WEBVTT

1 00:00:01.020 --> 00:00:02.850 <v ->All right, I'm very excited</v>

2 00:00:02.850 --> 00:00:04.170 to introduce our speaker for today.

3 00:00:04.170 --> 00:00:05.880 We have Dr. Meghan Short.

4 00:00:05.880 --> 00:00:07.800 Dr. Short has completed fellowships

5 00:00:07.800 --> 00:00:09.870 at the Glenn Biggs Institute for Alzheimer's

6 00:00:09.870 --> 00:00:11.940 and Neurodegenerative Diseases,

7 00:00:11.940 --> 00:00:14.310 and at Harvard's Huttenhower Lab.

8 00:00:14.310 --> 00:00:16.500 Currently, Dr. Short is an assistant professor

9 00:00:16.500 --> 00:00:17.610 at Tufts University.

10 00:00:17.610 --> 00:00:19.593 Let's give a warm welcome to Dr. Short.

11 00:00:31.200 --> 00:00:33.250 <v ->Hi, everyone, Thank you for being here.</v>

12 00:00:34.110 --> 00:00:34.943 Can you all hear me, okay?

13 00:00:34.943 --> 00:00:36.453 <v ->Sign in if you're registered.</v>

14 00:00:38.280 --> 00:00:40.860 <v ->All right, so, today, I'm going to talk about a project</v>

15 00:00:40.860 --> 00:00:43.410 that I worked on as part of my postdoc

16 00:00:43.410 --> 00:00:45.930 down at UT Health San Antonio

17 00:00:45.930 --> 00:00:48.930 with the Glenn Biggs Institute for Alzheimer's

18 00:00:48.930 --> 00:00:51.000 and Neurodegenerative Diseases,

19 00:00:51.000 --> 00:00:53.990 and I wanted to talk about this as a...

20 00:00:56.070 --> 00:00:58.140 None of the sort of methods that I'm gonna talk about

21 00:00:58.140 --> 00:01:00.810 in this talk are particularly new.

22 00:01:00.810 --> 00:01:03.750 This wasn't sort of a methods development project.

23 00:01:03.750 --> 00:01:07.890 So the sort of main network method I'll talk about

24 00:01:07.890 --> 00:01:10.200 is about a decade old at this point, at least,

25 00:01:10.200 --> 00:01:12.570 but what's nice about it is that

26 00:01:12.570 --> 00:01:14.700 with increasing availability

27 00:01:14.700 --> 00:01:16.800 of high dimensional biomedical data,

28 00:01:16.800 --> 00:01:19.530 it's sort of seeing more use cases,

29 00:01:19.530 --> 00:01:21.750 and it's not something that, at least, I learned about

30 00:01:21.750 --> 00:01:24.270 in my graduate program in biostatistics,

31 00:01:24.270 --> 00:01:25.830 but it's something that I thought

32 00:01:25.830 --> 00:01:27.750 would be good to talk about today

33 00:01:27.750 --> 00:01:29.350 since it's such a useful method.

34 00:01:31.860 --> 00:01:34.800 So let's see if I advance.

35 00:01:34.800 --> 00:01:36.360 There we go.

36 00:01:36.360 --> 00:01:39.390 So I'll start just by giving a quick introduction.

37 00:01:39.390 --> 00:01:43.320 I know that when I was in grad school, I always wanted,

38 00:01:43.320 --> 00:01:44.400 I thought it was interesting

39 00:01:44.400 --> 00:01:45.870 to hear about people's career paths

40 00:01:45.870 --> 00:01:47.343 as I was considering my own.

41 00:01:48.330 --> 00:01:52.980 So I started in biology as a field.

42 00:01:52.980 --> 00:01:55.503 I studied salt marsh ecology as an undergrad,

43 00:01:56.370 --> 00:01:57.570 and then by the end of undergrad,

44 00:01:57.570 --> 00:02:00.240 I was interested in getting more into sort of a human,

45 00:02:00.240 --> 00:02:02.490 more directly human-focused environment,

46 00:02:02.490 --> 00:02:04.650 and so I considered public health.

47 00:02:04.650 --> 00:02:05.760 I learned about statistics

48 00:02:05.760 --> 00:02:07.530 as part of my research in undergrad

49 00:02:07.530 --> 00:02:10.830 and wanted to continue with that so I participated in SIBS,

50 00:02:10.830 --> 00:02:13.597 which is a program that you may be aware of,

51 00:02:13.597 --> 00:02:16.221 and that was my first intro to biostat.

52 00:02:16.221 --> 00:02:19.260 I was a graduate student at Boston University.

53 00:02:19.260 --> 00:02:20.790 I had fortune of working

54 00:02:20.790 --> 00:02:22.110 with the Framingham Heart Study,

55 00:02:22.110 --> 00:02:23.520 which is where the data comes from

56 00:02:23.520 --> 00:02:25.590 that I'll be talking to you about today,

57 00:02:25.590 --> 00:02:26.880 which is a really interesting study,

58 00:02:26.880 --> 00:02:29.460 and I'll get more details on in the few slides.

59 00:02:29.460 --> 00:02:30.720 That was sort of my introduction

60 00:02:30.720 --> 00:02:32.673 to working with epidemiological data.

61 00:02:33.750 --> 00:02:35.610 After grad school, I continued on,

62 00:02:35.610 --> 00:02:38.130 again, to UT Health San Antonio,

63 00:02:38.130 --> 00:02:41.850 and then following that to postdoc at Harvard

64 00:02:41.850 --> 00:02:46.850 looking at developing methods for microbiome analysis.

65 00:02:47.310 --> 00:02:48.720 So if you have any interest in that,

66 00:02:48.720 --> 00:02:50.910 feel free to approach me,

67 00:02:50.910 --> 00:02:53.160 although I'm not gonna talk about that today,

68 00:02:54.270 --> 00:02:56.910 and then as of March this year,

69 00:02:56.910 --> 00:03:00.480 I started as an assistant professor at Tufts Medicine

70 00:03:00.480 --> 00:03:02.370 where I'm working on a variety of projects

71 00:03:02.370 --> 00:03:06.210 but a lot related to sort of omics data

72 00:03:06.210 --> 00:03:08.433 and aging and longevity.

73 00:03:12.150 --> 00:03:15.090 So I'll start today's talk with a bit of motivation

74 00:03:15.090 --> 00:03:18.450 for why network-based analyses we're a good fit

75 00:03:18.450 --> 00:03:22.863 for looking at sort of the proteome in Alzheimer's disease.

76 00:03:24.000 --> 00:03:26.700 So first of all, Alzheimer's disease

77 00:03:26.700 --> 00:03:29.520 is a very prevalent condition.

78 00:03:29.520 --> 00:03:32.700 Many of you may be like me and know some family members

79 00:03:32.700 --> 00:03:36.030 or people who have been affected by it.

80 00:03:36.030 --> 00:03:39.870 It's very common and expect it to be more so

81 00:03:39.870 --> 00:03:43.860 as populations age, and it's a leading cause of mortality,

82 00:03:43.860 --> 00:03:46.830 disability, and poor health among seniors,

83 00:03:46.830 --> 00:03:49.080 and one interesting feature of this disease

84 00:03:49.080 --> 00:03:51.960 is that precursors of it can appear years to decades

85 00:03:51.960 --> 00:03:54.750 before symptoms manifest.

86 00:03:54.750 --> 00:03:57.600 So those precursors can include indicators

87 00:03:57.600 --> 00:03:59.913 that are visible on brain MRIs,

88 00:04:00.900 --> 00:04:04.623 performance on neurocognitive testing, changes in gait,

89 00:04:05.460 --> 00:04:08.070 even changes in sense of smell,

90 00:04:08.070 --> 00:04:13.070 and cerebral spinal fluid markers, such as tau and amyloid.

91 00:04:17.460 --> 00:04:21.060 Because of this, there's interest in being able to find

92 00:04:21.060 --> 00:04:23.550 plasma biomarkers for Alzheimer's disease

93 00:04:23.550 --> 00:04:25.320 and related dementias.

94 00:04:25.320 --> 00:04:28.323 ADRD is a acronym we'll be using sort of throughout.

95 00:04:29.940 --> 00:04:32.750 Because since there are indicators

96 00:04:32.750 --> 00:04:34.680 of sort of pre-disease development

97 00:04:34.680 --> 00:04:37.320 in years to decades before being able to detect those,

98 00:04:37.320 --> 00:04:40.380 either earlier or in a less invasive or expensive way,

99 00:04:40.380 --> 00:04:44.430 is very useful,

100 00:04:44.430 --> 00:04:49.430 and so when I say invasive, I mentioned CSF markers,

101 00:04:49.650 --> 00:04:53.160 such as how an amyloid can predict dementia,

102 00:04:53.160 --> 00:04:55.560 but that involves doing a lumbar puncture

103 00:04:55.560 --> 00:04:59.733 versus something like a blood draw, which is easier to do.

104 00:05:00.990 --> 00:05:04.410 Another good aspect of trying to find biomarkers

105 00:05:04.410 --> 00:05:07.440 is that you can get a sense of biological processes

106 00:05:07.440 --> 00:05:10.290 that are involved in disease development,

107 00:05:10.290 --> 00:05:13.440 and that can hopefully lead to either preventative

108 00:05:13.440 --> 00:05:15.183 or therapeutic interventions.

109 00:05:19.260 --> 00:05:21.240 What makes this difficult?

110 00:05:21.240 --> 00:05:24.390 So in my case, I was looking at proteins.

111 00:05:24.390 --> 00:05:27.120 There are thousands and thousands to select from,

112 00:05:27.120 --> 00:05:29.610 and you get sort of this inherent trade off

113 00:05:29.610 --> 00:05:32.610 between trying to control a false positive rate

114 00:05:32.610 --> 00:05:36.000 for all these multiple tests that you may be performing,

115 00:05:36.000 --> 00:05:38.940 but if you effectively control the false positive rate,

116 00:05:38.940 --> 00:05:42.330 you're going to likely end up with low statistical power.

117 00:05:42.330 --> 00:05:44.640 There's this trade off between...

118 00:05:44.640 --> 00:05:47.220 It's sort of a needle in a haystack.

