nature portfolio

Peer Review File

Intracellular Spatial Transcriptomic Analysis Toolkit (InSTAnT)

Editorial Note: This manuscript has been previously reviewed at another journal that is not operating a transparent peer review scheme. This document only contains reviewer comments and rebuttal letters for versions considered at Nature Communications.

Reviewers' Comments:

Reviewer #1:
Remarks to the Author:
The author introduces a software package InSTAnT, which can be used to characterize the intro-cellular RNA interaction measured with spatial transcriptomics techniques such as MERFISH, seqFISH, and Xenium. The package utilizes a binomial test to identify pairs of co-localized RNAs, based on intra-cellular RNA distribution, and provides a P-value to determine if two genes are close within a defined distance 'd' in a given cell. In addition, other analysis modules enable users to examine the results from various perspectives. The Conditional Poisson Binomial (CPB) test, for instance, identifies representative gene pairs across all cells, while the Differential Colocalization module and Spatial modulation provide insight into differences in colocalization at the cell-type and tissue-level. The author conducted validation experiments on three additional datasets and compared their approach with another method, which demonstrated high accuracy and good repeatability. Although the manuscript is well-structured overall, I have some concerns regarding the experiments and results sections, which are listed below.
Comments
1. In the process of calculating FPR, the gene pairs under the real data are TN, and the new gene pairs found under the random data are FN. In addition, any 'blank' in a gene pair is considered as FN. The accuracy of this process heavily relies on the level of randomness, and the methodology doesn't specify if the FPR calculation is an average derived from multiple random datasets.

2. The researchers discovered that gene pairs were detectable when the distance between them was 1um, but not when it was 4um. They attributed this to varying RNA interaction mechanisms. However, I believe this issue may be correlated to gene expression within the cell, as their expression is less pronounced and the number of overlapping gene pairs is less affected by increasing distance. This suggests that the phenomenon may be linked to gene expression rather than biological mechanisms. Additionally, it may be possible to calculate the ideal distance d for genes with different expression levels using the null distribution. Could you please elaborate on the influence that gene expression values exert on your approach?

3. To determine cell-type specific colocalization, the authors performed pp-tests on each cell and tallied the results based on cell type. They then identified the gene pair that was specific to a certain cell type and categorized it based on whether either gene is a marker for that cell type. The CPB test appears to be a more appropriate method for identifying cell-type specific gene pairs. Have you considered utilizing the CPB test across each cell type to identify these gene pairs?

4. When performing spatial modulation tests, the p_global and p_local are calculated from the data, why is it said in L922 that they and w are calculated according to the maximum likelihood.

5. When the Xenium data is used to verify the spatial differences at the tissue level, the author used the cell-type-specific method instead of the spatial modulation mentioned below. This approach, which disregards the spatial location of the tissue, means that the results cannot conclusively confirm the presence of gene-pair differences at this level. To illustrate the tissue-level spatial variations of colocalization patterns, the 'spatial modulation' method would be more effective.

6. The figure 5.h. can not show the spatial differences at the tissue level. It may be better to use other samples to draw this figure.

Reviewer #2:
Remarks to the Author:
As a disclaimer, I was one of the original reviewers of the manuscript. I was impressed with the rigour, curtsey, and and quality of the provided answers. The authors now provide careful comparison with the exciting literature and new exciting data/analysis that permit a reader to obtain a comprehensive impression about InSTAnT. I therefore support the publication of this study.

I have some smaller comments concerning the actual implementation, or more precisely the documentation, which could potentially help to increase a more widespread use of this package. The Python code seems well documented and a detailed description for the analysis of publicly available data is provided.

- As a new user, I would appreciate to also find some guidance for the required input data-format, e.g. for expression matrix and cell/nuclear segmentation. One of the emerging standards in the domain is anndata (scverse), could an import wrapper be provided?

- Related to above. The universe of tools for the analysis of spatial OMICs data is growing rapidly and becomes hard to navigate. Personally, I think the establishment of standard frameworks to regroup these tools is beneficial for new users, to be able to easily apply different approaches to their data. Many new tools, e.g. BENTO, are now compatible with the scverse framework. Are the provided results compatible with this analysis framework, or could they be made compatible?

