AbSciCon 2019

1
00:00:00,790 --> 00:00:07,320
[Music]

2
00:00:13,150 --> 00:00:09,270
[Applause]

3
00:00:15,520 --> 00:00:13,160
hi I'm olive I'm undergrad at Carleton

4
00:00:18,609 --> 00:00:15,530
College working with Mika Anderson and

5
00:00:20,019 --> 00:00:18,619
my talk is about the silver pen genome

6
00:00:23,589 --> 00:00:20,029
and its evolution in deep sea

7
00:00:25,690 --> 00:00:23,599
hydrothermal vents so I would just like

8
00:00:27,370 --> 00:00:25,700
to reiterate how important it is to

9
00:00:29,589 --> 00:00:27,380
study hydrothermal vents system

10
00:00:32,019 --> 00:00:29,599
hydrothermal vents are one of the most

11
00:00:35,440 --> 00:00:32,029
ancient continuously inhabited ecosystem

12
00:00:37,150 --> 00:00:35,450
on earth and some research researchers

13
00:00:39,700 --> 00:00:37,160

believe that hydrothermal vent was the

14
00:00:41,770 --> 00:00:39,710
place libraries and if that's true them

15
00:00:45,099 --> 00:00:41,780
live diverse into these two lineages as

16
00:00:49,660 --> 00:00:45,109
we know bacteria and archaea and they

17
00:00:51,700 --> 00:00:49,670
invented new metabolic pathways and also

18
00:00:53,439 --> 00:00:51,710
they out the bacteria also and and

19
00:00:55,630 --> 00:00:53,449
everywhere like they also occupy new

20
00:00:57,430 --> 00:00:55,640
niches so that they spread spread out

21
00:01:05,140 --> 00:00:57,440
throughout the entire earth until now

22
00:01:08,200 --> 00:01:05,150
and so along that idea I would like to

23
00:01:10,240 --> 00:01:08,210
study microbial evolution and I think

24
00:01:12,640 --> 00:01:10,250
there is no more interesting place to

25
00:01:14,470 --> 00:01:12,650
study it and then in hydrothermal vent

26
00:01:17,170 --> 00:01:14,480
because this one of the key habitats in

27
00:01:18,880 --> 00:01:17,180
life's earliest stages and I have a few

28
00:01:21,250 --> 00:01:18,890
questions about microbial evolution and

29
00:01:23,200 --> 00:01:21,260
deep sea hydrothermal vents the first

30
00:01:26,860 --> 00:01:23,210
one what is the genome make variation

31
00:01:29,350 --> 00:01:26,870
that exists in this system and then why

32
00:01:33,010 --> 00:01:29,360
what drives this evolution is it

33
00:01:37,540 --> 00:01:33,020
challenge or or drift or is it necessity

34
00:01:41,430 --> 00:01:37,550
which is selection and lastly how do the

35
00:01:44,380 --> 00:01:41,440
microbes in hydrothermal vents diversify

36
00:01:46,720 --> 00:01:44,390
so first most of this talk is going to

37
00:01:48,580 --> 00:01:46,730
be about population genetics and when

38
00:01:51,280 --> 00:01:48,590

I'm thinking about population genetics I

39
00:01:54,280 --> 00:01:51,290
always think of the four phenomena or

40
00:01:56,560 --> 00:01:54,290
evolutionary forces such as selection

41
00:01:58,600 --> 00:01:56,570
which is the necessity mutation

42
00:02:03,520 --> 00:01:58,610
migration drift and which is the chance

43
00:02:05,050 --> 00:02:03,530
and how it creates an in tracks to give

44
00:02:06,300 --> 00:02:05,060
rise to this variation that we see in a

45
00:02:10,410 --> 00:02:06,310
population

46
00:02:13,690 --> 00:02:10,420
unlike in multicellular organisms in

47
00:02:16,000 --> 00:02:13,700
microbes the variation can take a whole

48
00:02:20,289 --> 00:02:16,010
new meaning so not only that their

49
00:02:22,500 --> 00:02:20,299
sequence can be diverse the genomic

50
00:02:26,610 --> 00:02:22,510
content can also vary between this

51
00:02:29,369 --> 00:02:26,620
and a specific species so some genes are

52
00:02:31,649 --> 00:02:29,379
shared by the whole species which is the

53
00:02:34,080 --> 00:02:31,659
core genes and some genes are not which

54
00:02:36,330 --> 00:02:34,090
is the accessory genes and there are

55
00:02:38,820 --> 00:02:36,340
multiple hypotheses how this accessory

56
00:02:41,100 --> 00:02:38,830
genes are acquired including horse no

57
00:02:44,729 --> 00:02:41,110
gene transfer and also large-scale