119 00:05:47.220 --> 00:05:49.950 Another thing that has tended to be true

120 00:05:49.950 --> 00:05:53.370 is that there is not very good replicability across studies.

121 00:05:53.370 --> 00:05:57.330 So one study may find 20 biomarkers

122 00:05:57.330 --> 00:05:58.860 and maybe one or two of them

123 00:05:58.860 --> 00:06:01.230 may replicate in a different study.

124 00:06:01.230 --> 00:06:04.563 So there's a lot of noise that ends up coming through.

125 00:06:07.950 --> 00:06:10.350 The approach that I took in this project

126 00:06:10.350 --> 00:06:12.960 was to use network analysis

127 00:06:12.960 --> 00:06:17.040 to analyze the protein data,

128 00:06:17.040 --> 00:06:19.920 and the motivation there is to try and capture

129 00:06:19.920 --> 00:06:23.400 subtle but consistent variation in groups of proteins.

130 00:06:23.400 --> 00:06:26.733 I'll refer to them as modules during this talk.

131 00:06:27.915 --> 00:06:30.060 In then just a few things, so first of all,

132 00:06:30.060 --> 00:06:31.620 it reduces the dimensionality

133 00:06:31.620 --> 00:06:34.890 of the statistical testing problem that you have.

134 00:06:34.890 --> 00:06:37.050 So rather than testing each protein individually

135 00:06:37.050 --> 00:06:40.187 and having to adjust for all of those multiple tests,

136 00:06:40.187 --> 00:06:43.080 you can sort of reduce the space

137 00:06:43.080 --> 00:06:45.660 to a smaller number of tests

138 00:06:45.660 --> 00:06:49.230 where the proteins within each group being tested

139 00:06:49.230 --> 00:06:51.130 are inter-correlated with one another,

140 00:06:51.990 --> 00:06:54.660 and unlike other dimensionality reduction methods,

141 00:06:54.660 --> 00:06:56.640 something like a principle components analysis

142 00:06:56.640 --> 00:06:59.490 that you may have maybe familiar with,

143 00:06:59.490 --> 00:07:02.651 the network method has sort of a benefit of looking

144 00:07:02.651 --> 00:07:05.580 not just at, say, correlations

145 00:07:05.580 --> 00:07:08.550 or relationships between pairs of proteins,

146 00:07:08.550 --> 00:07:11.070 but, also, at sort of the correlational neighborhood

147 00:07:11.070 --> 00:07:12.750 of what common neighbors

148 00:07:12.750 --> 00:07:14.613 those proteins share in the network.

149 00:07:18.270 --> 00:07:22.230 Another benefit of or sort of way

150 00:07:22.230 --> 00:07:23.347 that we try to get around some of the pitfalls

151 00:07:23.347 --> 00:07:28.347 of proteomic analysis is by focusing on biological pathways

152 00:07:29.130 --> 00:07:31.890 instead of on individual proteins themselves.

153 00:07:31.890 --> 00:07:35.670 So within groups of proteins that we find to be of interest

154 00:07:35.670 --> 00:07:39.750 or possibly associated with dementia outcomes,

155 00:07:39.750 --> 00:07:43.200 we use a tool called over-representation analysis,

156 00:07:43.200 --> 00:07:44.670 which I'll talk about later,

157 00:07:44.670 --> 00:07:48.030 but it essentially tries to pinpoint biological pathways

158 00:07:48.030 --> 00:07:50.790 that may be overrepresented by the proteins

159 00:07:50.790 --> 00:07:54.468 that are found to be associated with the outcome,

160 00:07:54.468 --> 00:07:56.280 and the hope there is to find,

161 00:07:56.280 --> 00:08:01.230 to get sort of insights that are more robust across studies

162 00:08:01.230 --> 00:08:03.000 and, hopefully, address some of the issues

163 00:08:03.000 --> 00:08:04.113 with replicability.

164 00:08:07.830 --> 00:08:11.280 Okay, so that's sort of the motivation for this study,

165 00:08:11.280 --> 00:08:13.773 and, now, I'll talk a little bit about the data.

166 00:08:18.030 --> 00:08:19.140 The data for this study

167 00:08:19.140 --> 00:08:21.720 comes from the Framingham Heart Study,

168 00:08:21.720 --> 00:08:23.880 which has been going on for a very long time.

169 00:08:23.880 --> 00:08:28.880 It started in 1948 in a town of Framingham, Massachusetts,

170 00:08:29.190 --> 00:08:30.660 and at the time they enrolled,

171 00:08:30.660 --> 00:08:33.510 they reached out to two-thirds of the population of the town

172 00:08:33.510 --> 00:08:35.940 to try and enroll them in this epidemiological study.

173 00:08:35.940 --> 00:08:38.043 It was one of the first ones of its kind,

174 00:08:39.030 --> 00:08:42.390 and people would come in for exams every few years,

175 00:08:42.390 --> 00:08:44.640 and they would take all of this information about them,

176 00:08:44.640 --> 00:08:47.413 and then follow them for outcomes.

177 00:08:47.413 --> 00:08:49.320 Cardiovascular outcomes was really

178 00:08:49.320 --> 00:08:52.533 the sort of outcome of interest when it first started.

179 00:08:53.490 --> 00:08:56.730 Over the years, they've then enrolled offspring

180 00:08:56.730 --> 00:08:59.010 of the original cohort participants

181 00:08:59.010 --> 00:09:02.130 as well as grandchildren and third generation,

182 00:09:02.130 --> 00:09:05.880 and then as sort of the demographics

183 00:09:05.880 --> 00:09:08.640 of Framingham have changed over the years,

184 00:09:08.640 --> 00:09:10.140 if you're only enrolling descendants

185 00:09:10.140 --> 00:09:11.790 of people who live there in 1948,

186 00:09:11.790 --> 00:09:13.020 you're not gonna capture that.

187 00:09:13.020 --> 00:09:15.450 So they also have been enrolling omni cohorts

188 00:09:15.450 --> 00:09:18.897 to reflect sort of more diverse populations (indistinct).

189 00:09:20.910 --> 00:09:23.100 Again, they were sort of aiming

190 00:09:23.100 --> 00:09:25.050 towards identifying risk factors

191 00:09:25.050 --> 00:09:28.560 and etiologies of cardiovascular disease,

192 00:09:28.560 --> 00:09:30.780 but as those populations age,

193 00:09:30.780 --> 00:09:34.304 brain health and cognition is also an important outcome,

194 00:09:34.304 --> 00:09:38.850 and so they've measured sort of cognitive outcomes

195 00:09:38.850 --> 00:09:41.370 and incidents of dementia as well, and, of course,

196 00:09:41.370 --> 00:09:44.133 those things are also related to cardiovascular.

197 00:09:48.210 --> 00:09:50.850 For our study in particular,

198 00:09:50.850 --> 00:09:53.130 we were using the offspring cohort,

199 00:09:53.130 --> 00:09:55.470 and at their examination cycle five,

200 00:09:55.470 --> 00:09:59.520 which was in the early 90s, they collected blood samples,

201 00:09:59.520 --> 00:10:02.880 and froze the plasma from those samples,

202 00:10:02.880 --> 00:10:06.300 and years later, when they sort of had

203 00:10:06.300 --> 00:10:10.500 these broader proteomic analysis assays available,

204 00:10:10.500 --> 00:10:13.680 they measured the plasma proteome,

205 00:10:13.680 --> 00:10:16.773 I'll talk about the methods for that on the next slide,

206 00:10:17.940 --> 00:10:20.550 but they did this in about 1,900 participants

207 00:10:20.550 --> 00:10:23.820 who were approximately aged 55 when the blood was drawn.

208 00:10:23.820 --> 00:10:26.250 So this is sort of a middle-aged cohort,

209 00:10:26.250 --> 00:10:28.120 generally, cognitively healthy

210 00:10:29.100 --> 00:10:30.873 and a little more than half women.

211 00:10:32.640 --> 00:10:35.490 The main outcomes that we looked at in this study

212 00:10:35.490 --> 00:10:40.490 are MRI-based measures, so brain MRIs were taken

213 00:10:41.310 --> 00:10:45.120 about 10 years or so, five to 10 years

214 00:10:45.120 --> 00:10:50.120 after the initial blood draws, and those had...

215 00:10:50.730 --> 00:10:54.060 The sort of outcomes that I looked at there are

216 00:10:54.060 --> 00:10:57.390 total brain volume as well as the volume of the hippocampus

217 00:10:57.390 --> 00:11:00.750 and then a measure called white matter hyperintensities,

218 00:11:00.750 --> 00:11:05.133 which is sort of a measure of vascular injury in the brain,

219 00:11:06.300 --> 00:11:10.200 and a reason to look at those outcomes is that

220 00:11:10.200 --> 00:11:12.780 I mentioned there are sort of precursors of dementia

221 00:11:12.780 --> 00:11:16.140 or risk factors for dementia that can be identified on MRI,

222 00:11:16.140 --> 00:11:17.690 those are some of the big ones.

223 00:11:19.080 --> 00:11:21.660 Especially since we had a middle-aged cohort,

224 00:11:21.660 --> 00:11:23.907 you may not see a lot of incident dementia,

225 00:11:23.907 --> 00:11:26.520 and so being able to detect proteins

226 00:11:26.520 --> 00:11:29.400 that are associated with some of those precursors

227 00:11:29.400 --> 00:11:32.283 is a way of getting at this issue.

228 00:11:33.840 --> 00:11:35.640 We did also look at incident dementia.

229 00:11:35.640 --> 00:11:37.380 So we had about 20 years of follow-up,

230 00:11:37.380 --> 00:11:39.570 which is one of the strengths of this,

231 00:11:39.570 --> 00:11:42.300 looking in this particular sample,

232 00:11:42.300 --> 00:11:45.930 and we had 128 incidences of dementia

233 00:11:45.930 --> 00:11:47.820 of which 94 of them were classified

234 00:11:47.820 --> 00:11:49.413 as Alzheimer's type dementia.

235 00:11:53.190 --> 00:11:55.260 We also had a replication cohort.

