Comparison of type VI secretion system between two strains

IM

June 11, 2018

Comparative genomics as a tool to understand the evolution of bacterial systems.

Identification of all components of the Type VI secretion system in two strains of Vibrio cholerae. The Type VI ss is organized in three different gene clusters. A main cluster, then two different auxiliar clusters.

After identification of all the components, we proceeded for a SNP analysis of each gene in the system between the two strains.

Add libraries

```
# Libraries install.packages('genoPlotR')
library("genoPlotR")
library(RColorBrewer)
library(tidyr)
# install.packages('grid')
library("ggplot2")
library("stringr")
library(grid)
library(pheatmap)
library(plyr)
library(ape)
library(ggtree)
library(PopGenome)
options(digits = 2)
library(RColorBrewer)
library(classInt)
library(ggpubr)
library(corrplot)
```

Components of the TypeVI secretion system

After identification of the components on the two different strains, we compared the clusters between the two strains.

Main cluster

```
# Set working directory
setwd("~/Documents/Melanie/TypeVISS_O395_A1552/Coordinates_blast/")
################################ import blast output xml
filesToProcess <- dir(pattern = "*\\.xml$")  #files to pro# if event 3 merged
filesToProcess <- filesToProcess[grep("Large", filesToProcess, invert = F)]  ##<- to modify
```

```
listOfFiles <- lapply(filesToProcess, function(x) tryCatch(read.table(x,
    header = F, stringsAsFactors = F, sep = c("\t", ",")), error = function(e) cbind.data.frame(V1 = "NA
    V2 = "NA", V3 = "NA", V4 = "NA", V5 = "NA", V6 = "NA", V7 = 0,
    V8 = 0, V9 = "NA", V10 = "NA", V11 = "NA", V12 = "NA")))

# Format of raw data .xml
head(listOfFiles[[2]])
```

```
##        V1          V2  V3  V4 V5 V6 V7  V8     V9    V10   V11  V12
## 1 VCA0105 NC_012583.1 100  94  0  0  1  94 115173 115454 8e-55  183
## 2 VCA0106 NC_012583.1 100 334  1  0  1 334 115441 116442 0e+00  683
## 3 VCA0107 NC_012583.1 100 110  0  0 59 168 117052 117381 2e-86  213
## 4 VCA0108 NC_012583.1 100 492  0  0  1 492 117425 118900 0e+00 1022
## 5 VCA0109 NC_012583.1  97 145  1  1  1 145 118906 119328 6e-89  284
## 6 VCA0110 NC_012583.1 100 589  0  0  1 589 119337 121103 0e+00 1189
```

```
colnam <- c("Gene", "Chr", "Identity", "Length", "MissMatch", "Gap",
    "QStart", "Qend", "Start", "End", "Eval", "BitScore")
listOfFiles <- lapply(listOfFiles, setNames, nm = colnam)


# prepare guide

guid <- ldply(Map(cbind, iso = gsub("blast_LargeCluster_", "", gsub(".xml",
    "", filesToProcess)), xmin = lapply(listOfFiles, function(x) min(x$Start)),
    xmax = lapply(listOfFiles, function(x) max(x$End)), stringsAsFactors = F),
    data.frame)
guid <- cbind.data.frame(guid, ymin = (1:length(guid$iso)) - 0.01,
    ymax = (1:length(guid$iso)) + 0.01)


plotlines <- ggplot(aes(xmin = xmin, xmax = xmax, ymin = ymin - 0.001,
    ymax = ymax + 0.001), data = guid) + geom_rect() + scale_y_continuous(limits = c(0,
    3)) + annotate("text", x = -2000, y = (1:length(unique(guid$.id))),
    label = unique(guid$.id))


# CDS
listOfFiles <- lapply(listOfFiles, function(x) cbind.data.frame(Gene = x$Gene,
    Chr = x$Chr, Start = x$Start - min(x$Start) + 1, End = x$End -
        min(x$Start) + 1))
df2 <- NULL
for (i in 1:length(listOfFiles)) {
    df2 <- rbind.data.frame(df2, listOfFiles[[i]])
}

df2 <- cbind.data.frame(df2, ISO = c(rep("A1552", length(listOfFiles[[1]][,
    1])), rep("O395", length(listOfFiles[[1]][, 1]))), ymin = c(rep(1,
    length(listOfFiles[[1]][, 1])), rep(2, length(listOfFiles[[1]][,
    1]))) - 0.05, ymax = c(rep(1, length(listOfFiles[[1]][, 1])),
    rep(2, length(listOfFiles[[1]][, 1]))) + 0.05)
```

