College of Medicine and he is a member of the Pelotonia Institute for Immuno-Oncology of Ohio State University as a candidate and a member. His research focuses on integrative analysis on synaptic and genomic data with biomedical real data. So welcome back Dongjun Chung. Although it’s all virtual. I hope someday we can see in person. So today I will discuss our recent project about the SPRUCE and MAPLE: Bayesian Multivariate Mixture Models for Spatial Transcriptomics Data.
And also a great amount of the data has been generated using the single cell genomic technology. Single cell genomic experiments provide high-dimensional data at the cell level. By doing so, it allows to investigate cellular heterogeneity within each subject or the patient which was not possible previously with the bulk of genomic data. Which means that genomic data collected at the tissue level. So some kind of standard visualization of the single cell genomic data is called a UMAP. This UMAP shows the distribution of the different clusters in the tumor, including the different immune cell type. And in this way, we can interrogate different types of the immune cell composition. And also there, we can look at what kind of general feature imaged for each cell cluster. One of the recent (numbles) is the emergence of the high-throughput spatial transcriptomics or the HST technology. So, with the emergence of the HST technology, we do not only look at the gene expression in the cell level or the close-to-cell level.
We can now also notice that there are cross pointing spatial information. The figure at the bottom shows one example. And here it shows the mouse brain tissue, and each cell cone. Here cross pointer to one spot which is a group of the smaller... small number of like two to ten at most. And color here indicate expression level of different gene. So left one cross point to the Hpca gene. Right one cross point to the Ttr gene, for example. And with the HST data, we can do a lot of interesting science to improve the parity in current medication. So for example, we can now look at the spatial information of the tissue architecture at the transcriptomics level. And then we can also investigate the cell-cell communication with the spatial information in our hand. So at the figure at the bottom left shows the UMAP. And here, the different color indicates a different cell cluster. And if you look at the figure on the right, then you can see that there are a cluster in a meaningful way on the tissue.
So in this way, we do not look at the different cell types within a tissue. But also look at their spatial information at the same time.

And there’s many exciting applications of the HST experiment, including the neuroscience.

Including the brain cancer study such as the immuno-oncology and the developmental biology which looks at the changes of the cellular composition across the different stage of the development.

And here I specifically discuss the application in the cancer, especially the tumor microenvironment.

And with the spatial information, we can now study their location of the immune cell and the tumor cell in the tumor tissue.

We can also interrogate implication of distance on the tissue and their corresponding density.

And we can also study the distribution of the immune regulator.

And finally, the special spacial patterns such as the tertiary lymphoid structure.

Then from the statistical point of view, how the HST data look like.

The first observation is in the HST data spatial structure, in the tissue architecture in a meaningful way.

So as you discussed earlier,
we can see a similar type of the cell cluster often located in the close proximity in the tissue.

And even after we exclude such kind of cell competition in the spatial location,

we can start to see some spatial pattern in the patient

So the figure on the top shows the expression pattern of the three genes, PCP4, MBP and MTC01. After regressing out, with respect to the cell clusters.

As you can see, even after considering the cell cluster patterns,

you can start to see some interesting spatial patterns.

That the figure at the bottom shows the distribution of each gene for each cell cluster.

And you can see that sometimes it’s asymmetric but also often we can see non-symmetry in vascular distribution for each gene.

So these are some of the key features of the HST data we want to consider.

Gene expression outcomes feature complex correlation such as the spatial correlation,

and also gene-gene correlation,
which mainly effects the biological pathway. Spatial structure can be cellular clustering entity expression patterns. And gene expression densities, often feature skewness and or heavy tears due to outlier cell spots. So ideally we seek to provide a model for identifying the tissue architecture while accommodating these challenging features. So, especially during the last two years, several statistical methods have been proposed to model HST data. And still many of them are network-based approaches. Partially because the stragglers; the very famous packages for the single cell genomic data analysis. And network-based approach has been proven to be powerful in this context. So based on that multiple network-based approach have been proposed including the Giotto, Seurat and stLearn. Because in the statistical model, recently BayesSpace was proposed by the group of the at the Fred Hutchinson. And essentially, it uses a multivariate-t mixture model to cluster cell spots.
It implement spatial smoothing of clusters via a Pott’s model prior on cluster labels. And interestingly, they try to predict sub-spots to increase the resolution.

