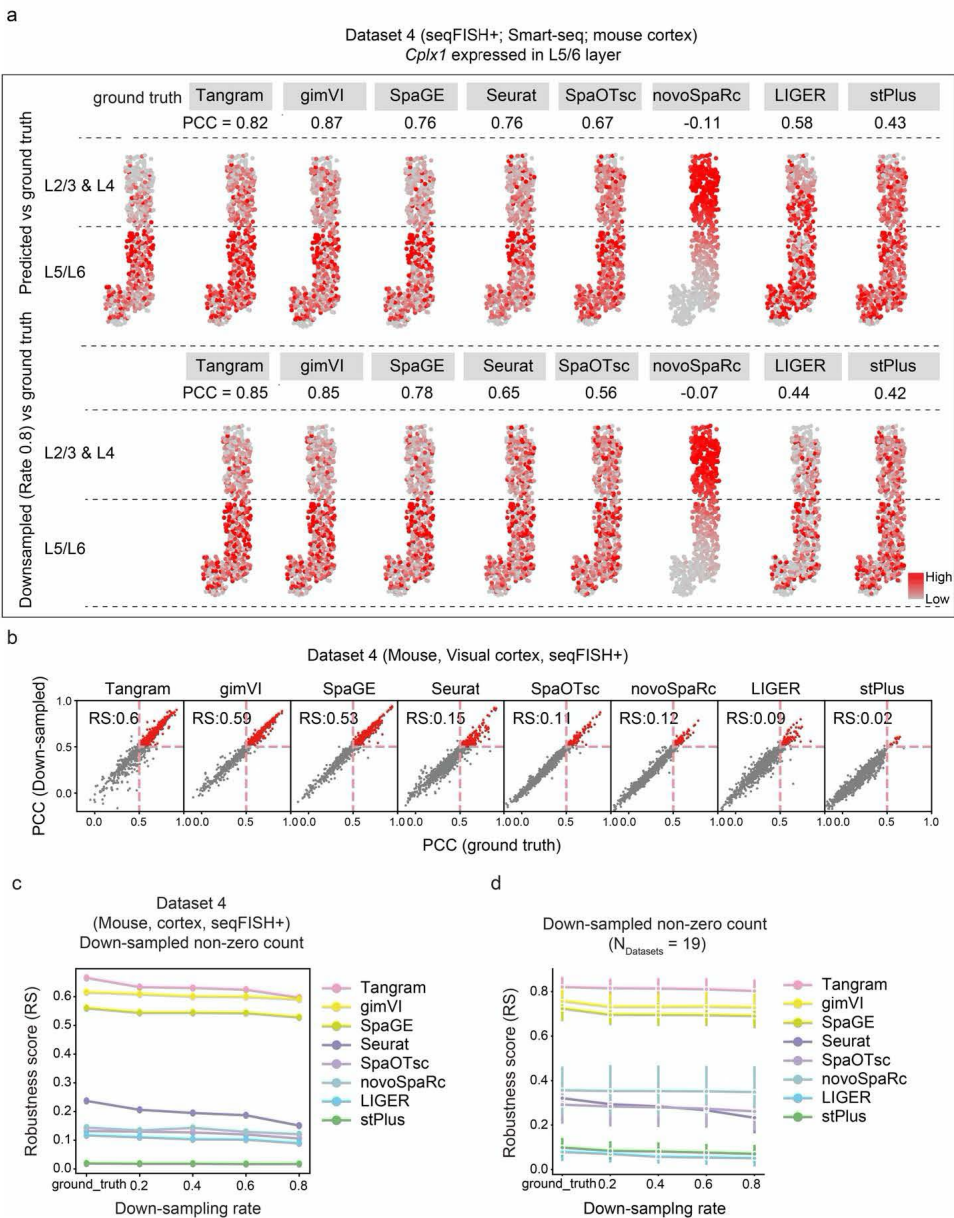


Revised **Figure 3. Comparing the accuracy of the eight integration methods for sparse spatial expression matrices down-sampled from the original datasets using *Splatter*.** **a**, PCC of the spatial distribution of transcripts predicted from the original data and down-sampled data from dataset 4. The PCC values of the red-colored transcripts are greater than 0.5 for both the original and the down-sampled data. The proportion of the red-colored transcripts in all transcripts was defined as the “robustness score” (RS). **b**, RS values of the eight integration methods when processing sparse expression matrices down-sampled from dataset 4 at different down-sampling rates. **c**, RS values of the eight integration methods when processing the sparse expression matrices of the down-sampled datasets. The original datasets (used to generate the down-sampled datasets) capture >1000 genes from >100 spots, and the sparsity of the expression matrices is <0.7. The confidence interval represented by the error bar is 90%.

