

```

ArchaeasHeatPlot <- read.table("chapter3/ArchaeasHeatPlot", header=TRUE, sep="\t")
ArchaeasTaxa <- read.table("chapter3/ArchaeasTaxa", header=TRUE, sep="\t")

## Adding order variable
ArchaeasHeatPlot$order<-c(1:nrow(ArchaeasHeatPlot))

#sorting RastId it accordig to order variable
ArchaeasHeatPlot$RastId <- with(ArchaeasHeatPlot,reorder(ArchaeasHeatPlot$RastId, ArchaeasHeatPlot$order))

# Merging heatplot and taxonomy table into one table
HP_Archaeas_Taxa<-merge(ArchaeasHeatPlot,ArchaeasTaxa,by.x = "RastId",by.y = "RastId")
HP_Archaeas_Taxa.m<- melt(HP_Archaeas_Taxa)

## Using RastId, Name, SuperPhylum, Phylum, Class, Order, Family as id variables

HP_Archaeas_Taxa.m<- ddply(HP_Archaeas_Taxa.m, .(variable), transform,rescale=scale(value)) ##

small<-HP_Archaeas_Taxa.m[which(HP_Archaeas_Taxa.m$RastId=='389007' | HP_Archaeas_Taxa.m$RastId=='390211')

#Este Escogiendo que variables quiero
small<-small[small$variable %in% c("Phosphoglycerate_dehydrogenase", "Isopropyl_malate_synthase", "Alanine_transaminase")]
## use http://colorbrewer2.org/ to find optimal divergent color palette (or set own)

small.m<- ddply(small, .(variable),summarize, mean = round(mean(value), 2),sd = round(sd(value), 2),exp
rownames(small.m)<-small.m$variable

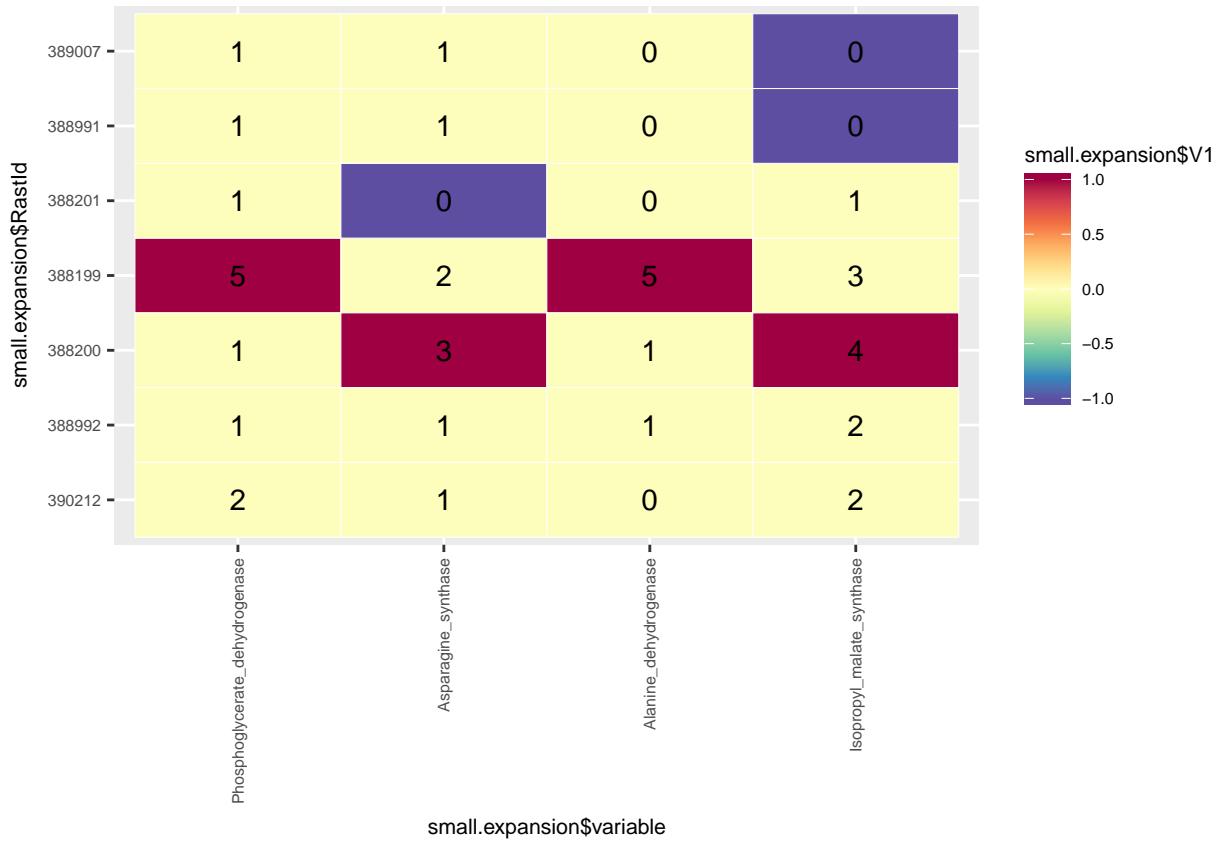
color_exp<-function(x){
  # Pendiente pasarle un dataframe en lugar de tener fijo small.m como variable global
  result=0
  expansion<-NULL
  reduction<-NULL
  local_value=x[1,"value"]
  met_family<-x[1,"variable"]
  Rast=x[1,1]
  #print ("rast",Rast)
  # print(paste(x,"family",met_family,"value",local_value,"Rast",Rast))
  expansion<-small.m$expansion [which(small.m$variable==met_family)]
  reduction<-small.m$reduction [which(small.m$variable==met_family)]
  if (local_value>=expansion){result= 1}
  else if (local_value<=reduction){result= -1}
  return (result)
}

small.expansion<-adply(small,1,color_exp)

#small.expansion$V1
#color_palette <- colorRampPalette(c("#3794bf", "#FFFFFF", "#df8640"))

ggplot(small.expansion, aes(small.expansion$variable, small.expansion$RastId,label=round(small.expansion$mean, 2)))

