Duke UNIVERSITY

AbGradCon 2021
Sept. 14-17

A comparative genomic analysis of the 250-million-year-old halophilic archaeon *Halosimplex carlsbadense* with related species

Avery Fulford, Dr. Amy Schmid
Department of Biology
Duke University

avery.fulford@duke.edu

1
00:00:04,789 --> 00:00:02,950
hello my name is avery fulford and i

2
00:00:06,869 --> 00:00:04,799
recently graduated with a bachelor of

3
00:00:08,870 --> 00:00:06,879
science from duke university

4
00:00:10,950 --> 00:00:08,880
i'm currently working at nasa with dr

5
00:00:13,749 --> 00:00:10,960
heather graham before i apply to phd

6
00:00:15,669 --> 00:00:13,759
programs in microbiology this fall

7
00:00:17,430 --> 00:00:15,679
today i'll be discussing my senior

8
00:00:19,590 --> 00:00:17,440
thesis entitled

9
00:00:20,790 --> 00:00:19,600
a comparative genomic analysis of the

10
00:00:23,269 --> 00:00:20,800
250

11
00:00:25,509 --> 00:00:23,279
million year old halophilic archaeon

12
00:00:26,870 --> 00:00:25,519
halo simplex carlsbadency with related

13
00:00:29,109 --> 00:00:26,880

species

14
00:00:31,750 --> 00:00:29,119
my mentor for this project is dr amy

15
00:00:34,549 --> 00:00:31,760
schmidt

16
00:00:36,870 --> 00:00:34,559
so in the year 2000 microbiologist

17
00:00:38,470 --> 00:00:36,880
vreeland at all extracted salt crystals

18
00:00:41,110 --> 00:00:38,480
from a halite formation

19
00:00:43,270 --> 00:00:41,120
deep underground in new mexico these

20
00:00:44,630 --> 00:00:43,280
salt crystals contained pockets of fluid

21
00:00:46,790 --> 00:00:44,640
which are believed to be

22
00:00:50,310 --> 00:00:46,800
the rare liquid remnants of a permian

23
00:00:51,510 --> 00:00:50,320
ocean which evaporated 250 million years

24
00:00:53,590 --> 00:00:51,520
ago

25
00:00:55,029 --> 00:00:53,600
as seen in figure 1 these halite

26
00:00:57,830 --> 00:00:55,039
crystals are pictured

27
00:00:59,830 --> 00:00:57,840
and also the fluid inclusions are marked

28
00:01:01,670 --> 00:00:59,840
with the letter i

29
00:01:03,110 --> 00:01:01,680
using extensive methods to prevent

30
00:01:05,429 --> 00:01:03,120
sample contamination

31
00:01:06,710 --> 00:01:05,439
vreeland at all swabbed petri dishes

32
00:01:08,950 --> 00:01:06,720
with these brines

33
00:01:10,950 --> 00:01:08,960
and found that four microorganisms which

34
00:01:13,350 --> 00:01:10,960
were living within the brines were able

35
00:01:15,109 --> 00:01:13,360
to grow and form colonies

36
00:01:17,670 --> 00:01:15,119
because the microbes lived in brines

37
00:01:18,870 --> 00:01:17,680
enclosed in salt crystals for these 250

38
00:01:20,710 --> 00:01:18,880

million years

39
00:01:23,429 --> 00:01:20,720
they are quite possibly the longest

40
00:01:25,190 --> 00:01:23,439
living organisms study to date

41
00:01:27,749 --> 00:01:25,200
while there has been some controversy

42
00:01:30,789 --> 00:01:27,759
surrounding one of the bacterial samples

43
