WEBVTT

1 00:00:00.825 --> 00:00:03.748 <v ->College of Medicine and he is a member</v>

2 00:00:03.748 --> 00:00:07.927 of the Pelotonia Institute for Immuno-Oncology

3 00:00:07.927 --> 00:00:11.527 of Ohio State University as a candidate and a member.

4 00:00:12.406 --> 00:00:14.514 His research focuses on

5 00:00:14.514 --> 00:00:16.681 (mumbles)

6 00:00:18.163 --> 00:00:21.570 for integrative analysis on synaptic and genomic data

7 00:00:21.570 --> 00:00:23.737 with biomedical real data.

8 00:00:25.820 --> 00:00:29.140 So welcome back Dongjun Chung.

9 00:00:29.140 --> 00:00:31.967 (audience member claps)

10 00:00:31.967 --> 00:00:32.800 <v ->Okay.</v>

11 00:00:34.270 --> 00:00:36.640 Thank you Wei, for the kind introduction

12 00:00:36.640 --> 00:00:38.880 and it's so great to come back.

13 00:00:38.880 --> 00:00:40.523 Although it's all virtual.

14 00:00:42.750 --> 00:00:44.873 I hope someday we can see in person.

15 00:00:45.840 --> 00:00:50.840 So today I will discuss our recent project

16 00:00:51.560 --> 00:00:56.560 about the SPRUCE and MAPLE: Bayesian Multivariate

17 00:00:57.010 --> 00:01:00.840 Mixture Models for Spatial Transcriptomics Data.

18 00:01:00.840 --> 00:01:03.290 Oh, by the way, can you hear me well?

19 00:01:03.290 --> 00:01:04.634 <v ->Ah yes, we can hear you.</v>

20 00:01:04.634 --> 00:01:05.717 <v ->Okay, great.</v>

21 00:01:07.431 --> 00:01:11.500 So, let me start us from some quick introduction

22 00:01:11.500 --> 00:01:14.610 about the single cell genomics.

23 00:01:15.890 --> 00:01:16.960 So in some sense,

24 00:01:16.960 --> 00:01:21.420 we can say that the last decade was the era of single cell

25 00:01:21.420 --> 00:01:23.517 genomic experiments.

26 00:01:23.517 --> 00:01:26.380 So it changed science in many ways.

27 00:01:26.380 --> 00:01:31.380 And also a great amount of the data has been generated

28 00:01:32.050 --> 00:01:34.523 using the single cell genomic technology.

29 00:01:35.601 --> 00:01:38.340 Single cell genomic experiments

30 00:01:38.340 --> 00:01:42.141 provide high-dimensional data at the cell level.

31 00:01:42.141 --> 00:01:43.660 By doing so,

32 00:01:43.660 --> 00:01:48.350 it allows to investigate cellular heterogeneity

33 00:01:48.350 --> 00:01:51.570 within each subject or the patient

34 00:01:51.570 --> 00:01:54.250 which was not possible previously

35 00:01:54.250 --> 00:01:56.000 with the bulk of genomic data.

36 00:01:56.000 --> 00:02:00.393 Which means that genomic data collected at the tissue level.

37 00:02:03.984 --> 00:02:07.151 So some kind of standard visualization

38 00:02:08.583 --> 00:02:12.892 of the single cell genomic data is called a UMAP.

39 00:02:12.892 --> 00:02:13.725 And here,

40 00:02:13.725 --> 00:02:18.330 this UMAP shows the distribution of the different clusters

41 00:02:18.330 --> 00:02:19.780 in the tumor,

42 00:02:19.780 --> 00:02:24.780 including the different immune cell type.

43 00:02:24.890 --> 00:02:27.020 And in this way,

44 00:02:27.020 --> 00:02:29.040 we can interrogate different types

45 00:02:29.040 --> 00:02:31.583 of the immune cell composition.

46 00:02:32.445 --> 00:02:33.460 And also there,

47 00:02:33.460 --> 00:02:36.740 we can look at what kind of general feature

48 00:02:36.740 --> 00:02:38.823 imaged for each cell cluster.

49 00:02:40.390 --> 00:02:43.271 One of the recent (mumbles)

50 00:02:43.271 --> 00:02:45.820 is the emergence of the high-throughput

51 00:02:45.820 --> 00:02:50.690 spatial transcriptomics or the HST technology.

52 00:02:50.690 --> 00:02:55.690 So, with the emergence of the HST technology,

53 00:02:56.060 --> 00:02:59.390 we do not only look at the gene expression

54 00:02:59.390 --> 00:03:03.230 in the cell level or the close-to-cell level.

55 00:03:03.230 --> 00:03:05.880 We can now also notice that there are cross pointing

56 00:03:05.880 --> 00:03:07.283 spatial information.

57 00:03:08.350 --> 00:03:11.710 The figure at the bottom shows one example.

58 00:03:11.710 --> 00:03:16.260 And here it shows the mouse brain tissue,

59 00:03:16.260 --> 00:03:18.520 and each cell cone.

60 00:03:18.520 --> 00:03:21.270 Here cross pointer to one spot

61 00:03:21.270 --> 00:03:24.130 which is a group of the smaller...

62 00:03:24.130 --> 00:03:29.020 small number of like two to ten at most.

63 00:03:29.020 --> 00:03:34.020 And color here indicate expression level of different gene.

64 00:03:34.410 --> 00:03:39.030 So left one cross point to the Hpca gene.

65 00:03:39.030 --> 00:03:43.693 Right one cross point to the Ttr gene, for example.

66 00:03:47.850 --> 00:03:49.600 And with the HST data,

67 00:03:49.600 --> 00:03:52.710 we can do a lot of interesting science

68 00:03:53.660 --> 00:03:56.890 to improve the parity in current medication.

69 00:03:56.890 --> 00:03:58.680 So for example,

70 00:03:58.680 --> 00:04:01.430 we can now look at the spatial information

71 00:04:01.430 --> 00:04:06.430 of the tissue architecture at the transcriptomics level.

72 00:04:07.330 --> 00:04:09.430 And then we can also investigate

73 00:04:09.430 --> 00:04:12.960 the cell-cell communication with the spatial information

74 00:04:12.960 --> 00:04:13.833 in our hand.

75 00:04:14.710 --> 00:04:19.250 So at the figure at the bottom left shows the UMAP.

76 00:04:19.250 --> 00:04:20.083 And here,

77 00:04:20.083 --> 00:04:24.140 the different color indicates a different cell cluster.

78 00:04:24.140 --> 00:04:26.710 And if you look at the figure on the right,

79 00:04:26.710 --> 00:04:29.880 then you can see that there are a cluster

80 00:04:29.880 --> 00:04:32.550 in a meaningful way on the tissue.

81 00:04:32.550 --> 00:04:36.470 So in this way, we do not look at the different cell types

82 00:04:36.470 --> 00:04:37.540 within a tissue.

83 00:04:37.540 --> 00:04:42.540 But also look at their spatial information at the same time.

84 00:04:46.597 --> 00:04:49.200 And there's many exciting applications

85 00:04:49.200 --> 00:04:54.200 of the HST experiment, including the neuroscience.

86 00:04:56.570 --> 00:05:01.320 Including the brain cancer study such as the immuno-oncology

87 00:05:02.180 --> 00:05:04.140 and the developmental biology

88 00:05:04.140 --> 00:05:07.723 which looks at the changes of the cellular composition

89 00:05:07.723 --> 00:05:10.563 across the different stage of the development.

90 00:05:11.540 --> 00:05:16.220 And here I specifically discuss the application

91 00:05:16.220 --> 00:05:19.453 in the cancer, especially the tumor microenvironment.

92 00:05:20.310 --> 00:05:22.310 And with the spatial information,

93 00:05:22.310 --> 00:05:26.910 we can now study their location of the immune cell

94 00:05:26.910 --> 00:05:29.823 and the tumor cell in the tumor tissue.

95 00:05:30.666 --> 00:05:35.384 We can also interrogate implication of distance

96 00:05:35.384 --> 00:05:38.143 on the tissue and their corresponding density.

97 00:05:39.000 --> 00:05:42.320 And we can also study the distribution

98 00:05:42.320 --> 00:05:44.990 of the immune regulator.