58
00:02:47,699 --> 00:02:44,739
deletion but I'm more interested in how

59
00:02:51,030 --> 00:02:47,709
this accessory genes are maintained

60
00:02:53,729 --> 00:02:51,040
throughout his evolution so there has

61
00:02:56,640 --> 00:02:53,739
been some literature debate on this

62
00:02:59,520 --> 00:02:56,650
either it is true selection which is the

63
00:03:03,509 --> 00:02:59,530

necessity or some paper also says that

64
00:03:07,080 --> 00:03:03,519
it is true drift which is chance or a

65
00:03:10,740 --> 00:03:07,090
little bit of both and to tie it back to

66
00:03:12,270 --> 00:03:10,750
our extreme engine environment I would

67
00:03:16,140 --> 00:03:12,280
like to study the span genome evolution

68
00:03:18,509 --> 00:03:16,150
in the deep sea a dream event and these

69
00:03:21,449 --> 00:03:18,519
are my two study sites the first one is

70
00:03:23,849 --> 00:03:21,459
the MIT caiman rice which is in the

71
00:03:27,000 --> 00:03:23,859
Caribbean Sea and these samples were

72
00:03:31,050 --> 00:03:27,010
taken in 2012 2013 by Joey Bruce group

73
00:03:33,390 --> 00:03:31,060
and the second one is the actual which

74
00:03:37,110 --> 00:03:33,400
is close to the hoenn de Fuca plate and

75
00:03:40,620 --> 00:03:37,120
these samples were taken in 2013 2015

76
00:03:43,099 --> 00:03:40,630
but also by Julie Hoover's lab so

77
00:03:46,830 --> 00:03:43,109
because a lot of these microbes were on

78
00:03:49,140 --> 00:03:46,840
Uncle Teva bull's eye we turned to

79
00:03:50,699 --> 00:03:49,150
metagenomic sequencing and I followed

80
00:03:53,490 --> 00:03:50,709
like the basic metagenomics against

81
00:03:57,750 --> 00:03:53,500
workflow so that assembly mapping and

82
00:04:00,020 --> 00:03:57,760
all that stuff and finally I did the

83
00:04:04,770 --> 00:04:00,030
pinning which is the interesting part

84
00:04:08,069 --> 00:04:04,780
and basically I've been my context based

85
00:04:12,720 --> 00:04:08,079
on the GC content and coverage mostly

86
00:04:14,640 --> 00:04:12,730
and then I also found like what taxa

87
00:04:17,940 --> 00:04:14,650
they belong to one talk set of pins

88
00:04:19,949 --> 00:04:17,950

belong to and for this purpose I'm

89
00:04:22,950 --> 00:04:19,959
mostly interested in the most abundant

90
00:04:25,560 --> 00:04:22,960
axon or at the genus level that I found

91
00:04:28,020 --> 00:04:25,570
in my samples which is sulfur ovum so

92
00:04:30,659 --> 00:04:28,030
Bravo miss sulfur oxidizing bacteria in

93
00:04:34,080 --> 00:04:30,669
the theatre among fans and from the bin

94
00:04:36,420 --> 00:04:34,090
recover meta silver atom genomes that I

95
00:04:39,540 --> 00:04:36,430
had I created a pan genome profile

96
00:04:44,700 --> 00:04:39,550
which is what you see here so each of

97
00:04:49,740 --> 00:04:44,710
this layer is a software from genome and

98
00:04:51,480 --> 00:04:49,750
then each of this sorry each of this bar

99
00:04:53,490 --> 00:04:51,490
represents like the gin grip so if the

100
00:05:01,620 --> 00:04:53,500
gin group is there then the bar X is on

101
00:05:03,360 --> 00:05:01,630
that layer and vice versa so now I'm

102
00:05:05,580 --> 00:05:03,370
interested in like what kind of genes

103
00:05:09,540 --> 00:05:05,590
there are in the cell phone pen genome

104
00:05:12,090 --> 00:05:09,550
so and also I'm interested in how those

105
00:05:16,560 --> 00:05:12,100
functions are distributed across the

106
00:05:19,620 --> 00:05:16,570
gene frequency so first here that each

107
00:05:22,230 --> 00:05:19,630
data point is the gin grip and the color

108
00:05:24,540 --> 00:05:22,240
bar basically means the gene function

109
00:05:26,610 --> 00:05:24,550
and on the x axis I have the gene

110
00:05:29,250 --> 00:05:26,620
frequency from gene containing only one

111
00:05:32,460 --> 00:05:29,260
genome to the one in 22 genome which is

112
00:05:34,499 --> 00:05:32,470
the core genome basically and on the XY

113
00:05:36,810 --> 00:05:34,509