Table to use for ploting the two clusters

```
head(df2)
```

```
##      Gene                        Chr Start  End   ISO ymin ymax
## 1 VCA0105 Vibrio_cholerae_A1552_ch2     1  282 A1552 0.95  1.1
## 2 VCA0106 Vibrio_cholerae_A1552_ch2   269 1270 A1552 0.95  1.1
## 3 VCA0107 Vibrio_cholerae_A1552_ch2  1705 2208 A1552 0.95  1.1
## 4 VCA0108 Vibrio_cholerae_A1552_ch2  2252 3727 A1552 0.95  1.1
## 5 VCA0109 Vibrio_cholerae_A1552_ch2  3733 4167 A1552 0.95  1.1
## 6 VCA0110 Vibrio_cholerae_A1552_ch2  4176 5942 A1552 0.95  1.1
```

Linear comparison between the two strains

```
plotlines + geom_rect(aes(xmin = Start, xmax = End, ymin = ymin, ymax = ymax),
    data = df2, color = "black", fill = "black", alpha = 2/4) + annotate("text",
    x = df2$Start + 500, y = df2$ymax + 0.08, label = df2$Gene, angle = 45,
    hjust = 0) + coord_cartesian(xlim = c(min(df2$Start) - 3000, max(df2$End) +
    3000))
```

