Resist Subtype

James Hunter

August 27, 2015

This is a copy of resistsubtype.R from April 2014 The data come from resistsubtipo.xlsx, one of the original files. This file is obviously derived from Anderson's original data, with the resistances being converted into Sim/Não categories. At the end of this, I will replicate the bar graph from "load Juncao db" file from earlier this week.

Legend: c2 = 2 class resistance c3 = 3 class resistance individual class names followed by "yn" indicate that the variables have been classified as boolean - resistance exists or not in a patient.

```r
## Set up R data frame based on resistsubtipo.csv
resistst <- read.csv("resistsubtipo.csv", header = TRUE,
                     stringsAsFactors = FALSE,
                     na.strings = c("*", "?"))
## Set up resistance types as 1/0
resistst$trnyn <- ifelse(resistst$trn > 0, 1, 0)
resistst$trnnyn <- ifelse(resistst$trnn > 0, 1, 0)
resistst$ipyn <- ifelse(resistst$ip > 0, 1, 0)


## Measure 3 class resistance
# using numeric variables
resistst$c3 <- ifelse(resistst$trn > 0 & resistst$trnn > 0
                      & resistst$ip > 0, 1, 0)
## measure 2 class resistance
# using numeric variables
resistst$c2 <- ifelse(with(resistst,
                           ((trn > 0) + (trnn > 0) + (ip >0)) == 2),
                           1, 0)
## Create names for state of resistances
resNames <- c("Absent", "Present")

## Work only on complete cases (1009 with missing data,
## of which 994 have stBlast = "BF")

resistcomp <- resistst[complete.cases(resistst),]

#resiststn <- resistst
#resiststn$c3 <- ifelse(resiststn$c3 == 1, "Present", "Absent")
#resiststn$c2 <- ifelse(resiststn$c2 == 1, "Present", "Absent")
## Create tables resistcomp data frame
## Whole data frame
with(resistcomp, table(ano, reg))
```

```
##       reg
## ano    Centro Oeste_Norte Nordeste Sudeste  Sul
##   2001                  0        0      45    0
##   2002                  0       22     419    0
##   2003                215       79     854    0
##   2004                239      176    2322   22
##   2005                292      241    2681  172
##   2006                413      278    2918  318
##   2007                155      162    1399  192
```

```
with(resistcomp, table(ano, st))
```

```
##       st
## ano      AF    B   BC  BCF   BF    C   CF    F    X
##   2001    0   34    0    1    7    0    0    1    2
##   2002    0  387    0    0   22    4    0   21    7
##   2003    0  969    1    0   82    9    0   70   17
##   2004    0 2350   14    0  191   57    0  134   13
##   2005    0 2783   25    0  249   69    3  229   28
##   2006    1 3119   39    0  291  194    1  225   57
##   2007    0 1505   28    2  152   91    0  120   10
```

```
resc3 <- with(resistcomp, table(ano, c3))
resc3
```

```
##       c3
## ano       0    1
##   2001   45    0
##   2002  423   18
##   2003 1123   25
##   2004 2702   57
##   2005 3299   87
##   2006 3832   95
##   2007 1862   46
```

```
resc2 <- with(resistcomp, table(ano, c2))
resc2
```

```
##       c2
## ano       0    1
##   2001   42    3
##   2002  368   73
##   2003  984  164
##   2004 2338  421
##   2005 2924  462
##   2006 3402  525
##   2007 1652  256
```

```
restrn <- with(resistcomp, table(ano, trnyn))
restrn
```

```
##       trnyn
## ano       0    1
##   2001   42    3
##   2002  383   58
##   2003 1033  115
##   2004 2452  307
##   2005 2978  408
##   2006 3497  430
##   2007 1679  229
```

```
restrnn <- with(resistcomp, table(ano, trnnyn))
restrnn
```

```
##       trnnyn
## ano       0    1
##   2001   37    8
##   2002  327  114
##   2003  857  291
##   2004 1979  780
##   2005 2392  994
##   2006 2726 1201
##   2007 1379  529
```

```
resip <- with(resistcomp, table(ano, ipyn))
resip
```

```