axis I have the proportion of that gene

114
00:05:39,800 --> 00:05:36,820
function across a column so the most

115
00:05:42,990 --> 00:05:39,810
important trend here is that R is this R

116
00:05:44,700 --> 00:05:43,000
increase in proportion for translation

117
00:05:47,400 --> 00:05:44,710
coenzyme metabolism and amino acid

118
00:05:50,100 --> 00:05:47,410
metabolism functions across the gene

119
00:05:52,890 --> 00:05:50,110
frequency and the takeaway here is that

120
00:05:54,659 --> 00:05:52,900
the housekeeping functions are basically

121
00:05:56,790 --> 00:05:54,669
more enriched in the core genome versus

122
00:06:00,120 --> 00:05:56,800
the accessory genome which makes sense

123
00:06:02,879 --> 00:06:00,130
but the opposite is also true for the

124
00:06:05,760 --> 00:06:02,889
environment related signaling genes such

125
00:06:08,100 --> 00:06:05,770
as signal transduction and so on so here

126
00:06:10,560 --> 00:06:08,110
I like to point out that the accessory

127
00:06:13,649 --> 00:06:10,570
genome acquire acquisition and

128
00:06:16,110 --> 00:06:13,659
maintenance seems to be not random based

129
00:06:17,790 --> 00:06:16,120
on functions and so that this kind of

130
00:06:23,520 --> 00:06:17,800
like points out toward the selection

131
00:06:25,500 --> 00:06:23,530
rather than the chance case then if been

132
00:06:27,120 --> 00:06:25,510
genome evolution is really driven by

133
00:06:29,189 --> 00:06:27,130
selection I would like to know what kind

134
00:06:31,980 --> 00:06:29,199
of selective pressure exists in this

135
00:06:33,600 --> 00:06:31,990
environment and I would also like to

136
00:06:37,350 --> 00:06:33,610
know like if there is any local

137
00:06:39,420 --> 00:06:37,360
adaptation of this pen genomes so here I

138
00:06:41,730 --> 00:06:39,430

realized that there are two environments

139
00:06:45,529 --> 00:06:41,740
that my samples came from the mid chemin

140
00:06:48,540 --> 00:06:45,539
rise vent and the actual vent on so and

141
00:06:49,800 --> 00:06:48,550
they're really separated by the

142
00:06:53,810 --> 00:06:49,810
continent so

143
00:06:56,610 --> 00:06:53,820
separate and so I calculated the

144
00:06:58,620 --> 00:06:56,620
proportion for each unit calculated the

145
00:07:02,010 --> 00:06:58,630
proportion of that being found in only

146
00:07:05,760 --> 00:07:02,020
actual genome and then I sorted them

147
00:07:09,090 --> 00:07:05,770
from lowest to highest and this are

148
00:07:10,860 --> 00:07:09,100
basically the least represented genes in

149
00:07:12,420 --> 00:07:10,870
the actual genome so they're mostly

150
00:07:16,740 --> 00:07:12,430
represented and only made came in rice

151
00:07:19,260 --> 00:07:16,750
genomes and most of just genes belong to

152
00:07:22,620 --> 00:07:19,270
the blue category which is the ion

153
00:07:24,659 --> 00:07:22,630
transport metabolism categories and most

154
00:07:26,400 --> 00:07:24,669
interestingly they're also mostly

155
00:07:29,070 --> 00:07:26,410
related to phosphate uptake and

156
00:07:31,260 --> 00:07:29,080
regulation which means that phosphate

157
00:07:33,930 --> 00:07:31,270
related genes are more represented than

158
00:07:36,960 --> 00:07:33,940
in the mid cameron rice genomes versus

159
00:07:39,510 --> 00:07:36,970
the actual genomes this is interesting

160
00:07:41,909 --> 00:07:39,520
because in the atlantic ocean where meat

161
00:07:44,700 --> 00:07:41,919
came in rice is the phosphate content is

162
00:07:47,610 --> 00:07:44,710
lower than the passive sorry in divisive

163
00:07:49,800 --> 00:07:47,620

than in the pacific ocean to me this

164
00:07:51,900 --> 00:07:49,810
means that microbes that live in mate

165
00:07:54,570 --> 00:07:51,910
caiman rice could potentially have to

166
00:07:57,600 --> 00:07:54,580
innovate due to this phosphate like an

167
00:07:58,890 --> 00:07:57,610
environment by maintaining the accessory

168
00:08:02,190 --> 00:07:58,900