99 00:05:44.990 --> 00:05:48.785 And finally, the special spacial patterns

100 00:05:48.785 --> 00:05:52.202 such as the tertiary lymphoid structure.

101 00:05:56.113 --> 00:05:59.101 Then from the statistical point of view,

102 00:05:59.101 --> 00:06:01.351 how the HST data look like.

103 00:06:05.259 --> 00:06:10.010 The first observation is in the HST data spatial structure,

104 00:06:10.010 --> 00:06:13.220 in the tissue architecture in a meaningful way.

105 00:06:13.220 --> 00:06:15.550 So as you discussed earlier,

106 00:06:15.550 --> 00:06:19.179 we can see a similar type of the cell cluster

107 00:06:19.179 --> 00:06:24.179 often located in the close proximity in the tissue.

108 00:06:26.800 --> 00:06:31.800 And even after we exclude such kind of cell competition

109 00:06:32.260 --> 00:06:34.600 in the spatial location,

110 00:06:34.600 --> 00:06:38.390 we can start to see some spatial pattern in the patient

111 00:06:38.390 --> 00:06:39.970 on the tissue.

112 00:06:39.970 --> 00:06:43.390 So the figure on the top shows the expression pattern

113 00:06:43.390 --> 00:06:44.963 of the three genes,

114 00:06:46.011 --> 00:06:47.483 PCP4, MBP and MTC01.

115 00:06:49.874 --> 00:06:54.203 After regressing out, with respect to the cell clusters.

116 00:06:55.907 --> 00:06:58.031 And as you can see, even after considering

117 00:06:58.031 --> 00:06:59.590 the cell cluster patterns,

118 00:06:59.590 --> 00:07:03.533 you can start to see some interesting spatial patterns.

119 00:07:04.540 --> 00:07:08.840 That the figure at the bottom shows the distribution

120 00:07:08.840 --> 00:07:13.840 of each gene for each cell cluster.

121 00:07:13.920 --> 00:07:17.270 And you can see that sometimes it's asymmetric

122 00:07:17.270 --> 00:07:21.660 but also often we can see non-symmetry

123 00:07:21.660 --> 00:07:24.513 in vascular distribution for each gene.

124 00:07:26.910 --> 00:07:30.050 So these are some of the key features

125 00:07:30.050 --> 00:07:33.160 of the HST data we want to consider

126 00:07:33.160 --> 00:07:36.620 in the modeling of the HST data.

127 00:07:36.620 --> 00:07:39.200 So if I profile pick somebody,

128 00:07:39.200 --> 00:07:43.310 Gene expression outcomes feature complex correlation

129 00:07:43.310 --> 00:07:45.530 such as the spatial correlation,

130 00:07:45.530 --> 00:07:48.860 and also gene-gene correlation,

131 00:07:48.860 --> 00:07:52.327 which mainly effects the biological pathway.

132 00:07:52.327 --> 00:07:54.642 Spatial structure can be

133 00:07:54.642 --> 00:07:56.230 (mumbles)

134 00:07:56.230 --> 00:07:59.902 cellular clustering entity expression patterns.

135 00:07:59.902 --> 00:08:01.800 And gene expression densities,

136 00:08:01.800 --> 00:08:05.840 often feature skewness and or heavy tears

137 00:08:05.840 --> 00:08:08.133 due to outlier cell spots.

138 00:08:09.454 --> 00:08:13.410 So ideally we seek to provide a model

139 00:08:13.410 --> 00:08:16.400 for identifying the tissue architecture

140 00:08:16.400 --> 00:08:18.923 while accommodating these challenging features.

141 00:08:24.120 --> 00:08:28.040 So, especially during the last two years,

142 00:08:28.040 --> 00:08:32.000 several statistical methods have been proposed

143 00:08:32.000 --> 00:08:34.171 to model HST data.

144 00:08:34.171 --> 00:08:38.870 And still many of them are network-based approaches.

145 00:08:38.870 --> 00:08:43.147 Partially because the stragglers; the very famous packages

146 00:08:43.147 --> 00:08:45.313 for the single cell genomic data analysis.

147 00:08:46.420 --> 00:08:48.530 And network-based approach has been proven

148 00:08:49.430 --> 00:08:51.360 to be powerful in this context.

149 00:08:51.360 --> 00:08:55.554 So based on that multiple network-based approach

150 00:08:55.554 --> 00:09:00.554 have been proposed including the Giotto, Seurat and stLearn.

151 00:09:03.370 --> 00:09:06.120 Because in the statistical model,

152 00:09:07.190 --> 00:09:11.843 recently BayesSpace was proposed by the group of the

153 00:09:11.843 --> 00:09:13.360 (mumbles)

154 00:09:13.360 --> 00:09:15.310 at the Fred Hutchinson.

155 00:09:15.310 --> 00:09:16.450 And essentially,

156 00:09:16.450 --> 00:09:21.140 it uses a multivariate-t mixture model

157 00:09:21.140 --> 00:09:23.830 to cluster cell spots.

158 00:09:23.830 --> 00:09:27.414 It implement spatial smoothing of clusters

159 00:09:27.414 --> 00:09:32.300 via a Pott's model prior on cluster labels.

160 00:09:32.300 --> 00:09:33.980 And interestingly,

161 00:09:33.980 --> 00:09:38.980 they try to predict sub-spots to increase the resolution.

162 00:09:40.890 --> 00:09:43.760 In spite of such interesting features,

163 00:09:43.760 --> 00:09:46.523 it has also some number of drawbacks.

164 00:09:47.380 --> 00:09:48.529 For example,

165 00:09:48.529 --> 00:09:52.210 it assumes the symmetry of the gene expression densities,

166 00:09:52.210 --> 00:09:56.223 and it also relies on the approximate inference.

167 00:09:57.560 --> 00:10:02.560 And here our goal is to develop a statistical model

168 00:10:02.930 --> 00:10:05.530 that overcome these limitations

169 00:10:05.530 --> 00:10:10.530 and also provide the optimal tissue architecture prediction

170 00:10:10.570 --> 00:10:14.600 using the HST data which we call SPRUCE

171 00:10:17.720 --> 00:10:20.300 or the spatial random effects-based clustering

172 00:10:20.300 --> 00:10:21.683 of the single cell data.

173 00:10:30.240 --> 00:10:32.083 So this is our SPRUCE model.

174 00:10:35.404 --> 00:10:39.750 So here we use the i as the index for the cell spot

175 00:10:39.750 --> 00:10:41.123 in the tissue sample.

176 00:10:42.260 --> 00:10:45.010 And then we denote y i

177 00:10:45.010 --> 00:10:48.873 as the length of gene expression vector for spot i.

178 00:10:50.020 --> 00:10:55.020 And based on the y i, we also may find a mixture model

179 00:10:55.670 --> 00:10:57.323 of the form.

180 00:10:58.600 --> 00:11:02.690 So here we assume the k number of the mixture component.

181 00:11:02.690 --> 00:11:03.990 or the cell spot clusters.

182 00:11:05.650 --> 00:11:09.670 Theta k indicates the set of the parameters

183 00:11:09.670 --> 00:11:12.163 specific to mixture component k.

184 00:11:13.105 --> 00:11:17.490 Pi k is the probability of the spot i

185 00:11:17.490 --> 00:11:19.343 belonging to the component k.

186 00:11:22.380 --> 00:11:26.920 We further introduce z1 to zn,

187 00:11:26.920 --> 00:11:30.910 which are the latent mixture component indicators

188 00:11:30.910 --> 00:11:32.675 for each spot.

189 00:11:32.675 --> 00:11:37.018 And zi can have the value between one to k.

190 00:11:37.018 --> 00:11:39.710 And as I mentioned earlier,

191 00:11:39.710 --> 00:11:42.266 can you see the gene-gene correlation

192 00:11:42.266 --> 00:11:46.550 are key features of the HST data?

193 00:11:46.550 --> 00:11:50.861 So to account for skewness and gene-gene correlation,

194 00:11:50.861 --> 00:11:55.861 we assume a multivariate skew-normal distribution.

195 00:11:55.961 --> 00:11:58.670 Where is the parameters?

