syn-genome_DataAnalysis

Larsen ML

19 August, 2019

```
# setwd("C:/Users/meglarse/GitHub/larsen-dissertation/analyses/chpt3_Genetic/")

#Load Packages
require(phylobase);require(picante)
```

```
## Loading required package: phylobase
```

```
## Loading required package: picante
```

```
## Loading required package: ape
```

```
##
## Attaching package: 'ape'
```

```
## The following object is masked from 'package:phylobase':
##
##     edges
```

```
## Loading required package: vegan
```

```
## Loading required package: permute
```

```
## Loading required package: lattice
```

```
## This is vegan 2.5-5
```

```
## Loading required package: nlme
```

```
require(ape)
```

#Project Summary

This project was designed to examine the impact of nutrient limitation on genomic adaptation to phage predation.
Specifically, I evaluated mutational biases including the number and type of mutation, genomic location of each mutation as well as other metrics including the ratio of transversions to transitions and the ration of nonsynonymous to synonymous mutations.

All statistical analyses for this project were performed with R version 3.6.1 (2019-07-05).

** COllaborators**

JE Barrick, The University of Texas at Austin **JT Lennon**, Indiana University

Project Questions 1. What are the genome mutations conferring resistance to phage infection?
- How many mutations are there per strain? in/del, SNP, S vs NS
- What are the mutation observed in each strain?
2. Are all completely resistant phenotypes genetically similar?
3. Does nutrient limitation affect observed mutations in candidate resistance genes?

Data collection Data for this analysis was collected from whole genome sequences of senstivive and resistant *Synechococcus* strains. Short read genomic sequences from both the Illumina Hi-Seq and Mi-Seq platforms were scaffolded onto an NCBI *Synechococcus* WH7803 reference strain using the computational program *breseq*. Previous analyses were completed in Excel.