genes that they got through horizontal

169
00:08:04,170 --> 00:08:02,200
gene transfer and this result is

170
00:08:08,730 --> 00:08:04,180
actually pretty similar to what more in

171
00:08:12,150 --> 00:08:08,740
common found and the prochlorococcus in

172
00:08:15,240 --> 00:08:12,160
the surface ocean and just like i would

173
00:08:18,500 --> 00:08:15,250
just like to try out throw it out there

174
00:08:21,779 --> 00:08:18,510
because i also found a lot of arsenate

175
00:08:25,170 --> 00:08:21,789
related genes in the mid commander eyes

176
00:08:27,810 --> 00:08:25,180
compared compared to actual genomes and

177
00:08:34,490 --> 00:08:27,820
this is also about the prochlorococcus

178
00:08:37,920 --> 00:08:34,500
paper found in the surface ocean and i

179
00:08:40,140 --> 00:08:37,930
we also looked at the PNP s ratio which

180
00:08:43,920 --> 00:08:40,150
kind of like suggests the strength of

181
00:08:46,200 --> 00:08:43,930
evolution on each gene I looked at um so

182
00:08:48,540 --> 00:08:46,210
basically if the P NP r--'s ratio is

183
00:08:51,240 --> 00:08:48,550
higher than 1 then just somewhat

184
00:08:54,990 --> 00:08:51,250
adaptive evolution or positive evolution

185
00:08:57,840 --> 00:08:55,000
and if the Pampas ratio is closer to 0

186
00:09:00,120 --> 00:08:57,850
then negative selection or purifying

187
00:09:03,010 --> 00:09:00,130
selection which is more of conservation

188
00:09:06,400 --> 00:09:03,020

then change happens on that gene

189
00:09:08,980 --> 00:09:06,410
what I found here is that accessory

190
00:09:11,650 --> 00:09:08,990
genes tend to have higher bnps ratios

191
00:09:13,780 --> 00:09:11,660
than the core genes and some of these

192
00:09:17,140 --> 00:09:13,790
genes also have been passed ratio higher

193
00:09:17,860 --> 00:09:17,150
than one but I'm not really entirely

194
00:09:21,130 --> 00:09:17,870
sure

195
00:09:23,140 --> 00:09:21,140
statistically here because this were not

196
00:09:25,870 --> 00:09:23,150
you I didn't really calculate defense

197
00:09:28,870 --> 00:09:25,880
base ratios using a maximum likelihood

198
00:09:32,440 --> 00:09:28,880
modal but just like kind of like point

199
00:09:35,350 --> 00:09:32,450
estimate for each gym so like this

200
00:09:38,020 --> 00:09:35,360
higher bnps ratio for accessory gene

201
00:09:40,420 --> 00:09:38,030
might be due to either adaptive

202
00:09:44,100 --> 00:09:40,430
evolution on these genes or that that

203
00:09:46,000 --> 00:09:44,110
this genes are less likely to undergo

204
00:09:53,410 --> 00:09:46,010
negative selection or purifying

205
00:09:55,030 --> 00:09:53,420
selection so in conclusion we could have

206
00:09:58,630 --> 00:09:55,040
saw that man genome evolution is

207
00:10:00,520 --> 00:09:58,640
selective and we had several evidence of

208
00:10:02,830 --> 00:10:00,530
this the first one is that different

209
00:10:05,230 --> 00:10:02,840
gene categories were enriched in core

210
00:10:08,050 --> 00:10:05,240
versus accessory genomes which was the

211
00:10:10,300 --> 00:10:08,060
first one and we also saw some local

212
00:10:12,610 --> 00:10:10,310
adaptation potentially due to phosphate

213
00:10:15,870 --> 00:10:12,620

on content difference between the mid

214
00:10:18,970 --> 00:10:15,880
came in rice in the actual environment

215
00:10:21,070 --> 00:10:18,980
and then we also saw that there was

216
00:10:23,530 --> 00:10:21,080
higher probability of adaptive evolution

217
00:10:26,290 --> 00:10:23,540
in the accessory genome compared to the

218
00:10:28,590 --> 00:10:26,300
core genome or that there was some

219
00:10:31,300 --> 00:10:28,600
different evolutionary scheme of

220
00:10:34,300 --> 00:10:31,310
accessory genome compared to the court

221
00:10:36,580 --> 00:10:34,310
genome and finally if though I didn't

222
00:10:38,950 --> 00:10:36,590
really mention it here we saw some

223
00:10:42,130 --> 00:10:38,960
evidence of gene specific sweeps which

224
00:10:46,480 --> 00:10:42,140