NMETH-AS46278B
Point-by-point-response

Li et al., (Qu) p.40

sFigure 7



Revised Supplementary Figure 7. Comparing the accuracy of the eight integration methods for sparse expression matrices down-sampled from the original datasets using Scuttle. **a**, Spatial distribution of *Cplx1* expression in dataset 4 (seqFISH+; Smart-seq; mouse cortex), predicted from the original data and down-sampled data (down-sampling rate = 0.8). **b**, PCC of the spatial distribution of transcripts predicted from the original data and down-sampled data from dataset 4. The PCC values of the red-colored transcripts are greater than 0.5 for both the original and the down-sampled data. The proportion of the red-colored transcripts in all transcripts was defined as the “robustness score” (RS). **c**, RS values of the eight integration methods when processing sparse expression matrices down-sampled from dataset 4 at different down-sampling rates. **d**, RS values of the eight integration methods when processing the sparse expression matrices of the down-sampled datasets. The original datasets (used to generate the down-sampled datasets) capture >1000 genes from >100 spots, and the sparsity of the expression matrices is <0.7. The confidence interval represented by the error bar is 90%.

7) The evaluation of computational efficiency (Supplementary Figure 7, which should be a table instead) is not sufficient. Why not do this assessment for all datasets? The number of shared genes and the difficulty in aligning will have an effect (not only the number of cells/spots).

Response: We thank the reviewer for focusing our attention on this concern. In our revised manuscript, we assessed the computational efficiency of eight integration methods for all 45 datasets in predicting spatial distribution of transcript (revised **Supplemental Table 4**). For the largest dataset (dataset 40, with 19,522 spots in its spatial transcriptomics data and 26,252 cells its scRNA-seq data) and down-sampling approach for comparing the computational resources consumed by the integration methods in our previous submission. As suggested, we have now assessed the computational efficiency for all datasets. Notably, Seurat and LIGER took less than 10 minutes of computer time for all of the 45 paired datasets, and Tangram and LIGER consumed less than 32GB of computer memory.

We then assessed the impacts of various data attributes (including the number of cells in scRNA-seq data, the number of spots in spatial data, and the number of genes used for training) on the computational resources consumed by those eight integration methods. By down-sampling the number of cells and the number of spots in dataset 40, and the number of training genes in dataset 6, we found that Seurat was invariably the most computationally efficient method among the eight integration methods for the prediction of the spatial distribution of undetected transcripts (revised **Supplementary Figure 9a-c**).

To compare the computational resources consumed by those ten integration methods that can deconvolute the cell types of spots, we used a large simulated dataset that contains 10000 cells,

NMETH-AS46278B
Point-by-point-response

Li et al., (Qu) p.42

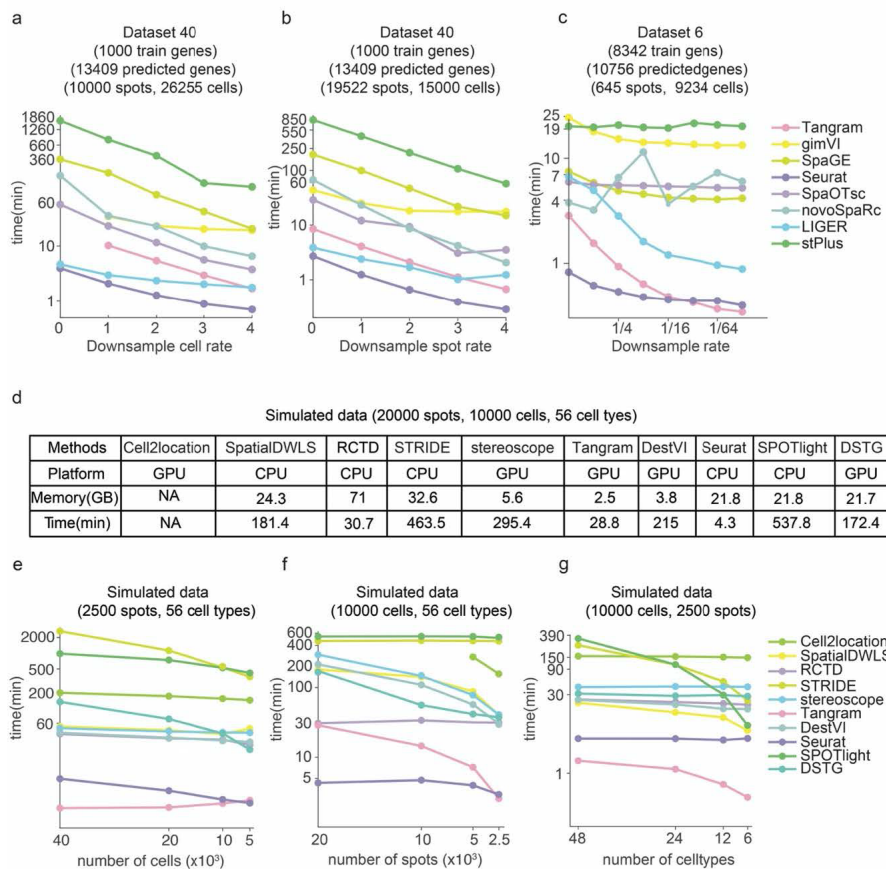
20000 spots, and 56 cell types. For this dataset, Cell2locations reported memory errors on our GPU platform. Seurat and Tangram took less than 30 minutes of computer time, and Stereoscope, Tangram, and DestVI consumed less than 8GB of computer memory (revised **Supplementary Figure 9d**). We then evaluated the impacts of the number of cells in scRNA-seq data, the number of spots in spatial data, and the number of cell types on computing time consumed by those ten integration methods, and found that Tangram and Seurat are the top 2 ranking most efficient methods for processing cell type deconvolution of spots (revised **Supplementary Figure 9e-g**).