00:01:33,109 --> 00:01:30,799
called sample 293 another microbe of the

44
00:01:35,990 --> 00:01:33,119
four extracted has been widely accepted

45
00:01:38,390 --> 00:01:36,000
as a new species and genus named halo

46
00:01:42,149 --> 00:01:38,400
simplex carlsbadency

47
00:01:44,789 --> 00:01:42,159
this organism is a halophilic archaeon

48
00:01:46,870 --> 00:01:44,799
so to define these terms a halophila is

49
00:01:47,910 --> 00:01:46,880
an organism adapted to highly saline

50
00:01:50,230 --> 00:01:47,920
conditions

51
00:01:51,590 --> 00:01:50,240
and an archaeon is a microorganism in

52
00:01:54,389 --> 00:01:51,600
the third domain of life

53
00:01:56,789 --> 00:01:54,399
the archaea as opposed to the bacteria

54
00:01:58,709 --> 00:01:56,799
or the eukaryota

55
00:02:01,190 --> 00:01:58,719
the overall goal of this study was to

56
00:02:03,270 --> 00:02:01,200
identify the physiological adaptations

57
00:02:05,109 --> 00:02:03,280
that enabled h-crossbarency

58
00:02:06,389 --> 00:02:05,119
to survive in the harsh environment of

59
00:02:09,430 --> 00:02:06,399
the salt crystal

60
00:02:11,589 --> 00:02:09,440
over 250 million years

61
00:02:12,710 --> 00:02:11,599
we achieved this by aligning its genome

62
00:02:15,350 --> 00:02:12,720
to the genomes of

63
00:02:16,550 --> 00:02:15,360

17 other halophiles to identify

64
00:02:18,550 --> 00:02:16,560
homologous genes

65
00:02:21,830 --> 00:02:18,560
defined as genes inherited by two

66
00:02:23,910 --> 00:02:21,840
species from their common

67
00:02:25,670 --> 00:02:23,920
homologous genes that are highly similar

68
00:02:28,229 --> 00:02:25,680
between two different species

69
00:02:29,670 --> 00:02:28,239
are said to be highly conserved which

70
00:02:34,150 --> 00:02:29,680
suggests that they encode an

71
00:02:37,430 --> 00:02:35,750
to move on to a description of our

72
00:02:38,869 --> 00:02:37,440
methodology we used several

73
00:02:40,949 --> 00:02:38,879
computational tools

74
00:02:41,910 --> 00:02:40,959
to align and compare the genome of h

75
00:02:43,589 --> 00:02:41,920
cosby density

76
00:02:45,030 --> 00:02:43,599
to the genomes of three groups of

77
00:02:47,110 --> 00:02:45,040
organisms

78
00:02:48,869 --> 00:02:47,120
the first comparison group included two

79
00:02:50,550 --> 00:02:48,879
recently discovered species within the

80
00:02:53,509 --> 00:02:50,560
halo simplex genus

81
00:02:55,750 --> 00:02:53,519
h rubrum and h pelagicum the second

82
00:02:58,070 --> 00:02:55,760
comparison group was composed of the 13

83
00:03:00,630 --> 00:02:58,080
most closely related species of archaea

84
00:03:02,869 --> 00:03:00,640
to h crossbedensi which were identified

85
00:03:05,030 --> 00:03:02,879
in a phylogenetic tree created by becker

86
00:03:07,110 --> 00:03:05,040
at all in 2014

87
00:03:09,350 --> 00:03:07,120
this tree is replicated at left in

88
00:03:11,750 --> 00:03:09,360

figure 2 in which you can see h cars by