kind of points a selection that happens

225
00:10:49,990 --> 00:10:46,490
in these microbial populations and some

226
00:10:53,080 --> 00:10:50,000
more of bigger-picture conclusions we

227
00:10:56,920 --> 00:10:53,090
saw that necessity was the key factor

228
00:10:59,980 --> 00:10:56,930
and pan genome evolution in hydrothermal

229
00:11:01,750 --> 00:10:59,990
vent compactive chance and third is

230
00:11:04,960 --> 00:11:01,760
still an open question how important

231
00:11:06,580 --> 00:11:04,970
necessary was in early life the genomes

232
00:11:09,160 --> 00:11:06,590
of easy today have been molded by

233
00:11:12,790 --> 00:11:09,170
evolution from geum's back then so we

234
00:11:15,220 --> 00:11:12,800
would infer some connection through that

235
00:11:16,870 --> 00:11:15,230
and finally I would also point out that

236
00:11:18,370 --> 00:11:16,880
when genome variation

237
00:11:22,840 --> 00:11:18,380
the importance of study pen genomic

238
00:11:25,329 --> 00:11:22,850

variation in addition to just single

239
00:11:27,340 --> 00:11:25,339
point polymorphisms because pen genomic

240
00:11:30,249 --> 00:11:27,350
variation is really widespread and it

241
00:11:31,780 --> 00:11:30,259
also takes into account the one

242
00:11:34,389 --> 00:11:31,790
evolutionary force which this original

243
00:11:36,280 --> 00:11:34,399
gene transfer that is not really taken

244
00:11:40,030 --> 00:11:36,290
into account by just single point

245
00:11:42,550 --> 00:11:40,040
polymorphism and by that I would like to

246
00:11:45,610 --> 00:11:42,560
thank the Andersen lab at Carleton and

247
00:11:53,829 --> 00:11:45,620
all the crews that did the same thing

248
00:11:56,740 --> 00:11:53,839
and the funding thank you okay do we

249
00:12:43,980 --> 00:11:56,750
have any questions for all of you can

250
00:12:52,329 --> 00:12:50,139
yeah so yes so the question was the

251
00:12:55,720 --> 00:12:52,339
connection livened extremo file and

252
00:12:58,210 --> 00:12:55,730
genome evolution part yeah like when I

253
00:13:00,610 --> 00:12:58,220
was getting to the project I didn't

254
00:13:02,740 --> 00:13:00,620
really care about extra part because I

255
00:13:05,319 --> 00:13:02,750
was really looking into the evolution of

256
00:13:08,379 --> 00:13:05,329
pan genomes and so it wasn't necessarily

257
00:13:11,170 --> 00:13:08,389
like into just like specific to a pen

258
00:13:13,650 --> 00:13:11,180
genome and I I realized that like

259
00:13:16,749 --> 00:13:13,660
studying XML files not probably the best

260
00:13:19,059 --> 00:13:16,759
like place to do like you know fusion

261
00:13:27,350 --> 00:13:19,069
study but yeah that was the data that I

262
00:13:34,050 --> 00:13:30,420
great talk I'm I'm all I'm a postdoc at

263
00:13:36,390 --> 00:13:34,060

Ames drift the chance drifts is usually

264
00:13:39,150 --> 00:13:36,400
a pretty strong function of population

265
00:13:41,550 --> 00:13:39,160
sides yeah I wonder if there's a way for

266
00:13:47,490 --> 00:13:41,560
in your data set to estimate population

267
00:13:55,380 --> 00:13:47,500
size yeah so I so first we kind of like

268
00:13:58,110 --> 00:13:55,390
had some like the time series data set

269
00:14:00,800 --> 00:13:58,120
as well and we kind of like I guess like

270
00:14:05,010 --> 00:14:00,810
I saw that there was like a decrease of

271
00:14:09,480 --> 00:14:05,020
like the coverage from year to year

272
00:14:12,390 --> 00:14:09,490
but I don't really know like the the was

273
00:14:14,220 --> 00:14:12,400
it called the absolute population size

274
00:14:16,590 --> 00:14:14,230
or like the aphid population side for

275
00:14:18,290 --> 00:14:16,600
the populations that I have so yeah

276
00:14:21,450 --> 00:14:18,300
probably

277
00:14:26,520 --> 00:14:21,460
Dhar like some ways to estimate the

278
00:14:30,000 --> 00:14:26,530
using just coverage but yeah sure thank

279
00:14:30,620 --> 00:14:30,010
you very much thank you everyone for