##       ipyn
## ano       0    1
##   2001   27   18
##   2002  216  225
##   2003  625  523
##   2004 1579 1180
##   2005 2087 1299
##   2006 2394 1533
##   2007 1195  713
```

```
with(resistcomp, table(st, c3))
```

```
##      c3
## st        0     1
##   AF      1     0
##   B   10860   287
##   BC    104     3
##   BCF     3     0
##   BF    970    24
##   C     418     6
##   CF      4     0
##   F     794     6
##   X     132     2
```

```
with(resistcomp, table(st, c2))
```

```
##      c2
## st       0    1
##   AF     0    1
##   B   9542 1605
##   BC    91   16
##   BCF    3    0
##   BF   861  133
##   C    382   42
##   CF     4    0
##   F    711   89
##   X    116   18
```

```
## prepare %'s of c2 for insertion in resist - only years 2001 - 2007
resc2prop <- prop.table(resc2, margin = 1)
resc2prop
```

```
##       c2
## ano             0          1
##   2001 0.93333333 0.06666667
##   2002 0.83446712 0.16553288
##   2003 0.85714286 0.14285714
##   2004 0.84740848 0.15259152
##   2005 0.86355582 0.13644418
##   2006 0.86631016 0.13368984
##   2007 0.86582809 0.13417191
```

```
require(gmodels)
```

```
## Loading required package: gmodels
```

```
CrossTable(resistcomp$ano, resistcomp$c2, prop.chisq = FALSE, format = "SPSS")
```

```
## 
##    Cell Contents
## |-------------------------|
## |                   Count |
## |             Row Percent |
## |          Column Percent |
## |           Total Percent |
## |-------------------------|
## 
## Total Observations in Table:  13614 
## 
##                 | resistcomp$c2 
## resistcomp$ano |        0  |        1  | Row Total | 
## ---------------|-----------|-----------|-----------|
##           2001 |       42  |        3  |       45  | 
##                |   93.333% |    6.667% |    0.331% | 
##                |    0.359% |    0.158% |           | 
##                |    0.309% |    0.022% |           | 
## ---------------|-----------|-----------|-----------|
##           2002 |      368  |       73  |      441  | 
##                |   83.447% |   16.553% |    3.239% | 
##                |    3.143% |    3.834% |           | 
##                |    2.703% |    0.536% |           | 
## ---------------|-----------|-----------|-----------|
##            2003 |       984  |       164  |      1148  |
##                 |    85.714% |    14.286% |     8.432% |
##                 |     8.403% |     8.613% |            |
##                 |     7.228% |     1.205% |            |
## ----------------|------------|------------|------------|
##            2004 |      2338  |       421  |      2759  |
##                 |    84.741% |    15.259% |    20.266% |
##                 |    19.966% |    22.111% |            |
##                 |    17.173% |     3.092% |            |
## ----------------|------------|------------|------------|
##            2005 |      2924  |       462  |      3386  |
##                 |    86.356% |    13.644% |    24.871% |
##                 |    24.970% |    24.265% |            |
##                 |    21.478% |     3.394% |            |
## ----------------|------------|------------|------------|
##            2006 |      3402  |       525  |      3927  |
##                 |    86.631% |    13.369% |    28.845% |
##                 |    29.052% |    27.574% |            |
##                 |    24.989% |     3.856% |            |
## ----------------|------------|------------|------------|
##            2007 |      1652  |       256  |      1908  |
##                 |    86.583% |    13.417% |    14.015% |
##                 |    14.108% |    13.445% |            |
##                 |    12.135% |     1.880% |            |
## ----------------|------------|------------|------------|
##    Column Total |     11710  |      1904  |     13614  |
##                 |    86.014% |    13.986% |            |
## ----------------|------------|------------|------------|
##
##
```