Reviewer #3:
Remarks to the Author:
In this manuscript, Kumar and colleagues introduce a package for the analysis of RNA colocalization patterns within image-based single-cell transcriptomics data. As I was asked to supplement the review of others, I will keep my comments brief.

The topic is an important one, and the contribution provides one additional way to capture elements of intracellular organization with RNAs. This work should eventually be published.

However, I had a few comments that the authors may wish to consider in a potential revision to this work.

First, the authors provide a statistical framework to identify pairs of genes that co-occur within a distance d more than random chance, given a specific null model of RNA distribution. From this framework they can assign a p-value for each pair of potentially interacting RNAs. While this approach provides for the ability to find statistically significant co-enrichment or co-occurrence, the authors appear to not provide a measure of the magnitude of the co-occurrence. For example, given the null model for RNA distribution and given the abundance of each of the pair of RNAs, what would be the expected frequency of d-co-occurrence and, importantly, what is the degree of enrichment (or de-enrichment) relative to this null frequency? I recommend the authors consider adding such a feature

to their method as one could have minute enrichments that are statistically significant, yet because the enrichment is so minute it may be difficult to draw biological significance from the enrichment.

Second, other reviewers noted that the authors did not originally acknowledge that measures of intracellular organization had been developed and leveraged previously, in particular in the context of MERFISH and seqFISH measurements. They have now added a comparison to a previous, unnamed, approach introduced with MERFISH. However, I found that comparison very difficult to follow and, in particular, to draw from the authors limited description an understanding of why they claim this previous approach has 100% false-positive rate. This proposed poor performance does not seem consistent with the clear biological consistency of the function of the gene patterns identified in this previous work. I think that more nuance should be taken in how previous work is contrasted with the current method.

Response to Reviewer #1

Comment: The author introduces a software package InSTAnT, which can be used to characterize the intro-cellular RNA interaction measured with spatial transcriptomics techniques such as MERFISH, seqFISH, and Xenium. The package utilizes a binomial test to identify pairs of co-localized RNAs, based on intra-cellular RNA distribution, and provides a P-value to determine if two genes are close within a defined distance 'd' in a given cell. In addition, other analysis modules enable users to examine the results from various perspectives. The Conditional Poisson Binomial (CPB) test, for instance, identifies representative gene pairs across all cells, while the Differential Colocalization module and Spatial modulation provide insight into differences in colocalization at the cell-type and tissue-level. The author conducted validation experiments on three additional datasets and compared their approach with another method, which demonstrated high accuracy and good repeatability. Although the manuscript is well-structured overall, I have some concerns regarding the experiments and results sections, which are listed below.
Comments
1. In the process of calculating FPR, the gene pairs under the real data are TN, and the new gene pairs found under the random data are FN. In addition, any 'blank' in a gene pair is considered as FN. The accuracy of this process heavily relies on the level of randomness, and the methodology doesn't specify if the FPR calculation is an average derived from multiple random datasets.

Response: The gene pairs obtained with InSTAnT on real data comprise true positives (TP) and false positives (FP). The gene pairs found under the random data are assumed to be false positives (FP). As is commonly done, we take a ratio of these two counts to estimate FPR. Randomization is done for each cell once, and FPR is obtained by aggregating across cells. We have now clarified this explicitly in manuscript (Methods, L759-766). Since thousands of cells are independently randomized (gene labels of transcripts in an individual cell are shuffled), this procedure makes use of an extensive level of randomization.

Comment: 2. The researchers discovered that gene pairs were detectable when the distance between them was 1um, but not when it was 4um. They attributed this to varying RNA interaction mechanisms. However, I believe this issue may be correlated to gene expression within the cell, as their expression is less pronounced and the number of overlapping gene pairs is less affected by increasing distance. This suggests that the phenomenon may be linked to gene expression rather than biological mechanisms. Additionally, it may be possible to calculate the ideal distance d for genes with different expression levels using the null distribution. Could you please elaborate on the influence that gene expression values exert on your approach?