We have also followed the reviewer's suggestion to change the **Supplementary Figure 7** in our previous submission to the revised **Supplementary Figure 9**, and provide **Supplemental Table 4** to show the computational efficiency for all 45 datasets in the revised main text.

NMETH-AS46278B
Point-by-point-response

Li et al., (Qu) p.43

sFigure 9



Revised **Supplementary Figure 9. Computer resources consumed by each integration method.** **a-c**, The impact of the number of cells in scRNA-seq data (a), the number of spots in spatial data (b), and the number of genes used for training (c), on computational resources consumed by the integration methods that can predict the spatial distribution of undetected transcripts. **d**, The computer time and memory spent by the integration methods that can deconvolute cell types of histological spots, when processing a simulated dataset which contains 20000 spots in its spatial transcriptomics data and 10000 cells in its scRNA-seq data. **e-g**, The impacts of the number of cells in scRNA-seq data (e), the number of spots in spatial data (f), and the number of the cell types (g) on computational resources consumed by the integration methods that can deconvolute cell types of histological spots.

Minor comments:

- Please explicitly define what you mean by sparsity.

NMETH-AS46278B
Point-by-point-response

Li et al., (Qu) p.44

Response: Thanks for bringing this ambiguity to our attention. Specifically, ‘sparsity’ as used in the manuscript is defined as the percentage of zero elements in the expression matrix. We have added a sentence to the **Results** section of the revised manuscript (Line 243-245, highlighted in yellow) to formally define this.

“The sparsity of the spatial transcriptomics and scRNA-seq data is defined as the percentage of zero elements in the expression matrix.”

- There are no details of how novoSpaRc was run. Did the authors run it in fully de novo mode? In the novoSpaRc paper, this did not work well in brain tissues and using reference genes is recommended.

Response: To ensure the fairness of the benchmarking study, we trained novoSpaRc using the expression and spatial information of the reference gene set, just as we did for the other methods. We have added a paragraph in the **Methods** section of the revised manuscript (Line 779-781, highlighted in yellow) to clearly state this.

“We trained novoSpaRc using the expression and spatial information of the training gene set, just as we did for the other methods, to ensure the fairness of the benchmarking study.”

- The choices made while simulating the “multi-cell spot problem” in terms of the grid size and how many pixels to group are not motivated at all. Why these choices? And how do they relate to the resolution of the original data and the current vizium/slide-seq resolutions. Please elaborate.

Response: We devised a procedure for coarse-grained spatial transcriptomic data by referring to SpatialDWLS (Dong, et al., *Genome Biology*, 2021). The somatosensory cortex seqFISH+ data were obtained from <https://github.com/CaiGroup/seqFISH-PLUS>. To simulate spot-like data, we defined a square with 500×500 pixels (~51.5 μm) as one spot-like region. The size of a simulated spot is close to a spot in 10X Visium data (i.e., 55 μm). We have added content to the **Methods** section of the revised manuscript (Line 841-846, highlighted in yellow) to present this.

“In order to obtain multi-cell-per-spot datasets with known cell compositions at each spot, we gridded single-cell resolution spatial transcriptomics datasets 4 and 10 to simulate datasets with potentially ambiguous cell type assignments per spot. For dataset 4 (seqFISH+; Smart-seq; mouse cortex), we defined a square with 500×500 pixels (~51.5 μm) as one spot-like region to grid the seqFISH+ slide, referring to the coarse-graining procedure introduced by SpatialDWLS”

Decision Letter, first revision:

Subject: AIP Decision on Manuscript NMETH-AS46278B

Message:

Our ref: NMETH-AS46278B

13th Feb 2022

Dear Dr. Qu,

Thank you for submitting your revised manuscript "Benchmarking spatial and single-cell transcriptomics integration methods for transcript distribution prediction and cell type deconvolution" (NMETH-AS46278B). It has now been seen by the original referees and their comments are below. The reviewers find that the paper has improved in revision, and therefore we'll be happy in principle to publish it in Nature Methods, pending minor revisions to satisfy the referees' final requests and to comply with our editorial and formatting guidelines.