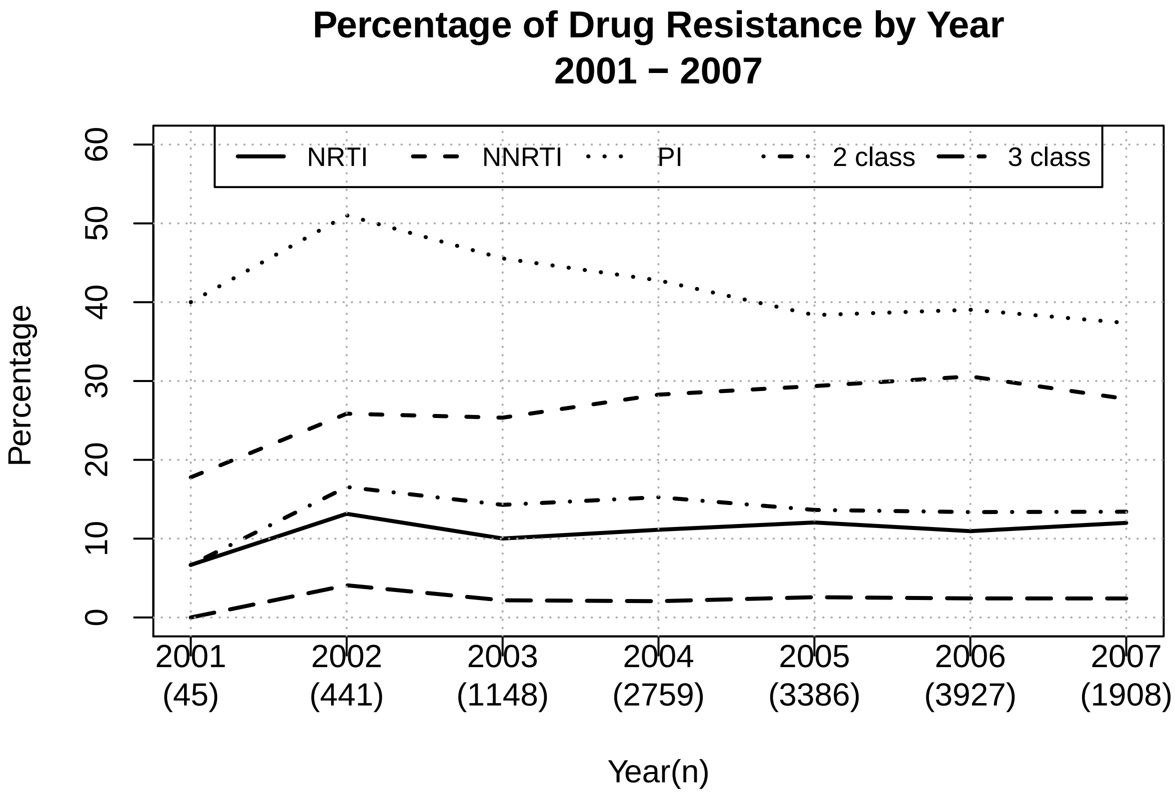
```r
CrossTable(resistcomp$ano, resistcomp$c3, prop.chisq = FALSE, format = "SPSS")
```

```
##
##    Cell Contents
## |-------------------------|
## |                   Count |
## |             Row Percent |
## |          Column Percent |
## |           Total Percent |
## |-------------------------|
##
## Total Observations in Table:  13614
##
##                 | resistcomp$c3
## resistcomp$ano |         0  |         1  | Row Total |
## ----------------|------------|------------|------------|
##            2001 |        45  |         0  |        45  |
##                 |   100.000% |     0.000% |     0.331% |
##                 |     0.339% |     0.000% |            |
##                 |     0.331% |     0.000% |            |
## ----------------|------------|------------|------------|
##            2002 |       423  |        18  |       441  |
##                 |    95.918% |     4.082% |     3.239% |
##                |    3.184% |    5.488% |           |
##                |    3.107% |    0.132% |           |
## ---------------|-----------|-----------|-----------|
##           2003 |     1123  |       25  |     1148  |
##                |   97.822% |    2.178% |    8.432% |
##                |    8.453% |    7.622% |           |
##                |    8.249% |    0.184% |           |
## ---------------|-----------|-----------|-----------|
##           2004 |     2702  |       57  |     2759  |
##                |   97.934% |    2.066% |   20.266% |
##                |   20.337% |   17.378% |           |
##                |   19.847% |    0.419% |           |
## ---------------|-----------|-----------|-----------|
##           2005 |     3299  |       87  |     3386  |
##                |   97.431% |    2.569% |   24.871% |
##                |   24.831% |   26.524% |           |
##                |   24.232% |    0.639% |           |
## ---------------|-----------|-----------|-----------|
##           2006 |     3832  |       95  |     3927  |
##                |   97.581% |    2.419% |   28.845% |