196 00:11:58.670 --> 00:12:03.140 So first one indicates the main vector for spot i,

197 00:12:03.140 --> 00:12:07.400 and alpha k indicates gene-specific skewness parameters

198 00:12:07.400 --> 00:12:09.700 for mixture component k.

199 00:12:09.700 --> 00:12:14.637 And omega k is the gg scale matrix that captures correlation

200 00:12:14.637 --> 00:12:17.993 among the gene expression feature in the component k.

201 00:12:23.810 --> 00:12:27.910 And then we further represent MSN distribution

202 00:12:27.910 --> 00:12:31.290 using a convenient conditional representation.

203 00:12:31.290 --> 00:12:35.740 We use mu k for the mean of component k,

204 00:12:35.740 --> 00:12:38.420 phi i for the spatial effect,

205 00:12:38.420 --> 00:12:43.420 and t i and ksi k for the component-specific skewness

206 00:12:44.050 --> 00:12:45.103 of each gene.

207 00:12:47.134 --> 00:12:49.563 Epsilon i for the multivariate normal error.

208 00:12:53.235 --> 00:12:57.600 And then in order to further accommodate spatial dependence,

209 00:12:57.600 --> 00:12:59.860 we used the multivariate intrinsic

210 00:12:59.860 --> 00:13:01.820 conditionally autoregressive,

211 00:13:01.820 --> 00:13:05.150 or the CAR prior for phi i.

212 00:13:05.150 --> 00:13:06.473 So essentially,

213 00:13:07.920 --> 00:13:11.087 given all the spots except for spot i,

214 00:13:12.450 --> 00:13:16.710 we might suggest pi i as the normal distribution

215 00:13:16.710 --> 00:13:19.653 with the mean of its neighbors.

216 00:13:21.253 --> 00:13:26.140 And with the covariance matrix denoted as the lambda.

217 00:13:32.960 --> 00:13:35.300 And as you can see earlier,

218 00:13:35.300 --> 00:13:39.870 we see the two different levels of the spatial patterns.

219 00:13:39.870 --> 00:13:43.840 One for the spatial pattern of defect clustering.

220 00:13:43.840 --> 00:13:46.220 And another one is the spatial pattern

221 00:13:46.220 --> 00:13:48.030 of the gene expression.

222 00:13:48.030 --> 00:13:53.030 So for the spatial pattern of the cell clusters,

223 00:13:53.297 --> 00:13:57.720 we want to allow the probability of pi

224 00:13:57.720 --> 00:14:00.730 of belonging to each mixture component.

225 00:14:00.730 --> 00:14:03.400 Also to vary spatially as well.

226 00:14:03.400 --> 00:14:05.340 So in order to do so,

227 00:14:05.340 --> 00:14:09.180 we extend model I showed previously

228 00:14:09.180 --> 00:14:11.890 using the pi i k,

229 00:14:11.890 --> 00:14:13.980 which is the i specific.

230 00:14:13.980 --> 00:14:18.170 And then here we modeled this one as the sigmoid

231 00:14:18.170 --> 00:14:19.733 of the two parameters.

232 00:14:20.993 --> 00:14:23.270 And then part one in the interceptor

233 00:14:23.270 --> 00:14:27.690 for the baseline propensity of the membership

234 00:14:27.690 --> 00:14:31.940 into component k shared by all cell spots.

235 00:14:31.940 --> 00:14:35.380 And second term indicates the spatial random effects

236 00:14:35.380 --> 00:14:40.053 allowing the variation about the intersect.

237 00:14:42.400 --> 00:14:43.320 And again,

238 00:14:43.320 --> 00:14:46.030 to introduce the spatial association

239 00:14:46.030 --> 00:14:48.610 into the component membership model,

240 00:14:48.610 --> 00:14:52.303 we further assume the univariate intrinsic CAR prior.

241 00:14:53.236 --> 00:14:55.320 As you can see here.

242 00:14:55.320 --> 00:14:59.713 And here the one computational challenges,

243 00:15:00.850 --> 00:15:02.863 if you're interested, is format.

244 00:15:04.386 --> 00:15:05.500 Then it do not allow us to...

245 00:15:05.500 --> 00:15:09.770 It do not provide the closed form posterior distribution,

246 00:15:09.770 --> 00:15:12.340 which prevent Gibbs sampler.

247 00:15:12.340 --> 00:15:16.600 And in order to address this computation challenge,

248 00:15:16.600 --> 00:15:19.660 we extended our model

249 00:15:19.660 --> 00:15:24.660 using the results from the Polson et al in 2013, Jasa

250 00:15:25.470 --> 00:15:30.300 on Polya-Gamma data augmentation to allow for Gibbs sampling

251 00:15:30.300 --> 00:15:32.643 of the mixing weight model parameters.

252 00:15:33.510 --> 00:15:34.343 And essentially,

253 00:15:34.343 --> 00:15:38.280 we could assume that this can be represented

254 00:15:38.280 --> 00:15:41.810 as the Polya-Gama Data Augmentation.

255 00:15:41.810 --> 00:15:43.420 And by doing so,

256 00:15:43.420 --> 00:15:47.403 everything can be implemented as the Gibbs sampler.

257 00:15:49.220 --> 00:15:53.140 In the case of the further outliers or heavy-tails,

258 00:15:53.140 --> 00:15:55.680 we can even further extend the model

259 00:15:55.680 --> 00:15:58.680 to the multivariate skew-t distribution

260 00:15:58.680 --> 00:16:00.325 that you can see here.

261 00:16:00.325 --> 00:16:02.850 Which can be very easily implemented

262 00:16:02.850 --> 00:16:04.523 given the existing model.

263 00:16:06.539 --> 00:16:09.700 To complete our model specification,

264 00:16:09.700 --> 00:16:13.690 we use the weekly specified prior,

265 00:16:13.690 --> 00:16:15.610 and then the quantity of prior.

266 00:16:15.610 --> 00:16:18.720 And by using this conjugate prior,

267 00:16:18.720 --> 00:16:22.670 we can do everything using the fully Gibbs sampler

268 00:16:22.670 --> 00:16:23.840 of the closed form

269 00:16:23.840 --> 00:16:26.053 which provide the best computation.

270 00:16:28.720 --> 00:16:31.303 And some additional consideration.

271 00:16:33.040 --> 00:16:33.950 So here,

272 00:16:33.950 --> 00:16:38.100 the one question is the optimal number of the k

273 00:16:38.100 --> 00:16:40.980 worked in number of disparate clusters.

274 00:16:40.980 --> 00:16:42.470 So for the proposal,

275 00:16:42.470 --> 00:16:46.130 we use the product of the model selection approaches,

276 00:16:46.130 --> 00:16:48.950 and specifically we use the WAIC,

277 00:16:48.950 --> 00:16:51.723 or the widely applicable information criterion.

278 00:16:54.521 --> 00:16:56.820 In the patient mixture it's very common

279 00:16:56.820 --> 00:16:59.950 to observe the label switching program.

280 00:16:59.950 --> 00:17:03.200 So to protect against the label switching issue

281 00:17:03.200 --> 00:17:08.200 in the MCMC sampler, we use the canonical projection of z

282 00:17:08.300 --> 00:17:12.580 using the Peng and Cavalho, in 2016.

283 00:17:12.580 --> 00:17:16.700 And finally for the actual implementation,

284 00:17:16.700 --> 00:17:18.690 we use the Rccp

285 00:17:18.690 --> 00:17:21.833 to further improve the computation efficiency.

286 00:17:27.090 --> 00:17:32.090 We implement the proposed model as on our package SPRUCE,

287 00:17:33.270 --> 00:17:37.280 and it's currently available from our data page.

288 00:17:38.366 --> 00:17:39.199 Here.

289 00:17:40.409 --> 00:17:43.992 And then the figure shows our digital page.

290 00:17:45.069 --> 00:17:47.652 When we developed our software,

291 00:17:49.220 --> 00:17:53.081 one of the popular software to pre-processing

292 00:17:53.081 --> 00:17:55.248 and analyzing the HST data

293 00:17:56.536 --> 00:17:58.453 is the Seurat workflow.

294 00:17:59.661 --> 00:18:01.700 So when you develop our software,

295 00:18:01.700 --> 00:18:05.432 we provide integration with the Seurat workflow

296 00:18:05.432 --> 00:18:10.326 so that our software can be embedded

297 00:18:10.326 --> 00:18:12.180 as part of the (mumbles) flow.