Response: The reviewer is right that the gene expression in a cell influences our PP test: the sensitivity of co-localization detection is greater when it involves larger expression levels. (In the Binomial test used by the PP test, the number of trials is the product of cellular gene expression of the two genes.) This is true irrespective of the value of the “d” parameter (1 um or 4 um). We have reported on the influence of gene expression on our approach, through Supplementary Figure 2, noting that certain highly

expressed genes feature among *d*-colocalized pairs far more frequently when using the Poisson Binomial test to aggregate evidence across cells. Indeed, this observation was the motivation for a significant feature of our methodology in the CPB test, where we sought to reduce the confounding effect of gene expression (Figure 2b) by using a gene-dependent null model.

Furthermore, the test also has varying sensitivity at different values of the "d" parameter, since larger numbers of proximal transcript pairs can be observed at larger values of "d", and this is suggested also by the much stronger p-values seen with "d" = 4 um than with "d" = 1 um in Figure 3f. We have now added the following clarification in the relevant text: "These results illustrate scale-dependence of the colocalization phenomenon and suggests multiple types of underlying biological relationships, though some part of the exclusivity is likely to be due to varying sensitivity of the test at different d values."

At the same time, we believe that the cases where we observe certain gene pairs as significantly colocalized only at one "d" value and not the other (Figure 3f) do largely reflect varying biological mechanisms, since those observations are based on the CPB test, and addresses the gene expression-dependence of the PP test. We have observed different types of biological relationships when using different values of "d". For instance, small d (< 1 um) was used to find an enrichment of RNA-RNA physical interactions (predicted using RNAPlex) among colocalized pairs, a medium "d" (=2 um) was used for the analysis that led us to identify MALAT1-SRRM2 as colocalized in nuclear speckles, while d=4um in u2-os data identified RNAs that tend to be around perinuclear space or ER (which are larger features). We have now added the above observation to the manuscript (Discussions, L470-L475).

Comment: 3. To determine cell-type specific colocalization, the authors performed pp-tests on each cell and tallied the results based on cell type. They then identified the gene pair that was specific to a certain cell type and categorized it based on whether either gene is a marker for that cell type. The CPB test appears to be a more appropriate method for identifying cell-type specific gene pairs. Have you considered utilizing the CPB test across each cell type to identify these gene pairs?

Response: Thanks for the comment. We did try using the CPB test for each cell type separately; however, it did not yield cell type-specific gene pairs. An extension of the CPB test for differential colocalization specific to a cell type is not straightforward. To put it simply, the CPB test asks if colocalization is observed surprisingly often in a set of cells, while what we sought in cell type-specific colocalization is the explicit *differential* question "is colocalization observed more often in one set of cells than others?".

Comment: 4. When performing spatial modulation tests, the p_global and p_local are calculated from the data, why is it said in L922 that they and w are calculated according to the maximum likelihood.

Response: All the parameters (p_global, p_local and w) are calculated by maximizing the log likelihood calculated from the data. So the two statements are consistent with each other.

Comment: 5. When the Xenium data is used to verify the spatial differences at the tissue level, the author used the cell-type-specific method instead of the spatial modulation mentioned below. This approach, which disregards the spatial location of the tissue, means that the results cannot conclusively confirm the presence of gene-pair differences at this level. To illustrate the tissue-level spatial variations of colocalization patterns, the 'spatial modulation' method would be more effective.

Response: The reviewer is correct that the Xenium data set can be analyzed using the spatial modulation method as well. Our use of differential colocalization method on the Xenium data identifies colocalization patterns that differ among three regions of the hippocampus – dentate gyrus (DG) and areas CA3 and CA1. These three regions are spatially separated (Figure 5a) and have been functionally studied in the literature, and this motivated us to use differential colocalization method to probe region-specific colocalization patterns. Our goal in this work was not to provide a comprehensive analysis of every data set but to showcase different functionalities of InSTAnT for extracting useful insights. Note: the spatial modulation function identifies spatial patterns in an unbiased manner, i.e., without any predefined knowledge of spatial regions; further follow-up is necessary to interpret such spatial patterns.

Comment: 6. The figure 5.h. can not show the spatial differences at the tissue level. It may be better to use other samples to draw this figure.

Response: We apologize for the lack of clarity regarding this figure. Cells in which the genes Gad1 and Syt2 form a proximal pair (PP test) are shown in blue in Figure 5h. It is easy to notice that those cells tend to be clustered at the four corners of the visualized sample, thus exhibiting a non-random spatial distribution. We have now updated the legend to remove the confusion. This is the kind of spatial pattern discerned by the spatial modulation function, which is why we have chosen to present it in the figure.