Among other revisions, please follow both reviewers' suggestions to improve the discussion section.

We are now performing detailed checks on your paper and will send you a checklist detailing our editorial and formatting requirements in about a week. Please do not upload the final materials and make any revisions until you receive this additional information from us.

TRANSPARENT PEER REVIEW

Nature Methods offers a transparent peer review option for new original research manuscripts submitted from 17th February 2021. We encourage increased transparency in peer review by publishing the reviewer comments, author rebuttal letters and editorial decision letters if the authors agree. Such peer review material is made available as a supplementary peer review file. Please state in the cover letter 'I wish to participate in transparent peer review' if you want to opt in, or 'I do not wish to participate in transparent peer review' if you don't. Failure to state your preference will result in delays in accepting your manuscript for publication.

Please note: we allow redactions to authors' rebuttal and reviewer comments in the interest of confidentiality. If you are concerned about the release of confidential data, please let us know specifically what information you would like to have removed. Please note that we cannot incorporate redactions for any other reasons. Reviewer names will be published in the peer review files if the reviewer signed the comments to authors, or if reviewers explicitly agree to release their name. For more information, please refer to our <https://www.nature.com/documents/nr-transparent-peer-review.pdf> target="new">FAQ page.

Thank you again for your interest in Nature Methods Please do not hesitate to contact me if you have any questions.

Sincerely,

Lin Tang, PhD
Senior Editor
Nature Methods

ORCID

IMPORTANT: Non-corresponding authors do not have to link their ORCIDs but are encouraged to do so. Please note that it will not be possible to add/modify ORCIDs at proof. Thus, please let your co-authors know that if they wish to have their ORCID added to the paper they must follow the procedure described in the following link prior to acceptance:

<https://www.springernature.com/gp/researchers/orcid/orcid-for-nature-research>

Reviewer #1 (Remarks to the Author):

I am extremely impressed by the authors' efforts in adding 9 new methods (and updated versions of existing methods), increasing the number of evaluated datasets by 3-fold, and adding 32 simulated datasets to this benchmarking changes. In my opinion, these and other changes have substantially improved the manuscript, and the benchmarking is a valuable contribution to the community.

I have only a two minor comments.

1. While the amount of benchmarking is impressive, I still find that the manuscript remains vague in its conclusions.

The authors claim to "provide example guidance" (Lines 40 and 364) to help researchers select an appropriate method and refer to their GitHub tutorial. However, based on my reading the GitHub page doesn't seem to provide guidance (advice) on the selection of methods, but rather a pipeline to run many methods. Related, the phrase "example guidance" is strange – the authors can provide "guidance" (advice) or can provide "examples", but I am not sure what is meant by the combination of these words.

Further, the authors claim in the introduction:

"Our findings can help researchers choose appropriate integration methods for their datasets and also raise interesting questions about how various processing and dataset-specific attributes influence the integration performance of these essential tools for spatial transcriptomics research." (Lines 104-106)

However, the Discussion mentions only the 3-top performing methods in each category and discusses the issue of data sparsity. While I realize that the authors might be reluctant to endorse a few methods, it would be very helpful if there was a summary table or flowchart which provided direction to the user on what analyses are possible, which methods are available for each analysis task, and some information/ranking on these methods' performance (perhaps with references to the figures in the manuscript that support this guidance). Currently, the guidance seems to be spread across many dense figures and supplementary figures. A clear summary table/figure would substantially increase the value of the benchmarking.

Similarly, the "interesting questions..." section above could be better summarized in the Discussion/Conclusion. The only relevant item I see in the Discussion is regarding data sparsity. Are there other interesting questions for future work?

2. In response to my previous comment #6: I appreciate the reviewer's response that the upcoming generation of spatial sequencing technologies will not yet achieve single-cell resolution. I suggest including these details in the Discussion to further emphasize the continuing need for these integration methods, and thus the value of the authors' benchmarking.

Reviewer #2 (Remarks to the Author):

I would like to thank the authors for addressing all my comments. I find the revised manuscript more comprehensive and I commend the authors for taking the effort to fairly compare the method and refine their text to reflect the work done. I have two additional remarks on the revised version:

- I find the comparison of the methods for cell-cell communication quite interesting. Specially the low discordance between the different methods. It will be good to make a point on this either by including the resulting data or a discussion point of why benchmarking cell-cell communication is difficult. Do the authors expect the low agreement to stem from differences in the databases used by these methods to infer ligand-receptor interactions?