```



```
#####
## Trying to sort the heatplot
## Reading heatplot table and taxa information and saving it on data.frame data structure
ArchaeasHeatPlot <- read.table("chapter3/ArchaeasHeatPlot", header=TRUE, sep="\t")
ArchaeasTaxa <- read.table("chapter3/ArchaeasTaxa", header=TRUE, sep="\t")
hm.palette <- colorRampPalette(rev(brewer.pal(11, 'Spectral'))), space='Lab')
## Adding order variable
ArchaeasHeatPlot$order<-c(1:nrow(ArchaeasHeatPlot))
#sorting RastId it accordig to order variable
ArchaeasHeatPlot$RastId <- with(ArchaeasHeatPlot,reorder(ArchaeasHeatPlot$RastId, ArchaeasHeatPlot$order))
# Merging heatplot and taxonomy table into one table
HP_Archaeas_Taxa<-merge(ArchaeasHeatPlot,ArchaeasTaxa,by.x = "RastId",by.y = "RastId")
# Melting information leting as variables just enzymatic families and copy number
HP_Archaeas_Taxa.m<- melt(HP_Archaeas_Taxa)

## Using RastId, Name, SuperPhylum, Phylum, Class, Order, Family as id variables

# Cleaning data
HP_Archaeas_Taxa.m<- ddply(HP_Archaeas_Taxa.m, .(variable), transform,value) ##
HP_Archaeas_Taxa.m<- ddply(HP_Archaeas_Taxa.m, .(variable), transform,rescale=scale(value)) ##

HP_Archeas.calcs<- ddply(HP_Archaeas_Taxa.m, .(variable),summarize, mean = round(mean(value), 2),sd = round(sd(value), 2))

rownames(HP_Archeas.calcs)<-HP_Archeas.calcs$variable
#####
#####3
color_exp<-function(x){
```

```

# Pendiente pasarle un dataframe en lugar de tener fijo small.m como variable global
result=0
expansion<-NULL
reduction<-NULL
local_value=x[1,"value"]
met_family<-x[1,"variable"]
Rast=x[1,1]
#print ("rast",Rast)
# print(paste(x,"family",met_family,"value",local_value,"Rast",Rast))
expansion<-HP_Archeas.calcs$expansion [which(small.m$variable==met_family)]
reduction<-HP_Archeas.calcs$reduction [which(small.m$variable==met_family)]
if (local_value>=expansion){result= 1}
else if (local_value<=reduction){result= -1}
return (result)
}
small[ !small$variable %in% c("Contigs", "Size","TOTAL"), ]