89
00:03:13,670 --> 00:03:11,760
density highlighted in yellow

90
00:03:15,910 --> 00:03:13,680
each node or fork in the branches

91
00:03:16,550 --> 00:03:15,920
represents the last common ancestor

92
00:03:18,149 --> 00:03:16,560
shared by

93
00:03:21,110 --> 00:03:18,159
two groups of species before they

94
00:03:22,470 --> 00:03:21,120
diverged confidence values for each node

95
00:03:24,229 --> 00:03:22,480
are written below

96
00:03:26,149 --> 00:03:24,239
these are called bootstrap values which

97
00:03:27,190 --> 00:03:26,159
indicate how many times out of 100

98
00:03:28,630 --> 00:03:27,200
repetitions

99
00:03:30,789 --> 00:03:28,640
the same node was recorded when

100
00:03:33,910 --> 00:03:30,799
repeating the phylogenetic construction

101
00:03:35,589 --> 00:03:33,920
on a re-sampled set of the same data

102
00:03:37,350 --> 00:03:35,599
the third comparison group we chose

103
00:03:39,350 --> 00:03:37,360
included sample 293

104
00:03:41,110 --> 00:03:39,360
a bacterium which is another of the four

105
00:03:44,070 --> 00:03:41,120
samples extracted by relin

106
00:03:46,229 --> 00:03:44,080
at all in 2000 we assembled its genome

107
00:03:49,990 --> 00:03:46,239
from short nucleotide sequences

108
00:03:52,070 --> 00:03:50,000
using spades software

109
00:03:53,350 --> 00:03:52,080
okay for my results section i'm first

110
00:03:55,030 --> 00:03:53,360
going to talk about the genomic

111
00:03:57,350 --> 00:03:55,040
alignment and comparison

112
00:03:59,030 --> 00:03:57,360
of h cos badensi and the two other halo

113
00:04:00,630 --> 00:03:59,040

simplex species

114
00:04:03,350 --> 00:04:00,640
in this alignment we found that there

115
00:04:05,190 --> 00:04:03,360
was one long homologous sequence which

116
00:04:07,350 --> 00:04:05,200
was highly conserved among all three

117
00:04:09,110 --> 00:04:07,360
halo simplex species

118
00:04:11,750 --> 00:04:09,120
this was much longer than the other

119
00:04:14,710 --> 00:04:11,760
homologous sequences identified

120
00:04:16,710 --> 00:04:14,720
this long homologous region encodes 70

121
00:04:19,430 --> 00:04:16,720
labeled genes of known function

122
00:04:22,230 --> 00:04:19,440
as well as 93 hypothetical genes whose

123
00:04:24,230 --> 00:04:22,240
function has not been categorized yet

124
00:04:26,390 --> 00:04:24,240
we grouped the 70 labeled genes into

125
00:04:28,469 --> 00:04:26,400
categories based on gene function

126
00:04:29,510 --> 00:04:28,479
and compared the number of genes in each

127
00:04:31,990 --> 00:04:29,520
functional group

128
00:04:33,350 --> 00:04:32,000
in the longest homologous region between

129
00:04:34,629 --> 00:04:33,360
halo simplex species

130
00:04:36,870 --> 00:04:34,639
to the number of genes in each

131
00:04:39,430 --> 00:04:36,880
functional group in the entire h

132
00:04:40,870 --> 00:04:39,440
carlsbad c genome this can be seen in

133
00:04:43,030 --> 00:04:40,880
figure 3.