- I find the discussion section relatively thin. For instance, I would encourage the authors to discuss why certain methods work better/faster than others. Do certain types of methods work consistently better than others? The top 3 performing methods are quite different methodologically (e.g. SpaGE relies on linear alignment while gimVI relies on non-linear alignments). This shall direct future development in this field towards areas for improvement and or interesting methodologies to follow upon.

Responses to reviewers

Reviewer #1:

I am extremely impressed by the authors' efforts in adding 9 new methods (and updated versions of existing methods), increasing the number of evaluated datasets by 3-fold, and adding 32 simulated datasets to this benchmarking changes. In my opinion, these and other changes have substantially improved the manuscript, and the benchmarking is a valuable contribution to the community.

I have only a two minor comments.

1. While the amount of benchmarking is impressive, I still find that the manuscript remains vague in its conclusions. The authors claim to “provide example guidance” (Lines 40 and 364) to help researchers select an appropriate method and refer to their GitHub tutorial. However, based on my reading the GitHub page doesn't seem to provide guidance (advice) on the selection of methods, but rather a pipeline to run many methods. Related, the phrase “example guidance” is strange – the authors can provide “guidance” (advice) or can provide “examples”, but I am not sure what is meant by the combination of these words.

Further, the authors claim in the introduction:

“Our findings can help researchers choose appropriate integration methods for their datasets and also raise interesting questions about how various processing and dataset-specific attributes influence the integration performance of these essential tools for spatial transcriptomics research.” (Lines 104-106)

However, the Discussion mentions only the 3-top performing methods in each category and discusses the issue of data sparsity. While I realize that the authors might be reluctant to endorse a few methods, it would be very helpful if there was a summary table or flowchart

which provided direction to the user on what analyses are possible, which methods are available for each analysis task, and some information/ranking on these methods' performance (perhaps with references to the figures in the manuscript that support this guidance). Currently, the guidance seems to be spread across many dense figures and supplementary figures. A clear summary table/figure would substantially increase the value of the benchmarking.

Similarly, the “interesting questions...” section above could be better summarized in the Discussion/Conclusion. The only relevant item I see in the Discussion is regarding data sparsity. Are there other interesting questions for future work?

Response: We thank the reviewer for their positive assessment of our revised manuscript as well as important comments to further improve our paper. We have made the following modification of the paper to address the reviewer's additional concerns:

- i) We have changed the “example guidance” in the **Abstract** and **Discussion** sections to “a benchmark pipeline” in the revised manuscript.

- ii) We have added a new supplementary table (**revised Supplementary Table 5**) to summarize the properties and performance of all the benchmarked methods in the revised manuscript.

a				b						Scalability (min)							
Method				Accuracy						Downsampled spots (cell*spot)				Downsampled cells (cell*spot)			
Properties				Adjust Per metric													
Gene prediction				PCC	SSIM	RMSE	JS	AS	RS								
Language	GPU	Methodology															
Tangram	Python	✓	deep learning							~6.5k*2.5k	~6.5k*5k	~6.5k*10k	~6.5k*20k	~3.2k*10k	~6.5k*10k	~13k*10k	~26k*10k
gimVI	Python	✓	probabilistic model							1.12	2.1	4.09	8.48	2.96	5.43	10.2	NA
SpaGE	Python	×	joint embedding-based							17.7	18.3	25.2	43.5	20.2	23	34	NA
Seurat	R	×	joint embedding-based							21.9	46.8	99	195	41.9	83.9	208	366
SpaOTsc	Python	×	optimal Transport							0.39	0.65	1.24	2.71	0.89	1.28	2.09	3.96
novoSpaRc	Python	×	optimal Transport							3.07	9.3	12.1	28.9	5.63	11.5	22.9	56.1
LIGER	R	×	joint embedding-based							4.21	8.64	23.3	66.9	9.9	22.8	35.1	186
stPlus	Python	✓	joint embedding-based							1.02	1.7	2.4	3.89	2.04	2.37	2.98	4.67
										107	210	422	834	176	421	828	1819
Deconvolution																	
Cell2location	Python	✓	probabilistic model							~10k*2.5k	~10k*5k	~10k*10k	~10k*20k	~5k*2.5k	~10k*2.5k	~20k*2.5k	~40k*2.5k
SpatialDWLS	R	×	weighted least squares							157	274	NA	NA	148	157	173	197
RCTD	R	×	probabilistic model							40.4	87.8	142	181	48.6	39.3	45.7	52.8
STRIDE	Python	×	topic model (LDA)							31.4	31.7	33.8	30.7	25.9	31.6	33.4	39.3
DestVI	Python	✓	probabilistic & deep learning							462	465	470	463	367	529	1029	2189
Stereoscope	Python	✓	probabilistic model							30.5	57.1	109	215	28.8	30.6	34.1	41.2
Tangram	Python	✓	deep learning							40.9	77.8	146	295	41.5	42	43.4	49.5
Seurat	R	×	joint embedding-based							2.56	7.21	14.5	28.8	2.95	2.61	2.24	2.18
SPOTlight	R	×	topic model (NMF)							2.94	3.97	4.7	4.3	2.66	3.05	4.3	6.87
DSTG	Python	✓	deep learning							513	538	541	538	427	523	708	910
										37.4	41.6	56	172	21.4	40.4	70.8	138