298 00:18:12.180 --> 00:18:14.177 So for example,

299 00:18:14.177 --> 00:18:18.971 the data can be loaded into our using the Seurat,

300 00:18:18.971 --> 00:18:22.690 and then people can apply the pre-processing

301 00:18:22.690 --> 00:18:24.323 using the Seurat workflow.

302 00:18:25.532 --> 00:18:26.365 And then that objective

303 00:18:26.365 --> 00:18:31.140 can be fed into the SPRUCE analysis workflow.

304 00:18:31.140 --> 00:18:34.360 And then the output from the SPRUCE can, again,

305 00:18:34.360 --> 00:18:38.646 fit into the Seurat workflow for the visualization

306 00:18:38.646 --> 00:18:40.580 and downstream analysis

307 00:18:46.385 --> 00:18:48.718 So first for the simulation,

308 00:18:49.942 --> 00:18:54.067 the first for the simulation is about the...

309 00:18:54.067 --> 00:18:55.510 Has the two purposes.

310 00:18:55.510 --> 00:18:59.079 So first one is to assess the validity

311 00:18:59.079 --> 00:19:01.293 of the parameter estimation algorithm.

312 00:19:02.320 --> 00:19:04.870 And second is to quantify the effect

313 00:19:04.870 --> 00:19:07.923 of ignoring skewness and spatial information.

314 00:19:09.250 --> 00:19:13.189 So in order to make our simulation more realistic,

315 00:19:13.189 --> 00:19:17.718 we use the sagittal mouse brain data as the tissue shape

316 00:19:17.718 --> 00:19:20.020 and the spot location.

317 00:19:20.020 --> 00:19:22.800 And we simulated the full clusters

318 00:19:22.800 --> 00:19:26.630 from the multivariate skew-normal distribution

319 00:19:26.630 --> 00:19:28.163 with the 16 genes.

320 00:19:31.032 --> 00:19:32.510 We considered the 26...

321 00:19:34.620 --> 00:19:37.530 2696 spots.

322 00:19:37.530 --> 00:19:40.620 And then we considered three models,

323 00:19:40.620 --> 00:19:43.700 including the multivariate normal,

324 00:19:43.700 --> 00:19:45.040 multivariate skew-normal,

325 00:19:45.040 --> 00:19:48.690 and with no skew-normal with no spatial.

326 00:19:48.690 --> 00:19:51.480 So first one shows the implication

327 00:19:51.480 --> 00:19:54.930 of inadequate study of skewness and spatial.

328 00:19:54.930 --> 00:19:57.510 Second shows the implication

329 00:19:57.510 --> 00:20:00.440 of ignoring the spatial structure.

330 00:20:00.440 --> 00:20:03.453 And the final was our proposed model.

331 00:20:05.040 --> 00:20:07.539 And here the top left figure,

332 00:20:07.539 --> 00:20:10.790 shows the true cluster labels.

333 00:20:10.790 --> 00:20:12.130 And top of right shows

334 00:20:12.130 --> 00:20:17.130 the UMAP reduction of the gene expression pattern.

335 00:20:17.720 --> 00:20:22.070 And as you can see, we can make the orange and the green,

336 00:20:22.070 --> 00:20:24.294 which is far away from each other,

337 00:20:24.294 --> 00:20:25.660 similar in the gene expression,

338 00:20:25.660 --> 00:20:29.970 so that it can be more challenging in the prediction.

339 00:20:29.970 --> 00:20:34.770 And we really test the performance of each model

340 00:20:34.770 --> 00:20:38.638 using the ARI where the very close one

341 00:20:38.638 --> 00:20:40.683 indicates the better performance.

342 00:20:41.910 --> 00:20:45.530 And as you can see here, when we ignore

343 00:20:47.308 --> 00:20:50.550 the skewness and the spatial pattern,

344 00:20:50.550 --> 00:20:52.383 there is the big loss of the ARI.

345 00:20:55.182 --> 00:20:57.013 And by considering the skewness,

346 00:20:57.013 --> 00:20:59.980 we gain some but still that there is being lost.

347 00:20:59.980 --> 00:21:03.950 And by further considering the spatial pattern,

348 00:21:03.950 --> 00:21:06.807 we can improve the high level of the ARI.

349 00:21:10.770 --> 00:21:14.160 And for the real data application,

350 00:21:14.160 --> 00:21:16.943 we consider the two applications.

351 00:21:18.160 --> 00:21:19.690 So,

352 00:21:19.690 --> 00:21:24.690 to compare the performance of the SPRUCE to existing tools,

353 00:21:25.630 --> 00:21:28.880 we used the 10X Visium human brain data

354 00:21:29.740 --> 00:21:33.423 from the Maynard et al, 2021, Nature Neuroscience.

355 00:21:36.340 --> 00:21:40.000 Here at the rehab we have about the 3000 spots.

356 00:21:40.000 --> 00:21:45.000 And one of the good aspect of this data is

357 00:21:45.050 --> 00:21:48.130 It's very well annotated.

358 00:21:48.130 --> 00:21:50.900 So, the author,

359 00:21:50.900 --> 00:21:54.490 using his expert knowledge,

360 00:21:54.490 --> 00:21:59.490 they annotated the 3000 spots into the 5 brain layers.

361 00:21:59.630 --> 00:22:03.443 Including the white matter and the frontal cortex layers.

362 00:22:04.848 --> 00:22:06.030 And as I mentioned earlier,

363 00:22:06.030 --> 00:22:10.510 we use the standard Seurat pre-processing pipeline,

364 00:22:10.510 --> 00:22:15.510 including the normalization of using the sc transform

365 00:22:15.880 --> 00:22:20.080 and also selection of the most variable genes

366 00:22:20.080 --> 00:22:22.306 using the existing pipeline.

367 00:22:22.306 --> 00:22:26.543 We consider the top 16 most variable genes.

368 00:22:28.912 --> 00:22:33.670 And we also consider the three other existing algorithms

369 00:22:33.670 --> 00:22:38.263 including BayesSpace, stLearn, Seurat and Giotto

370 00:22:40.100 --> 00:22:42.460 as the computing algorithms.

371 00:22:42.460 --> 00:22:45.883 And we use the default parameters for each of them.

372 00:22:49.318 --> 00:22:50.568 Here it shows the regions

373 00:22:51.872 --> 00:22:54.490 and top left figure shows the manual annotation

374 00:22:54.490 --> 00:22:57.640 provided by the author in the paper.

375 00:22:57.640 --> 00:23:02.640 And you can see the nice, five spatial clusters

376 00:23:02.905 --> 00:23:05.070 from inside out.

377 00:23:05.070 --> 00:23:07.590 And also there you can see

378 00:23:07.590 --> 00:23:11.271 that there is one, narrow cell cluster

379 00:23:11.271 --> 00:23:14.593 corresponding to the number four.

380 00:23:15.775 --> 00:23:18.392 Here we showed the real data for the SPRUCE,

381 00:23:18.392 --> 00:23:22.533 BayesSpace, stLearn, Seurat and the Giotto.

382 00:23:23.810 --> 00:23:28.260 And in this case, the network-based approaches,

383 00:23:28.260 --> 00:23:32.087 including the stLearn, Seurat and the Giotto,

384 00:23:32.087 --> 00:23:37.087 all showed a lower performance compared to those algorithms.

385 00:23:38.074 --> 00:23:41.620 The BayesSpace showed relatively higher performance

386 00:23:41.620 --> 00:23:44.963 about the ARI of 0.55.

387 00:23:46.350 --> 00:23:49.240 SPRUCE further improved the performance

388 00:23:49.240 --> 00:23:51.570 compared to the BayesSpace.

389 00:23:51.570 --> 00:23:54.830 And one thing I noted here is the...

390 00:23:57.796 --> 00:24:00.130 The narrowed cell cluster,

391 00:24:00.130 --> 00:24:02.633 could it be identified by the SPRUCE?

392 00:24:04.015 --> 00:24:05.003 Which is interesting.

393 00:24:06.090 --> 00:24:08.333 And as the second example.

394 00:24:09.557 --> 00:24:12.620 So first one is the more labeled data.