Response to Reviewer #2:

Comment: As a disclaimer, I was one of the original reviewers of the manuscript. I was impressed with the rigour, curtsey, and and quality of the provided answers. The authors now provide careful comparison with the exciting literature and new exciting data/analysis that permit a reader to obtain a comprehensive impression about InSTAnT. I therefore support the publication of this study.

Response: We thank the reviewer for their favorable assessment of the revised manuscript, and also their insightful and constructive critique of the original submission, which truly helped improve the manuscript.

Comment: I have some smaller comments concerning the actual implementation, or more precisely the documentation, which could potentially help to increase a more widespread use of this package. The Python code seems well documented and a detailed description for the analysis of publicly available data is provided.

- As a new user, I would appreciate to also find some guidance for the required input data-format, e.g. for expression matrix and cell/nuclear segmentation. One of the emerging standards in the domain is anndata (scverse), could an import wrapper be provided?

Response: Thanks very much for the suggestion. We have now updated the codebase and the documentation where the input data is anndata and results from different analyses are saved in anndata object.

Comment: - Related to above. The universe of tools for the analysis of spatial OMICs data is growing rapidly and becomes hard to navigate. Personally, I think the establishment of standard frameworks to regroup these tools is beneficial for new users, to be able to easily apply different approaches to their data. Many new tools, e.g. BENTO, are now compatible with the scverse framework. Are the provided results compatible with this analysis framework, or could they be made compatible?

Response: Thanks for the comments. We have now incorporated the comments and released our tool as a python package sc-instant (pip install sc-instant) that is compatible with scverse.

Response to Reviewer #3:

Comment: In this manuscript, Kumar and colleagues introduce a package for the analysis of RNA colocalization patterns within image-based single-cell transcriptomics data. As I was asked to supplement the review of others, I will keep my comments brief.

The topic is an important one, and the contribution provides one additional way to capture elements of intracellular organization with RNAs. This work should eventually be published.

Response: We thank the reviewer for their favorable assessment of our work, and for their time and effort in providing us with suggestions for improvement.

Comment: However, I had a few comments that the authors may wish to consider in a potential revision to this work. First, the authors provide a statistical framework to identify pairs of genes that co-occur within a distance d more than random chance, given a specific null model of RNA distribution. From this framework they can assign a p-value for each pair of potentially interacting RNAs. While this approach provides for the ability to find statistically significant co-enrichment or co-occurrence, the authors appear to not provide a measure of the magnitude of the co-occurrence. For example, given the null model for RNA distribution and given the abundance of each of the pair of RNAs, what would be the expected frequency of d-co-occurrence and, importantly, what is the degree of enrichment (or de-enrichment) relative to this null frequency? I recommend the authors consider adding such a feature to their method as one could have minute enrichments that are statistically significant, yet because the enrichment is so minute it may be difficult to draw biological significance from the enrichment.

Response: We thank the reviewer for their suggestion, and have added the requested feature to InSTAnT's output.

A gene pair's colocalization is detected by InSTAnT at two levels – first, at the level of individual cells using the PP test, and then at the level of a collection of cells using the CPB test. The PP test is at its core a Binomial test, with N = number of pairs of transcripts (of the given gene pair), p = probability of a transcript pair being within distance "d" (under the null model learnt from that cell), and k = number of transcript pairs observed with distance d. However, the null expectation of Np can be quite small ($<< 1$) in many cases, especially for cells with relatively few transcripts, and the enrichment factor is not very reliable in such cases. A more reliable way to assess the "degree of enrichment" (as the reviewer puts it) is through the CPB test, where we now report the expected number of cells (under the null model) alongside the observed number of cells and of course the CPB test p-value. (See Supplementary Table1, Supplementary Table2, Supplementary Table5.)

Comment: Second, other reviewers noted that the authors did not originally acknowledge that measures of intracellular organization had been developed and leveraged previously, in particular in the context of MERFISH and seqFISH measurements. They have now added a comparison to a previous, unnamed, approach introduced with MERFISH. However, I found that comparison very difficult to follow

and, in particular, to draw from the authors limited description an understanding of why they claim this previous approach has 100% false-positive rate. This proposed poor performance does not seem consistent with the clear biological consistency of the function of the gene patterns identified in this previous work. I think that more nuance should be taken in how previous work is contrasted with the current method.