Adjust Score

0
0.2
0.4
0.6
0.8
1

Revised Supplementary Table 5: Summary of the integration methods in predicting the spatial distribution of transcripts and/or deconvolving cell types of spots.

iii) In the **Discussion** section of the revised manuscript, we have added a paragraph to describe the theoretical basis of those well-performed integration methods (Line 363-370, highlighted in yellow), which reads as follows:

“Methods constructed based on probabilistic models combined with negative binomial or Poisson distributions, such as gimVI, Cell2location and RCTD are generally better performed in

predicting the spatial distribution of transcripts or deconvolving cell types of spots. Deep learning algorithm was also applied in several integration methods, in which Tangram is one of the best performed methods in predicting spatial distribution of the undetected transcripts. Technically, Tangram employed nonconvex optimization in the model and selected only the optimal subset of scRNA-seq observations in the loss function. A combination of these measures may help improve the predictive power of the tools.”

Additionally, we have added a paragraph discussing the difficulty of benchmarking methods for predicting cell-cell interactions in the **Discussion** section of the revised manuscript (Line 379-388, highlighted in yellow), which reads as follows:

“Another potential application of spatial transcriptomics is to predict ligand-receptor interactions between two cell types that are spatially close to each other. Many analytical tools have been developed in this regard, such as SpaOTsc, Giotto, CellChat, NicheNet, ICELLNET, and SingleCellSignalR. However, the vast discrepancies in the results provided from different methods precludes the possibility of informative comparison of these methods in predicting cell-cell interactions. For instance, only a small proportion (<5%) of the predicted ligand-receptor interactions were shared in the results for >3 methods (<https://github.com/QuKunLab/SpatialBenchmarking/tree/main/FigureData>). Benchmark comparisons thereby rely on more experimental validation of cell-cell interactions in the future.”

2. In response to my previous comment #6: I appreciate the reviewer’s response that the upcoming generation of spatial sequencing technologies will not yet achieve single-cell resolution. I suggest including these details in the Discussion to further emphasize the continuing need for these integration methods, and thus the value of the authors’ benchmarking.

Response: We thank the reviewer for this suggestion, and we have added the following paragraph to emphasize the continuing need for these integration methods and the value of our benchmarking analysis in the **Discussion** section of the revised manuscript (Line 405-413, highlighted in yellow), which reads as follows:

“The upcoming new versions of 10X Visium and BGI Stereo-seq can detect transcriptomic

information for spots with diameters much smaller than cell size. However, each spot will not correspond exactly to a single cell based on these anticipated technologies in the near future. Moreover, consider that there is a very large amount of publicly available spatial transcriptomics datasets that are highly valuable to a plethora of research communities, and a large number of academic studies using spatial transcriptomics technologies that are now underway, we contend there will still be challenging and strong interest for the integration of spatial transcriptomics and scRNA-seq data to determine the spatial distribution of cells or undetected transcripts."

Reviewer #2:

I would like to thank the authors for addressing all my comments. I find the revised manuscript more comprehensive and I commend the authors for taking the effort to fairly compare the method and refine their text to reflect the work done. I have two additional remarks on the revised version:

1. I find the comparison of the methods for cell-cell communication quite interesting. Specially the low discordance between the different methods. It will be good to make a point on this either by including the resulting data or a discussion point of why benchmarking cell-cell communication is difficult. Do the authors expect the low agreement to stem from differences in the databases used by these methods to infer ligand-receptor interactions?

Response: We appreciate the reviewer's positive assessment of our revised manuscript. Since we used exactly the same spatial transcriptomics datasets and the same ligand-receptor database for all these methods (including SpaOTsc, Giotto, CellChat, NicheNet, ICELLNET, and SingleCellSignalR) to benchmark their performance for cell-cell communication, we believe that the low agreement of these methods was not caused by the differences of the database/datasets.

Guided by the reviewer, we have added a paragraph to explain the difficulties of benchmarking cell-cell communication in the **Discussion** section of the revised manuscript (Line 379-388, highlighted in yellow), which reads as follows:

“Another potential application of spatial transcriptomics is to predict ligand-receptor interactions between two cell types that are spatially close to each other. Many analytical tools have been developed in this regard, such as SpaOTsc, Giotto, CellChat, NicheNet, ICELLNET, and SingleCellSignalR. However, the vast discrepancies in the results provided from different methods precludes the possibility of informative comparison of these methods in predicting cell-cell interactions. For instance, only a small proportion (<5%) of the predicted ligand-receptor interactions were shared in the results for >3 methods (<https://github.com/QuKunLab/SpatialBenchmarking/tree/main/FigureData>). Benchmark comparisons thereby rely on more experimental validation of cell-cell interactions in the future.”