395 00:24:12.620 --> 00:24:17.250 We can compare our prediction to the existing annotation.

396 00:24:17.250 --> 00:24:21.170 And to further demonstrate the application of the SPRUCE

397 00:24:22.174 --> 00:24:26.290 to unlabeled data, we analyze the publicly available

398 00:24:26.290 --> 00:24:30.890 human invasive ductal carcinoma breast tissue.

399 00:24:30.890 --> 00:24:33.633 Again using the 10 X Visium platform.

400 00:24:35.900 --> 00:24:38.420 And we essentially followed the similar workflow

401 00:24:38.420 --> 00:24:43.420 and we identify the top 16 most spatially variable genes.

402 00:24:44.544 --> 00:24:49.544 And those included several tumor associated antigens,

403 00:24:49.650 --> 00:24:53.847 TAA, in creating the GFRA1 and CXCL14.

404 00:24:56.470 --> 00:25:00.250 And also that there is the tumor suppressive gene,

405 00:25:00.250 --> 00:25:02.823 like MALAT1.

406 00:25:04.430 --> 00:25:09.430 And we use the SPRUCE to identify the 5 sub regions

407 00:25:09.600 --> 00:25:11.493 using these 16 features.

408 00:25:12.479 --> 00:25:16.370 This shows the 16 most variable genes.

409 00:25:16.370 --> 00:25:21.370 And you can see that there are very clear spatial patterns.

410 00:25:22.430 --> 00:25:27.430 For example the CXCL14 and GFRA1,

411 00:25:27.840 --> 00:25:30.350 expel on the right bottom side.

412 00:25:30.350 --> 00:25:35.350 While the MALAT1 express higher in the top left side.

413 00:25:38.400 --> 00:25:41.540 And this is the cluster prediction

414 00:25:41.540 --> 00:25:43.883 made by the SPRUCE algorithm.

415 00:25:45.670 --> 00:25:47.760 And you can see that it identified

416 00:25:47.760 --> 00:25:52.283 the cluster too, which it highly coincide with the CLCX14

417 00:25:54.859 --> 00:25:57.192 and GFRAI1 with a study on.

418 00:25:59.048 --> 00:26:01.200 (mumbles)

419 00:26:01.200 --> 00:26:03.693 What the cell cluster 1,

420 00:26:05.336 --> 00:26:09.230 Is the MALAT1

421 00:26:09.230 --> 00:26:11.493 which is more tumor suppressor.

422 00:26:12.774 --> 00:26:16.945 So here we can see that the SPRUCE can identify

423 00:26:16.945 --> 00:26:20.080 the different group of the tissue architecture,

424 00:26:20.080 --> 00:26:25.080 such as the tumor suppressor and then tumor related

425 00:26:25.354 --> 00:26:27.521 (mumbles)

426 00:26:32.947 --> 00:26:36.800 And we can also easily look at there,

427 00:26:36.800 --> 00:26:39.520 within cluster expression pattern

428 00:26:39.520 --> 00:26:41.093 and gene-gene correlation.

429 00:26:42.710 --> 00:26:44.110 As you could see earlier,

430 00:26:44.110 --> 00:26:48.020 on cell cluster 2 which equals 0.2 to the right

431 00:26:48.963 --> 00:26:52.493 higher than the GFRA1 and CXCL14.

432 00:26:52.493 --> 00:26:57.039 One, which is the cross point here is the high-end MALAT1

433 00:26:57.039 --> 00:26:58.000 and so on.

434 00:26:58.000 --> 00:27:02.470 And also, in the case of cell cluster 2,

435 00:27:02.470 --> 00:27:05.500 there's a very strong gene-gene correlation pattern.

436 00:27:05.500 --> 00:27:10.023 So we just support the proposed model that considered

437 00:27:11.120 --> 00:27:14.430 spatial pattern and also gene-gene correlation

438 00:27:14.430 --> 00:27:15.263 simultaneously.

439 00:27:19.800 --> 00:27:20.633 So,

440 00:27:20.633 --> 00:27:24.550 so far I discussed the method

441 00:27:25.717 --> 00:27:29.783 for our SPRUCE and its application.

442 00:27:32.510 --> 00:27:37.510 And that we essentially expanded our work a little bit more

443 00:27:37.510 --> 00:27:38.510 to the MAPLE,

444 00:27:38.510 --> 00:27:42.173 which is the multi-sample spatial transcriptomics model

445 00:27:43.967 --> 00:27:48.967 Why we care about the multi-sample analysis of HST data?

446 00:27:49.280 --> 00:27:52.910 So currently most algorithms are designed in a way

447 00:27:52.910 --> 00:27:56.630 that it can more focus on a single sample.

448 00:27:56.630 --> 00:27:59.102 But even intuitively,

449 00:27:59.102 --> 00:28:03.460 joint analysis of the multiple HST data

450 00:28:03.460 --> 00:28:05.840 can potentially boost the signal

451 00:28:05.840 --> 00:28:08.980 by sharing the information amongst samples.

452 00:28:08.980 --> 00:28:13.170 And also the joint analysis of the different samples

453 00:28:13.170 --> 00:28:18.120 can allow the differentiation analysis of the HST data.

454 00:28:18.120 --> 00:28:23.120 So very often, each tissue is not our main interest.

455 00:28:23.980 --> 00:28:27.400 But we also want to compare tissue architecture

456 00:28:27.400 --> 00:28:29.540 between the different samples.

457 00:28:29.540 --> 00:28:34.510 For example, between the disease group versus the controls,

458 00:28:34.510 --> 00:28:38.661 responders versus the non responders to 13 treatments,

459 00:28:38.661 --> 00:28:41.100 such as the cancer immuno-therapy.

460 00:28:41.100 --> 00:28:45.633 So to offset this limitation, we proposed MAPLE.

461 00:28:46.550 --> 00:28:50.080 And actually our existing SPRUCE framework

462 00:28:50.080 --> 00:28:53.463 already allows this one naturally.

463 00:28:54.796 --> 00:28:56.997 So, simply what it can do is

464 00:28:56.997 --> 00:29:01.290 instead of now analyzing each sample individually,

465 00:29:01.290 --> 00:29:04.720 we can jointly analyze all the samples together.

466 00:29:04.720 --> 00:29:06.260 And then by doing so,

467 00:29:06.260 --> 00:29:07.940 we can share information

468 00:29:07.940 --> 00:29:12.940 about the modeling of each cell spot cluster,

469 00:29:12.960 --> 00:29:17.060 and also their spatial pattern.

470 00:29:17.060 --> 00:29:21.920 But by introducing the sample-level covariate exp xi

471 00:29:21.920 --> 00:29:23.723 in the cell type composition,

472 00:29:27.380 --> 00:29:28.790 we can see the impact

473 00:29:28.790 --> 00:29:31.817 of the different sample-level covariate.

474 00:29:33.320 --> 00:29:36.823 Which I show more in detail in the coming slides.

475 00:29:41.460 --> 00:29:44.970 So the first application is the same mouse brain data,

476 00:29:44.970 --> 00:29:47.230 the human brain data...

477 00:29:47.230 --> 00:29:49.310 Sorry this should be the mouse brain,

478 00:29:49.310 --> 00:29:53.033 and here we see the two anterior parts,

479 00:29:53.900 --> 00:29:55.600 which look very similar.

480 00:29:55.600 --> 00:29:57.400 And then as you can see here,

481 00:29:57.400 --> 00:30:00.807 when we jointly analyze the two sample

482 00:30:00.807 --> 00:30:04.380 cross pointing to the same part of the brain.

483 00:30:04.380 --> 00:30:08.210 It nicely identifies the cross pointing part

484 00:30:08.210 --> 00:30:09.830 between the two sample.

485 00:30:09.830 --> 00:30:13.682 Like one in the end, three on the top,

486 00:30:13.682 --> 00:30:15.853 five at the bottom and so on.

487 00:30:17.120 --> 00:30:20.950 And because this is the Bayesag framework,

488 00:30:20.950 --> 00:30:24.640 it can also provide uncertainty measures

489 00:30:24.640 --> 00:30:27.510 about our clustering prediction.