Response: Thank you for the suggestion. For the reviewer's benefit, we first outline the method of Chen et al. which is the unnamed approach noted in the reviewer's comment, and is the only existing statistical method for assessing co-localization in a *collection of cells*. Their method divides each cell into four equal bins (compartments), and counts each gene's transcripts in these bins. For any gene pair, it calculates the correlation between their respective transcript counts across these four bins. It then averages the resulting correlation coefficient across all cells. This average is the final measure of a gene pair's colocalization. Note that (a) the spatial resolution of this approach is quite low (about a quarter of a cells area/volume), and (b) correlation coefficients are calculated from four samples at a time, leading to unreliable estimates, which are then averaged. We have updated the manuscript for better clarification (Supplementary Methods, L582-584). Moreover, unlike InsTAnT, this approach does not explicitly deal with confounding effects of gene expression variation and covariation.

Secondly, we outline how we estimate the false positive rate (FPR) of a method's findings, for comparison between our approach and that of X et al.

1) We count how many gene pairs are found to be significant by a method *M*, at some significance level α (a p-value for InSTAnT, a correlation value for Chen et al.). Let this count be denoted by $D_M(\alpha)$.
2) Then we randomize the entire data set: for each cell, we shuffle the gene labels of all transcripts in that cell, so that the locations of transcripts remain unchanged, the transcript counts of genes remain unchanged, but any *gene-specific* spatial patterns of transcript distribution are destroyed. This is done *independently for each cell, for the thousands of cells in that data set.*
3) Next, we run the method *M* on the randomized data set from step (2) and count the number of gene pairs significant at level α. Let this count be denoted by $R_M(\alpha)$.
4) We argue that $D_M(\alpha)$ counts both true positives and false positives, while $R_M(\alpha)$ estimates the number of false positives. Thus, we report the ratio $R_M(\alpha)/D_M(\alpha)$ as the estimated FPR at significance level α. We repeat such estimation at varying values of α.

The above procedure (steps 1-5) is performed with the method M being either InSTAnT or a competing method, thus estimating FPR at varying significance thresholds of discovery by the method; these are then compared between methods. This procedure is described in Methods (L758-L764). It was using this intuitive approach that we observed the FPR estimates reported in Figure 2a, Figure 2b, and the estimated false positive counts for the method of Chen et al. in Supplementary Figure 15.

We would like to note that the reported FPR estimates and comparisons are on U2OS cells while the original analysis of Chen et al. was done on fibroblast cells. Thus, the unexpectedly high FPRs could reflect the fact that their method does not generalize to other data sets such as the one analyzed here.

To dive more deeply into the method of Chen et al., we analyzed the (average) correlations reported by it for all gene pairs, on the U2OS data set, and performed Hierarchical clustering of genes (rows and columns). The results are shown below (Figure R1), the left panel corresponding to the U2OS data set and the right panel corresponding to a randomized version of this data set (randomized as in Step 2 above). We note that clusters/modules of genes can be found by the method (e.g., three diagonally located blocks in the left panel), just as their original analysis on fibroblast data had found. However, we note that cluster patterns can also be seen in the results on randomized data, where there should not be any patterns. This casts doubts on the procedure of using gene modules identified by the above-mentioned clustering approach as evidence of a method's success. We believe a more direct estimation of false positive rates, as done in our evaluations, is important for robust assessment. Note: the same procedure applied to gene pair colocalization strengths calculated by InSTAnT yields strong cluster patterns (Figure R2, left) on the U2OS data but *not* on the randomized data (Figure R2, right).

We have presented the comparison to Chen et al. through the following passage of text, with the red font indicating newly added text: "We compared the CPB test with the only alternative method for aggregating colocalization information across cells, the bin-based approach of Chen et al., which yielded estimated FPR ~ 100% and reproducibility < 10% (Supplementary Figure 15). These comparisons underscore the importance of InSTAnT's rigorous statistical testing procedures for reliable detection of spatial patterns. The greatly improved specificity of InSTAnT in our evaluations may also be in part due to its higher resolution of spatial proximity. We also note that the reported FPR estimates and comparisons are on U2OS cells while the original analysis of Chen et al. was done on fibroblast cells. Thus, the unexpectedly high FPRs of the method of Chen et al. could reflect the fact that their method does not generalize to other data sets such as the one analyzed here."