2. I find the discussion section relatively thin. For instance, I would encourage the authors to discuss why certain methods work better/faster than others. Do certain types of methods work consistently better than others? The top 3 performing methods are quite different methodologically (e.g. SpaGE relies on linear alignment while gimVI relies on non-linear alignments). This shall direct future development in this field towards areas for improvement and or interesting methodologies to follow upon.

Response: We thank the reviewer for this suggestion. We have added a new supplementary table (**revised Supplementary Table 5**) to summarize the properties (including methodological basis) and performance of all the benchmarked methods in the revised manuscript.

a Method				b Accuracy						Scalability (min)							
Properties				Adjust Per metric						Downsampled spots (cell*spot)				Downsampled cells (cell*spot)			
Gene prediction				PCC	SSIM	RMSE	JS	AS	RS								
Language	GPU	Methodology								~6.5k*2.5k	~6.5k*5k	~6.5k*10k	~6.5k*20k	~3.2k*10k	~6.5k*10k	~13k*10k	~26k*10k
Tangram	Python	✓	deep learning							1.12	2.1	4.09	8.48	2.96	5.43	10.2	NA
gimVI	Python	✓	probabilistic model							17.7	18.3	25.2	43.5	20.2	23	34	NA
SpaGE	Python	×	joint embedding-based							21.9	46.8	99	195	41.9	83.9	208	366
Seurat	R	×	joint embedding-based							0.39	0.65	1.24	2.71	0.89	1.28	2.09	3.96
SpaOTsc	Python	×	optimal Transport							3.07	9.3	12.1	28.9	5.63	11.5	22.9	56.1
novoSpaRc	Python	×	optimal Transport							4.21	8.64	23.3	66.9	9.9	22.8	35.1	186
LIGER	R	×	joint embedding-based							1.02	1.7	2.4	3.89	2.04	2.37	2.98	4.67
stPlus	Python	✓	joint embedding-based							107	210	422	834	176	421	828	1819
Deconvolution										~10k*2.5k	~10k*5k	~10k*10k	~10k*20k	~5k*2.5k	~10k*2.5k	~20k*2.5k	~40k*2.5k
Cell2location	Python	✓	probabilistic model							157	274	NA	NA	148	157	173	197
SpatialDWLS	R	×	weighted least squares							40.4	87.8	142	181	48.6	39.3	45.7	52.8
RCTD	R	×	probabilistic model							31.4	31.7	33.8	30.7	25.9	31.6	33.4	39.3
STRIDE	Python	×	topic model (LDA)							462	465	470	463	367	529	1029	2189
DestVI	Python	✓	probabilistic & deep learning							30.5	57.1	109	215	28.8	30.6	34.1	41.2
Stereoscope	Python	✓	probabilistic model							40.9	77.8	146	295	41.5	42	43.4	49.5
Tangram	Python	✓	deep learning							2.56	7.21	14.5	28.8	2.95	2.61	2.24	2.18
Seurat	R	×	joint embedding-based							2.94	3.97	4.7	4.3	2.66	3.05	4.3	6.87
SPOTlight	R	×	topic model (NMF)							513	538	541	538	427	523	708	910
DSTG	Python	✓	deep learning							37.4	41.6	56	172	21.4	40.4	70.8	138

Adjust Score

0 0.2 0.4 0.6 0.8 1

Revised Supplementary Table 5: Summary of the integration methods in predicting the spatial distribution of transcripts and/or deconvolving cell types of spots.

Guided by the reviewer, we have extended our content of the **Discussion** section to discuss the potential methodological reasons why several integration methods outperformed other methods in the **Discussion** section of the revised manuscript (Line 363-370, highlighted in yellow), which reads as follows:

“Methods constructed based on probabilistic models combined with negative binomial or Poisson distributions, such as gimVI, Cell2location and RCTD are generally better performed in predicting the spatial distribution of transcripts or deconvolving cell types of spots. Deep learning algorithm was also applied in several integration methods, in which Tangram is one of the best performed methods in predicting spatial distribution of the undetected transcripts. Technically, Tangram employed nonconvex optimization in the model and selected only the optimal subset of scRNA-seq observations in the loss function. A combination of these measures may help improve the predictive power of the tools.”