490 00:30:27.510 --> 00:30:30.940 And as you can see usually there is more uncertain

491 00:30:30.940 --> 00:30:34.520 about the clustering prediction

492 00:30:34.520 --> 00:30:37.850 around the boundary between different cell clusters.

493 00:30:37.850 --> 00:30:40.070 Which kind of makes sense,

494 00:30:40.070 --> 00:30:43.190 because we expect that maybe cell type

495 00:30:43.190 --> 00:30:47.510 might be more mixed together in the same cell spot.

496 00:30:47.510 --> 00:30:50.190 Also, there are some cell clusters

497 00:30:50.190 --> 00:30:52.890 with the higher level of the uncertainty

498 00:30:52.890 --> 00:30:55.640 of which we are still trying to understand more

499 00:30:55.640 --> 00:30:56.493 at this point.

500 00:30:58.180 --> 00:31:01.450 And this kind of the figure is the...

501 00:31:01.450 --> 00:31:04.673 what utility of this kind of joint analysis.

502 00:31:05.510 --> 00:31:08.840 So, for the identifier with T,

503 00:31:08.840 --> 00:31:13.650 we set the first cell cluster as the reference.

504 00:31:13.650 --> 00:31:16.370 And then here we see the two (mumbles)

505 00:31:16.370 --> 00:31:20.180 The top one shows the intercept,

506 00:31:20.180 --> 00:31:24.740 and then we can interpret this one as the relative size

507 00:31:24.740 --> 00:31:26.470 of each cell cluster.

508 00:31:26.470 --> 00:31:28.770 So then compared to the one,

509 00:31:28.770 --> 00:31:31.283 we can say three and the six are larger.

510 00:31:32.230 --> 00:31:35.910 So the three and the six are larger, compared to the one.

511 00:31:35.910 --> 00:31:38.514 Why the four is the smaller,

512 00:31:38.514 --> 00:31:40.480 well just smaller compared to the one.

513 00:31:40.480 --> 00:31:44.840 So this is what it can see by eye

514 00:31:44.840 --> 00:31:47.347 from the tissue prediction region.

515 00:31:48.372 --> 00:31:51.770 But good thing is that this model allows us to quantify,

516 00:31:51.770 --> 00:31:53.143 what you see by eye.

517 00:31:54.520 --> 00:31:57.450 And what is more interesting is the second one.

518 00:31:57.450 --> 00:31:58.283 So this one,

519 00:31:58.283 --> 00:32:01.530 is about the difference between the two sample.

520 00:32:01.530 --> 00:32:02.363 So again,

521 00:32:03.870 --> 00:32:06.654 so basically if it's higher,

522 00:32:06.654 --> 00:32:11.654 then it means that certain tissue spot cluster

523 00:32:11.800 --> 00:32:14.250 getting bigger in the second sample.

524 00:32:14.250 --> 00:32:17.718 And if it's lower immune state is a kind of smaller

525 00:32:17.718 --> 00:32:20.270 in the second sample and so on.

526 00:32:20.270 --> 00:32:21.192 So in this way,

527 00:32:21.192 --> 00:32:26.192 we can quantify the change of the tissue architecture

528 00:32:26.320 --> 00:32:28.003 between different cell clusters.

529 00:32:30.330 --> 00:32:35.320 And another interesting example is this one.

530 00:32:35.320 --> 00:32:39.431 So here, the image of 2D to anterior samples,

531 00:32:39.431 --> 00:32:44.210 we now also look at the posterior sample as well.

532 00:32:44.210 --> 00:32:47.950 So because this is two parts of the brain

533 00:32:47.950 --> 00:32:49.980 anterior and the posterior,

534 00:32:49.980 --> 00:32:53.060 the issue is kind of continuous between two.

535 00:32:53.060 --> 00:32:54.430 And as you can see here,

536 00:32:54.430 --> 00:32:58.933 cell cluster three is connected to the posterior side here.

537 00:32:59.974 --> 00:33:04.720 Cell cluster one is connected to here and so on.

538 00:33:04.720 --> 00:33:08.050 And then this kind of pattern is not clear

539 00:33:08.050 --> 00:33:12.220 if you analyze each data independently.

540 00:33:12.220 --> 00:33:15.610 And our MAPLE framework nicely captures

541 00:33:15.610 --> 00:33:17.750 such kind of sharing pattern.

542 00:33:17.750 --> 00:33:19.770 And also the difference pattern

543 00:33:19.770 --> 00:33:22.950 between the different samples, interestingly.

544 00:33:22.950 --> 00:33:24.937 So at this point,

545 00:33:24.937 --> 00:33:27.350 we are working on more simulation study

546 00:33:27.350 --> 00:33:29.440 and the real data analysis

547 00:33:29.440 --> 00:33:32.540 to further show the performance

548 00:33:32.540 --> 00:33:36.043 and understand the properties of the MAPLE at this point.

549 00:33:38.650 --> 00:33:43.650 So then I can't summarize my presentation today.

550 00:33:44.630 --> 00:33:49.297 So the high throughput spatial transcriptomics, or HST,

551 00:33:50.290 --> 00:33:53.680 provides unprecedented opportunities

552 00:33:53.680 --> 00:33:57.430 to investigate novel biological hypotheses,

553 00:33:57.430 --> 00:34:01.513 such as the tumor microenvironment and certain structure

554 00:34:04.816 --> 00:34:08.190 about the human brain and Alzheimer,

555 00:34:08.190 --> 00:34:09.720 and so on.

556 00:34:09.720 --> 00:34:12.700 And here we propose SPRUCE,

557 00:34:12.700 --> 00:34:15.640 a Bayesian multivariate mixture model

558 00:34:15.640 --> 00:34:17.733 for HST data analysis.

559 00:34:19.460 --> 00:34:22.190 SPRUCE has multiple strengths

560 00:34:23.290 --> 00:34:25.860 including the novel combination

561 00:34:25.860 --> 00:34:28.500 of the skewed normal density,

562 00:34:28.500 --> 00:34:31.137 Polya-Gamma data augmentation,

563 00:34:31.137 --> 00:34:33.043 and spatial random effect.

564 00:34:34.750 --> 00:34:36.850 Altogether, it allows to

565 00:34:36.850 --> 00:34:41.040 precisely infer spatially correlated mixture component

566 00:34:41.040 --> 00:34:43.293 membership probabilities.

567 00:34:44.365 --> 00:34:48.829 In our simulation study and real data analysis,

568 00:34:48.829 --> 00:34:52.820 we could see that SPRUCE outperforms the existing method,

569 00:34:52.820 --> 00:34:56.160 in the tissue architecture identification.

570 00:34:56.160 --> 00:35:01.160 And finally our recent extension of the MAPLE

571 00:35:01.270 --> 00:35:04.530 allows the joint clustering and differential analysis

572 00:35:04.530 --> 00:35:06.933 of multiple HST data.

573 00:35:08.548 --> 00:35:12.815 So at this point SPRUCE is on the review in,

574 00:35:12.815 --> 00:35:14.970 (mumbles)

575 00:35:14.970 --> 00:35:17.020 in the biometrics.

576 00:35:17.020 --> 00:35:21.033 Cross pointing manuscript is available in the bio archive.

577 00:35:22.040 --> 00:35:25.240 And there are multiple ongoing work

578 00:35:25.240 --> 00:35:28.230 regarding the HST data modeling

579 00:35:28.230 --> 00:35:29.163 in our lab.

580 00:35:30.418 --> 00:35:34.580 So we are actually currently working on further improving

581 00:35:34.580 --> 00:35:36.700 the SPRUCE and the MAPLE

582 00:35:36.700 --> 00:35:39.350 by incorporating other characteristics

583 00:35:39.350 --> 00:35:44.350 of the HST data, such as the relationships among cells.

584 00:35:44.360 --> 00:35:45.861 For example,

585 00:35:45.861 --> 00:35:50.000 we know that there are some likened and receptor,

586 00:35:50.000 --> 00:35:50.833 for example.

587 00:35:50.833 --> 00:35:55.140 Which we expect that they interact with each other

588 00:35:55.140 --> 00:35:57.390 in their cell structure.

589 00:35:57.390 --> 00:36:00.950 And then by incorporating different prior information,

590 00:36:00.950 --> 00:36:04.163 we can further improve the SPRUCE and MAPLE.