Contents

Mutational analysis

What are the genome mutations conferring resistance to phage infection?

```
# read in selected data files for analysis
breseq <- read.csv("./data/breseq.compare.csv", header = T)      #breseq produced compare file

# Prior to this point, the mutation annotation data has been moved to a separate file.

breseq = breseq[,-c(2,39,40,41)]

# restructure data for analysis
# rename the strains
N <- 6
strains = as.matrix(breseq[1,-1])
strain.names <- sapply(strsplit(strains, ""), function(x, n) paste(head(x, n), collapse = ""), N)


# Remove unnessary rows
breseq <- breseq[-which(breseq[,1]=="position"),]

# Convert matrix into binary data file.
# Predicted mutation goes from 100% to 1
# empty cell to 0
# "?" to an NA suggesting that there isn't enough information to predict a mutation
breseq <- as.matrix(breseq)
breseq[breseq == "100%"] <- 1
breseq[breseq == "?"] <- 0
breseq[breseq == ""] <- 0

#rename the columns and rows
breseq <- breseq[-1,]

rownames(breseq) <- breseq[,1]
breseq <- breseq[,-1]
colnames(breseq) = strain.names

# remove rRNA hits
#breseq <- breseq[-c(which(rownames(breseq) == "2,019,461"):which(rownames(breseq) == "2,020,083"),

# Remove rows with no predicted mutations
class(breseq) = "numeric"
breseq <- breseq[-c(which(rowSums(breseq)==35)),]
#rowSums(breseq)

# separate the mutation data
mut.num <- seq(1,nrow(breseq),1)

# add mutation numbers and remove from the breseq data file
```

```
breseq.mutno <- cbind(mut.num,rownames(breseq))

# transform the matrix to get strains in the rows
breseq <- t(breseq)
breseq.data <- breseq[-c(which(rowSums(breseq)==0)),]

print(rownames(breseq[which(rowSums(breseq)==0),]))
```

```
## NULL
```

```
#merge the meta data
#breseq <- merge(strains, breseq, by = "strain.name", all = T)
```

```
strains <- read.csv("./data/strains.csv",header = T)

breseq.all <- as.data.frame(merge(strains,breseq, by.x = "strain_ID",by.y = 0))

breseq.minus <- breseq.all[breseq.all$trt=="Ph-",]
breseq.minus <- breseq.minus[,-c(which(colSums(breseq.minus[,6:117]) ==0))]
breseq.minus.muts = colnames(breseq.minus[3:41])
breseq.plus <- breseq.all[breseq.all$trt=="Ph+",]
breseq.plus <- breseq.plus[,-c(which(colSums(breseq.plus[,6:117]) ==0))]
breseq.plus.muts = colnames(breseq.plus[5:89])

both <- intersect(breseq.minus.muts,breseq.plus.muts)
phage.minus.only <- setdiff(breseq.minus.muts,breseq.plus.muts)
phage.plus.only <- setdiff(breseq.plus.muts,breseq.minus.muts)
```

```
breseq.N.plus <- breseq.plus[breseq.plus$lim == "N",]
breseq.N.plus.muts <- colnames(breseq.N.plus[5:ncol(breseq.N.plus)])
breseq.N.minus <- breseq.minus[breseq.minus$lim == "N",]
breseq.N.minus.muts <- colnames(breseq.N.minus[3:ncol(breseq.N.minus)])
breseq.P.plus <-breseq.plus[breseq.plus$lim == "P",]
breseq.P.plus.muts <- colnames(breseq.P.plus[5:ncol(breseq.P.plus)])
breseq.P.minus <- breseq.minus[breseq.minus$lim == "P",]
breseq.P.minus.muts <- colnames(breseq.P.minus[3:ncol(breseq.P.minus)])

# Mutations unique to phage exposed
setdiff(breseq.N.plus.muts,breseq.N.minus.muts)