Final Decision Letter:

Subject: Decision on Nature Methods submission NMETH-AS46278C

Message:

30th Mar 2022

Dear Professor Qu,

I am pleased to inform you that your Analysis, "Benchmarking spatial and single-cell transcriptomics integration methods for transcript distribution prediction and cell type deconvolution", has now been accepted for publication in Nature Methods. Your paper is tentatively scheduled for publication in our June print issue, and will be published online prior to that. The received and accepted dates will be 10th Jun 2021 and 30th Mar 2022. This note is intended to let you know what to expect from us over the next month or so, and to let you know where to address any further questions.

Acceptance is conditional on the data in the manuscript not being published elsewhere, or announced in the print or electronic media, until the embargo/publication date. These restrictions are not intended to deter you from presenting your data at academic meetings and conferences, but any enquiries from the media about papers not yet scheduled for publication should be referred to us.

In approximately 10 business days you will receive an email with a link to choose the appropriate publishing options for your paper and our Author Services team will be in touch regarding any additional information that may be required.

Please note that Nature Methods is a Transformative Journal (TJ). Authors may publish their research with us through the traditional subscription access route or make their paper immediately open access through payment of an article-processing charge (APC). Authors will not be required to make a final

decision about access to their article until it has been accepted. Find out more about Transformative Journals

Authors may need to take specific actions to achieve compliance with funder and institutional open access mandates. If your research is supported by a funder that requires immediate open access (e.g. according to Plan S principles) then you should select the gold OA route, and we will direct you to the compliant route where possible. For authors selecting the subscription publication route, the journal's standard licensing terms will need to be accepted, including self-archiving policies. Those licensing terms will supersede any other terms that the author or any third party may assert apply to any version of the manuscript.

You will not receive your proofs until the publishing agreement has been received through our system.

If you have any questions about our publishing options, costs, Open Access requirements, or our legal forms, please contact ASJournals@springernature.com

Your paper will now be copyedited to ensure that it conforms to Nature Methods style. Once proofs are generated, they will be sent to you electronically and you will be asked to send a corrected version within 24 hours. It is extremely important that you let us know now whether you will be difficult to contact over the next month. If this is the case, we ask that you send us the contact information (email, phone and fax) of someone who will be able to check the proofs and deal with any last-minute problems.

If, when you receive your proof, you cannot meet the deadline, please inform us at rjsproduction@springernature.com immediately.

Once your manuscript is typeset and you have completed the appropriate grant of rights, you will receive a link to your electronic proof via email with a request to make any corrections within 48 hours. If, when you receive your proof, you cannot meet this deadline, please inform us at rjsproduction@springernature.com immediately.

Once your paper has been scheduled for online publication, the Nature press office will be in touch to confirm the details.

If you have posted a preprint on any preprint server, please ensure that the preprint details are updated with a publication reference, including the DOI and a URL to the published version of the article on the journal website.

Content is published online weekly on Mondays and Thursdays, and the embargo is set at 16:00 London time (GMT)/11:00 am US Eastern time (EST) on the day of publication. If you need to know the exact publication date or when the news embargo will be lifted, please contact our press office after you have submitted your proof corrections. Now is the time to inform your Public Relations or Press Office about your paper, as they might be interested in promoting its publication. This will allow them time to prepare an accurate and satisfactory press release. Include your manuscript tracking number NMETH-AS46278C and the name of the journal, which they will need when they contact our office.

About one week before your paper is published online, we shall be distributing a press release to news organizations worldwide, which may include details of your work. We are happy for your institution or funding agency to prepare its own press release, but it must mention the embargo date and Nature Methods. Our Press Office will contact you closer to the time of publication, but if you or your Press Office have any inquiries in the meantime, please contact press@nature.com.

To assist our authors in disseminating their research to the broader community, our SharedIt initiative provides you with a unique shareable link that will allow anyone (with or without a subscription) to read the published article. Recipients of the link with a subscription will also be able to download and print the PDF.

As soon as your article is published, you will receive an automated email with your shareable link.

You can now use a single sign-on for all your accounts, view the status of all your manuscript submissions and reviews, access usage statistics for your published articles and download a record of your refereeing activity for the Nature journals.

Nature Research journals encourage authors to share their step-by-step experimental protocols on a protocol sharing platform of their choice. Nature Research's Protocol Exchange is a free-to-use and open resource for protocols; protocols deposited in Protocol Exchange are citable and can be linked from the published article. More details can be found at www.nature.com/protocolexchange/about.

Please note that you and any of your coauthors will be able to order reprints and single copies of the issue containing your article through Nature Research Group's reprint website, which is located at <http://www.nature.com/reprints/author-reprints.html>. If there are any questions about reprints please send an email to author-reprints@nature.com and someone will assist you.

Please feel free to contact me if you have questions about any of these points. Thank you very much again for publishing your work at Nature Methods.

Best regards,

Lin Tang, PhD
Senior Editor
Nature Methods