591 00:36:05.610 --> 00:36:09.633 We are also working on the other statistical models

592 00:36:09.633 --> 00:36:14.130 for somewhat relevant, but different tasks.

593 00:36:14.130 --> 00:36:15.252 For example,

594 00:36:15.252 --> 00:36:18.820 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 combination

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 communications.

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 experimentally

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 genomics,

619 00:37:48.230 --> 00:37:52.580 proteomics and the T-cell receptor at the same time.

620 00:37:52.580 --> 00:37:53.580 And very soon,

621 00:37:53.580 --> 00:37:56.780 everything are expected to be combined

622 00:37:56.780 --> 00:37:59.663 as the spatial transcriptomic structure.

623 00:38:00.630 --> 00:38:03.430 We are working on the direction

624 00:38:03.430 --> 00:38:06.020 to develop the statistical model

625 00:38:06.020 --> 00:38:09.733 for integration of the HST data with other matched data.

626 00:38:12.769 --> 00:38:15.867 So I would like to acknowledge my research team at OSU.

627 00:38:17.870 --> 00:38:22.560 Carter Allen is the main driver this project,

628 00:38:22.560 --> 00:38:25.310 and also my pitch assistant

629 00:38:26.883 --> 00:38:31.870 Qin Ma and Yuzhou Chang is my close collaborator

630 00:38:31.870 --> 00:38:36.410 for the HST data modeling project.

631 00:38:36.410 --> 00:38:37.798 And Zihai Li,

632 00:38:37.798 --> 00:38:41.600 who is the director of the Immuno-Oncology Institute

633 00:38:41.600 --> 00:38:44.143 and also the expert in cancer.

634 00:38:46.070 --> 00:38:48.730 Won Chang at the University of Cincinnati

635 00:38:48.730 --> 00:38:51.523 who are the spatial statistics expert,

636 00:38:52.907 --> 00:38:56.370 and MUSC collaborator Brian Neelon

637 00:38:56.370 --> 00:38:57.833 and my grant support.

638 00:38:59.650 --> 00:39:02.920 So, and this is the end of my presentation,

639 00:39:02.920 --> 00:39:05.480 and you can find my manuscript

640 00:39:05.480 --> 00:39:09.660 and the software from the link here.

641 00:39:09.660 --> 00:39:11.773 If you have any questions and comment,

642 00:39:12.686 --> 00:39:16.969 please let me know by email at chung.911@osu.edu.

643 00:39:16.969 --> 00:39:19.636 So thank you for your attention.

644 00:39:28.588 --> 00:39:29.838 <v ->So thank you.</v>

645 00:39:31.862 --> 00:39:35.693 Do we have any questions from the audience in the classroom,

646 00:39:35.693 --> 00:39:38.110 or from the audience on zoom?

647 00:39:42.661 --> 00:39:44.917 <v ->Can I ask a question?</v>

648 00:39:44.917 --> 00:39:46.250 Can you hear me?

649 00:39:46.250 --> 00:39:47.260 <v ->Yes, mm-hm.</v>

650 00:39:47.260 --> 00:39:48.760 <v ->Right, Dongjun welcome back.</v>

651 00:39:49.600 --> 00:39:51.720 Great work, it's a nice presentation.

652 00:39:51.720 --> 00:39:53.194 I'm just wondering, like,

653 00:39:53.194 --> 00:39:55.760 when you do this from your own experience

654 00:39:55.760 --> 00:39:57.200 on the cell clustering,

655 00:39:57.200 --> 00:40:00.410 how much the spatial information contributes

656 00:40:00.410 --> 00:40:02.593 to the clustering.

657 00:40:02.593 --> 00:40:03.426 <v ->Sure.</v>

658 00:40:09.017 --> 00:40:09.900 So,

659 00:40:09.900 --> 00:40:12.067 (mumbles)

660 00:40:14.598 --> 00:40:16.132 If you're here,

661 00:40:16.132 --> 00:40:18.450 so if you look at the Seurat workflow,

662 00:40:18.450 --> 00:40:21.710 you can see there's a still lot of the, kind of,

663 00:40:21.710 --> 00:40:24.783 local boundary between different cell spot clusters.

664 00:40:27.730 --> 00:40:31.830 And when you analyze the same data using the SPRUCE,

665 00:40:31.830 --> 00:40:33.740 you can see much cleaner boundary.

666 00:40:33.740 --> 00:40:36.439 And often it will coincide with the

667 00:40:36.439 --> 00:40:39.740 expert analogy annotation.

668 00:40:39.740 --> 00:40:44.293 So given that there is the significant contribution,

669 00:40:46.210 --> 00:40:49.230 of course even the gene expression,

670 00:40:49.230 --> 00:40:52.460 we still get some big picture, as you can see here.

671 00:40:52.460 --> 00:40:56.673 But spatial information provide much cleaner prediction

672 00:40:56.673 --> 00:40:59.263 about the tissue architecture in general.

673 00:41:01.330 --> 00:41:02.163 <v ->I see.</v>

674 00:41:02.163 --> 00:41:03.747 And also the skewness.

675 00:41:04.950 --> 00:41:07.740 Do you estimate that or that's like your heart

676 00:41:07.740 --> 00:41:09.240 was persuaded by the skewness?

677 00:41:12.049 --> 00:41:14.443 <v ->You mean which one?</v>

678 00:41:14.443 --> 00:41:16.000 <v ->On k model.</v>

679 00:41:16.000 --> 00:41:19.120 Your model to specify, the k model you have there.

680 00:41:19.120 --> 00:41:20.440 I missed that part.

681 00:41:20.440 --> 00:41:22.863 Like, do you need to specify the skewness?

682 00:41:24.396 --> 00:41:26.440 <v ->Or learn from data.</v>

683 00:41:26.440 --> 00:41:27.273 <v ->Oh, I see.</v>

684 00:41:27.273 --> 00:41:28.870 But from the data, how skew?

685 00:41:28.870 --> 00:41:30.600 I mean, just in terms of how stable

686 00:41:30.600 --> 00:41:32.573 that alpha k can be estimated.

687 00:41:35.790 --> 00:41:38.240 <v ->So maybe I can answer it in two different ways.</v>

688 00:41:39.720 --> 00:41:42.663 So if there is this skewness in the data, I think yes.

689 00:41:43.655 --> 00:41:47.650 So we'll say it depends on how processed the data as well.

690 00:41:48.878 --> 00:41:50.910 So usually there's three different approaches

691 00:41:50.910 --> 00:41:55.910 to model the HST data in closed spatial embedding gene.

692 00:41:56.830 --> 00:41:58.103 And so you can see here,

693 00:41:58.103 --> 00:42:00.853 who are the people using the principle components?

694 00:42:01.854 --> 00:42:04.596 Who are the people use the team learning

695 00:42:04.596 --> 00:42:05.953 as the embedding step?

696 00:42:08.179 --> 00:42:12.040 If you use the team learning or the PCA

697 00:42:12.040 --> 00:42:15.580 it's more likely symmetry in the real data.

698 00:42:15.580 --> 00:42:19.940 If you consider the spatial embedding gene,

699 00:42:19.940 --> 00:42:23.123 we often hope to have the skewness, as you can see here.

700 00:42:24.410 --> 00:42:29.410 And then regarding your question, overall it works well.

701 00:42:30.434 --> 00:42:32.993 I don't have the exact quantification, but it works well.

702 00:42:34.220 --> 00:42:36.830 Especially stably in most cases.

703 00:42:36.830 --> 00:42:38.770 <v ->Yeah, I read the spatial Bayes paper.</v>

704 00:42:38.770 --> 00:42:41.160 They seem to be working on the principle components, right?

705 00:42:41.160 --> 00:42:42.940 They do not work on individual genes, right?.

706 00:42:42.940 --> 00:42:43.773 <v ->No, yeah.</v>

707 00:42:43.773 --> 00:42:45.500 They base this on the PCA.

708 00:42:45.500 --> 00:42:47.180 <v ->Yeah, that's why it's completely puzzling me</v>

709 00:42:47.180 --> 00:42:48.013 while you're doing that.

710 00:42:48.013 --> 00:42:49.210 But anyway, yeah.

711 00:42:49.210 --> 00:42:50.043 Thank you.