##                |   28.842% |   28.963% |           |
##                |   28.147% |    0.698% |           |
## ---------------|-----------|-----------|-----------|
##           2007 |     1862  |       46  |     1908  |
##                |   97.589% |    2.411% |   14.015% |
##                |   14.015% |   14.024% |           |
##                |   13.677% |    0.338% |           |
## ---------------|-----------|-----------|-----------|
##   Column Total |     13286 |      328  |    13614  |
##                |   97.591% |    2.409% |           |
## ---------------|-----------|-----------|-----------|
##
##
```

```
## axis data with year and n
years <- min(resistcomp$ano) : max(resistcomp$ano)
n <- rep(NA, length(years))
for (i in 1 :length(years)){
   n[i] <- resip[i,1] + resip[i,2]
}
tickvec2 <- rep(NA, length(years))
for (i in 1:length(years)) {
   tickvec2[i] <- paste(years[i],
                         "\n(",n[i],")", sep = "")
}

## Mount data frame with final proportion data
tables <- c("restrn", "restrnn", "resip", "resc2", "resc3")
resistprop <- as.data.frame(years)
resistprop$trn <-  100 * prop.table(restrn, margin = 1)[,2]
resistprop$trnn <-  100 * prop.table(restrnn, margin = 1)[,2]
resistprop$ip <-  100 * prop.table(resip, margin = 1)[,2]
resistprop$c2 <-  100 * prop.table(resc2, margin = 1)[,2]
resistprop$c3 <-  100 * prop.table(resc3, margin = 1)[,2]
```

```
## Graph of resistances
plot(resistprop$trn ~ resistprop$years, type = "l",
     lwd = 2, xaxt = "n", ylim = c(0, 60),
     main = "Percentage of Drug Resistance by Year\n2001 - 2007",
     xlab = "Year(n)",
     ylab = "Percentage ")
lines(resistprop$years, resistprop$trnn, lwd = 2, lty = 2)
lines(resistprop$years, resistprop$ip, lwd = 2, lty = 3)
lines(resistprop$years, resistprop$c2, lwd = 2, lty = 4)
lines(resistprop$years, resistprop$c3, lwd = 2, lty = 5)
legend("top",
       legend = c("NRTI", "NNRTI", "PI", "2 class", "3 class"),
       lwd = 2,
       lty = seq(1,5), cex = 0.8,
       ncol = 5)
axis(1, at = resistprop$years, labels = tickvec2)
grid(col = "darkgrey")
```