```

	RastId	Name		
## 1	388199	Haloferax sp ATCC BAA-644ATCC BAA-644 AOLF01		
## 2	388200	Candidatus Methanoperedens nitroreducensANME-2d JMIY01		
## 3	388201	Marine group II euryarchaeote REDSEA-S40_B11N13 LURX01		
## 551	388991	Uncultured Acidilobus sp MG AYMA01		
## 552	388992	Candidate divison MSBL1 archaeon SCGC-AAA259005 LHXV01		
## 567	389007	Uncultured Acidilobus sp OSP8 AYMC01		
## 690	390212	Archaeoglobus fulgidus DSM 8774 DSM 8774 CP006577.1		
## 14017	388199	Haloferax sp ATCC BAA-644ATCC BAA-644 AOLF01		
## 14018	388200	Candidatus Methanoperedens nitroreducensANME-2d JMIY01		
## 14019	388201	Marine group II euryarchaeote REDSEA-S40_B11N13 LURX01		
## 14567	388991	Uncultured Acidilobus sp MG AYMA01		
## 14568	388992	Candidate divison MSBL1 archaeon SCGC-AAA259005 LHXV01		
## 14583	389007	Uncultured Acidilobus sp OSP8 AYMC01		
## 14706	390212	Archaeoglobus fulgidus DSM 8774 DSM 8774 CP006577.1		
## 63073	388199	Haloferax sp ATCC BAA-644ATCC BAA-644 AOLF01		
## 63074	388200	Candidatus Methanoperedens nitroreducensANME-2d JMIY01		
## 63075	388201	Marine group II euryarchaeote REDSEA-S40_B11N13 LURX01		
## 63623	388991	Uncultured Acidilobus sp MG AYMA01		
## 63624	388992	Candidate divison MSBL1 archaeon SCGC-AAA259005 LHXV01		
## 63639	389007	Uncultured Acidilobus sp OSP8 AYMC01		
## 63762	390212	Archaeoglobus fulgidus DSM 8774 DSM 8774 CP006577.1		
## 68329	388199	Haloferax sp ATCC BAA-644ATCC BAA-644 AOLF01		
## 68330	388200	Candidatus Methanoperedens nitroreducensANME-2d JMIY01		
## 68331	388201	Marine group II euryarchaeote REDSEA-S40_B11N13 LURX01		
## 68879	388991	Uncultured Acidilobus sp MG AYMA01		
## 68880	388992	Candidate divison MSBL1 archaeon SCGC-AAA259005 LHXV01		
## 68895	389007	Uncultured Acidilobus sp OSP8 AYMC01		
## 69018	390212	Archaeoglobus fulgidus DSM 8774 DSM 8774 CP006577.1		
	SuperPhylum	Phylum	Class	Order
## 1	Euryarchaeota	Euryarchaeota	Halobacteria	Haloferacales
## 2	Euryarchaeota	Euryarchaeota	Methanomicrobia	Methanosarcinales
## 3	Euryarchaeota	Euryarchaeota	unclassified	unclassified
## 551	TACK group	Crenarchaeota	Thermoprotei	Acidilobales
## 552	Euryarchaeota	Euryarchaeota	unclassified	unclassified
## 567	TACK group	Crenarchaeota	Thermoprotei	Acidilobales

```

## 690 Euryarchaeota Euryarchaeota Archaeoglobi Archaeoglobales
## 14017 Euryarchaeota Euryarchaeota Halobacteria Haloferacales
## 14018 Euryarchaeota Euryarchaeota Methanomicrobia Methanosarcinales
## 14019 Euryarchaeota Euryarchaeota unclassified unclassified
## 14567 TACK group Crenarchaeota Thermoprotei Acidilobales
## 14568 Euryarchaeota Euryarchaeota unclassified unclassified
## 14583 TACK group Crenarchaeota Thermoprotei Acidilobales
## 14706 Euryarchaeota