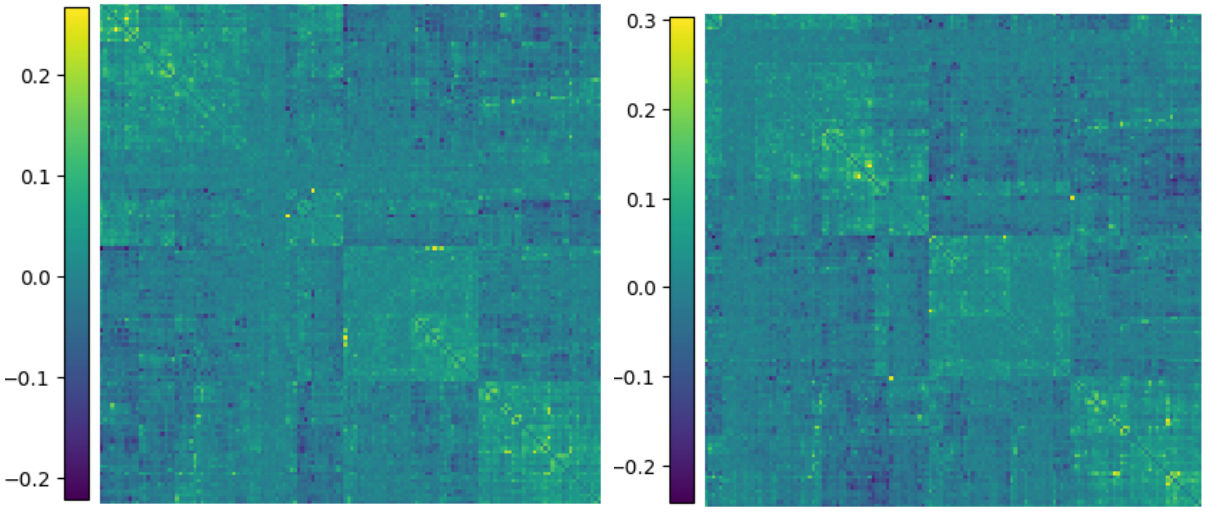

Table R1. Gene pair colocalization matrix based on the method of Chen et al., with rows and columns clustered, on U2OS data (left) and randomized version of the same data (right). Note: diagonal elements set to 0 rather than 1, for better visualization of heat map.