236 00:11:55.260 --> 00:11:57.690 I mentioned the importance replication,

237 00:11:57.690 --> 00:12:00.000 and so we worked with collaborators

238 00:12:00.000 --> 00:12:03.930 at the University of Washington and their cohort study

239 00:12:03.930 --> 00:12:05.550 called the Cardiovascular Health Study,

240 00:12:05.550 --> 00:12:08.610 which has sites, I think, four different sites around the US

241 00:12:08.610 --> 00:12:13.290 and has measures of the same proteomic platform

242 00:12:13.290 --> 00:12:16.053 and same outcomes that we're looking at in the study.

243 00:12:19.410 --> 00:12:22.530 The assay that we used to measure proteins

244 00:12:22.530 --> 00:12:24.180 is called SOMAScan.

245 00:12:24.180 --> 00:12:26.670 It's by this company called SomaLogic.

246 00:12:26.670 --> 00:12:29.430 They use these single-stranded DNA aptamers

247 00:12:29.430 --> 00:12:31.320 that are designed to specifically bind

248 00:12:31.320 --> 00:12:34.818 to different proteins, and you can sort of tag them

249 00:12:34.818 --> 00:12:37.724 that way and measure their concentrations.

250 00:12:37.724 --> 00:12:42.120 In our sample, the assay had 1,300 proteins,

251 00:12:42.120 --> 00:12:44.580 which that's even sort of becoming dated now.

252 00:12:44.580 --> 00:12:46.470 I think the latest version

253 00:12:46.470 --> 00:12:48.300 has something like 7,000 proteins.

254 00:12:48.300 --> 00:12:50.580 So there's a lot that can be measured with this,

255 00:12:50.580 --> 00:12:55.203 but there is some sort of bias towards, I think,

256 00:12:56.850 --> 00:12:59.190 molecules that sort of have some evidence

257 00:12:59.190 --> 00:13:01.080 of being important in cardiovascular disease.

258 00:13:01.080 --> 00:13:04.743 So it's not an entirely sort of agnostic choice of proteins,

259 00:13:05.793 --> 00:13:07.893 but it does get a pretty wide range.

260 00:13:10.620 --> 00:13:14.730 Okay, so that's a description of the data,

261 00:13:14.730 --> 00:13:16.920 and, now, I want to dig in a bit

262 00:13:16.920 --> 00:13:19.083 to the network methods that we used.

263 00:13:20.010 --> 00:13:24.390 So this is sort of a graphical abstract

264 00:13:24.390 --> 00:13:26.253 from their original paper,

265 00:13:28.080 --> 00:13:29.460 describing this weighted gene

266 00:13:29.460 --> 00:13:31.413 correlation network analysis method.

267 00:13:32.310 --> 00:13:34.410 So that's what WGCNA stands for.

268 00:13:34.410 --> 00:13:37.159 I put gene in parentheses because they've started

269 00:13:37.159 --> 00:13:40.290 dropping that from the name when it gets used elsewhere

270 00:13:40.290 --> 00:13:42.330 because, originally, it was developed

271 00:13:42.330 --> 00:13:45.180 for gene expression data, but it's been found to have use

272 00:13:45.180 --> 00:13:48.240 in other high dimensional data sets as well,

273 00:13:48.240 --> 00:13:52.380 and so in our case, we're using it to analyze proteins,

274 00:13:52.380 --> 00:13:56.463 but the language here makes reference to gene expression.

275 00:13:57.450 --> 00:14:00.621 So just broadly, what this method does

276 00:14:00.621 --> 00:14:04.050 is you get a co-expression network,

277 00:14:04.050 --> 00:14:07.710 and I'll sort of give details on the next few slides,

278 00:14:07.710 --> 00:14:09.720 but the idea is that the network is based

279 00:14:09.720 --> 00:14:13.500 on co-occurrence or correlation in your sample.

280 00:14:13.500 --> 00:14:16.800 So there’s not really information coming from outside.

281 00:14:16.800 --> 00:14:18.810 You’re not even considering your outcome at all.

282 00:14:18.810 --> 00:14:21.417 It’s just looking at the space of the proteins

283 00:14:21.417 --> 00:14:24.123 and which proteins are correlated with one another.

284 00:14:25.620 --> 00:14:29.040 Once you’ve identified this sort of network matrix,

285 00:14:29.040 --> 00:14:32.100 you use a hierarchical clustering algorithm

286 00:14:32.100 --> 00:14:34.350 to define modules.

287 00:14:34.350 --> 00:14:37.320 It’s a little small here, but I’ll show a a bigger example.

288 00:14:37.320 --> 00:14:39.240 Basically, you have a dendrogram,

289 00:14:39.240 --> 00:14:41.730 and you see that if sort of proteins

290 00:14:41.730 --> 00:14:44.063 are on this x-axis of this figure here.

291 00:14:44.063 --> 00:14:46.383 I’ll do the mouse for people who are online.

292 00:14:47.760 --> 00:14:51.150 You get these sort of bands or groups of proteins

293 00:14:51.150 --> 00:14:53.340 that are highly correlated with one another

294 00:14:53.340 --> 00:14:55.743 and not correlated with other proteins.

295 00:14:57.840 --> 00:15:00.693 So that is where those sort of protein groups come from.

296 00:15:01.590 --> 00:15:05.790 Once you have those, you can use a numerical summary

297 00:15:05.790 --> 00:15:09.780 of each protein group as sort of a feature or a predictor

298 00:15:09.780 --> 00:15:12.750 in a regression or some sort of analysis

299 00:15:12.750 --> 00:15:15.210 to try and relate the modules or groups

300 00:15:15.210 --> 00:15:16.440 to external information.

301 00:15:16.440 --> 00:15:20.160 So that’s how we relate our protein groups

302 00:15:20.160 --> 00:15:22.743 to dementia outcomes in this study.

303 00:15:23.880 --> 00:15:25.200 There's also the possibility

304 00:15:25.200 --> 00:15:27.900 of looking at relationships between modules.

305 00:15:27.900 --> 00:15:31.560 So I mentioned the modules in the network

306 00:15:31.560 --> 00:15:33.210 are highly inter-correlated

307 00:15:33.210 --> 00:15:35.580 within the proteins within themselves,

308 00:15:35.580 --> 00:15:38.362 but there may also be some correlation between modules,

309 00:15:38.362 --> 00:15:41.100 and that could be important to look at as well,

310 00:15:41.100 --> 00:15:44.340 and then within modules, you may have

311 00:15:44.340 --> 00:15:47.340 tens or hundreds of proteins, and so trying to figure out

312 00:15:47.340 --> 00:15:49.500 which proteins within those modules

313 00:15:49.500 --> 00:15:51.696 are driving any associations you see

314 00:15:51.696 --> 00:15:54.870 is sort of a final step that can be

315 00:15:54.870 --> 00:15:57.060 useful for getting sort of biological meaning

316 00:15:57.060 --> 00:15:58.443 out of these associations.

317 00:16:02.070 --> 00:16:03.240 So that's a broad overview.

318 00:16:03.240 --> 00:16:07.890 This is sort of a more graphical abstract from our study,

319 00:16:07.890 --> 00:16:10.510 and I'll sort of go through bit by bit

320 00:16:11.430 --> 00:16:13.683 the different pieces of the analysis.

321 00:16:14.610 --> 00:16:17.580 So, again, this WGCNA step is sort of the first step

322 00:16:17.580 --> 00:16:19.950 of getting from this protein expression matrix

323 00:16:19.950 --> 00:16:23.760 where you have sort of your proteins by participants,

324 00:16:23.760 --> 00:16:27.420 and using the sort of correlations in your sample

325 00:16:27.420 --> 00:16:30.753 to come up with these modules of co-expressed proteins.

326 00:16:33.300 --> 00:16:35.040 The first step in doing that

327 00:16:35.040 --> 00:16:38.760 is to make a pairwise correlation or similarity matrix.

328 00:16:38.760 --> 00:16:40.293 So if you have n proteins,

329 00:16:40.293 --> 00:16:42.510 then that becomes an n by n matrix

330 00:16:42.510 --> 00:16:44.670 where each cell is describing

331 00:16:44.670 --> 00:16:47.130 the similarity or correlation

332 00:16:47.130 --> 00:16:51.273 between protein i and protein j in your sample.

333 00:16:52.290 --> 00:16:53.610 You then use this to create

334 00:16:53.610 --> 00:16:56.340 what's called an adjacency matrix, which is,

335 00:16:56.340 --> 00:16:58.290 I'll talk about more in the next slide,

336 00:16:58.290 --> 00:17:00.940 but is sort of a more networky way

337 00:17:02.190 --> 00:17:05.226 of describing the association between proteins,

338 00:17:05.226 --> 00:17:08.070 and then a topological overlap matrix,

339 00:17:08.070 --> 00:17:09.990 which then takes into account

340 00:17:09.990 --> 00:17:12.510 not only the correlation between proteins

341 00:17:12.510 --> 00:17:15.750 but their shared neighborhood, and then, again,

342 00:17:15.750 --> 00:17:18.693 that is what is used to cluster the proteins.

343 00:17:22.860 --> 00:17:24.900 So to get into a bit more detail

344 00:17:24.900 --> 00:17:28.143 about sort of the network construction,

345 00:17:30.060 --> 00:17:32.910 again, you described the network as an n by n matrix

346 00:17:32.910 --> 00:17:36.180 with the number of nodes or genes, proteins, et cetera,

347 00:17:36.180 --> 00:17:39.210 and, in our case, we use to describe the similarity,

348 00:17:39.210 --> 00:17:41.163 just a simple correlation,

349 00:17:42.060 --> 00:17:43.860 absolute value of the correlation,

350 00:17:43.860 --> 00:17:46.233 between a given node i and j.

351 00:17:48.046 --> 00:17:51.420 The adjacency is then a measure of whether or how strongly

352 00:17:51.420 --> 00:17:53.310 the nodes are connected in the network.

353 00:17:53.310 --> 00:17:55.830 So the idea being that

354 00:17:55.830 --> 00:17:57.870 nodes that have very high correlations

355 00:17:57.870 --> 00:17:59.730 are particularly interesting.