134
00:04:44,870 --> 00:04:43,040
in other words the green bars represent

135
00:04:46,870 --> 00:04:44,880
the percentage of all genes in the h

136
00:04:50,230 --> 00:04:46,880
cosmetic genome which are involved

137
00:04:52,710 --> 00:04:50,240
in stress response nucleotides so

138
00:04:55,590 --> 00:04:52,720

nucleotide synthesis dna repair and

139
00:04:58,230 --> 00:04:55,600
sometimes dna replication etc

140
00:05:00,150 --> 00:04:58,240
and membrane transport the blue bars

141
00:05:02,230 --> 00:05:00,160
represent the percentage of all genes in

142
00:05:03,350 --> 00:05:02,240
the longest homologous region among halo

143
00:05:05,029 --> 00:05:03,360
simplex species

144
00:05:06,469 --> 00:05:05,039
which are involved in the same three

145
00:05:08,310 --> 00:05:06,479
categories

146
00:05:09,990 --> 00:05:08,320
while there are many more variables we

147
00:05:11,350 --> 00:05:10,000
studied i've included the three with

148
00:05:13,029 --> 00:05:11,360
the greatest difference between the

149
00:05:14,390 --> 00:05:13,039
percentage of genes involved in each

150
00:05:17,029 --> 00:05:14,400
functional category

151
00:05:19,430 --> 00:05:17,039
in the entire halo simplex genome and

152
00:05:21,670 --> 00:05:19,440
the longest homologous region

153
00:05:23,350 --> 00:05:21,680
to summarize figure 2 suggests that the

154
00:05:24,950 --> 00:05:23,360
longest homologous region between the

155
00:05:27,029 --> 00:05:24,960
halo simplex species

156
00:05:28,070 --> 00:05:27,039
is enriched in stress response

157
00:05:30,390 --> 00:05:28,080
nucleotides

158
00:05:31,909 --> 00:05:30,400
and membrane transport mechanisms which

159
00:05:33,510 --> 00:05:31,919
could mean that these functions are

160
00:05:35,749 --> 00:05:33,520
important for survival

161
00:05:37,670 --> 00:05:35,759
in hypersaline conditions this is

162
00:05:42,150 --> 00:05:37,680
because these organisms are highly

163
00:05:45,670 --> 00:05:43,830

now i'm going to address the second

164
00:05:46,550 --> 00:05:45,680
genomic alignment and comparison we

165
00:05:49,189 --> 00:05:46,560
performed

166
00:05:51,350 --> 00:05:49,199
between h carlsbademc and its 13 most

167
00:05:52,629 --> 00:05:51,360
closely related species of halophilic

168
00:05:55,110 --> 00:05:52,639
archaea

169
00:05:57,029 --> 00:05:55,120
this alignment pinpointed 32 homologous

170
00:05:59,110 --> 00:05:57,039
genes within 20 clusters

171
00:06:01,590 --> 00:05:59,120
which encode highly conserved proteins

172
00:06:04,870 --> 00:06:01,600
involved in oxidative stress response

173
00:06:06,550 --> 00:06:04,880
dna replication and respiration

174
00:06:08,390 --> 00:06:06,560
oxidative stress is caused by

175
00:06:10,070 --> 00:06:08,400
hypersaline conditions

176
00:06:11,590 --> 00:06:10,080
it's defined as a disturbance in the

177
00:06:12,309 --> 00:06:11,600
balance between the production of

178
00:06:14,390 --> 00:06:12,319
reactive

179
00:06:16,790 --> 00:06:14,400
oxygen radicals and the cell's

180
00:06:18,629 --> 00:06:16,800
antioxidant defenses

181
00:06:21,110 --> 00:06:18,639
now for the sake of time i'm not going

182
00:06:22,230 --> 00:06:21,120
to elaborate on all 32 genes involved in

183
00:06:25,029 --> 00:06:22,240
this alignment

184
00:06:27,189 --> 00:06:25,039
however two interesting genes are long b

185
00:06:28,790 --> 00:06:27,199
and sophie because of their function

186
00:06:31,029 --> 00:06:28,800
which is highly specialized in

187
00:06:33,670 --> 00:06:31,039
halophilic archaea

188
00:06:34,230 --> 00:06:33,680

lawn b removes misfolded proteins from

189
00:06:35,990 --> 00:06:34,240
cells

190
00:06:37,749 --> 00:06:36,000
which are damaged in hypersaline

191
00:06:41,189 --> 00:06:37,759
conditions due to

192
00:06:44,230 --> 00:06:41,199
oxidative stress suffie repairs

193
00:06:46,390 --> 00:06:44,240
cellular iron sulfur clusters which are

194
00:06:48,309 --> 00:06:46,400
also damaged by the oxidative stress of

195
00:06:50,309 --> 00:06:48,319
highly saline brines

196
00:06:52,469 --> 00:06:50,319
these iron sulfur clusters are

197
00:06:53,670 --> 00:06:52,479
incredibly important for the cell

198
00:06:55,909 --> 00:06:53,680
because they are required for the

199
00:06:57,589 --> 00:06:55,919
function of proteins involved in a wide

200
00:06:59,270 --> 00:06:57,599
range of activities

201
00:07:01,510 --> 00:06:59,280
such as electron transport in

202
00:07:04,150 --> 00:07:01,520
respiratory chain complexes

203
00:07:04,629 --> 00:07:04,160
regulatory sensing and most importantly

204
00:07:08,469 --> 00:07:04,639
for this

205
00:07:10,550 --> 00:07:08,479
analysis dna repair

206
00:07:11,670 --> 00:07:10,560
finally the third genomic comparison we

207
00:07:14,790 --> 00:07:11,680
performed was between

208
00:07:16,790 --> 00:07:14,800
hrasbadensi and bacterial sample 293

209
00:07:18,550 --> 00:07:16,800
which again was another of the four

210
00:07:20,790 --> 00:07:18,560
ancient samples extracted

211
00:07:22,309 --> 00:07:20,800
from salt crystals by relent at all in

212
00:07:26,790 --> 00:07:22,319
2000.