```

```
##  [1] "12,864"    "101,693"   "102,841"   "102,850"   "103,283"
##  [6] "103,845"   "104,094"   "104,119"   "104,881"   "104,891"
## [11] "104,908"   "106,218"   "106,237"   "157,239"   "163,807"
## [16] "209,236"   "239,024"   "239,739"   "239,935"   "240,165"
## [21] "240,719"   "284,356"   "414,594"   "451,666"   "498,871"
## [26] "504,192"   "535,535"   "535,604"   "565,198"   "568,385"
## [31] "758,238"   "771,126"   "771,129"   "771,136"   "771,138"
## [36] "771,150"   "771,183"   "771,186"   "771,189"   "771,193"
```

```
## [41] "771,201"   "771,207"   "771,210"   "771,222"   "771,225"
## [46] "771,228"   "771,231"   "771,237"   "771,240"   "771,246"
## [51] "771,249"   "771,273"   "771,297"   "917,144"   "917,611"
## [56] "951,302"   "1,019,327" "1,019,857" "1,020,535" "1,061,867"
## [61] "1,135,395" "1,175,012" "1,268,916" "1,268,922" "1,271,925"
## [66] "1,272,000" "1,372,875" "1,379,294" "1,407,417" "1,500,963"
## [71] "1,760,010" "1,891,124" "1,922,186" "1,945,019" "2,006,492"
## [76] "2,019,461" "2,019,464" "2,019,467" "2,019,473"
```

```
# mutations unique to unexposed
setdiff(breseq.N.minus.muts,breseq.N.plus.muts)
```

```
##  [1] "103,376"   "104,894"   "104,895"   "152,770"   "154,970"
##  [6] "185,296"   "240,454"   "256,217"   "337,494"   "340,700"
## [11] "534,877"   "771,291"   "771,303"   "821,170"   "934,026"
## [16] "1,272,008" "1,272,988" "1,328,915" "1,565,247" "1,618,650"
## [21] "1,640,030" "1,703,838" "1,723,445"
```

```
# Mutations unique to phage exposed
setdiff(breseq.P.plus.muts,breseq.P.minus.muts)
```

```
##  [1] "12,864"    "101,693"   "102,841"   "102,850"   "103,283"
##  [6] "103,845"   "104,094"   "104,119"   "104,881"   "104,891"
## [11] "104,908"   "106,218"   "106,237"   "157,239"   "163,807"
## [16] "209,236"   "239,024"   "239,739"   "239,935"   "240,165"
## [21] "240,719"   "284,356"   "414,594"   "451,666"   "498,871"
## [26] "504,192"   "535,535"   "535,604"   "565,198"   "568,385"
## [31] "758,238"   "771,126"   "771,129"   "771,136"   "771,138"
## [36] "771,150"   "771,183"   "771,186"   "771,189"   "771,193"
## [41] "771,201"   "771,207"   "771,210"   "771,222"   "771,225"
## [46] "771,228"   "771,231"   "771,237"   "771,240"   "771,246"
## [51] "771,249"   "771,273"   "771,297"   "917,144"   "917,611"
## [56] "951,302"   "1,019,327" "1,019,857" "1,020,535" "1,061,867"
## [61] "1,135,395" "1,175,012" "1,268,916" "1,268,922" "1,271,925"
## [66] "1,272,000" "1,372,875" "1,379,294" "1,407,417" "1,500,963"
## [71] "1,760,010" "1,891,124" "1,922,186" "1,945,019" "2,006,492"
## [76] "2,019,461" "2,019,464" "2,019,467" "2,019,473"
```

```
# Mutations unique to phage exposed
setdiff(breseq.P.minus.muts,breseq.P.plus.muts)
```

```
##  [1] "103,376"   "104,894"   "104,895"   "152,770"   "154,970"
##  [6] "185,296"   "240,454"   "256,217"   "337,494"   "340,700"
## [11] "534,877"   "771,291"   "771,303"   "821,170"   "934,026"
## [16] "1,272,008" "1,272,988" "1,328,915" "1,565,247" "1,618,650"
## [21] "1,640,030" "1,703,838" "1,723,445"
```