356 00:17:59.730 --> 00:18:01.830 Nodes that have moderate to low correlations

357 00:18:01.830 --> 00:18:03.450 are probably not informative

358 00:18:03.450 --> 00:18:06.690 is sort of the the underlying idea,

359 00:18:06.690 --> 00:18:11.690 and so if you look at sort of this figure here,

360 00:18:12.390 --> 00:18:15.180 the correlation or similarity is on the x-axis,

361 00:18:15.180 --> 00:18:18.991 and then the adjacency is on the y, and so if you use

362 00:18:18.991 --> 00:18:22.260 what's called an unweighted network approach,

363 00:18:22.260 --> 00:18:25.350 you pick a threshold value, here, it's 0.8,

364 00:18:25.350 --> 00:18:28.260 and you say that anything with a similarity less than 0.8

365 00:18:28.260 --> 00:18:31.110 is considered to not be a connection in the network,

366 00:18:31.110 --> 00:18:32.870 and everything greater than 0.8

367 00:18:32.870 --> 00:18:34.320 is considered to be a connection.

368 00:18:34.320 --> 00:18:36.483 So it's sort of a binary yes or no.

369 00:18:38.100 --> 00:18:41.850 What WGCNA does that was novel

370 00:18:41.850 --> 00:18:44.610 was to introduce a weighting

371 00:18:44.610 --> 00:18:49.050 where sort of the downside of this unweighted metric is that

372 00:18:49.050 --> 00:18:52.080 if you have a correlation of 0.79,

373 00:18:52.080 --> 00:18:55.080 that could be useful to know, but it counts as a zero.

374 00:18:55.080 --> 00:18:56.613 So you're losing information,

375 00:18:57.480 --> 00:18:59.790 and so what the weighted network does

376 00:18:59.790 --> 00:19:03.330 is it uses a sort of power transformation

377 00:19:03.330 --> 00:19:06.720 to get from sort of the straight correlation

378 00:19:06.720 --> 00:19:08.430 shown in this red line,

379 00:19:08.430 --> 00:19:12.090 and sort of depending on this power value that you use,

380 00:19:12.090 --> 00:19:15.673 you weight more or less towards the higher correlations

381 00:19:15.673 --> 00:19:19.980 in your network, and when you fit this model

382 00:19:19.980 --> 00:19:23.880 or when you sort of build the network, your choice of data

383 00:19:23.880 --> 00:19:27.030 is sort of one of the parameters that you choose going in,

384 00:19:27.030 --> 00:19:29.940 and there's ways to sort of measure

385 00:19:29.940 --> 00:19:32.103 which gives the best fit to the data.

386 00:19:37.980 --> 00:19:41.490 So then once you have your sort of unweighted

387 00:19:41.490 --> 00:19:44.550 or weighted adjacency matrix,

388 00:19:44.550 --> 00:19:47.790 then is the part where you account for shared neighbors.

389 00:19:47.790 --> 00:19:51.543 So this is this topological overlap matrix that is created,

390 00:19:52.434 --> 00:19:56.613 so, basically, this measure omega of connectedness.

391 00:19:57.960 --> 00:20:00.810 The equation, I don't find super sort of intuitive,

392 00:20:00.810 --> 00:20:03.120 but the components are...

393 00:20:03.120 --> 00:20:05.370 This is the sum, so u are, basically,

394 00:20:05.370 --> 00:20:07.110 all of the nodes other than i and j

395 00:20:07.110 --> 00:20:09.930 that you're looking at the connectedness between,

396 00:20:09.930 --> 00:20:11.490 and so you're summing up

397 00:20:11.490 --> 00:20:15.240 the sort of common connection strength between i and u

398 00:20:15.240 --> 00:20:18.120 and j and u as a product.

399 00:20:18.120 --> 00:20:21.690 So if I and J both have a strong connection

400 00:20:21.690 --> 00:20:25.953 to this other node, then that's adding to this term l,

401 00:20:27.240 --> 00:20:28.890 and then these k terms here

402 00:20:28.890 --> 00:20:32.010 are just the individual connections between, no,

403 00:20:32.010 --> 00:20:34.290 each sort of the node i of interest

404 00:20:34.290 --> 00:20:35.840 and other nodes in the network,

405 00:20:36.870 --> 00:20:41.460 but I find sort of the easiest or most intuitive explanation

406 00:20:41.460 --> 00:20:45.930 from this original paper shows that for the unweighted case,

407 00:20:45.930 --> 00:20:49.560 omega is equal to one if the node with fewer connections

408 00:20:49.560 --> 00:20:51.360 has all of its neighbors,

409 00:20:51.360 --> 00:20:52.920 also, has connections of the other node.

410 00:20:52.920 --> 00:20:55.350 So the connections of node i

411 00:20:55.350 --> 00:20:58.530 are a subset of the connections of node j,

412 00:20:58.530 --> 00:21:00.750 and, also, i and j are directly connected.

413 00:21:00.750 --> 00:21:02.520 So that’s sort of the most interconnected

414 00:21:02.520 --> 00:21:03.920 that those two nodes can be,

415 00:21:04.770 --> 00:21:07.740 and then the least interconnected they can be

416 00:21:07.740 --> 00:21:09.690 is if they are not connected to one another,

417 00:21:09.690 --> 00:21:10.920 and they don’t share any neighbors.

418 00:21:10.920 --> 00:21:13.200 So that would be sort of the zero case.

419 00:21:15.510 --> 00:21:17.970 So this a value can either take on

420 00:21:17.970 --> 00:21:19.950 the unweighted or the weighted case,

421 00:21:19.950 --> 00:21:22.710 and in our sample with WGCNA,

422 00:21:22.710 --> 00:21:26.250 we’re using those sort of weighted network connections

423 00:21:26.250 --> 00:21:27.990 that just adds more information

424 00:21:27.990 --> 00:21:30.693 into this topological overlap matrix.

425 00:21:36.107 --> 00:21:36.940 Okay.

426 00:21:39.150 --> 00:21:44.150 So, now, once you have the topological overlap matrix,

427 00:21:44.850 --> 00:21:47.970 again, this measure of sort of interconnectedness

428 00:21:47.970 --> 00:21:49.743 accounting for shared neighbors,

429 00:21:50.670 --> 00:21:53.250 then you can use hierarchical clustering

430 00:21:53.250 --> 00:21:56.700 to divide those proteins

431 00:21:56.700 --> 00:21:59.373 into groups based on their similarity,

432 00:22:00.390 --> 00:22:03.480 and this is the results from our analysis.

433 00:22:03.480 --> 00:22:06.060 So sort of on the x-axis,

434 00:22:06.060 --> 00:22:09.030 you have the different proteins, you have the dendrogram,

435 00:22:09.030 --> 00:22:11.130 which represents the hierarchical clustering

436 00:22:11.130 --> 00:22:13.950 of the topological overlap matrix,

437 00:22:13.950 --> 00:22:18.950 and then you have this dynamic tree cut algorithm

438 00:22:19.867 --> 00:22:22.020 which then defines these clusters

439 00:22:22.020 --> 00:22:26.010 which are shown in colors on the bottom based on the tree.

440 00:22:26.010 --> 00:22:28.380 So you see this huge branch down here.

441 00:22:28.380 --> 00:22:30.030 That's gonna be this black cluster.

442 00:22:30.030 --> 00:22:32.343 There's this other cluster over here in green,

443 00:22:33.360 --> 00:22:36.660 and so there's, again, a few more parameters

444 00:22:36.660 --> 00:22:40.110 that you can use to decide how those cuts are made,

445 00:22:40.110 --> 00:22:42.660 and, in some cases, you can sort of merge branches

446 00:22:42.660 --> 00:22:45.420 that have correlation with one another,

447 00:22:45.420 --> 00:22:47.670 and my general advice

448 00:22:47.670 --> 00:22:49.290 for when you're doing this on real data

449 00:22:49.290 --> 00:22:50.940 is to try different values

450 00:22:50.940 --> 00:22:52.500 and see how robust the network is

451 00:22:52.500 --> 00:22:56.490 to choosing different values because, in our case,

452 00:22:56.490 --> 00:22:59.370 it tended to be pretty consistent

453 00:22:59.370 --> 00:23:02.070 where we saw four modules pretty much regardless.

454 00:23:02.070 --> 00:23:03.600 I think if we merged,

455 00:23:03.600 --> 00:23:05.820 if we really cranked up one of the merging parameters,

456 00:23:05.820 --> 00:23:06.653 we would get to three,

457 00:23:06.653 --> 00:23:09.453 but other than that it sort of stayed put.

458 00:23:12.900 --> 00:23:13.733 Okay.

459 00:23:15.150 --> 00:23:17.850 So the next step is trying to get

460 00:23:17.850 --> 00:23:22.290 a numerical summary measure of the groups of proteins

461 00:23:22.290 --> 00:23:25.140 that we've identified from our network.

462 00:23:25.140 --> 00:23:28.380 So from these modules of co-expressed proteins,

463 00:23:28.380 --> 00:23:32.940 we then use, basically, a principle components analysis

464 00:23:32.940 --> 00:23:35.220 to get what we call an eigenprotein

465 00:23:35.220 --> 00:23:38.790 or it was called an eigen gene in the original paper.

466 00:23:38.790 --> 00:23:42.510 What it is is, essentially, a weighted sum

467 00:23:42.510 --> 00:23:46.530 of the values of each of the proteins in the module,

468 00:23:46.530 --> 00:23:50.220 and the weights correspond to sort of how well correlated

469 00:23:50.220 --> 00:23:52.560 that protein is with the overall module.

470 00:23:52.560 --> 00:23:55.500 So if a protein has a high weight in the module,

471 00:23:55.500 --> 00:23:58.410 it means that it's sort of the most interconnected

472 00:23:58.410 --> 00:24:02.733 in the module or sort of best represents the overall module.

473 00:24:03.900 --> 00:24:06.000 So each person is going to have

474 00:24:06.000 --> 00:24:10.173 an eigenprotein value for each module,

475 00:24:16.020 --> 00:24:18.180 and when we look at the sort of weights

476 00:24:18.180 --> 00:24:22.314 within each of the modules, so just to sort of orient us,

477 00:24:22.314 --> 00:24:27.000 on the x-axis are each of the module eigen genes

478 00:24:27.000 --> 00:24:32.000 or eigenproteins, and then each sort of bar

479 00:24:33.630 --> 00:24:36.390 on the y is a different protein.

480 00:24:36.390 --> 00:24:38.880 In this case, we're only including

481 00:24:38.880 --> 00:24:41.970 proteins that fall into one of the four modules.