213
00:07:27,270 --> 00:07:26,800

surprisingly we found a 2545 base pair

214
00:07:29,749 --> 00:07:27,280
long

215
00:07:30,629 --> 00:07:29,759
homologous region between the genomes as

216
00:07:33,670 --> 00:07:30,639
well as seven

217
00:07:35,670 --> 00:07:33,680
isolated homologous genes this is a very

218
00:07:37,510 --> 00:07:35,680
high number of homologous regions for

219
00:07:38,070 --> 00:07:37,520
two species from separate domains of

220
00:07:40,790 --> 00:07:38,080
life

221
00:07:43,029 --> 00:07:40,800
the archaea and the bacteria this is

222
00:07:45,029 --> 00:07:43,039
suggestive of horizontal gene transfer

223
00:07:46,469 --> 00:07:45,039
which is the transfer of dna from one

224
00:07:49,670 --> 00:07:46,479
organism to another

225
00:07:51,270 --> 00:07:49,680
without inheritance from a shared parent

226
00:07:53,270 --> 00:07:51,280
it has been demonstrated before in the

227
00:07:55,270 --> 00:07:53,280
literature that hypersaline environments

228
00:07:58,309 --> 00:07:55,280
make some species more likely

229
00:08:00,309 --> 00:07:58,319
to accept dna from other organisms even

230
00:08:01,270 --> 00:08:00,319
if that organism is from another domain

231
00:08:02,950 --> 00:08:01,280
of life

232
00:08:05,189 --> 00:08:02,960
this is because of the broad scale

233
00:08:08,150 --> 00:08:05,199
competency exhibited by stressed

234
00:08:10,710 --> 00:08:08,160
cells the homologous genes between h

235
00:08:11,670 --> 00:08:10,720
crossbodency and sample 293 are

236
00:08:14,390 --> 00:08:11,680
represented

237
00:08:16,390 --> 00:08:14,400
though not to scale in figure four the

238
00:08:18,150 --> 00:08:16,400

blue lines in this figure represent the

239
00:08:19,749 --> 00:08:18,160
genomes of the two species

240
00:08:21,749 --> 00:08:19,759
and the colored boxes indicate

241
00:08:23,510 --> 00:08:21,759
homologous regions which are connected

242
00:08:25,430 --> 00:08:23,520
between the genomes by a line of the

243
00:08:27,350 --> 00:08:25,440
same color

244
00:08:29,749 --> 00:08:27,360
some homologous genes of interest are

245
00:08:30,710 --> 00:08:29,759
the five dna repair proteins in one

246
00:08:33,670 --> 00:08:30,720
cluster

247
00:08:34,469 --> 00:08:33,680
as shown in the teal box there's also

248
00:08:38,469 --> 00:08:34,479
the abc

249
00:08:40,550 --> 00:08:38,479
transporter glts and duf domain protein

250
00:08:42,149 --> 00:08:40,560
which all encode inter membrane

251
00:08:45,590 --> 00:08:42,159
transporters

252
00:08:50,470 --> 00:08:45,600
there's also ace f and pep f genes

253
00:08:54,630 --> 00:08:52,710
in conclusion gene conservation across

254
00:08:57,110 --> 00:08:54,640
this halo simplex clade

255
00:08:59,190 --> 00:08:57,120
closely related halophilic archaea and

256
00:09:01,590 --> 00:08:59,200
the halophilic sample 293

257
00:09:02,550 --> 00:09:01,600
suggests that dna repair membrane

258
00:09:05,030 --> 00:09:02,560
transport

259
00:09:06,070 --> 00:09:05,040
and oxidative stress response mechanisms

260
00:09:08,389 --> 00:09:06,080
are required

261
00:09:09,430 --> 00:09:08,399
for the survival of organisms entrapped

262
00:09:12,630 --> 00:09:09,440
in hypersaline

263
00:09:14,870 --> 00:09:12,640

conditions over long periods of time