In spite of such interesting features, it also has some number of drawbacks. For example, it assumes the symmetry of the gene expression densities, and it also relies on the approximate inference.

Our goal is to develop a statistical model that overcome these limitations and provide the optimal tissue architecture prediction using the HST data which we call SPRUCE or the spatial random effects-based clustering of the single cell data.

So here we use the i as the index for the cell spot and then we denote \( y_i \) as the length of gene expression vector for spot i. And based on the \( y_i \), we also may find a mixture model of the form.

So here we assume the k number of the mixture component. Theta k indicates the set of the parameters.
specific to mixture component $k$. We further introduce $z_1$ to $z_n$, which are the latent mixture component indicators.

For mixture component $k$, $\pi_k$ is the probability of spot $i$ belonging to component $k$. And as I mentioned earlier, can you see the gene-gene correlation are key features of the HST data?

So to account for skewness and gene-gene correlation, we assume a multivariate skew-normal distribution. Where is the parameters?

First one indicates the main vector for spot $i$, and $\alpha_k$ indicates gene-specific skewness parameters for mixture component $k$. And $\omega_k$ is the $gg$ scale matrix that captures correlation among the gene expression feature in the component $k$.

And then we further represent MSN distribution using a convenient conditional representation. We use $\mu_k$ for the mean of component $k,$ $\phi_i$ for the spatial effect, $\tau_i$ and $\kappa_k$ for the component-specific skewness of each gene. Epsilon $i$ for the multivariate normal error.
And then in order to further accommodate spatial dependence, we used the multivariate intrinsic conditionally autoregressive, or the CAR prior for phi i. So essentially, given all the spots except for spot i, we might suggest pi i as the normal distribution with the mean of its neighbors. And with the covariance matrix denoted as the lambda. And as you can see earlier, we see the two different levels of the spatial patterns. One for the spatial pattern of defect clustering. And another one is the spatial pattern of the gene expression. So for the spatial pattern of the cell clusters, we want to allow the probability of pi of belonging to each mixture component. Also to vary spatially as well. So in order to do so, we extend model I showed previously using the pi i k, which is the i specific. And then here we modeled this one as the sigmoid of the two parameters. And then part one in the interceptor for the baseline propensity of the membership into component k shared by all cell spots.
And second term indicates the spatial random effects allowing the variation about the intersect.

And again, to introduce the spatial association into the component membership model, we further assume the univariate intrinsic CAR prior. As you can see here. And here the one computational challenges, if you’re interested, is format. Then it do not allow us to... It do not provide the closed form posterior distribution, which prevent Gibbs sampler. And in order to address this computation challenge, we extended our model using the results from the Polson et al in 2013, Jasa on Polya-Gamma data augmentation to allow for Gibbs sampling of the mixing weight model parameters. And essentially, we could assume that this can be represented as the Polya-Gama Data Augmentation. And by doing so, everything can be implemented as the Gibbs sampler. In the case of the further outliers or heavy-tails, we can even further extend the model to the multivariate skew-t distribution.
That you can see here. Which can be very easily implemented given the existing model. To complete our model specification, we use the weekly specified prior, and then the quantity of prior. And by using this conjugate prior, we can do everything using the fully Gibbs sampler of the closed form which provide the best computation. So here, the one question is the optimal number of the \( k \) worked in number of disparate clusters. So for the proposal, we use the product of the model selection approaches, specifically we use the WAIC, or the widely applicable information criterion. In the patient mixture it’s very common to observe the label switching program. So to protect against the label switching issue in the MCMC sampler, we use the canonical projection of \( z \) using the Peng and Cavalho, in 2016. And finally for the actual implementation, we use the Rccp to further improve the computation efficiency. We implement the proposed model as on our package SPRUCE,
and it’s currently available from our data page. Here.