Visualize genomic mutations

```
# integrate circle plot script for all genomic mutations and highlight the ones from the candidat
# Ideally contains circle with gene sections and corresponding mutations added in
```

Breadth of resistance

In a previous study, we challenged host and phage strains in a time-shift experiment. We selected strains with unique profiles to examine the genetic diversity underlying the phenotypic patterns in resistance to phage isolated from the NL and PL environments.

Summary of major results

1. The ancestor WH7803 was sensitive to all challenged phages from each of the resource treatments.
2. Thirteen of the 28 isolates were completely resistant to all challenged phages.
3. Average resistance in the remaining strains ranged between 0.287 and 0.960.

- in every case, hosts were more resistant to PL phages than to NL phages (stats)
- boxplot for strains (color hosts by NL and PL, shapes by +Ph and -Ph)

```r
# Read in data file
inf.mat <- as.matrix(read.csv("./data/bor.csv"))

# Rearrange data file for the appropriate analyses
inf.mat <- t(inf.mat)
inf.strains <- inf.mat[8:nrow(inf.mat),6]
p.dat <- inf.mat[1:6,7:ncol(inf.mat)]
inf.mat <- inf.mat[-c(1:7),-c(1:6)]
rownames(inf.mat) <- inf.strains
class(inf.mat) <- "numeric"

# Subset the infection matrix for only the sequenced strains
index2 <- which(rownames(inf.mat)  %in% strain.names)
WH7803 <- inf.mat[rownames(inf.mat)=="WH7803",]
inf.mat2 <- inf.mat[index2,]
inf.mat2 <- rbind(inf.mat2,WH7803)
inf.mat2 <- inf.mat2[-c(27,31),-c(95,96)]
class(inf.mat2) = "numeric"
```

```r
# Function to calculate resistance
infect.prop<-function(x){
  x <- x[!is.na(x)] # remove NA's
  length(x[x == 1])/length(x)
    }

inf.dat <- matrix(NA,nrow = length(inf.strains), ncol=4)

# Calculate resistance
#rownames(inf.dat) <- inf.strains
inf.dat[,1] <- inf.strains
inf.dat [,2] <- round(apply(inf.mat,1,infect.prop),digits = 2)
inf.dat[1:85,3] <- round(apply(inf.mat[1:85,2:46],1,infect.prop),digits = 2)
inf.dat[88:225,3] <- apply(inf.mat[88:nrow(inf.mat),47:96],1,infect.prop)
inf.dat[1:85,4] <- apply(inf.mat[1:85,47:96],1,infect.prop)
inf.dat[88:225,4] <- apply(inf.mat[88:nrow(inf.mat),2:46],1,infect.prop)
names(inf.dat) <- c(NULL,NULL)

#Subset calculated  matrix with the sequenced strain names
WH7803 <- inf.dat[inf.dat[,1]=="WH7803"]
index <- which(inf.dat[,1]  %in% strain.names)
inf.dat2 <- inf.dat[index,]
inf.dat2 = inf.dat2[-c(27,31),]   # remove 12P2S5, 21P2S4 due to lack of inf data
inf.dat2 <- rbind(WH7803,inf.dat2)
rownames(inf.dat2) <- NULL
colnames(inf.dat2) = c("strain","Average","Home","Away")
#write.csv(inf.dat2,file = "./supporting-files/data/BOR.csv",row.names=FALSE)
rownames(inf.dat2) <- inf.dat2[,1]
inf.dat2 <- inf.dat2[,-1]
class(inf.dat2)="numeric"
```

```
#Note: Make two separate trait tables then merge data together in one graphic for the manuscript

require(ape); require(phylobase); require(RColorBrewer);require(adephylo)
```

```
## Loading required package: RColorBrewer
```

```
## Loading required package: adephylo
```

