Antibiotic resistance genes thta occur in Biosynthetic gene clusters

DATA

CARD database contains antibiotic resistance (AR) genes.
MIBiG database is composed by Biosynthetic gene clusters.
antiSMASHDB contains predicted BGCs.

Methodology

CARD gene families

Genes in CARD were classified into gene families using FastORTHO with default parameters.

Homology searches between CARD and MIBiG

Homology search was conducted using blast (e-vaule) in an all vs all comparison with query antibiotic resistance genes from The Comprehensive Antibiotic Resistance Database (CARD) against genes in BGCs from MIBiG database.

In the following sections we will refer to the following questions: How many BGCs contain an AR gene? ¿How many families of AR genes are present in BGCs? Which families are over represented? How many BGCs per AR gene/Family?

Protein families in CARD

The following figure shows the most populated families in CARD database.

The total number of antibiotic gene resistance families is 76 counting singletons. Without singletons there are 0 families.

These first five families, in a preliminar search corresponds to:

Family	Number of genes	Annotation
Family 0	647	β-lactamase 2
Family 1	302	β-lactamase
Family 2	283	Transpeptidase
Family 3	103	class B extended-spectrum β-lactamase
Family 4	55	MFS_1

BGC clases in MiBIG

The following figure shows the most populated classes in MiBIG database.

There are 1811 different BGCs reported at MIBiG distributed in several classes as indicated in the following table.

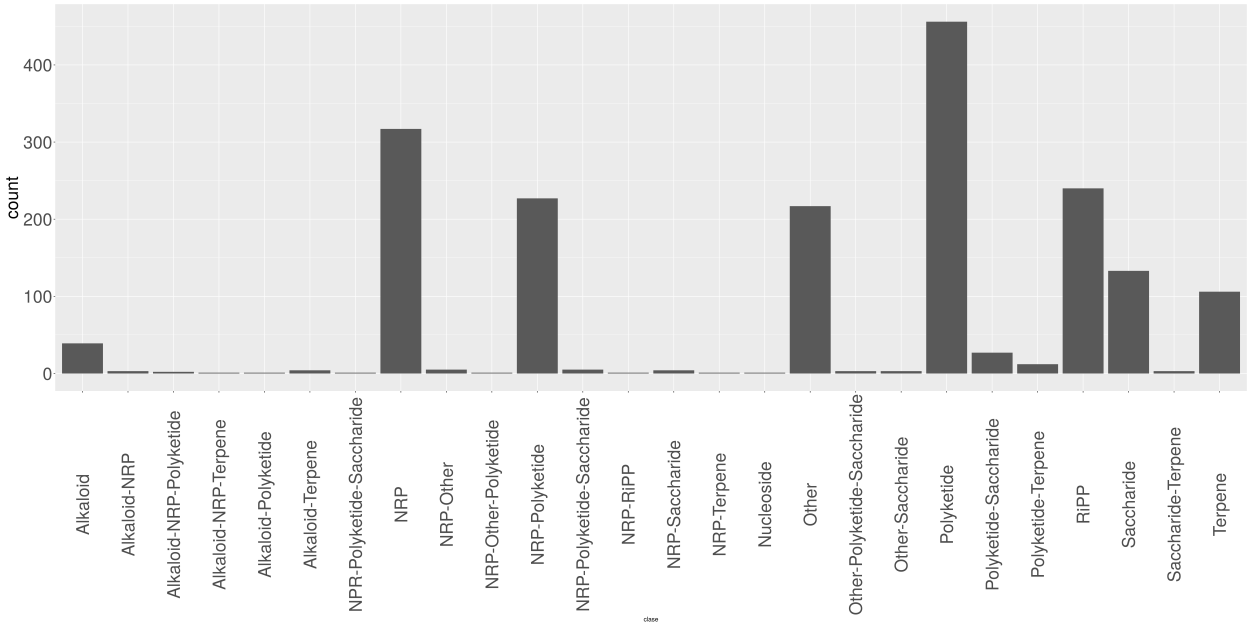


Figure 1: MIBiG classes

Elements	MIBiG class
39	Alkaloid
3	Alkaloid-NRP
2	Alkaloid-NRP-Polyketide
1	Alkaloid-NRP-Terpene
1	Alkaloid-Polyketide
4	Alkaloid-Terpene
1	NPR-Polyketide-Saccharide
317	NRP
5	NRP-Other
1	NRP-Other-Polyketide
227	NRP-Polyketide
5	NRP-Polyketide-Saccharide
1	NRP-RiPP
4	NRP-Saccharide
1	NRP-Terpene
1	Nucleoside
217	Other
3	Other-Polyketide-Saccharide
3	Other-Saccharide
456	Polyketide
27	Polyketide-Saccharide
12	Polyketide-Terpene
240	RiPP
133	Saccharide
3	Saccharide-Terpene
106	Terpene

CARD families and their interaction with BGCs in MIBiG

MIBiG is classified in 24 classes, each one contains a variable number of BGCs
$c_1 = \{c_1^1, c_2^1, c_3^1, ..., c_{n1}^1\}$

$c_2 = \{c_1^2, c_2^2, c_3^2, ..., c_{n1}^1\}$
.
.
.