And then the figure shows our digital page. When we developed our software, one of the popular software to pre-processing and analyzing the HST data is the Seurat workflow. So when you develop our software, we provide integration with the Seurat workflow so that our software can be embedded as part of the (mumbles) flow.

So for example, the data can be loaded into our using the Seurat, and then people can apply the pre-processing using the Seurat workflow. And then that objective can be fed into the SPRUCE analysis workflow. And then the output from the SPRUCE can, again, fit into the Seurat workflow for the visualization.

And downstream analysis So first for the simulation, the first for the simulation is about the... Has the two purposes. So first one is to assess the validity of the parameter estimation algorithm. And second is to quantify the effect of ignoring skewness and spatial information.
So in order to make our simulation more realistic,
we use the sagittal mouse brain data as the tissue shape
and the spot location.
And we simulated the full clusters from the multivariate skew-normal distribution
with the 16 genes.
We considered the 26... 2696 spots.
And then we considered three models,
including the multivariate normal,
multivariate skew-normal,
and with no skew-normal with no spatial.
So first one shows the implication of inadequate study of skewness and spatial.
Second shows the implication of ignoring the spatial structure.
And the final was our proposed model.
And here the top left figure,
shows the true cluster labels.
And top of right shows
the UMAP reduction of the gene expression pattern.
And as you can see, we can make the orange and the green,
which is far away from each other,
similar in the gene expression,
so that it can be more challenging in the prediction.
And we really test the performance of each model.
using the ARI where the very close one indicates the better performance. And as you can see here, when we ignore the skewness and the spatial pattern, there is the big loss of the ARI. And by considering the skewness, we gain some but still that there is being lost. And by further considering the spatial pattern, we can improve the high level of the ARI. And for the real data application, we consider the two applications. So, to compare the performance of the SPRUCE to existing tools, we used the 10X Visium human brain data from the Maynard et al, 2021, Nature Neuroscience. Here at the rehab we have about the 3000 spots. And one of the good aspect of this data is it’s very well annotated. So, the author, using his expert knowledge, they annotated the 3000 spots into the 5 brain layers. Including the white matter and the frontal cortex layers. And as I mentioned earlier, we use the standard Seurat pre-processing pipeline, including the normalization of using the sc transform and also selection of the most variable genes.
using the existing pipeline.

We consider the top 16 most variable genes.

And we also consider the three other existing algorithms including BayesSpace, stLearn, Seurat and Giotto as the computing algorithms.

And we use the default parameters for each of them.

Here it shows the regions and top left figure shows the manual annotation provided by the author in the paper.

You can see the nice, five spatial clusters from inside out.

And also there you can see that there is one, narrow cell cluster corresponding to the number four.

Here we showed the real data for the SPRUCE, BayesSpace, stLearn, Seurat and the Giotto.

And in this case, the network-based approaches, including the stLearn, Seurat and the Giotto, all showed a lower performance compared to those algorithms.

The BayesSpace showed relatively higher performance about the ARI of 0.55.

SPRUCE further improved the performance compared to the BayesSpace.

And one thing I noted here is the narrowed cell cluster,
could it be identified by the SPRUCE?
Which is interesting.

And as the second example.

So first one is the more labeled data.

We can compare our prediction to the existing annotation.

And to further demonstrate the application of the SPRUCE

to unlabeled data, we analyze the publicly available

human invasive ductal carcinoma breast tissue.

Again using the 10X Visium platform.

And we essentially followed the similar workflow

and we identify the top 16 most spatially variable genes.

And those included several tumor associated antigens,

in creating the GFRA1 and CXCL14.

And also that there is the tumor suppressive gene,

like MALAT1.

And we use the SPRUCE to identify the 5 sub regions

using these 16 features.

This shows the 16 most variable genes.

And you can see that there are very clear spatial patterns.

For example the CXCL14 and GFRA1,

expel on the right bottom side.

While the MALAT1 express higher in the top left side.

And this is the cluster prediction

made by the SPRUCE algorithm.
And you can see that it identified the cluster too, which it highly coincide with the CLCX14 and GFRA1 with a study on.

What the cell cluster 1, Is the MALAT1 which is more tumor suppressor.