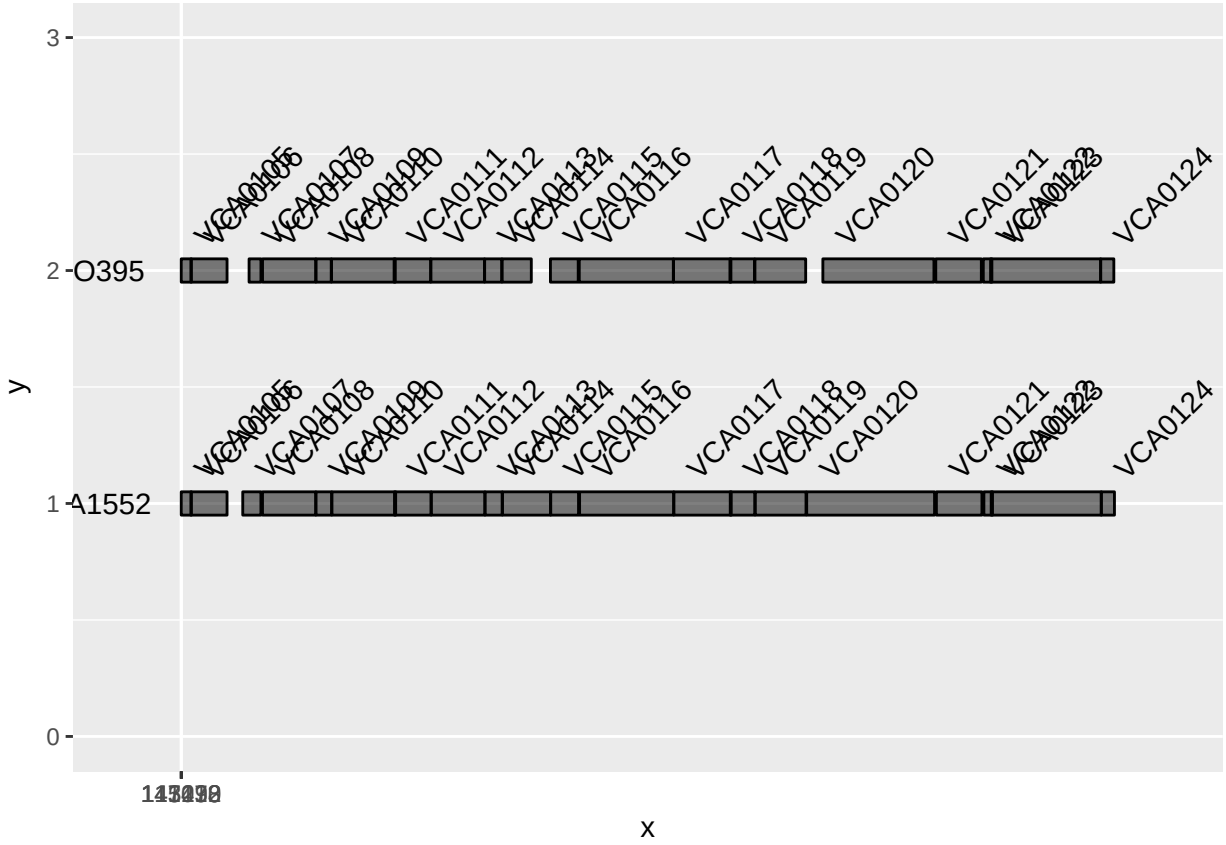

Figure 1: Comparison of the main cluster of the TypeVI SS

Analyse sequences SNP and Syn/Non-Syn modifications

Before starting the SNP analysis, align the sequences in bash and put the resulting alignement in a new folder.

```
# In bash for i in $(ls *A1552*.fa); do echo $i $(echo $i | sed
# 's/A1552/O395/' ); mkdir $(echo $i | cut -d'_' -f2); cat $i >
# $(echo $i | cut -d'_' -f2)/$(echo $i | cut -d'_' -f2).fa; cat
# $(echo $i | sed 's/A1552/O395/' ) >> $(echo $i | cut -d'_'
# -f2)/$(echo $i | cut -d'_' -f2).fa; clustalo -i $(echo $i | cut
# -d'_' -f2)/$(echo $i | cut -d'_' -f2).fa -o $(echo $i | cut
# -d'_' -f2)/$(echo $i | cut -d'_' -f2).fa --force ; done#2. align
```

SNP and Syn/Non-Syn analysis

```
dirs <- dir("~/Documents/Melanie/TypeVISS_O395_A1552/LargeCluster/",
    pattern = "^VC")
genes <- list()
for (x in dirs) {
    genes[[x]] <- readData(as.character(paste("~/Documents/Melanie/TypeVISS_O395_A1552/LargeCluster/",
        x, sep = "")), include.unknown = F)
}
```

```
## |            :            |            :            | 100 %
## |==================================================
## |            :            |            :            | 100 %
## |==================================================
## |            :            |            :            | 100 %
## |==================================================
## |            :            |            :            | 100 %
## |==================================================
## |            :            |            :            | 100 %
## |==================================================
## |            :            |            :            | 100 %
## |==================================================
## |            :            |            :            | 100 %
## |==================================================
## |            :            |            :            | 100 %
## |==================================================
## |            :            |            :            | 100 %
## |==================================================
## |            :            |            :            | 100 %
## |==================================================
## |            :            |            :            | 100 %
## |==================================================
## |            :            |            :            | 100 %
## |==================================================
## |            :            |            :            | 100 %
## |==================================================
## |            :            |            :            | 100 %
## |==================================================
## |            :            |            :            | 100 %
## |==================================================
```

```
## |            :            |            :            | 100 %
## |==================================================
## |            :            |            :            | 100 %
## |==================================================
## |            :            |            :            | 100 %
## |==================================================
## |            :            |            :            | 100 %
## |==================================================
## |            :            |            :            | 100 %
## |==================================================
```

```
# get summary statistics
genes_SNP <- unlist(lapply(genes, function(x) length(x@region.data@synonymous[[1]])))
nbSites <- unlist(lapply(genes, function(x) get.sum.data(x)[, 1]))
Syno <- lapply(genes, function(x) x@region.data@synonymous[[1]])


Syn <- unlist(lapply(Syno, function(x) length(x[x == TRUE])))
NonSyn <- unlist(lapply(Syno, function(x) length(x[x != TRUE])))

SeqComp <- cbind.data.frame(Sites = nbSites, SNP = genes_SNP, Syn = Syn,
    NonSyn = NonSyn)
```

Results

```
SeqComp
```

```
##         Sites SNP Syn NonSyn
## VCA0105   282   0   0      0
## VCA0106  1002   0   0      0
## VCA0107   504   1   1      0
## VCA0108  1476   9   9      0
## VCA0109   435   1   0      1
## VCA0110  1767  10  10      0
## VCA0111  1014  14  10      4
## VCA0112  1485   9   8      1
## VCA0113   474   1   1      0
## VCA0114  1332   5   5      0
## VCA0115   771   0   0      0
## VCA0116  2607   0   0      0
## VCA0117  1590   0   0      0
## VCA0118   681   1   0      1
## VCA0119  1407   0   0      0
## VCA0120  3543   0   0      0
## VCA0121  1263   0   0      0
## VCA0122   240   0   0      0
## VCA0123  3051   1   0      1
## VCA0124   366   0   0      0
```

Can repeat the same analysis on Auxiliar cluster 1 and 2. The analysis of SNP on single genes, can be done, by alligning the sequences and then placing them in a single folder.

We can observe that the main cluster of Type VISS betwen the two strains is very similar. however there are

some differences: VCA0114 and VCA0119 have different size between both strains.

VCA0110, VCA0111 and VCA0112 display more than 10 SNP's each, but only VCA0111 display 4 non-synonimous mutations. Suggesting that these three genes are less conserved than the other genes in the cluster. This difference could be the result of either genetic drift, or adaptive selection. Nevertheless, it is necessary to compare a higher number of strains to get biological insights about the conservation or divergenece of the different genes.