$c_{24} = \{c_1^{24}, c_2^{24}, c_3^{24}, ..., c_{n24}^{24}\}$

CARD is divided in 645 families each one with a variable number of genes
$f_1 = \{f_1^1, f_2^1, f_3^1, ..., f_{n1}^1\}$

$f_2 = \{f_1^2, f_2^2, f_3^2, ..., f_{n1}^1\}$
.
.
.

$f_{645} = \{f_1^{645}, f_2^{645}, f_3^{645}, ..., f_{n645}^{645}\}$
$A(f, c) = \frac{\#\ hits\ of\ c\ \ over\ f}{(\#f)\times(\#c)}$

Average of blast hits of the family f over the class c where normalized by the size of these sets, $\#f$ and $\#c$ represent the number of elements in family f and class c respectively.

The following analysis finds the most represented CARD families by MIBiG classes. An average about how many hits normalized by the family size and the number of BGCs elements of the class.

The following figure shows MIBiG Clases, and in each class stacks a CARD family average

The following figure shows the average of CARD families by MIBiG classes

```
CARD_BIG_family_averageHorizontal<-ggplot(AverageFamily, aes(x = FO_Family , y = average))+  geom_col(a
                                    face="bold",
                                    family="American Typewriter",
                                    color="tomato",
                                    lineheight=1.2),  # title
```

```
            axis.title.y=element_text(size=1),  # Y axis title
            axis.text.x=element_text(size=1,
                                     angle = 90,
                                     vjust=.5),  # X axis text
            axis.text.y=element_text(size=1))  # Y axis text
#ggsave(file="CARD_BIG_family_averageHorizontal.svg", plot=CARD_BIG_family_averageHorizontal, width=30,
ggsave(file="CARD_BIG_family_averageHorizontal_with_leyend.svg", plot=CARD_BIG_family_averageHorizontal

#bb_AverageFamily <- match_df(AverageFamily, "AAA50325.1|ARO:3003036|oleB", on="id")
## Now I want to substet the families with the highets average #3
# I look into inkscape to determine the 10 highest picks

list<-c("AAA50325.1|ARO:3003036|oleB","AAV85982.1|ARO:3000535|macB","APB03219.1|ARO:3003986|TaeA","CAD7
#APB03216.1|ARO:3003982|LlmA
### Now We substet this list
#AverageFamily$FO_Family[which(
#    AverageFamily$FO_Family == "AAA50325.1|ARO:3003036|oleB"|AverageFamily$FO_Family == "AAV85981.1|AR

bb_AverageFamily <- AverageFamily[AverageFamily$FO_Family %in% list, ]
CARD_BIG_family_averageHorizontalSubset<-  ggplot(bb_AverageFamily, aes(x = FO_Family , y = average))+
                                    face="bold",
                                    family="American Typewriter",
                                    color="tomato",
                                    lineheight=1.2),  # title
            axis.title.y=element_text(size=20),  # Y axis title
            axis.text.x=element_text(size=20,
                                     angle = 90,
                                     vjust=.5),  # X axis text
            axis.text.y=element_text(size=20))  # Y axis text
ggsave(file="CARD_BIG_family_averageHorizontal_with_leyend_Subset.svg", plot=CARD_BIG_family_averageHor
```

From 1811 different BGCs at MIBiG 1240 contains a CARD hit

The following figure shows which CARD families are the most populated classes in MiBIG database. and now an histogram by class with the number of BGCs with diferente frequecnies or CARd by color