So here we can see that the SPRUCE can identify the different group of the tissue architecture, such as the tumor suppressor and then tumor related.

And we can also easily look at there, within cluster expression pattern and gene-gene correlation.

As you could see earlier, on cell cluster 2 which equals 0.2 to the right higher than the GFRA1 and CXCL14.

One, which is the cross point here is the high-end MALAT1 and gene-gene correlation.

So, in the case of cell cluster 2, there’s a very strong gene-gene correlation pattern.

So we just support the proposed model that considered spatial pattern and also gene-gene correlation simultaneously.

So, so far I discussed the method for our SPRUCE and its application.
And that we essentially expanded our work a little bit more to the MAPLE, which is the multi-sample spatial transcriptomics model. Why we care about the multi-sample analysis of HST data?

So currently most algorithms are designed in a way that it can more focus on a single sample. But even intuitively, joint analysis of the multiple HST data can potentially boost the signal by sharing the information amongst samples. And also the joint analysis of the different samples can allow the differentiation analysis of the HST data.

So very often, each tissue is not our main interest. But we also want to compare tissue architecture between the different samples. For example, between the disease group versus the controls, responders versus the non responders to 13 treatments, such as the cancer immuno-therapy.

So to offset this limitation, we proposed MAPLE. And actually our existing SPRUCE framework already allows this one naturally.

So, simply what it can do is instead of now analyzing each sample individually,
we can jointly analyze all the samples together. And then by doing so, we can share information about the modeling of each cell spot cluster, and also their spatial pattern. But by introducing the sample-level covariate exp xi in the cell type composition, we can see the impact of the different sample-level covariate. Which I show more in detail in the coming slides. So the first application is the same mouse brain data, the human brain data... Sorry this should be the mouse brain, and here we see the two anterior parts, which look very similar. And then as you can see here, when we jointly analyze the two sample cross pointing to the same part of the brain. It nicely identifies the cross pointing part between the two sample. Like one in the end, three on the top, five at the bottom and so on. And because this is the Bayesag framework, it can also provide uncertainty measures about our clustering prediction. And as you can see usually there is more uncertain about the clustering prediction.
around the boundary between different cell clusters. Which kind of makes sense, because we expect that maybe cell type might be more mixed together in the same cell spot. Also, there are some cell clusters with the higher level of the uncertainty of which we are still trying to understand more at this point. And this kind of the figure is the... what utility of this kind of joint analysis. So, for the identifier with T, we set the first cell cluster as the reference. And then here we see the two (mumbles) The top one shows the intercept, and then we can interpret this one as the relative size of each cell cluster. So then compared to the one, we can say three and the six are larger. So the three and the six are larger, compared to the one. Why the four is the smaller, well just smaller compared to the one. So this is what it can see by eye from the tissue prediction region. But good thing is that this model allows us to quantify, what you see by eye. And what is more interesting is the second one.
So this one, 00:31:58.283 is about the difference between the two sample.

So again, 00:32:01.530 so basically if it’s higher,

then it means that certain tissue spot cluster

getting bigger in the second sample.

And if it’s lower immune state is a kind of smaller

in the second sample and so on.

So in this way, 00:32:21.192 we can quantify the change of the tissue architecture

between different cell clusters.

And another interesting example is this one.

So here, the image of 2D to anterior samples, 00:32:39.431 we now also look at the posterior sample as well.

So because this is two parts of the brain 00:32:47.950 anterior and the posterior,

the issue is kind of continuous between two.

And as you can see here,

cell cluster three is connected to the posterior side here.

Cell cluster one is connected to here and so on.

And then this kind of pattern is not clear 00:33:08.050 if you analyze each data independently.

And our MAPLE framework nicely captures

such kind of sharing pattern.

And also the difference pattern

between the different samples, interestingly.

So at this point,
we are working on more simulation study and the real data analysis to further show the performance and understand the properties of the MAPLE at this point.

So then I can’t summarize my presentation today.

So the high throughput spatial transcriptomics, or HST, provides unprecedented opportunities to investigate novel biological hypotheses, such as the tumor microenvironment and certain structure about the human brain and Alzheimer, and so on.