Euryarchaeota Archaeoglobi Archaeoglobales
## 63073 Euryarchaeota Euryarchaeota Halobacteria Haloferacales
## 63074 Euryarchaeota Euryarchaeota Methanomicrobia Methanosarcinales
## 63075 Euryarchaeota Euryarchaeota unclassified unclassified
## 63623 TACK group Crenarchaeota Thermoprotei Acidilobales
## 63624 Euryarchaeota Euryarchaeota unclassified unclassified
## 63639 TACK group Crenarchaeota Thermoprotei Acidilobales
## 63762 Euryarchaeota Euryarchaeota Archaeoglobi Archaeoglobales
## 68329 Euryarchaeota Euryarchaeota Halobacteria Haloferacales
## 68330 Euryarchaeota Euryarchaeota Methanomicrobia Methanosarcinales
## 68331 Euryarchaeota Euryarchaeota unclassified unclassified
## 68879 TACK group Crenarchaeota Thermoprotei Acidilobales
## 68880 Euryarchaeota Euryarchaeota unclassified unclassified
## 68895 TACK group Crenarchaeota Thermoprotei Acidilobales
## 69018 Euryarchaeota Euryarchaeota Archaeoglobi Archaeoglobales
## Family variable value rescale
## 1 Haloferacaceae Phosphoglycerate_dehydrogenase 5 1.72610114
## 2 Methanoperedenaceae Phosphoglycerate_dehydrogenase 1 -0.60015209
## 3 unclassified Phosphoglycerate_dehydrogenase 1 -0.60015209
## 551 Acidilobaceae Phosphoglycerate_dehydrogenase 1 -0.60015209
## 552 unclassified Phosphoglycerate_dehydrogenase 1 -0.60015209
## 567 Acidilobaceae Phosphoglycerate_dehydrogenase 1 -0.60015209
## 690 Archaeoglobaceae Phosphoglycerate_dehydrogenase 2 -0.01858878
## 14017 Haloferacaceae Asparagine_synthase 2 0.08780760
## 14018 Methanoperedenaceae Asparagine_synthase 3 0.89748609
## 14019 unclassified Asparagine_synthase 0 -1.53154939
## 14567 Acidilobaceae Asparagine_synthase 1 -0.72187090
## 14568 unclassified Asparagine_synthase 1 -0.72187090
## 14583 Acidilobaceae Asparagine_synthase 1 -0.72187090
## 14706 Archaeoglobaceae Asparagine_synthase 1 -0.72187090
## 63073 Haloferacaceae Alanine_dehydrogenase 5 1.53589524
## 63074 Methanoperedenaceae Alanine_dehydrogenase 1 -0.26825391
## 63075 unclassified Alanine_dehydrogenase 0 -0.71929120
## 63623 Acidilobaceae Alanine_dehydrogenase 0 -0.71929120
## 63624 unclassified Alanine_dehydrogenase 1 -0.26825391
## 63639 Acidilobaceae Alanine_dehydrogenase 0 -0.71929120
## 63762 Archaeoglobaceae Alanine_dehydrogenase 0 -0.71929120
## 68329 Haloferacaceae Isopropyl_malate_synthase 3 0.49494473
## 68330 Methanoperedenaceae Isopropyl_malate_synthase 4 1.23231138
## 68331 unclassified Isopropyl_malate_synthase 1 -0.97978855
## 68879 Acidilobaceae Isopropyl_malate_synthase 0 -1.71715520
## 68880 unclassified Isopropyl_malate_synthase 2 -0.24242191
## 68895 Acidilobaceae Isopropyl_malate_synthase 0 -1.71715520
## 69018 Archaeoglobaceae Isopropyl_malate_synthase 2 -0.24242191