which MiBiG classes are over represented as CARD hits !–> ### OLD CODE

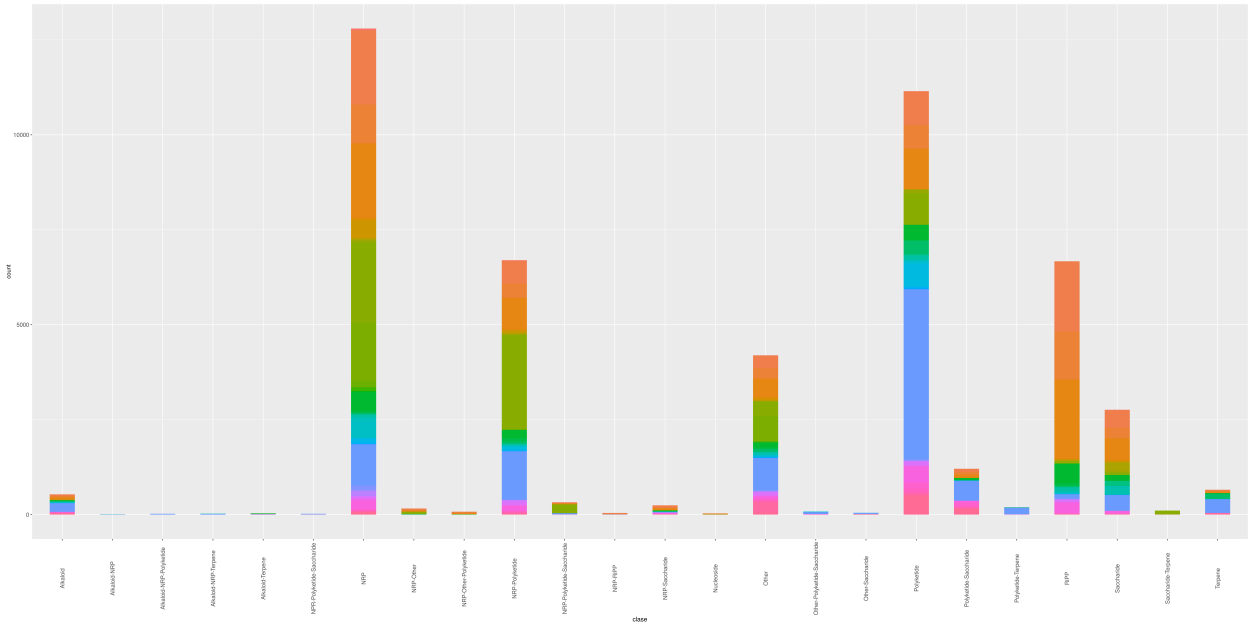
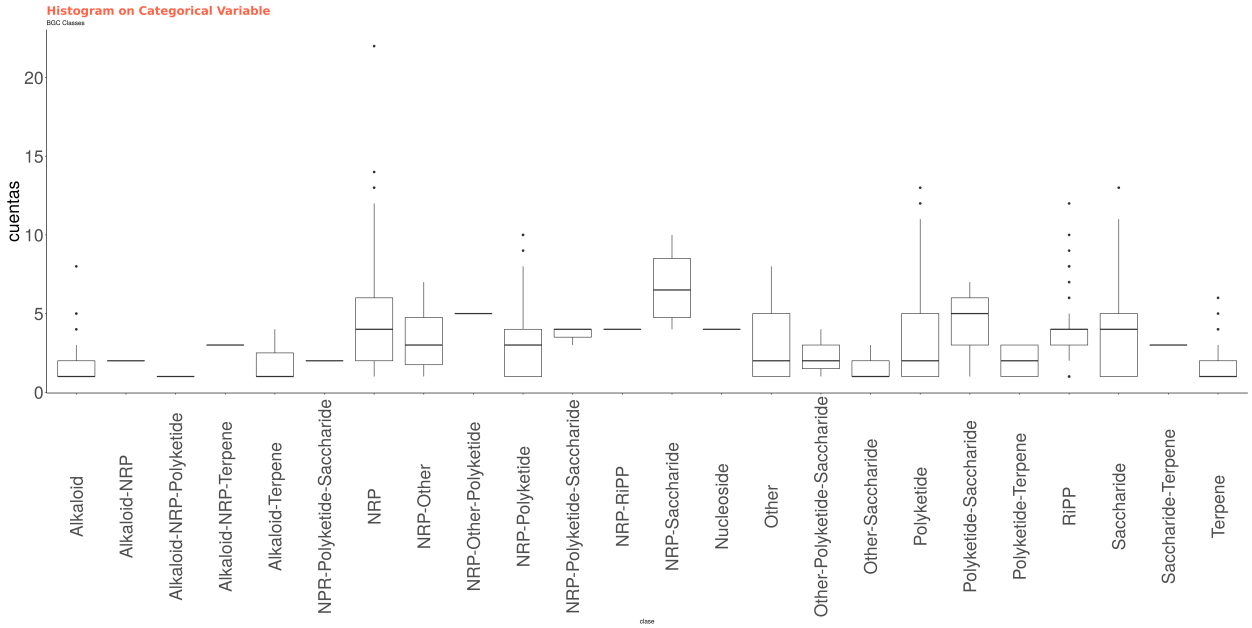


Figure 2: CARD families in MIBiG classes



aaaaaqui voy