482 00:24:41.970 --> 00:24:45.240 There were, also, if you notice on the last slide,

483 00:24:45.240 --> 00:24:47.910 plenty of proteins that didn't fall into any module

484 00:24:47.910 --> 00:24:50.940 and were sort of the extras, so to speak,

485 00:24:50.940 --> 00:24:53.670 and if you were to expand this down

486 00:24:53.670 --> 00:24:56.160 and include more rows with those,

487 00:24:56.160 --> 00:25:00.000 that would sort of show those, but for purposes of this,

488 00:25:00.000 --> 00:25:01.020 we're just including ones

489 00:25:01.020 --> 00:25:03.020 that fell into at least one of the four,

490 00:25:04.198 --> 00:25:08.520 and each of these bars represents a correlation

491 00:25:08.520 --> 00:25:10.620 between the individual protein

492 00:25:10.620 --> 00:25:12.363 and the overall eigenprotein.

493 00:25:13.260 --> 00:25:14.823 So for these blocks of red,

494 00:25:14.823 --> 00:25:16.950 it's sort of the higher weighted proteins

495 00:25:16.950 --> 00:25:20.580 that are within in this example module one,

496 00:25:20.580 --> 00:25:23.583 module two, three, and four, and then you can see,

497 00:25:24.450 --> 00:25:27.990 if you look sort of laterally from these proteins,

498 00:25:27.990 --> 00:25:30.060 it's the correlation of these proteins

499 00:25:30.060 --> 00:25:31.230 with the other modules.

500 00:25:31.230 --> 00:25:35.580 So the idea being we wanna see sort of blocks of red,

501 00:25:35.580 --> 00:25:37.500 and then not a lot of correlation

502 00:25:37.500 --> 00:25:40.440 between the blocks and other modules,

503 00:25:40.440 --> 00:25:42.093 which is what we see.

504 00:25:46.020 --> 00:25:49.590 All right, now that we've constructed our network,

505 00:25:49.590 --> 00:25:52.290 and we've come up with numerical summary measures

506 00:25:52.290 --> 00:25:55.500 for each of the protein groups that we've identified,

507 00:25:55.500 --> 00:25:58.500 that is sort of the input or the predictor

508 00:25:58.500 --> 00:26:01.860 for these associations with outcomes.

509 00:26:01.860 --> 00:26:04.080 So for the MRI measures, which, again,

510 00:26:04.080 --> 00:26:07.080 our total brain volume, hippocampal volume,

511 00:26:07.080 --> 00:26:08.790 and white matter hyperintensities,

512 00:26:08.790 --> 00:26:11.520 we use just a simple or, you know,

513 00:26:11.520 --> 00:26:14.100 linear regression with covariates,

514 00:26:14.100 --> 00:26:16.830 and then a Cox proportional hazards regression,

515 00:26:16.830 --> 00:26:20.310 we use to predict incident dementia

516 00:26:20.310 --> 00:26:23.163 and, specifically, Alzheimer's type dementia.

517 00:26:25.560 --> 00:26:27.720 These are the regression equations.

518 00:26:27.720 --> 00:26:29.910 Again, these eigenproteins are,

519 00:26:29.910 --> 00:26:31.740 they're sort of one for each module.

520 00:26:31.740 --> 00:26:34.620 So we'll run a separate regression analysis

521 00:26:34.620 --> 00:26:37.350 for modules one, two, three, and four.

522 00:26:37.350 --> 00:26:41.820 We adjust for age and age squared, sex education.

523 00:26:41.820 --> 00:26:44.220 APOE is a gene that confers a lot of risk

524 00:26:44.220 --> 00:26:45.270 for Alzheimer's disease.

525 00:26:45.270 --> 00:26:46.767 So it's associated with the outcomes,

526 00:26:46.767 --> 00:26:48.807 and we include it as a covariate,

527 00:26:48.807 --> 00:26:51.240 and then a measure of time lag

528 00:26:51.240 --> 00:26:53.040 between when the blood was sampled

529 00:26:53.040 --> 00:26:56.652 and when the MRI was taken to account for any differences

530 00:26:56.652 --> 00:26:59.733 between people or the time difference,

531 00:27:01.170 --> 00:27:05.790 and for dementia, it's slightly simpler regression equation.

532 00:27:05.790 --> 00:27:09.123 We only adjust for age, sex, and APOE status.

533 00:27:13.410 --> 00:27:16.590 All right, so next, I will show

534 00:27:16.590 --> 00:27:19.653 the results in the Framingham Heart Study.

535 00:27:20.670 --> 00:27:24.030 So from the four modules that we tested,

536 00:27:24.030 --> 00:27:26.580 there were two that we identified to have

537 00:27:26.580 --> 00:27:28.890 some association with outcomes.

538 00:27:28.890 --> 00:27:31.170 The first is module two.

539 00:27:31.170 --> 00:27:34.560 I gave it sort of a name clearance and synaptic maintenance,

540 00:27:34.560 --> 00:27:36.630 and I'll talk about how I arrived

541 00:27:36.630 --> 00:27:39.660 at that name for the module in a bit.

542 00:27:39.660 --> 00:27:42.093 It has 165 proteins in it.

543 00:27:43.830 --> 00:27:46.680 Some of the half weighted proteins sort of give an idea

544 00:27:46.680 --> 00:27:49.300 of which ones are sort of most highly weighted

545 00:27:51.120 --> 00:27:53.763 or sort of most correlated with the eigen protein.

546 00:27:56.160 --> 00:27:58.560 I'll talk about how we got to these

547 00:27:58.560 --> 00:28:00.510 in another slide as well,

548 00:28:00.510 --> 00:28:01.890 but, basically, this is from that

549 00:28:01.890 --> 00:28:03.540 over-representation analysis

550 00:28:03.540 --> 00:28:06.480 where you're trying to identify biological pathways

551 00:28:06.480 --> 00:28:09.116 that are important or overrepresented

552 00:28:09.116 --> 00:28:12.150 by proteins in those modules.

553 00:28:12.150 --> 00:28:14.250 So we have the Axon guidance pathway

554 00:28:14.250 --> 00:28:19.173 was most strongly associated with this module,

555 00:28:21.120 --> 00:28:24.510 and then in terms of relating to outcomes,

556 00:28:24.510 --> 00:28:25.710 total brain volume

557 00:28:25.710 --> 00:28:28.830 was the only significant association that we saw.

558 00:28:28.830 --> 00:28:33.462 So since this is a linear aggression,

559 00:28:33.462 --> 00:28:37.110 effect greater than zero means a positive association.

560 00:28:37.110 --> 00:28:39.930 So we see that for larger values

561 00:28:39.930 --> 00:28:42.180 of the eigenprotein for module two,

562 00:28:42.180 --> 00:28:44.310 we saw larger total brain volume.

563 00:28:44.310 --> 00:28:46.090 So it's sort of a protective effect

564 00:28:47.370 --> 00:28:50.913 since brain atrophy is what is the risk factor for dementia,

565 00:28:52.860 --> 00:28:54.570 and then for incident dementia,

566 00:28:54.570 --> 00:28:56.220 we did not see a significant effect

567 00:28:56.220 --> 00:28:58.320 after correcting our p-values

568 00:28:58.320 --> 00:29:00.240 using a Bonferroni correction.

569 00:29:00.240 --> 00:29:03.660 You'll notice that the confidence interval excludes one,

570 00:29:03.660 --> 00:29:04.860 which would be the null value,

571 00:29:04.860 --> 00:29:06.200 and that's just because that's based

572 00:29:06.200 --> 00:29:10.260 on the non-Bonferroni corrected value,

573 00:29:10.260 --> 00:29:14.160 but after testing for or adjusting for the four modules

574 00:29:14.160 --> 00:29:18.330 that we tested, we didn't see a significant association.

575 00:29:18.330 --> 00:29:21.990 It is nice at least that the direction of effect

576 00:29:21.990 --> 00:29:23.040 is what we would expect

577 00:29:23.040 --> 00:29:26.130 based on our total brain volume association,

578 00:29:26.130 --> 00:29:28.160 which is that higher values of M2

579 00:29:31.290 --> 00:29:36.183 correspond to sort of a lower incident dementia occurrence.

580 00:29:38.460 --> 00:29:40.830 The second module that we found to be associated

581 00:29:40.830 --> 00:29:43.650 with total brain volume was this M4,

582 00:29:43.650 --> 00:29:46.950 which I will call sort of an inflammation-related module.

583 00:29:46.950 --> 00:29:48.843 It had 42 proteins in it.

584 00:29:49.680 --> 00:29:52.200 The highlighted pathway there

585 00:29:52.200 --> 00:29:54.630 was cytokine-cytokine receptor interactions,

586 00:29:54.630 --> 00:29:57.490 so these sort of immune signaling molecules,

587 00:29:57.490 --> 00:30:00.030 and in this case, the association

588 00:30:00.030 --> 00:30:01.230 was in the opposite direction

589 00:30:01.230 --> 00:30:04.530 where higher values of this module for eigen-protein

590 00:30:04.530 --> 00:30:06.570 are associated with lower total brain volume.

591 00:30:06.570 --> 00:30:10.320 So it's sort of a risk conferring module

592 00:30:10.320 --> 00:30:13.706 and, again, similar to what we saw here, not a significant,

593 00:30:13.706 --> 00:30:17.077 sort of an annoyingly borderline association

594 00:30:17.077 --> 00:30:20.310 between this and dementia, but, again,

595 00:30:20.310 --> 00:30:23.760 the direction of effect is what we would expect

596 00:30:23.760 --> 00:30:27.273 based on our observed association with brain volume,

597 00:30:28.860 --> 00:30:31.290 and, also, I'll just mention that I standardize

598 00:30:31.290 --> 00:30:33.900 the eigenprotein so that the effect sizes

599 00:30:33.900 --> 00:30:36.810 correspond to a standard deviation increase in eigenprotein.

600 00:30:36.810 --> 00:30:38.730 So it's a little bit...

601 00:30:38.730 --> 00:30:40.440 One sort of drawback I would say

602 00:30:40.440 --> 00:30:43.020 of these methods is the interpretation

603 00:30:43.020 --> 00:30:46.680 since a standard deviation increase, in this case,

604 00:30:46.680 --> 00:30:49.230 depends entirely on the sample that you're using.