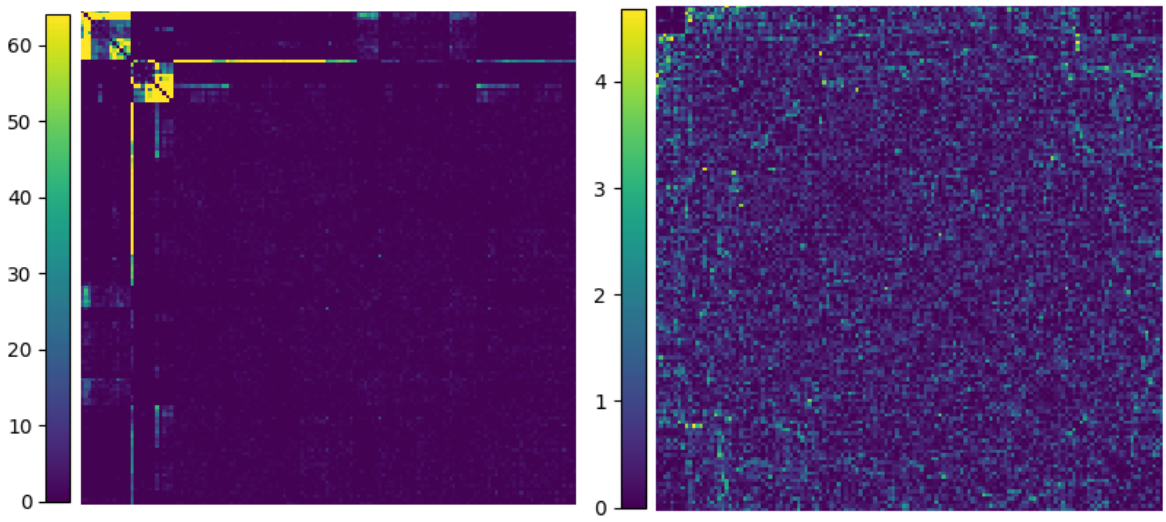

Table R2. Gene pair colocalization matrix based on CPB test in InSTAnT, with rows and columns clustered, on U2OS data (left) and randomized version of the same data (right). Hierarchical Clustering was performed with distance metric being Euclidean distances between vectors of negative logarithms of CPB p-values.

Reviewers' Comments:

Reviewer #1:
Remarks to the Author:
Accept

Reviewer #2:
Remarks to the Author:
The authors addressed all my comments. I recommend publication of this manuscript.

Reviewer #3:
Remarks to the Author:
In this revised manuscript, the authors describe InSTAnT a software package to explore spatial co-localization in image-based single-cell transcriptomic methods. I had minor comments for the reviewers upon my last review, and I feel the authors have addressed these comments partially in their response.

The only remaining sticking point that I have is the way in which they compare their work to the previous spatial analysis method presented in the original MERFISH paper by Chen et al. I raised a concern in my previous review that the authors produce an estimate of a false discovery rate (FDR) of 100% for that analysis method yet this did not seem to be accurate, as Chen et al describe groups of spatially colocalizing genes that make biological sense. Thus, a 100% FDR would appear far more likely to be an overestimate based on how this quantity was calculated.

I thank the authors for expanding on this point in their rebuttal; however, I do not think that they have addressed my central concern with their response. Now, thanks to their detailed response, I suspect that I may have found the issue.

First, returning to Chen et al, the authors in that work used their approach, albeit limited and rudimentary compared to InSTAnT, to identify genes enriched near the perinuclear region and enriched in the cellular periphery by identify spatially co-varying sets of genes and then visually examining the intracellular location of the two groups they found. Supporting the accuracy of this analysis, the authors found that the perinuclear group contained genes known to be translated at the endoplasmic reticulum while the peripheral group was enriched in genes associated with cytoskeletal function. These results are expected biologically, strongly arguing that the false discover rate of this method cannot be 100%.

However, I think the error arises in the way in which Kumar and colleagues are calculating FDR. Specifically, in the response letter they say that the first step of this calculation is to ‘count how many gene pairs are found to be significant by a method M, at some significance level alpha (a p-value for InSTAnT or a correlation value for Chen et al.)’.

I think the key error here is comparing a rigorously calculated p-value in one case and a correlation coefficient in the other. Specifically, correlation coefficient magnitude cannot be interpreted as statistical significance; thus, I suspect that many of the genes identified by this interpretation of Chen et al are indeed not statistically significant. Critically, a rereading of Chen et al makes clear that they made no effort to interpret their correlation coefficients in this fashion. Rather they made a, perhaps

heuristic, assessment of validity by looking for co-varying groups of genes.

Thus, without a measure of the p-value for a given correlation coefficient, I do not think it is possible to make a fair comparison between an FDR calculated via a p-value for one method and via interpretation of the magnitude of the correlation coefficient as a measure of significance for another. The lack of a specific significance for any given correlation pair may be why Chen et al do not appear to draw conclusions from any individual pairwise correlation coefficient.

So, in short, I don't think it is an accurate representation of Chen et al's method to state that it has a 100% FDR, as the biological results simply don't support this, as Chen et al drew no measure of statistical significance for any individual pairwise correlation coefficient, as it is not possible to use a correlation coefficient directly as a measure of significance (as Kumar and colleagues do here), and it is not a meaningful comparison between techniques if one leverages a rigorously calculated p-value for an FDR estimate while another uses this interpretation of correlation coefficient.

My recommendation would be for the authors to consider removing this estimate and instead highlight the clear qualitative advantages of InSTAnT over the method in Chen et al, including both the increased spatial resolution and the ability to define a precise p-value for each pair of genes.

Otherwise, I whole heartedly support the publication of this work.