```
## Loading required package: ade4
```

```
## Registered S3 method overwritten by 'spdep':
##   method   from
##   plot.mst ape
```

```
# Make the distance matrix with infection matrix
inf.mat3 <- inf.mat2[-32,]

# Due to the high number of zeros, must transform data
#inf.dist <- vegdist(decostand(inf.mat3, method = "log"), method = "gower", na.rm = TRUE)
inf.dist <- vegdist(inf.mat3, method = "gower", na.rm = TRUE)
nj.tree <- nj(inf.dist)

# Define the Outgroup
#outgroup <- match("WH7803", nj.tree$tip.label)

# Create a Rooted Tree {ape}
#nj.rooted <- root(nj.tree, outgroup, resolve.root = TRUE)
#nj.rooted$edge.length <- nj.rooted$edge.length + min(nj.rooted$edge.length)
#nj.rooted <- drop.tip(nj.rooted, "WH7803")

# Alternative clustering algorithm for distance matrix
hc <- hclust(inf.dist, method = "complete")
hcp <- as.phylo(hc)

par(mar=c(1,1,1,1) + 0.2)
mypalette <- colorRampPalette(brewer.pal(9, "YlOrRd"))
mypalette2 <- colorRampPalette(brewer.pal(9, "YlGnBu"))


mut.cts <- read.csv("./data/mut-cts.csv", header=T)
strain.list <- read.csv("./data/strains.csv")     #strain metadata

# Merge all the data together to make the tree
tree.dat <- merge(strain.list,mut.cts,by.x="strain_ID",by.y="strain")
tree.dat <- merge(inf.dat2,tree.dat,by.x = 0 ,by.y="strain_ID")
row.names(tree.dat) <- tree.dat[,1]
```

```
tree.dat2 <- tree.dat[-32,-c(1,8)]

pdf(file="./supporting-files/figures/fig-pub_traits-res.pdf",width = 3, height = 6)

x <- phylo4d(x = hcp, inf.dat2)
```

```
## Warning in formatData(phy = x, dt = tip.data, type = "tip", ...): The
## following names are not found in the tree: WH7803
```

```
table.phylo4d(x, treetype = "phylo", symbol = "colors", show.node = TRUE,
              cex.label = 0.35, scale = FALSE, use.edge.length = TRUE,
              edge.color = "black", edge.width = 1, box = FALSE, grid = TRUE,
              col= mypalette2(25),
              # col = gray.colors(15, start = 0.3, end = 0.9, gamma = 2.2, alpha = NULL),
              pch = 15, cex.symbol = 1.25,
              ratio.tree = 0.25, cex.legend = 1.25, center = FALSE)

dev.off()
```

```
## pdf
##   2
```

```
# Resistance + mutations
pdf(file="./supporting-files/figures/fig-pub_traits2.pdf",width = 5, height = 6)

x <- phylo4d(x = hcp, tree.dat2)
table.phylo4d(x, treetype = "phylo", symbol = "colors", show.node = TRUE,
              cex.label = 0.35, scale = FALSE, use.edge.length = TRUE,
              edge.color = "black", edge.width = 2, box = FALSE, grid = TRUE,
              col= mypalette2(50),
              # col = gray.colors(15, start = 0.3, end = 0.9, gamma = 2.2, alpha = NULL),
              pch = 15, cex.symbol = 1,
              ratio.tree = 0.20, cex.legend = 1, center = FALSE)

dev.off()
```

```
## pdf
##   2
```

```r
require(ade4)

#check dimensions of matrices
dim(breseq);dim(inf.mat2)
```

```
## [1]  36 206
```

```
## [1] 32 94
```

```r
rmv <- which(rownames(breseq)%in%setdiff(rownames(breseq),rownames(inf.mat2)))
which(rownames(inf.mat2)%in%setdiff(rownames(inf.mat2),rownames(breseq)))
```