712 00:42:50.043 --> 00:42:51.421 <v ->Yeah so, so...</v>

713 00:42:51.421 --> 00:42:53.988 (mumbles)

714 00:42:53.988 --> 00:42:55.050 so they mainly target the PCA.

715 00:42:55.050 --> 00:42:58.650 So they only can start the multivariate distribution.

716 00:42:58.650 --> 00:43:01.180 And also because of the same reason,

717 00:43:01.180 --> 00:43:04.980 their equivalence metrics means less density.

718 00:43:04.980 --> 00:43:05.813 <v ->I see.</v>

719 00:43:07.680 --> 00:43:09.222 Thank you.

720 00:43:09.222 --> 00:43:10.222 <v ->Thank you.</v>

721 00:43:22.460 --> 00:43:26.380 <v ->Do we have any questions from students in the classroom?</v>

722 00:43:31.690 --> 00:43:33.453 <v ->Wait, can I ask another question?</v>

723 00:43:35.740 --> 00:43:36.573 So, towards the end,

724 00:43:36.573 --> 00:43:37.850 you mentioned you tried

725 00:43:37.850 --> 00:43:41.033 to look at the cell-cell communication.

726 00:43:44.580 --> 00:43:46.080 That part.

727 00:43:46.080 --> 00:43:47.527 I'm very interested in that

728 00:43:49.464 --> 00:43:53.063 From our experience on the single cell spatial data are...

729 00:43:55.230 --> 00:43:58.190 Are you talking about you're learning from the single cell,

730 00:43:58.190 --> 00:43:59.540 or the spatial single cell?

731 00:44:02.298 --> 00:44:05.220 <v ->So, regarding the cell-cell communication</v>

732 00:44:05.220 --> 00:44:09.870 it's still very ongoing research at this point.

733 00:44:09.870 --> 00:44:12.960 I mean, not just our side but in general.

734 00:44:12.960 --> 00:44:16.876 Because most of the cell-cell communication prediction

735 00:44:16.876 --> 00:44:19.700 based on the database.

736 00:44:19.700 --> 00:44:22.480 So based on data, like on the receptor,

737 00:44:22.480 --> 00:44:24.550 pairing the database and checking

738 00:44:24.550 --> 00:44:28.170 their cross point on the expression in cross point spot

739 00:44:28.170 --> 00:44:29.460 of the cell.

740 00:44:29.460 --> 00:44:32.651 And then by checking that the cross pointing pair

741 00:44:32.651 --> 00:44:34.330 of the expression pattern

742 00:44:34.330 --> 00:44:36.290 between the like and the receptor.

743 00:44:36.290 --> 00:44:38.753 They want to model cell-cell communication.

744 00:44:39.720 --> 00:44:42.160 It's not perfect, as you know,

745 00:44:42.160 --> 00:44:44.880 because it's like a computer.

746 00:44:44.880 --> 00:44:47.617 If you look at the chip, it's almost like

747 00:44:47.617 --> 00:44:48.450 (mumbles)

748 00:44:48.450 --> 00:44:49.993 but more like motive analysis.

749 00:44:50.850 --> 00:44:52.565 So there's some limitation,

750 00:44:52.565 --> 00:44:57.214 but it's a more likely general limitation at this point.

751 00:44:57.214 --> 00:44:58.047 <v ->Yeah,</v>

752 00:44:58.047 --> 00:44:59.430 I'm asking because we've been looking

753 00:44:59.430 --> 00:45:01.700 at some of the spatial single cell data

754 00:45:01.700 --> 00:45:04.060 that were too noisy for the like

755 00:45:04.060 --> 00:45:05.873 and receptor gene expression levels.

756 00:45:07.200 --> 00:45:09.112 Just couldn't make it too far.

757 00:45:09.112 --> 00:45:10.710 (mumbles)

758 00:45:10.710 --> 00:45:12.730 But for a single cell, may be different?

759 00:45:12.730 --> 00:45:17.050 I mean, probably there'll be more that, like...

760 00:45:17.050 --> 00:45:18.320 <v ->Yeah, three already.</v>

761 00:45:18.320 --> 00:45:22.370 I mean, so if you go to high-resolution,

762 00:45:22.370 --> 00:45:23.613 it's a very noisy,

763 00:45:24.498 --> 00:45:28.100 so very often we need to do some simplification.

764 00:45:28.100 --> 00:45:31.200 Like looking at multi-modal or the cell cluster,

765 00:45:31.200 --> 00:45:32.333 rather than the cell.

766 00:45:34.100 --> 00:45:37.820 It's still very multiple experimental limitation,

767 00:45:37.820 --> 00:45:38.979 at this point.

768 00:45:38.979 --> 00:45:39.920 (mumbles)

769 00:45:39.920 --> 00:45:40.753 Thank you.

770 00:45:50.936 --> 00:45:55.269 (class teacher addresses classroom)

771 00:46:00.210 --> 00:46:02.530 <v ->On the data from multiple samples</v>

772 00:46:02.530 --> 00:46:04.940 So, if we have samples from...

773 00:46:05.847 --> 00:46:08.014 (mumbles)

774 00:46:17.627 --> 00:46:20.663 <v ->Oh yeah, that's a very good question.</v>

775 00:46:22.056 --> 00:46:22.889 So,

776 00:46:22.889 --> 00:46:27.760 actually we can answer in the two different ways.

777 00:46:28.800 --> 00:46:30.400 In some sense,

778 00:46:30.400 --> 00:46:33.390 good pre-processing is still important

779 00:46:35.362 --> 00:46:40.362 because it still depends on the expression patterns.

780 00:46:43.148 --> 00:46:45.910 But still regarding the differences

781 00:46:45.910 --> 00:46:48.060 between the different tissues.

782 00:46:48.060 --> 00:46:49.400 If there is a big difference,

783 00:46:49.400 --> 00:46:51.110 it can still detect the difference

784 00:46:51.110 --> 00:46:53.720 between the different sample.

785 00:46:53.720 --> 00:46:55.444 So, it can detect spots.

786 00:46:55.444 --> 00:46:58.552 But still like a main goal is more

787 00:46:58.552 --> 00:47:01.000 for the similar type of tissue.

788 00:47:01.000 --> 00:47:02.080 If it's too different,

789 00:47:02.080 --> 00:47:04.083 maybe it's different research project.

790 00:47:05.375 --> 00:47:06.710 So, for example,

791 00:47:06.710 --> 00:47:09.990 here our targets is more about, for example,

792 00:47:09.990 --> 00:47:12.960 like same breast tissue,

793 00:47:12.960 --> 00:47:17.562 but with a different responders and non-responders group,

794 00:47:17.562 --> 00:47:19.172 for example.

795 00:47:19.172 --> 00:47:23.560 Or like a cell-cell long tissue, but the tumor but not tumor

796 00:47:23.560 --> 00:47:24.393 and so on.

797 00:47:25.410 --> 00:47:27.493 If you like a human and mouse,

798 00:47:29.332 --> 00:47:32.600 then it might be somewhat different story,

799 00:47:32.600 --> 00:47:34.410 which might need much more work.

800 00:47:38.443 --> 00:47:41.526 <v ->Do we have any more questions here?</v>

801 00:47:57.568 --> 00:47:59.170 Okay, can we have all the questions

802 00:47:59.170 --> 00:48:01.313 from the audience on zoom?

803 00:48:21.176 --> 00:48:25.550 Okay, so it looks like we don't have any more questions.

804 00:48:25.550 --> 00:48:30.340 So Dr. Chung, thank you again for your nice presentation.

805 00:48:31.210 --> 00:48:33.860 Look forward to meeting in person sometime soon.

806 00:48:35.650 --> 00:48:38.247 <v ->And then thank you again Wei and Hongyou</v>

807 00:48:38.247 --> 00:48:39.540 for the invitation

808 00:48:39.540 --> 00:48:43.320 and it's a great come back, although virtually.

809 00:48:43.320 --> 00:48:45.980 And I hope to see you again.

810 00:48:45.980 --> 00:48:47.280 <v ->We'll come by in person.</v>

811 00:48:49.450 --> 00:48:50.700 <v ->Hopefully someday soon.</v>

812 00:48:52.820 --> 00:48:53.653 Okay, thank you.