Response to Reviewer #3

Comment: The only remaining sticking point that I have is the way in which they compare their work to the previous spatial analysis method presented in the original MERFISH paper by Chen et al. I raised a concern in my previous review that the authors produce an estimate of a false discovery rate (FDR) of 100% for that analysis method yet this did not seem to be accurate, as Chen et al describe groups of spatially colocalizing genes that make biological sense. Thus, a 100% FDR would appear far more likely to be an overestimate based on how this quantity was calculated.

I thank the authors for expanding on this point in their rebuttal; however, I do not think that they have addressed my central concern with their response. Now, thanks to their detailed response, I suspect that I may have found the issue.

First, returning to Chen et al, the authors in that work used their approach, albeit limited and rudimentary compared to InSTAnT, to identify genes enriched near the perinuclear region and enriched in the cellular periphery by identify spatially co-varying sets of genes and then visually examining the intracellular location of the two groups they found. Supporting the accuracy of this analysis, the authors found that the perinuclear group contained genes known to be translated at the endoplasmic reticulum while the peripheral group was enriched in genes associated with cytoskeletal function. These results are expected biologically, strongly arguing that the false discover rate of this method cannot be 100%.

However, I think the error arises in the way in which Kumar and colleagues are calculating FDR. Specifically, in the response letter they say that the first step of this calculation is to ‘count how many gene pairs are found to be significant by a method M, at some significance level alpha (a p-value for InSTAnT or a correlation value for Chen et al.)’.

I think the key error here is comparing a rigorously calculated p-value in one case and a correlation coefficient in the other. Specifically, correlation coefficient magnitude cannot be interpreted as statistical significance; thus, I suspect that many of the genes identified by this interpretation of Chen et al are indeed not statistically significant. Critically, a rereading of Chen et al makes clear that they made no effort to interpret their correlation coefficients in this fashion. Rather they made a, perhaps heuristic, assessment of validity by looking for co-varying groups of genes.

Thus, without a measure of the p-value for a given correlation coefficient, I do not think it is possible to make a fair comparison between an FDR calculated via a p-value for one method and via interpretation of the magnitude of the correlation coefficient as a measure of significance for another. The lack of a specific significance for any given correlation pair may be why Chen et al do not appear to draw conclusions from any individual pairwise correlation coefficient.

So, in short, I don’t think it is an accurate representation of Chen et al’s method to state that it has a 100% FDR, as the biological results simply don’t support this, as Chen et al drew no measure of statistical significance for any individual pairwise correlation coefficient, as it is not possible to use a correlation coefficient directly as a measure of significance (as Kumar and colleagues do here), and it is not a meaningful comparison between techniques if one

leverages a rigorously calculated p-value for an FDR estimate while another uses this interpretation of correlation coefficient.

My recommendation would be for the authors to consider removing this estimate and instead highlight the clear qualitative advantages of InSTAnT over the method in Chen et al, including both the increased spatial resolution and the ability to define a precise p-value for each pair of genes.

Otherwise, I whole heartedly support the publication of this work.

Response: We thank the reviewer for the suggestion, and have now updated the manuscript to remove the assessment of FPR of the method of Chen et al. Instead, we now highlight the qualitative advantages of InSTAnT, as follows:

"The only alternative approach for aggregating colocalization information across cells is the bin-based based approach of Chen et al.[13] (Methods). This approach calculates the correlation coefficient between transcript counts of a gene pair in four subcellular regions (bins), and aggregates correlations across cells. The low sample count (four) used in correlation calculation may result in less reliable colocalization quantification compared to the rigorous p-values of the CPB test. Furthermore, the coarse binning may result in missed colocalized pairs at finer spatial resolutions, e.g., ~4 micron, while InSTAnT robustly handles such resolution. These considerations underscore the importance of InSTAnT's rigorous statistical testing procedures for reliable detection of spatial patterns."

The following text has been removed:
"We compared the CPB test with the only alternative method for aggregating colocalization information across cells, the bin-based approach of Chen et al.13, which yielded estimated FPR ~ 100% and reproducibility < 10% (Supplementary Figure 15)."

Supplementary Figure 15 has also been removed.

Reviewers' Comments:

Reviewer #3:
Remarks to the Author:
Again, I thank the authors for a strong contribution to the literature and for a careful consideration of my one last point. The modifications that they have made remove all of my concerns, and I fully support the publication of the manuscript.