605 00:30:49.230 --> 00:30:52.240 So it's really just sort of a direction of effect

606 00:30:53.730 --> 00:30:54.680 more than anything.

607 00:30:56.460 --> 00:31:00.060 So to try and get at some of, get a better understanding

608 00:31:00.060 --> 00:31:03.390 of how these modules relate to our data

609 00:31:03.390 --> 00:31:05.550 or sort of what may be responsible

610 00:31:05.550 --> 00:31:08.160 for some of the associations we see,

611 00:31:08.160 --> 00:31:11.610 this is a map of the correlations

612 00:31:11.610 --> 00:31:14.520 between different demographic variables

613 00:31:14.520 --> 00:31:17.520 and each of the modules, and I mentioned that we have

614 00:31:17.520 --> 00:31:19.920 a replication cohort as well, the CHS.

615 00:31:19.920 --> 00:31:23.100 So these two bars, sort of the two columns,

616 00:31:23.100 --> 00:31:26.553 show the two different cohorts that were included.

617 00:31:27.510 --> 00:31:31.350 So I put blue arrows to show the covariates

618 00:31:31.350 --> 00:31:33.600 that were included in our regression model,

619 00:31:33.600 --> 00:31:35.490 and you can see that there are some correlations

620 00:31:35.490 --> 00:31:37.994 between, say, sex and the modules,

621 00:31:37.994 --> 00:31:41.610 not really anything with APOE carrier status,

622 00:31:41.610 --> 00:31:44.130 maybe some education associations,

623 00:31:44.130 --> 00:31:45.660 and some associations with age.

624 00:31:45.660 --> 00:31:49.290 So it's good that we adjusted for those in our models.

625 00:31:49.290 --> 00:31:52.740 However, you can also see there are a lot of other factors,

626 00:31:52.740 --> 00:31:54.450 cardiovascular risk factors,

627 00:31:54.450 --> 00:31:58.020 such as systolic blood pressure, BMI,

628 00:31:58.020 --> 00:32:01.770 fasting glucose that have associations with these modules.

629 00:32:01.770 --> 00:32:05.490 So we wanted to see if any of those could perhaps explain

630 00:32:05.490 --> 00:32:07.263 the associations that we saw.

631 00:32:10.350 --> 00:32:13.740 So I'm repeating sort of our standard model here

632 00:32:13.740 --> 00:32:16.203 was what I showed results from previously.

633 00:32:17.040 --> 00:32:18.840 The expanded model that we considered

634 00:32:18.840 --> 00:32:21.213 included a bunch of these risk factors,

635 00:32:22.590 --> 00:32:26.700 basically, something representing BMI,

636 00:32:26.700 --> 00:32:31.700 hypertension, sort of lipid dysregulation, and diabetes,

637 00:32:33.360 --> 00:32:35.643 and I also included smoking as well,

638 00:32:36.719 --> 00:32:40.380 and we also included a measure of kidney function,

639 00:32:40.380 --> 00:32:43.593 which can also be an indicator of cardiovascular disease.

640 00:32:45.120 --> 00:32:46.533 So for module two,

641 00:32:47.520 --> 00:32:50.400 I'm repeating the sort of effects we saw

642 00:32:50.400 --> 00:32:51.963 from the standard model here,

643 00:32:52.950 --> 00:32:55.650 and when you adjust for the expanded set of covariates,

644 00:32:55.650 --> 00:32:58.320 your effect is attenuated by half,

645 00:32:58.320 --> 00:33:01.170 and it's no longer significantly associated.

646 00:33:01.170 --> 00:33:04.320 So with that says, it's either you have

647 00:33:04.320 --> 00:33:08.490 a sort of confounding issue

648 00:33:08.490 --> 00:33:12.300 where the association you're seeing between these proteins

649 00:33:12.300 --> 00:33:15.660 and total brain volume is really just in effect

650 00:33:15.660 --> 00:33:19.590 of sort of poor cardiovascular health

651 00:33:19.590 --> 00:33:21.160 or better cardiovascular health

652 00:33:22.230 --> 00:33:24.870 or you may think that it might be

653 00:33:24.870 --> 00:33:26.370 some sort of mediation effect

654 00:33:26.370 --> 00:33:29.860 where perhaps the risk associated

655 00:33:31.290 --> 00:33:34.470 between the proteins and the sort of total brain volume

656 00:33:34.470 --> 00:33:35.370 could be mediated

657 00:33:35.370 --> 00:33:39.813 by some poor cardiovascular health outcomes,

658 00:33:41.430 --> 00:33:43.080 and then for module four,

659 00:33:43.080 --> 00:33:45.300 again, this sort of inflammation module,

660 00:33:45.300 --> 00:33:48.120 we don't see any real effect attenuation.

661 00:33:48.120 --> 00:33:49.410 Regardless of whether you adjust

662 00:33:49.410 --> 00:33:51.540 for cardiovascular factors or not,

663 00:33:51.540 --> 00:33:53.550 it's still associated with total brain volume,

664 00:33:53.550 --> 00:33:56.670 which suggests it's sort of different mechanism

665 00:33:56.670 --> 00:33:58.721 or lack of compounding between

666 00:33:58.721 --> 00:34:01.293 or based on cardiovascular health.

667 00:34:04.740 --> 00:34:07.921 Okay, so I mentioned

668 00:34:07.921 --> 00:34:11.550 in the sort of initial graphical abstract

669 00:34:11.550 --> 00:34:13.500 that once you find protein modules

670 00:34:13.500 --> 00:34:15.900 associated with your outcomes of interest,

671 00:34:15.900 --> 00:34:18.990 it can be good to look within the proteins of those modules

672 00:34:18.990 --> 00:34:20.820 to try and find sort of subsets

673 00:34:20.820 --> 00:34:25.530 or specific proteins that may be driving the associations.

674 00:34:25.530 --> 00:34:26.850 So for modules two and four,

675 00:34:26.850 --> 00:34:29.939 where we found associations with brain volume,

676 00:34:29.939 --> 00:34:34.180 we wanted to see if we removed proteins one at a time

677 00:34:35.040 --> 00:34:37.020 based on their sort of increasing weight,

678 00:34:37.020 --> 00:34:42.020 so remove the lowest weighted proteins in the modules first,

679 00:34:42.240 --> 00:34:45.750 what sort of happened to the strength of the associations.

680 00:34:45.750 --> 00:34:48.990 So these are both associations with total brain volume.

681 00:34:48.990 --> 00:34:52.620 It's sort of the p-value on the y-axis,

682 00:34:52.620 --> 00:34:56.819 and you can see that as you remove, say, from module two,

683 00:34:56.819 --> 00:34:59.460 the first 20 proteins or so,

684 00:34:59.460 --> 00:35:01.260 you're really not seeing a difference

685 00:35:01.260 --> 00:35:05.280 in the effect of the overall module with total brain volume,

686 00:35:05.280 --> 00:35:06.990 which suggests that those proteins

687 00:35:06.990 --> 00:35:10.620 aren't really impacting the association,

688 00:35:10.620 --> 00:35:15.420 whereas beyond that point, once you start removing proteins,

689 00:35:15.420 --> 00:35:17.400 the association becomes less strong,

690 00:35:17.400 --> 00:35:20.190 and so that's suggesting that those proteins

691 00:35:20.190 --> 00:35:24.720 may have more of an impact on sort of the overall module,

692 00:35:24.720 --> 00:35:28.590 and so for both of these modules, we identified the spot

693 00:35:28.590 --> 00:35:31.650 where sort of the based on the lowest p-value,

694 00:35:31.650 --> 00:35:33.910 which proteins were

695 00:35:35.190 --> 00:35:37.470 sort of the most important in the module.

696 00:35:37.470 --> 00:35:40.830 I wanna emphasize that we didn't use this to...

697 00:35:40.830 --> 00:35:43.800 So for things like dementia, if you were to run this,

698 00:35:43.800 --> 00:35:46.590 since we didn't see a strong association

699 00:35:46.590 --> 00:35:49.770 or a significant association beforehand,

700 00:35:49.770 --> 00:35:52.260 we didn't sort of use that to try and find a subset

701 00:35:52.260 --> 00:35:53.850 that we're significantly associated

702 00:35:53.850 --> 00:35:55.600 because I would call that cheating.

703 00:36:01.096 --> 00:36:05.010 Okay, so the last piece that I'll talk about

704 00:36:05.010 --> 00:36:09.210 in terms of teasing apart associations

705 00:36:09.210 --> 00:36:12.040 or sort of understanding protein within the modules

706 00:36:12.990 --> 00:36:15.780 is this functional enrichment

707 00:36:15.780 --> 00:36:19.650 or over-representation analysis within the modules.

708 00:36:19.650 --> 00:36:24.360 So based on the ones, sort of the significant modules

709 00:36:24.360 --> 00:36:27.093 or significantly associated modules with the outcomes,

710 00:36:28.080 --> 00:36:30.600 there is this software called STRING

711 00:36:30.600 --> 00:36:35.310 that does a few different things, but what I used it for

712 00:36:35.310 --> 00:36:38.490 is doing an over-representation analysis

713 00:36:38.490 --> 00:36:41.070 of biological pathways.

714 00:36:41.070 --> 00:36:45.090 So the idea is that there are annotation databases

715 00:36:45.090 --> 00:36:48.360 for proteins that sort of group them

716 00:36:48.360 --> 00:36:50.670 into biological functions

717 00:36:50.670 --> 00:36:52.830 or pathways that they're involved in,

718 00:36:52.830 --> 00:36:55.170 and the idea is that if you have a module

719 00:36:55.170 --> 00:36:57.660 that has more proteins than you would expect

720 00:36:57.660 --> 00:36:59.190 from a given pathway,

721 00:36:59.190 --> 00:37:02.010 then that's sort of the over-representation piece,

722 00:37:02.010 --> 00:37:04.770 and it indicates that that biological pathway

723 00:37:04.770 --> 00:37:07.620 might be important in whatever functions

724 00:37:07.620 --> 00:37:09.423 the module is carrying out.

725 00:37:12.030 --> 00:37:15.728 So this is just a screen grab of one example.

726 00:37:15.728 --> 00:37:18.060 So this is from module four.