```

##Bueno aqui voy

```
##Heat.expansion<-adply(HP_Archaeaes_Taxa.m[!HP_Archaeaes_Taxa.m$variable %in% c("Contigs", "Size", "TOTAL
```

```
#####
#####333
#####

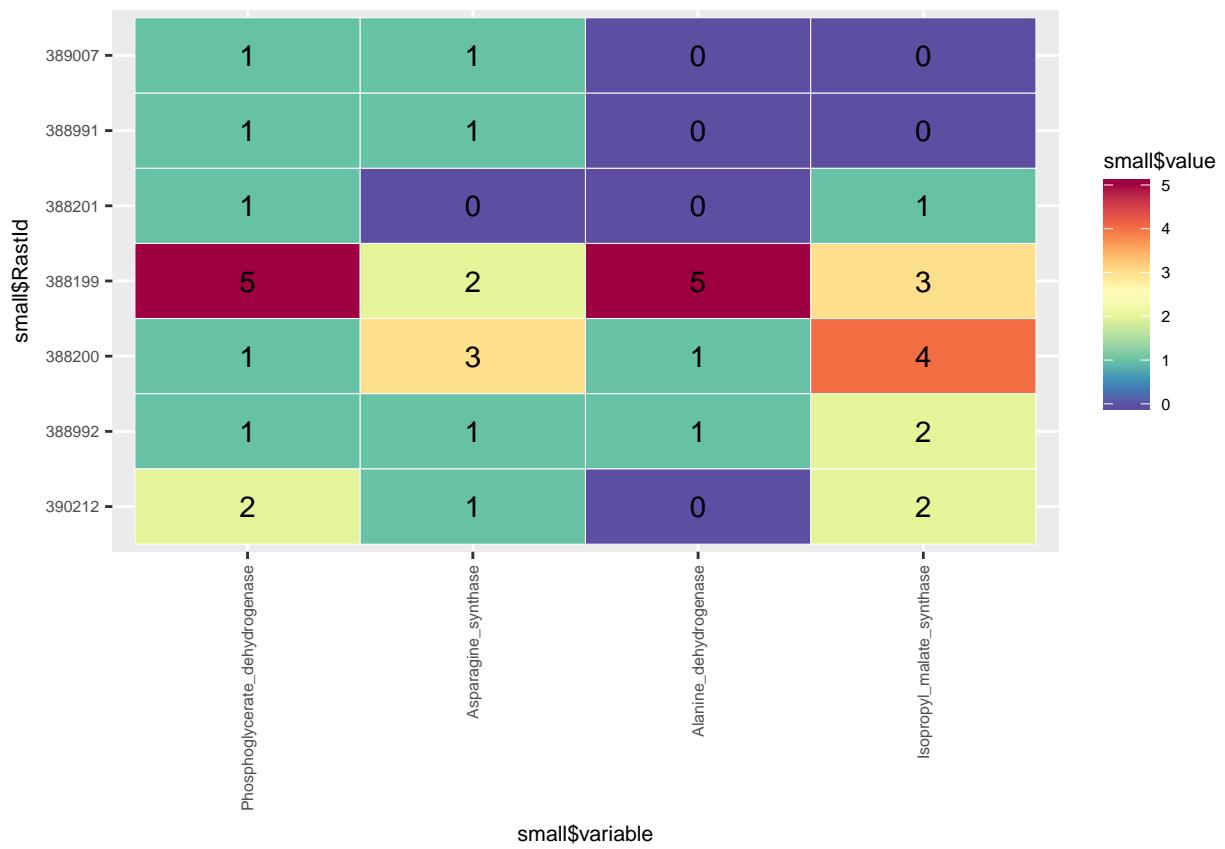
## geom tile with rescale rescala por column
# Graph! with rescale
```

```
ggplot(small, aes(small$variable, small$RastId,label=round(small$rescale,digits=1)))+ geom_tile(aes(fill =
```