Bar chart equal to earlier bar chart with data from this dataset - comparison of percentage of resistance with subtype

Measure class resistances in same manner as Anderson data

```
suppressPackageStartupMessages(library(dplyr))
resistcomp <- mutate(resistcomp, numclasses =
                      (trnyn >0) + (trnnyn > 0) + (ipyn > 0))
resistcomp <- mutate(resistcomp, noclass = numclasses == 0)
resistcomp <- mutate(resistcomp, oneclass = numclasses == 1)
```

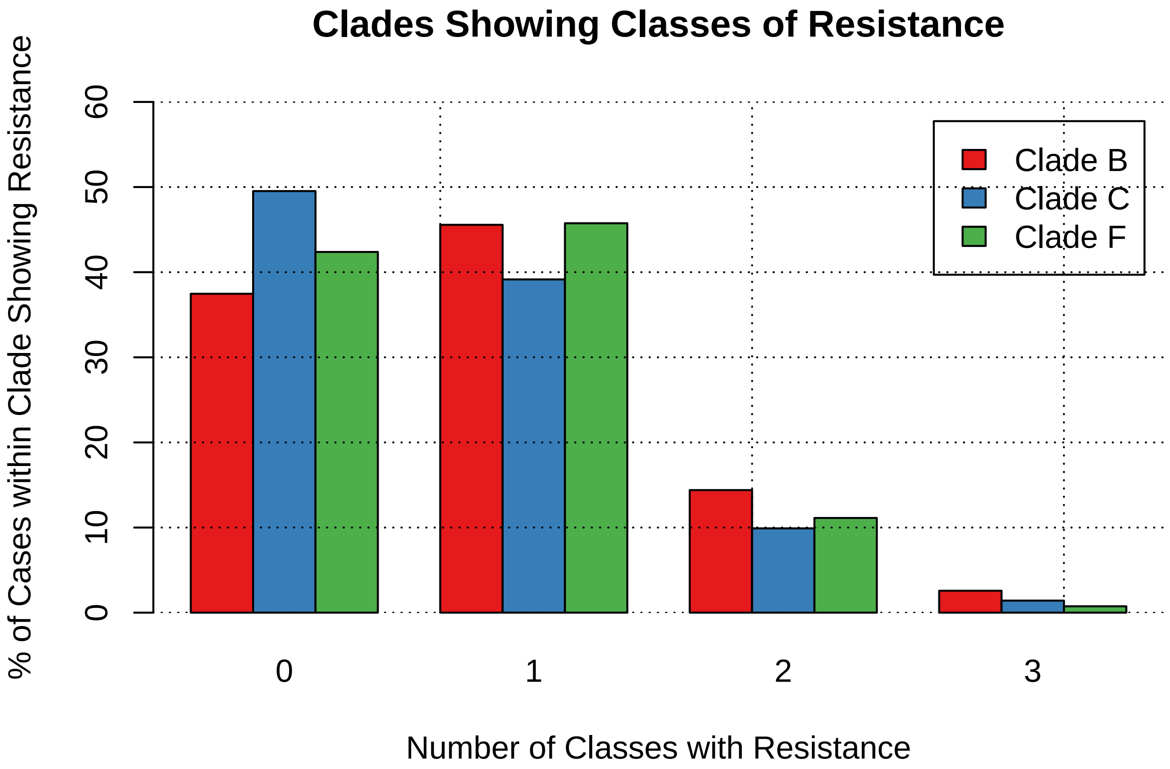
```r
resistcomp <- mutate(resistcomp, twoclass = numclasses == 2)
resistcomp <- mutate(resistcomp, threeclass = numclasses == 3)
table(resistcomp$numclasses) #all clades
```

```
##
##    0    1    2    3
## 5216 6166 1904  328
```