727 00:37:18.060 --> 00:37:22.320 So you can see the annotation database is over on the left.

728 00:37:22.320 --> 00:37:24.090 So KEGG is one of them.

729 00:37:24.090 --> 00:37:26.190 Gene Ontology is another,

730 00:37:26.190 --> 00:37:30.321 and so you have these sort of observed proteins,

731 00:37:30.321 --> 00:37:33.210 and then the background is sort of the total number

732 00:37:33.210 --> 00:37:35.550 of proteins that are in the pathway,

733 00:37:35.550 --> 00:37:38.760 and the idea being that if you were to grab, I don't know,

734 00:37:38.760 --> 00:37:41.250 however many proteins out of the background,

735 00:37:41.250 --> 00:37:44.550 like how many would you expect to be in this module

736 00:37:44.550 --> 00:37:48.840 due to chance, and do we have sort of over-representation

737 00:37:48.840 --> 00:37:51.030 compared to what we would expect?

738 00:37:51.030 --> 00:37:52.440 And so for module four,

739 00:37:52.440 --> 00:37:54.930 the cytokine-cytokine receptor interaction

740 00:37:54.930 --> 00:37:59.160 was the strongest overrepresented pathway,

741 00:37:59.160 --> 00:38:02.200 and then you can sort of look at these others that

742 00:38:03.240 --> 00:38:07.770 have some sort of false discovery rate greater than 0.05,

743 00:38:07.770 --> 00:38:10.830 and so I found the KEGG pathways, personally,

744 00:38:10.830 --> 00:38:12.330 to be the most informative.

745 00:38:12.330 --> 00:38:15.210 Gene Ontology tends to be a lot more specific,

746 00:38:15.210 --> 00:38:17.550 which may be more useful for targeting

747 00:38:17.550 --> 00:38:20.940 certain sort of therapeutic processes

748 00:38:20.940 --> 00:38:21.810 or something like that,

749 00:38:21.810 --> 00:38:24.840 but so depending on the scale that is important to you,

750 00:38:24.840 --> 00:38:26.973 you can sort of use different annotations.

751 00:38:30.780 --> 00:38:33.360 Okay, so the last thing I wanted to talk about,

752 00:38:33.360 --> 00:38:35.703 with the Framingham data in particular,

753 00:38:37.530 --> 00:38:39.540 was sort of getting back to our motivation

754 00:38:39.540 --> 00:38:41.940 for doing a network analysis in the first place.

755 00:38:42.780 --> 00:38:46.590 So the sort of contrast or comparator would be to do

756 00:38:46.590 --> 00:38:48.632 individual protein analyses where you're running

757 00:38:48.632 --> 00:38:52.530 a regression model for each protein that you're analyzing,

758 00:38:52.530 --> 00:38:55.192 and so we did that as a point of comparison.

759 00:38:55.192 --> 00:38:59.310 So for total brain volume, there were like a dozen proteins

760 00:38:59.310 --> 00:39:01.950 that were associated with total brain volume.

761 00:39:01.950 --> 00:39:04.080 One was associated with hippocampal volume,

762 00:39:04.080 --> 00:39:07.230 and two were associated with Alzheimer's disease

763 00:39:07.230 --> 00:39:09.843 at an FDR value of less than 0.1.

764 00:39:11.400 --> 00:39:14.130 So what was interesting,

765 00:39:14.130 --> 00:39:15.660 especially with the brain volume results,

766 00:39:15.660 --> 00:39:16.800 and, again, that was where we had seen

767 00:39:16.800 --> 00:39:19.140 associations with these modules,

768 00:39:19.140 --> 00:39:22.950 some of the proteins that were significantly associated

769 00:39:22.950 --> 00:39:27.933 were from module two and module four and others weren't.

770 00:39:28.860 --> 00:39:31.770 So what I get from that is a few things.

771 00:39:31.770 --> 00:39:33.900 One is that some proteins

772 00:39:33.900 --> 00:39:35.940 that are associated with the outcome

773 00:39:35.940 --> 00:39:38.820 are sort of individually associated

774 00:39:38.820 --> 00:39:41.010 but not sort of detectable

775 00:39:41.010 --> 00:39:43.860 within sort of a larger network of proteins

776 00:39:43.860 --> 00:39:46.328 that are associated with that outcome,

777 00:39:46.328 --> 00:39:48.390 and then the other is that

778 00:39:48.390 --> 00:39:51.042 for those that are within the modules,

779 00:39:51.042 --> 00:39:52.800 we would only be getting information

780 00:39:52.800 --> 00:39:55.710 about sort of a few of the proteins in the modules,

781 00:39:55.710 --> 00:39:58.893 whereas, as we see here,

782 00:40:00.150 --> 00:40:03.450 the associations tend or continue to get stronger

783 00:40:03.450 --> 00:40:05.670 with sort of looking at the broader network

784 00:40:05.670 --> 00:40:08.850 around sort of the most highly weighted proteins.

785 00:40:08.850 --> 00:40:10.410 So you're getting a bit more information

786 00:40:10.410 --> 00:40:12.510 about proteins that may be associated

787 00:40:12.510 --> 00:40:14.010 with total brain volume

788 00:40:14.010 --> 00:40:16.950 and maybe at some of the biological processes

789 00:40:16.950 --> 00:40:19.950 compared to if you're looking at things individually,

790 00:40:19.950 --> 00:40:21.900 but, again, because you're seeing associations

791 00:40:21.900 --> 00:40:23.280 that you don't catch with the modules,

792 00:40:23.280 --> 00:40:25.350 it's sort of important to look at both,

793 00:40:25.350 --> 00:40:27.660 and you get sort of complimentary information

794 00:40:27.660 --> 00:40:29.013 from the two approaches.

795 00:40:32.700 --> 00:40:34.383 So a caveat,

796 00:40:35.700 --> 00:40:36.960 I mentioned issues with lack

797 00:40:36.960 --> 00:40:39.870 with sort of difficulties in replication.

798 00:40:39.870 --> 00:40:41.610 We replicated this analysis

799 00:40:41.610 --> 00:40:44.310 in the Cardiovascular Health Study,

800 00:40:44.310 --> 00:40:47.490 and we did so by taking the same module,

801 00:40:47.490 --> 00:40:49.530 so module two and module four,

802 00:40:49.530 --> 00:40:52.080 taking the same weights from those proteins

803 00:40:52.080 --> 00:40:56.310 and applying them to the protein concentrations

804 00:40:56.310 --> 00:40:59.490 in the Cardiovascular Health Study.

805 00:40:59.490 --> 00:41:02.010 So we didn't do a network reconstruction or anything

806 00:41:02.010 --> 00:41:03.480 in the different study.

807 00:41:03.480 --> 00:41:06.990 We were just seeing if these modules replicated

808 00:41:06.990 --> 00:41:10.290 in their associations with outcomes in a different cohort.

809 00:41:10.290 --> 00:41:14.250 So in this case, it's really not seeing much

810 00:41:14.250 --> 00:41:18.480 in terms of association with both total brain volume

811 00:41:18.480 --> 00:41:21.810 and we also looked at dementia out of interest

812 00:41:21.810 --> 00:41:26.430 since things were sort of close in our cohort,

813 00:41:26.430 --> 00:41:29.853 but, really, we're not seeing much in terms of associations.

814 00:41:31.020 --> 00:41:32.730 Part of the reason for that,

815 00:41:32.730 --> 00:41:35.670 so there are not that many cohorts

816 00:41:35.670 --> 00:41:38.850 that are available that have a large proteomic panel

817 00:41:38.850 --> 00:41:40.650 with the same proteins that we were looking at

818 00:41:40.650 --> 00:41:44.670 as well as MRI and incident dementia outcomes,

819 00:41:44.670 --> 00:41:47.700 and, in this case, the demographics of the cohort

820 00:41:47.700 --> 00:41:50.520 are fairly different from (indistinct) Framingham.

821 00:41:50.520 --> 00:41:54.783 So about 20 years older on average.

822 00:41:55.890 --> 00:41:57.930 I'm just including the sort of first few rows

823 00:41:57.930 --> 00:42:00.810 of our table one, but you can see differences in education,

824 00:42:00.810 --> 00:42:03.180 systolic blood pressure, and the same is true

825 00:42:03.180 --> 00:42:05.940 of a lot of the other cardiovascular risk factors.

826 00:42:05.940 --> 00:42:08.280 So it's a very different cohort,

827 00:42:08.280 --> 00:42:10.320 and digging a bit into the literature

828 00:42:10.320 --> 00:42:12.990 about sort of proteins over the life course,

829 00:42:12.990 --> 00:42:15.510 it's not too surprising that we don't see

830 00:42:15.510 --> 00:42:18.600 the same associations, but it it does sort of,

831 00:42:18.600 --> 00:42:20.070 it's a good cautionary message

832 00:42:20.070 --> 00:42:22.590 about drawing conclusions too far

833 00:42:22.590 --> 00:42:24.600 based on sort of one set of data

834 00:42:24.600 --> 00:42:26.973 or one set of demographics.

835 00:42:29.730 --> 00:42:32.280 Just to put these results in context,

836 00:42:32.280 --> 00:42:35.580 so our module four included

837 00:42:35.580 --> 00:42:38.040 a lot of immune-related signaling molecules

838 00:42:38.040 --> 00:42:41.430 like interleukins, TNF receptor proteins,

839 00:42:41.430 --> 00:42:44.490 which are both types of cytokines, and have been associated

840 00:42:44.490 --> 00:42:47.310 with Alzheimer's disease previously,

841 00:42:47.310 --> 00:42:51.660 in particular, interleukin-1 beta was in our module four,

842 00:42:51.660 --> 00:42:53.250 and it had been found to be elevated

843 00:42:53.250 --> 00:42:56.070 in 80 cases in a meta-analysis.

844 00:42:56.070 --> 00:42:59.760 However, other biomarkers that have been sort of validated

845 00:42:59.760 --> 00:43:04.427 in other cohorts were not identified in our module.