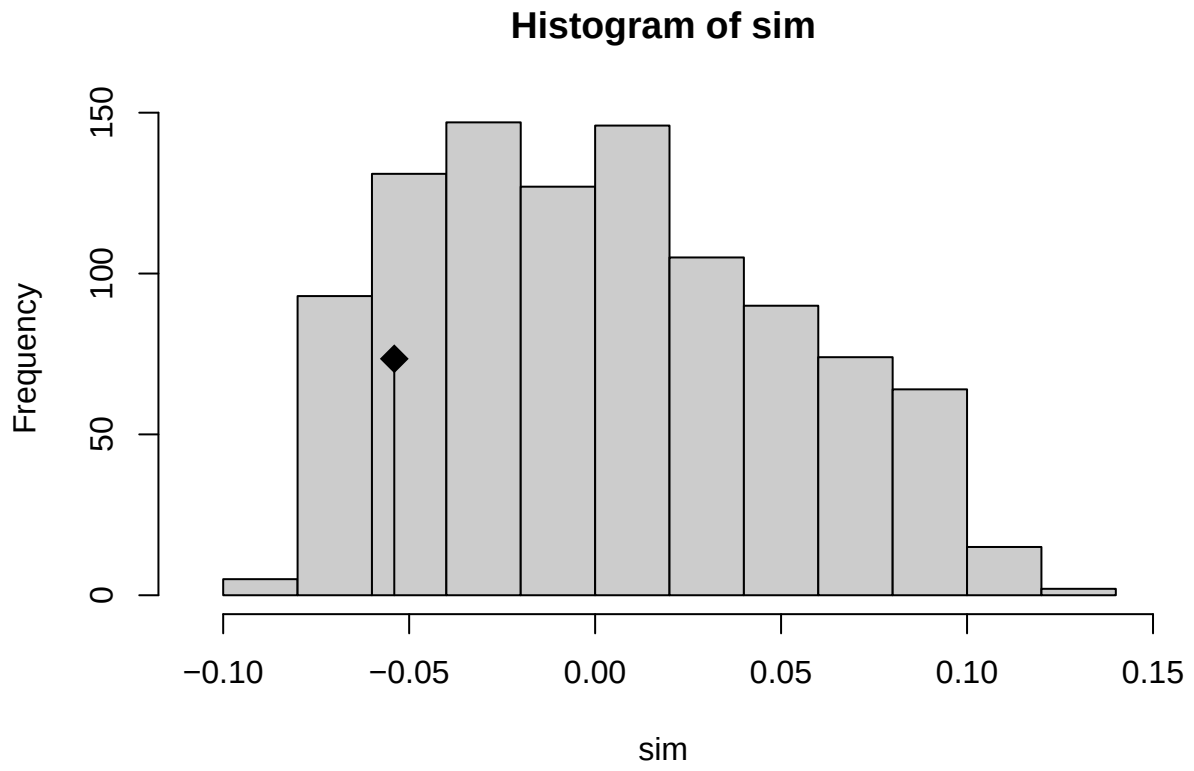
```
## [1] 32
```

```r
# rmv=c(22,28,31,33,34)
breseq = breseq[-rmv,]
inf.mat2= inf.mat2[-32,]
#inf.dat2 = inf.dat2[-32,]

# Make a distance matrix for the mutational data
breseq.dist <- vegdist(breseq, method="gower", na.rm = TRUE)
inf.dist <- vegdist(inf.mat2, method="gower", na.rm = TRUE)

# perform Mantel Test between two distance matrices
man.test <- mantel.randtest(breseq.dist, inf.dist, nrepet = 999)

plot(man.test)
```




```
cols = c(2:8)
dat <- tree.dat[-32,cols]
dat[,4] = droplevels(dat[,4])
adonis(breseq.dist ~ dat$Average*dat$lim*dat$trt, permutations = 999)
```

```
##
## Call:
## adonis(formula = breseq.dist ~ dat$Average * dat$lim * dat$trt,      permutations = 999)
##
## Permutation: free
## Number of permutations: 999
##
## Terms added sequentially (first to last)
##
##                              Df SumsOfSqs   MeanSqs F.Model      R2 Pr(>F)
## dat$Average                   1  0.008523 0.0085229 0.90830 0.03223  0.495
## dat$lim                       1  0.005378 0.0053782 0.57317 0.02034  0.768
## dat$trt                       1  0.012193 0.0121925 1.29937 0.04611  0.258
## dat$Average:dat$lim           1  0.009597 0.0095970 1.02276 0.03630  0.392
## dat$Average:dat$trt           1  0.003535 0.0035354 0.37677 0.01337  0.840
## dat$lim:dat$trt               1  0.001836 0.0018359 0.19566 0.00694  0.925
## dat$Average:dat$lim:dat$trt   1  0.007527 0.0075274 0.80221 0.02847  0.397
## Residuals                    23  0.215818 0.0093834         0.81623
## Total                        30  0.264407                   1.00000
```

```r
adonis(inf.dist ~ dat$Average*dat$lim*dat$trt, permutations = 999)
```

```
##
## Call:
## adonis(formula = inf.dist ~ dat$Average * dat$lim * dat$trt,      permutations = 999)
##
## Permutation: free
## Number of permutations: 999
##
## Terms added sequentially (first to last)
##
##                              Df SumsOfSqs MeanSqs F.Model      R2 Pr(>F)
## dat$Average                   1    0.8465 0.84654  9.8154 0.23462  0.004 **
## dat$lim                       1    0.1109 0.11094  1.2863 0.03075  0.293
## dat$trt                       1    0.0163 0.01630  0.1889 0.00452  0.679
## dat$Average:dat$lim           1    0.0316 0.03156  0.3660 0.00875  0.553
## dat$Average:dat$trt           1    0.2063 0.20634  2.3924 0.05719  0.126
## dat$lim:dat$trt               1    0.3643 0.36426  4.2235 0.10096  0.034 *
## dat$Average:dat$lim:dat$trt   1    0.0485 0.04852  0.5626 0.01345  0.466
## Residuals                    23    1.9837 0.08625         0.54978
## Total                        30    3.6081                 1.00000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```