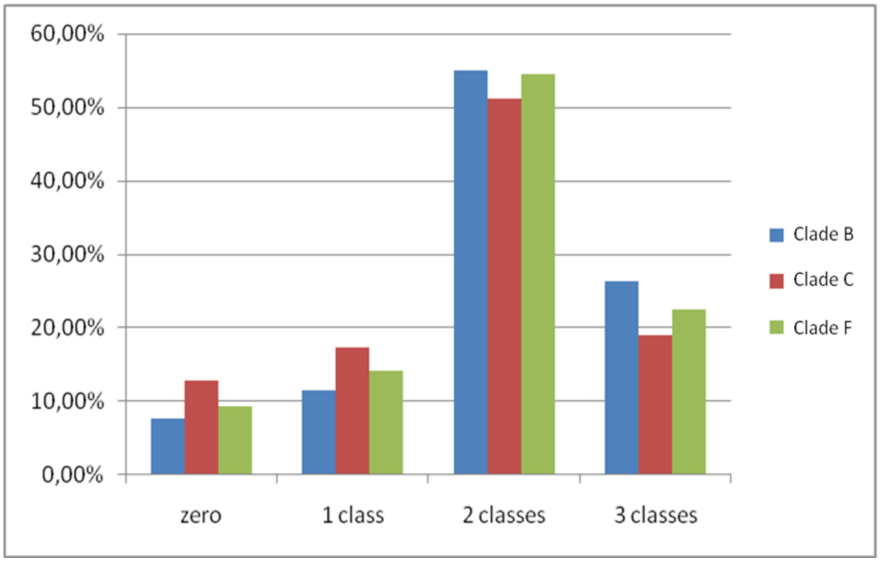
```r
# get reduced data set for BCF only
resistbcf <- filter(resistcomp, st %in% c("B", "C", "F"))
resistbcf$st <- factor(resistbcf$st, levels = c("B", "C", "F"))
subtype_table <- table(resistbcf$st)
resistbcftable <- table(resistbcf$st, factor(resistbcf$numclasses))
resistbcfprop <- 100 * prop.table(resistbcftable, 1) # in pct terms
resistbcfprop
```

```
##
##             0         1         2         3
##   B 37.462995 45.563829 14.398493  2.574684
##   C 49.528302 39.150943  9.905660  1.415094
##   F 42.375000 45.750000 11.125000  0.750000
```

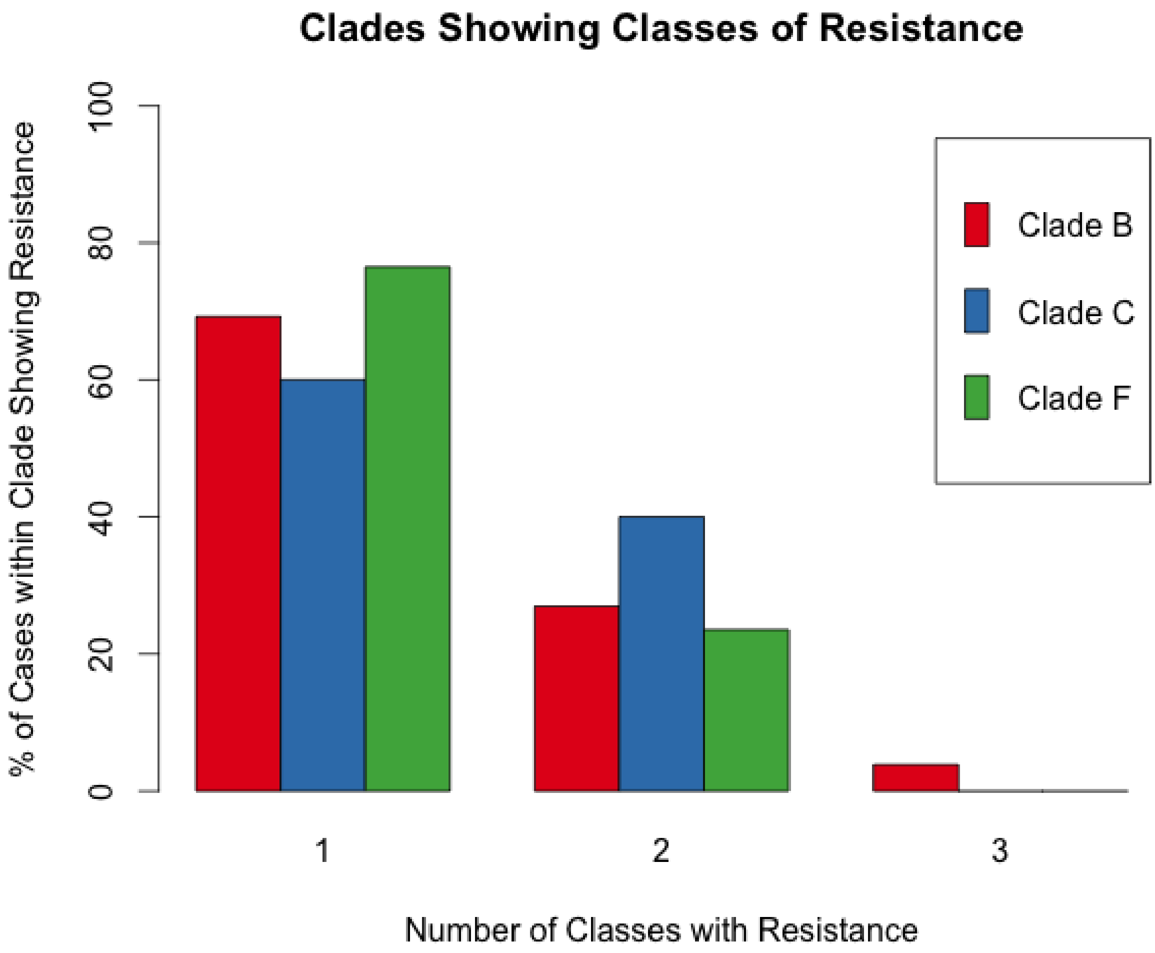
```r
library(RColorBrewer)
bcfresist <- barplot(resistbcfprop, beside = TRUE,
                col = brewer.pal(3, "Set1"),
                ylim = c(0, 60),
                xlab = "Number of Classes with Resistance",
                ylab = "% of Cases within Clade Showing Resistance",
                main = "Clades Showing Classes of Resistance",
                legend = c("Clade B", "Clade C", "Clade F"))
grid(col = "black")
```