846 00:43:07.590 --> 00:43:11.040 In module two, we saw Axon guidance pathway proteins

847 00:43:11.040 --> 00:43:13.470 including ephrins, netrins, and semaphorins,

848 00:43:13.470 --> 00:43:16.800 which have been associated with AD in previous work,

849 00:43:16.800 --> 00:43:20.430 and complement cascades are also have been associated

850 00:43:20.430 --> 00:43:22.470 with AD probably for the reason

851 00:43:22.470 --> 00:43:26.810 of inducing these immune cells called microglia

852 00:43:26.810 --> 00:43:29.980 in the brain to, basically, eat up

853 00:43:31.470 --> 00:43:35.100 cells in response to amyloid deposition.

854 00:43:35.100 --> 00:43:37.440 So there's some biologically plausible mechanisms

855 00:43:37.440 --> 00:43:40.030 that could be associated with these modules

856 00:43:41.640 --> 00:43:43.683 in Alzheimer's disease,

857 00:43:46.080 --> 00:43:48.750 and the last thing I'll say is talking about some sort

858 00:43:48.750 --> 00:43:50.790 of other ways of approaching this problem,

859 00:43:50.790 --> 00:43:53.910 so as I mentioned, the CHS cohort

860 00:43:53.910 --> 00:43:55.830 has different underlying characteristics,

861 00:43:55.830 --> 00:43:58.950 and so it may well have a different network structure.

862 00:43:58.950 --> 00:44:02.130 So one thing that could be good to do

863 00:44:02.130 --> 00:44:07.130 is to look at sort of consensus modules across the cohorts

864 00:44:07.140 --> 00:44:09.240 where you construct networks in each cohort,

865 00:44:09.240 --> 00:44:12.390 and then look at where the overlaps are,

866 00:44:12.390 --> 00:44:13.860 and you can get sort of a more,

867 00:44:13.860 --> 00:44:16.383 hopefully, more robust network across cohorts,

868 00:44:17.640 --> 00:44:20.310 and then there are other network-based approaches

869 00:44:20.310 --> 00:44:22.290 that can incorporate external information.

870 00:44:22.290 --> 00:44:24.060 So, again, our network approach

871 00:44:24.060 --> 00:44:27.450 was just based on correlation in our dataset,

872 00:44:27.450 --> 00:44:32.450 whereas other methods use sort of those annotation databases

873 00:44:32.670 --> 00:44:34.890 and that sort of thing to construct the networks

874 00:44:34.890 --> 00:44:39.090 and sort of decide how strong the similarities between nodes

875 00:44:39.090 --> 00:44:41.100 or the strength of connections will be.

876 00:44:41.100 --> 00:44:42.450 So that's another approach,

877 00:44:43.290 --> 00:44:44.760 and then the last thing I'll say is that

878 00:44:44.760 --> 00:44:48.150 I'm sort of still using this kind of method

879 00:44:48.150 --> 00:44:51.270 now in work with longevity and aging

880 00:44:51.270 --> 00:44:53.940 and trying to apply it to metabolomics,

881 00:44:53.940 --> 00:44:58.940 so metabolites data in cohorts related to those outcomes.

882 00:45:02.160 --> 00:45:03.720 So thank you all for being here.

883 00:45:03.720 --> 00:45:05.610 Thank you, my collaborators.

884 00:45:05.610 --> 00:45:09.060 This is the folks down at UT.

885 00:45:09.060 --> 00:45:10.637 I'll say that (indistinct).

886 00:45:10.637 --> 00:45:11.470 Thank you.

887 00:45:16.170 --> 00:45:18.330 <v ->Thank you for wonderful presentation.</v>

888 00:45:18.330 --> 00:45:19.500 We're open for questions.

889 00:45:19.500 --> 00:45:21.300 So let's start with people in the room.

890 00:45:21.300 --> 00:45:22.710 Any questions?

891 00:45:22.710 --> 00:45:24.570 <v ->Got one over here.</v> <v ->Perfect, thank you.</v>

892 00:45:24.570 --> 00:45:26.220 <v Audience>Yeah, so my research interest</v>

893 00:45:26.220 --> 00:45:28.200 is about the cancer, and, also,

894 00:45:28.200 --> 00:45:30.450 we're interested in your study.

895 00:45:30.450 --> 00:45:34.920 So I've got some technical issues about this project.

896 00:45:34.920 --> 00:45:36.480 So the first issue that,

897 00:45:36.480 --> 00:45:41.070 how do you do the normalization in your process?

898 00:45:41.070 --> 00:45:42.390 <v ->Yeah, great question.</v>

899 00:45:42.390 --> 00:45:44.160 So yeah, I totally glossed over

900 00:45:44.160 --> 00:45:45.610 all the pre-processing stuff.

901 00:45:46.740 --> 00:45:51.090 So before doing the network construction,

902 00:45:51.090 --> 00:45:53.970 I log transformed the protein concentrations

903 00:45:53.970 --> 00:45:55.920 to reduce stiffness.

904 00:45:55.920 --> 00:45:57.720 There was a standardization within,

905 00:45:57.720 --> 00:46:01.590 there were sort of two phases of runs of protein modules,

906 00:46:01.590 --> 00:46:05.700 so I sort of standardized within those batches,

907 00:46:05.700 --> 00:46:10.700 and then after that, I did a rank normalized

908 00:46:11.111 --> 00:46:15.663 or inverse normal rank transformation to sort of-

909 00:46:15.663 --> 00:46:17.036 (audience speaks indistinctly) <v ->What's that?</v>

910 00:46:17.036 --> 00:46:18.600 <v ->(indistinct) normalization?</v> <v ->Basically.</v>

911 00:46:18.600 --> 00:46:20.040 Yeah, yeah, yeah.

912 00:46:20.040 --> 00:46:22.980 So that was sort of the data pre-processing.

913 00:46:22.980 --> 00:46:24.633 So I think I, you know,

914 00:46:25.800 --> 00:46:27.720 I've thought about sort of the pros and cons

915 00:46:27.720 --> 00:46:30.780 of those things as well and I think my biggest qualm

916 00:46:30.780 --> 00:46:34.350 with the way that I did it is sort of interpretability,

917 00:46:34.350 --> 00:46:37.110 because, yeah, sort of what does it mean

918 00:46:37.110 --> 00:46:38.790 to be at one quantile versus another

919 00:46:38.790 --> 00:46:40.440 where you have this huge dynamic range

920 00:46:40.440 --> 00:46:42.330 of protein concentrations?

921 00:46:42.330 --> 00:46:44.100 <v Audience>So another question is that</v>

922 00:46:44.100 --> 00:46:46.230 I know that in your project,

923 00:46:46.230 --> 00:46:48.450 the modules identification is very important.

924 00:46:48.450 --> 00:46:50.883 So I wonder,

925 00:46:53.130 --> 00:46:54.210 you have talked a little bit

926 00:46:54.210 --> 00:46:56.600 about how to answer the modules,

927 00:46:56.600 --> 00:47:00.310 but so can you explain a little bit more

928 00:47:00.310 --> 00:47:05.223 about how you gonna bring modules from the data?

929 00:47:08.250 --> 00:47:10.590 <v ->I'm not sure, can you say a little bit more?</v>

930 00:47:10.590 --> 00:47:13.110 <v Audience>Yeah, so in your previous pages,</v>

931 00:47:13.110 --> 00:47:16.980 I think you talked a little bit about the clustering

932 00:47:16.980 --> 00:47:18.283 of the modules so that we know

933 00:47:18.283 --> 00:47:21.750 that there are four main modules.

934 00:47:21.750 --> 00:47:23.970 <v ->Yes.</v> <v ->In the whole dataset.</v>

935 00:47:23.970 --> 00:47:28.110 So what is the name of that algorithm

936 00:47:28.110 --> 00:47:30.712 and how it basically work?

937 00:47:30.712 --> 00:47:34.600 <v ->Yeah, so the clustering itself was done</v>

938 00:47:35.730 --> 00:47:40.530 using algorithm called H+.

939 00:47:40.530 --> 00:47:42.540 To be honest, I'm not too sure

940 00:47:42.540 --> 00:47:44.610 about sort of the details of it.

941 00:47:44.610 --> 00:47:47.563 It can use any dissimilarity measure,

942 00:47:47.563 --> 00:47:52.350 which, in our case, comes from the TOM matrix, but-

943 00:47:52.350 --> 00:47:55.140 <v Audience>So this is the algorithm that we separate</v>

944 00:47:55.140 --> 00:47:58.123 the whole proteins into four different modules

945 00:47:58.123 --> 00:48:00.330 so that we can analyze it one by one.

946 00:48:00.330 --> 00:48:01.440 <v ->Yeah, yeah, yeah, yeah.</v> <v ->Yeah,</v>

947 00:48:01.440 --> 00:48:05.230 so I also noticed that

948 00:48:07.290 --> 00:48:12.290 in the weighted protein expression network analysis,

949 00:48:13.320 --> 00:48:15.630 you talk about the beta values.

950 00:48:15.630 --> 00:48:17.560 <v ->Yes.</v> <v ->That you use that</v>

951 00:48:19.945 --> 00:48:22.705 like the soft threshold. <v ->Yeah.</v>

952 00:48:22.705 --> 00:48:27.510 <v Audience>To make the genes to be more important</v>

953 00:48:27.510 --> 00:48:31.110 if that is the thing that you wanna analyze.

954 00:48:31.110 --> 00:48:35.220 So in this process, I want to know how you would make sure

955 00:48:35.220 --> 00:48:39.053 the value of the data in this process.

956 00:48:39.053 --> 00:48:41.927 <v ->So sorry, we have to end 'cause it's 12:15.</v>

957 00:48:41.927 --> 00:48:43.830 I know others have classes and everything.

958 00:48:43.830 --> 00:48:45.568 Maybe you guys can discuss a little bit.

959 00:48:45.568 --> 00:48:47.580 <v ->Yeah, (indistinct), yeah.</v> <v ->Maybe if you have time.</v>

960 00:48:47.580 --> 00:48:49.140 Please, if you're registered,

961 00:48:49.140 --> 00:48:51.330 make sure you signed in on a sign in sheet.

962 00:48:51.330 --> 00:48:52.163 There's three of 'em.

963 00:48:52.163 --> 00:48:53.640 You only have to sign on one of them,

964 00:48:53.640 --> 00:48:56.640 and then one-fourth page reflections will be due

965 00:48:56.640 --> 00:48:58.872 before the next speaker's time to speak.

966 00:48:58.872 --> 00:49:02.039 (indistinct talking)