And here we propose SPRUCE, a Bayesian multivariate mixture model for HST data analysis.

SPRUCE has multiple strengths including the novel combination of the skewed normal density, Polya-Gamma data augmentation, and spatial random effect.

Altogether, it allows to precisely infer spatially correlated mixture component membership probabilities.

In our simulation study and real data analysis, we could see that SPRUCE outperforms the existing method, in the tissue architecture identification.

And finally our recent extension of the MAPLE
allows the joint clustering and differential analysis of multiple HST data. So at this point SPRUCE is on the review in, (mumbles) in the biometrics. Cross pointing manuscript is available in the bio archive. And there are multiple ongoing work regarding the HST data modeling in our lab. So we are actually currently working on further improving the SPRUCE and the MAPLE by incorporating other characteristics of the HST data, such as the relationships among cells. For example, we know that there are some likened and receptor, for example. Which we expect that they interact with each other in their cell structure. And then by incorporating different prior information, we can further improve the SPRUCE and MAPLE. We are also working on the other statistical models for somewhat relevant, but different tasks. For example, currently we are also working on the streamlining framework,
595 00:36:18.820 --> 00:36:20.973 especially the graph neural network,
596 00:36:21.918 --> 00:36:24.186 which is called RESEPT.
597 00:36:24.186 --> 00:36:27.019 And then using the gene framework,
598 00:36:27.853 --> 00:36:29.610 we tried to come up with good embedding
599 00:36:29.610 --> 00:36:32.163 of the HST gene expression pattern.
600 00:36:34.290 --> 00:36:37.510 Our current results show that such a combi-
nation
601 00:36:37.510 --> 00:36:41.420 of the stem learning and the statistical model
approach
602 00:36:41.420 --> 00:36:44.303 can provide nice prediction performance.
603 00:36:47.149 --> 00:36:50.437 For this proposal, we developed a framework
called RESEPT
604 00:36:51.970 --> 00:36:54.020 and cross pointing bio archive
605 00:36:54.877 --> 00:36:57.090 is also available publicly.
606 00:36:57.090 --> 00:36:59.350 And then cross pointing paper
607 00:36:59.350 --> 00:37:02.843 is now under revision in the nature communica-
tions.
608 00:37:05.850 --> 00:37:08.724 Regarding cell-cell communications,
609 00:37:08.724 --> 00:37:11.980 using network-based approaches has some
benefit
610 00:37:11.980 --> 00:37:15.930 because the cell-cell communication can be
nicely
611 00:37:15.930 --> 00:37:19.403 and naturally modeled using AGR network.
612 00:37:20.976 --> 00:37:24.620 So we have the parallel work called the the
Banyan
613 00:37:24.620 --> 00:37:26.970 to identify the cell-cell communication
614 00:37:26.970 --> 00:37:31.230 and tissue architecture using the network-
based approaches.
615 00:37:31.230 --> 00:37:36.230 And finally, there are the multiple effort ex-
perimentally
616 00:37:36.930 --> 00:37:40.573 to generate the spatial multimodal data.
617 00:37:41.670 --> 00:37:42.503 For example,
618 00:37:42.503 --> 00:37:47.503 the effect to seek such as the single cell ge-
genomics,
proteomics and the T-cell receptor at the same time.

And very soon,
everything are expected to be combined as the spatial transcriptomic structure.

We are working on the direction to develop the statistical model for integration of the HST data with other matched data.

So I would like to acknowledge my research team at OSU.

Carter Allen is the main driver this project, and also my pitch assistant Qin Ma and Yuzhou Chang is my close collaborator.

Zihai Li, who is the director of the Immuno-Oncology Institute and also the expert in cancer.

Won Chang at the University of Cincinnati who are the spatial statistics expert, and MUSC collaborator Brian Neelon and my grant support.

So, and this is the end of my presentation, and you can find my manuscript and the software from the link here.

If you have any questions and comment, please let me know by email at chung.911@osu.edu.

So thank you for your attention.

Do we have any questions from the audience in the classroom,
Can I ask a question?

Can you hear me?

Yes, mm-hm.

Right, Dongjun welcome back.