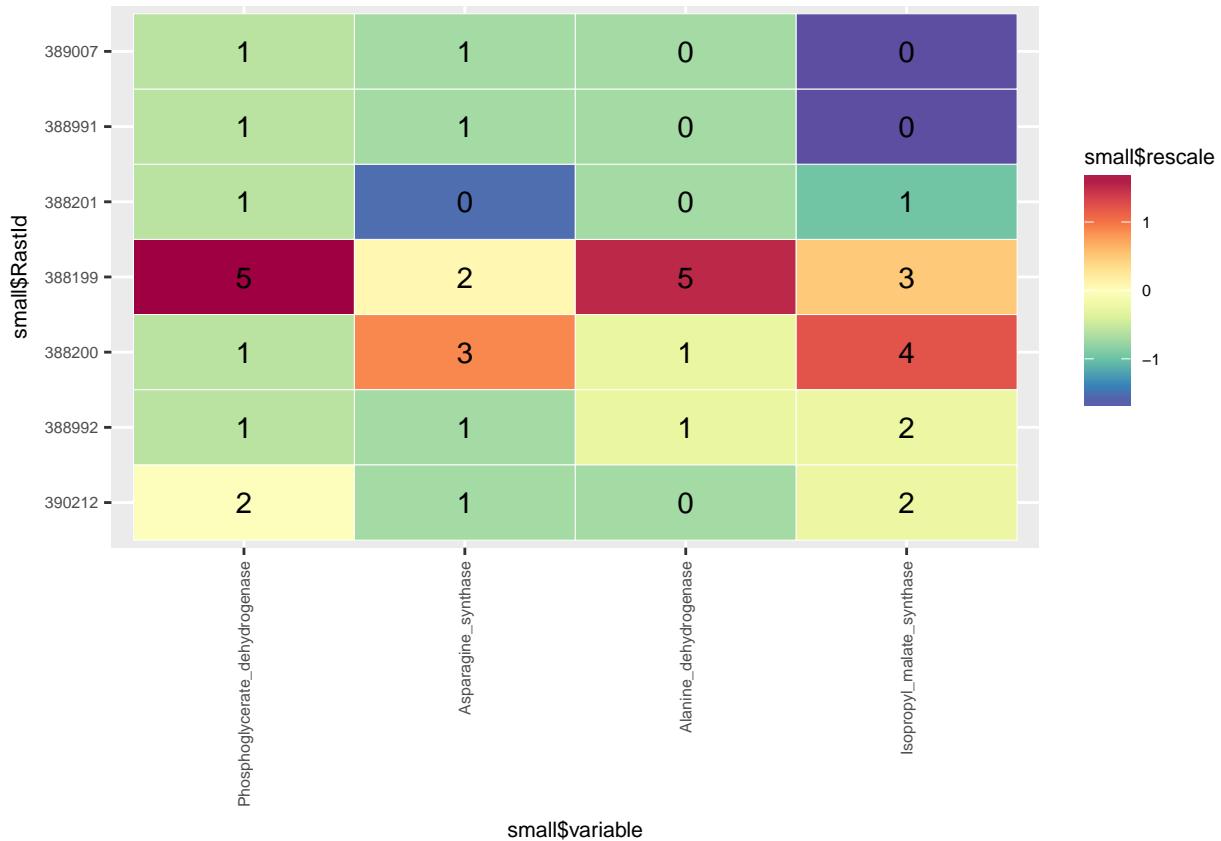
```
## Graph scaled according to whole matrix
```

```
ggplot(small, aes(small$variable, small$RastId,label=round(small$value,digits=1)))+ geom_tile(aes(fill =
```



```
## Graph coloured by columns, showing true values, not scaled
```

```
ggplot(small, aes(small$variable, small$RastId, label=round(small$value,digits=1)))+ geom_tile(aes(fill =
```



```
#####
### Color according to mean plus standar deviation
#####

##### trying to make heatplot colors on dataframe

small.m<- ddply(small, .(variable),summarize, mean = round(mean(value), 2),sd = round(sd(value), 2),exp
rownames(small.m)<-small.m$variable

#####
```