Now, the original graph:



Also, against the graph from earlier in the week (Anderson's spreadsheets):



Comparison of Anderson's Stanford Clade data with the Brazilian clade assignment.

```
load("juncao_data.RData") # load the Anderson data
# from the original plan2 data, get the three subtype classifications
subtypes <- data_frame(stA_Hmmer = plan2$subtipo_Hmmer,
                       stA_Blast = plan2$subtipo_blast,
                       stA_st = plan2$subtipo)
# get reduced set for B,C,F only
subtypes_bcf <- filter(subtypes, stA_st %in% c("B", "C", "F") &
                          stA_Hmmer %in% c("B", "C", "F") &
                          stA_Blast %in% c("B", "C", "F"))
subtypes_bcf$stA_Blast <- factor(subtypes_bcf$stA_Blast, levels = c("B", "C", "F"))
subtypes_bcf$stA_st <- factor(subtypes_bcf$stA_st, levels = c("B", "C", "F"))
subtypes_bcf$stA_Hmmer <- factor(subtypes_bcf$stA_Hmmer, levels = c("B", "C", "F"))
```

Table of proportions of clades in Anderson spreadsheets

```
stAtable <- table(subtypes_bcf$stA_st)
stAbcfprop <- 100 * stAtable/nrow(subtypes_bcf) # in pct terms
round(stAbcfprop,3)
```

```
##
##      B      C      F
## 85.714  7.143  7.143
```

Table of proportions of clades in resistsubtipo spreadsheet

```
resistbcf_table <- table(resistbcf$st)
resistbcf_prop <- 100 * resistbcf_table/nrow(resistbcf)
round(resistbcf_prop,3)
```

```
##
##      B      C      F
## 90.106  3.427  6.467
```

```
t.test(resistbcf_prop, stAbcfprop)
```

```
##
##  Welch Two Sample t-test
##
## data:  resistbcf_prop and stAbcfprop
## t = 1.8392e-16, df = 3.974, p-value = 1
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -107.5386  107.5386
## sample estimates:
## mean of x mean of y
##  33.33333  33.33333
```

T-test shows no difference in distributions of B, C, and F clades coming from the two sets of data.

Conclusion

Since we have been working consistently with the dataset shown in this report (resistcomp), I would suggest continuing with this and adding this version of the bar chart to the paper. It is consistent with the other results on clades and resistance we have been reporting as suggested by the t-test above. Given everyone's lack of familiarity with the data sets we received last week, I would suggest we focus on these we have been working with more directly.