264
00:09:16,630 --> 00:09:14,880
dna repair for instance can combat the

265
00:09:18,070 --> 00:09:16,640
damage that hypersailing conditions

266
00:09:20,150 --> 00:09:18,080
wreak on dna

267
00:09:22,949 --> 00:09:20,160
particularly through the rearrangement

268
00:09:25,110 --> 00:09:22,959
of bonds in adjacent pyrimidines

269
00:09:27,430 --> 00:09:25,120
membrane transport proteins such as

270
00:09:29,350 --> 00:09:27,440
potassium and sodium transporters

271
00:09:31,350 --> 00:09:29,360
protect against salty hypertonic

272
00:09:33,350 --> 00:09:31,360
conditions which can cause

273
00:09:35,110 --> 00:09:33,360
a cell to shrivel up and die through

274
00:09:37,269 --> 00:09:35,120
desiccation

275
00:09:39,829 --> 00:09:37,279
finally oxidative stress response

276
00:09:43,110 --> 00:09:39,839
mechanisms prevent protein and membrane

277
00:09:47,990 --> 00:09:45,509
we've identified three important avenues

278
00:09:49,750 --> 00:09:48,000
for future research on this topic

279
00:09:52,310 --> 00:09:49,760
the first is the identification of

280
00:09:54,710 --> 00:09:52,320
genomic sequences in h carlsbadency

281
00:09:56,230 --> 00:09:54,720
that did not align to related species

282
00:09:57,670 --> 00:09:56,240
because genes that are unique to this

283
00:10:01,269 --> 00:09:57,680
ancient organism

284
00:10:03,190 --> 00:10:01,279
could encode novel physiological traits

285
00:10:05,990 --> 00:10:03,200
future research could also study two

286
00:10:08,389 --> 00:10:06,000
hypothetical proteins that we identified

287
00:10:11,350 --> 00:10:08,399
on in our alignment of the 14 closely

288
00:10:13,750 --> 00:10:11,360

related species of halophilic archaea

289
00:10:15,829 --> 00:10:13,760
it is rare for a hypothetical protein to

290
00:10:16,550 --> 00:10:15,839
be so highly conserved amongst a large

291
00:10:18,389 --> 00:10:16,560
group

292
00:10:20,230 --> 00:10:18,399
which suggests that these proteins could

293
00:10:21,030 --> 00:10:20,240
encode an important physiological

294
00:10:24,630 --> 00:10:21,040
function

295
00:10:26,550 --> 00:10:24,640
that is not yet characterized finally

296
00:10:29,110 --> 00:10:26,560
future researchers could investigate the

297
00:10:32,150 --> 00:10:29,120
possibility of horizontal gene transfer

298
00:10:34,949 --> 00:10:32,160
between h crossbodency and sample 293

299
00:10:36,550 --> 00:10:34,959
using two methods of species specific

300
00:10:39,509 --> 00:10:36,560
genomic metrics

301
00:10:40,069 --> 00:10:39,519
these are species-specific codon use and

302
00:10:43,350 --> 00:10:40,079
gc

303
00:10:48,389 --> 00:10:45,750
to conclude i'd like to thank dr amy

304
00:10:50,389 --> 00:10:48,399
schmid and riley hackley my advisors

305
00:10:52,389 --> 00:10:50,399
as well as my honors thesis faculty

306
00:10:53,990 --> 00:10:52,399
reader dr tom mitchell olds

307
00:10:56,550 --> 00:10:54,000
for their invaluable feedback and

308
00:10:57,350 --> 00:10:56,560
support i'd also like to thank dr lauren

309
00:10:59,350 --> 00:10:57,360
luger

310
00:11:02,710 --> 00:10:59,360
and joey prinz and thank you for

311
00:11:05,910 --> 00:11:04,870
if you have any questions feel free to

312
00:11:09,350 --> 00:11:05,920
email me