Great work, it's a nice presentation.

I'm just wondering, like,

when you do this from your own experience

on the cell clustering,

how much the spatial information contributes

to the clustering.

Sure.

So,

(mumbles)

If you're here,

so if you look at the Seurat workflow,

you can see there's a still lot of the, kind of,

local boundary between different cell spot clusters.

And when you analyze the same data using the SPRUCE,

you can see much cleaner boundary.

And often it will coincide with the

expert analogy annotation.

So given that there is the significant contribution,

definitely even the gene expression,

we still get some big picture, as you can see here.

But spatial information provide much cleaner prediction

about the tissue architecture in general.
I see.

And also the skewness.

Do you estimate that or that’s like your heart was persuaded by the skewness?

You mean which one?

On k model.

Your model to specify, the k model you have there.

I missed that part.

Like, do you need to specify the skewness?

Or learn from data.

Oh, I see.

But from the data, how skew?

I mean, just in terms of how stable that alpha k can be estimated.

So maybe I can answer it in two different ways.

So if there is this skewness in the data, I think yes.

So we’ll say it depends on how processed the data as well.

So usually there’s three different approaches to model the HST data in closed spatial embedding gene.

And so you can see here,

who are the people using the principle components?

Who are the people use the team learning as the embedding step?

If you use the team learning or the PCA it’s more likely symmetry in the real data.

If you consider the spatial embedding gene,
we often hope to have the skewness, as you can see here.

And then regarding your question, overall it works well.

I don’t have the exact quantification, but it works well.

Especially stably in most cases.

Yeah, I read the spatial Bayes paper. They seem to be working on the principle components, right?

They do not work on individual genes, right?.

No, yeah.

They base this on the PCA.

Yeah, that’s why it’s completely puzzling me.

while you’re doing that.

But anyway, yeah.

They mainly target the PCA.

So they only can start the multivariate distribution.

And also because of the same reason,

their equivalence metrics means less density.

I see.</n>

Thank you.

Thank you.

Do we have any questions from students in the classroom?

Wait, can I ask another question?

So, towards the end,
you mentioned you tried to look at the cell-cell communication. That part. I’m very interested in that. From our experience on the single cell spatial data are...

Are you talking about you’re learning from the single cell, or the spatial single cell?

So, regarding the cell-cell communication it’s still very ongoing research at this point. I mean, not just our side but in general. Because most of the cell-cell communication prediction based on the database. So based on data, like on the receptor, pairing the database and checking their cross point on the expression in cross point of the cell. And then by checking that the cross pointing pair of the expression pattern between the like and the receptor. They want to model cell-cell communication. It’s not perfect, as you know, because it’s like a computer. If you look at the chip, it’s almost like

If you look at the chip, it’s almost like

So there’s some limitation,
but it’s a more likely general limitation at this point.

I’m asking because we’ve been looking at some of the spatial single cell data that were too noisy for the like receptor gene expression levels. Just couldn’t make it too far.

But for a single cell, may be different? I mean, probably there’ll be more that, like...

Like looking at multi-modal or the cell cluster, rather than the cell. It’s still very multiple experimental limitation, at this point.

So, if we have samples from...

On the data from multiple samples

So, actually we can answer in the two different ways.
good pre-processing is still important because it still depends on the expression patterns. But still regarding the differences between the different tissues. If there is a big difference, it can still detect the difference between the different sample. So, it can detect spots. But still like a main goal is more for the similar type of tissue. If it’s too different, maybe it’s different research project. So, for example, here our targets is more about, for example, like same breast tissue, but with a different responders and non-responders group, for example. Or like a cell-cell long tissue, but the tumor but not tumor and so on. If you like a human and mouse, then it might be somewhat different story, which might need much more work. Do we have any more questions here? Okay, can we have all the questions from the audience on zoom? Okay, so it looks like we don’t have any more questions. So Dr. Chung, thank you again for your nice presentation.
Look forward to meeting in person sometime soon.

And then thank you again Wei and Hongyou.

And it’s a great come back, although virtually.

And I hope to see you again.

We’ll come by in person.

Hopefully someday soon.

Okay, thank you.