## Archaea EvoMining Results

During the decade between 1970 and 1980, Archaea was recognized as new life domain, a kingdom different from Bacteria and Eucarya in an exciting first great application of 16S phylogeny[@woese\_phylogenetic\_1977,@woese\_are\_1981]. Main differences between this kingdoms are that Archaeal DNA is not arranged in a nucleus as in Eucarya and Archaeal cellular walls are not composed from peptidoglycans as in Bacteria. Archaeal proteins may be highly valuable to biotechnology industry for their great stability due to extreme temperature, PH and salt content conditions on Archeal habitats. Despite no Archaeal Natural products biosynthetic gene clusters (BGC's) has been reported on MiBIG, Archaea do have BGC's, some of them seems to be acquired by horizontal gene transfer (HGT) like methano nrps {search reference}. Other Archeal natural products known are archaeosins, Diketopiperazines, Acyl Homoserine Lactones, Exopolysaccharides, Carotenoids, Biosurfactants, Phenazines and Organic Solutes but this knowledge is not comparable to Bacterial BGC's knowledge[@charlesworth\_untapped\_2015].

Natural products biosynthetic gene clusters search is actually performed using either *high-confidence/low-novelty or low-confidence/high-novelty* bioinformatic approaches [@medema\_computational\_2015]. High confidence methods compares query sequences with previously known BGC's such as nrps or PKS, examples of this algorithms are antiSMASH and clusterfinder [@\_antismash\_????]. EvoMining searches on expansions from central metabolic pathways enzyme families, it has been classified as low confidence/high novelty method. EvoMining has proved useful on Actinobacteria phylum where its use lead to Arseno-compounds discovery [@cruz-morales\_phylogenomic\_2016]. Also on Actinobacteria antiSMASH analysis on 1245 genomes found 774 different classes of natural products, the same analysis on 876 Archaeal genomes, a full kingdom, identifies only 35 BGC's classes. So either Archaea does not have natural products BGC's or this are not yet known. Next paragraph deals with a possible approach about how natural products BGC's can be find.

Archaea resembled Bacteria in that Archaea uses horizontal gene transfer as a genic interchange mechanism, Archaeal genomes contains operons [@howland\_surprising\_2000] and in general there is introns absence{Reference to Computational Methods for Understanding Bacterial and Archaeal Genomes}. Archaeas do have introns, but they are mainly located on genes that encodes ribosomal and transfer RNA [@howland\_surprising\_2000]. General lack of introns allows automatic genome annotation, operons gene organization permits functional inference to a certain degree and HGT contribute to expansions on Archaeal genomes. Some phylum on Archaea has an open pangenome, and as we will show on this chapter some Archaea has central pathway expansions. Enzyme families from central pathways expansions, open pangenome and operon organization made EvoMining succesful on Actinobacteria, this lead us to think that evoMining is suitable to analize Archaeal genomes, even more since EvoMining is a method oriented to use evolution and its not entirelyy based on previous knowledge of BGC's sequences if evolutionary logic behave on Archaea as on bacteria, new BGC's classes may be be found on Archaea.

EvoMining is a trade off between conserved known central metabolic function and enough expansions divergence on sequence and on clusters to divergence

## Tables

Table 1: Families on Archaeabacteria

Factors	Correlation between Parents & Child
GenomeDB	876
Phylum	12
Order	23

# Are genome expansions driving metabolic evolution in the Kingdom Archaea?

Nelly Sélem, Pablo Cruz, Christian Martínez, Francisco Barona  
nelly.selem@cinvestav.mx



## Abstract

On 1977 Carl Woese based on phylogenetic analysis identifies Archaea as a separate kingdom from Bacteria. There are 850 Archaeal genomes available at NCBI, where it is possible to study their genomic expansions. Even though it is known Archaea produces some natural products they are poorly detected by homology-model based genome mining approaches. Therefore Archaea doesn't produce secondary metabolites, or they are produced by biosynthetic gene clusters really different from bacteria and Fungi. Evolutive genome mining strategies based on genome expansions may help to discover this different pathways.

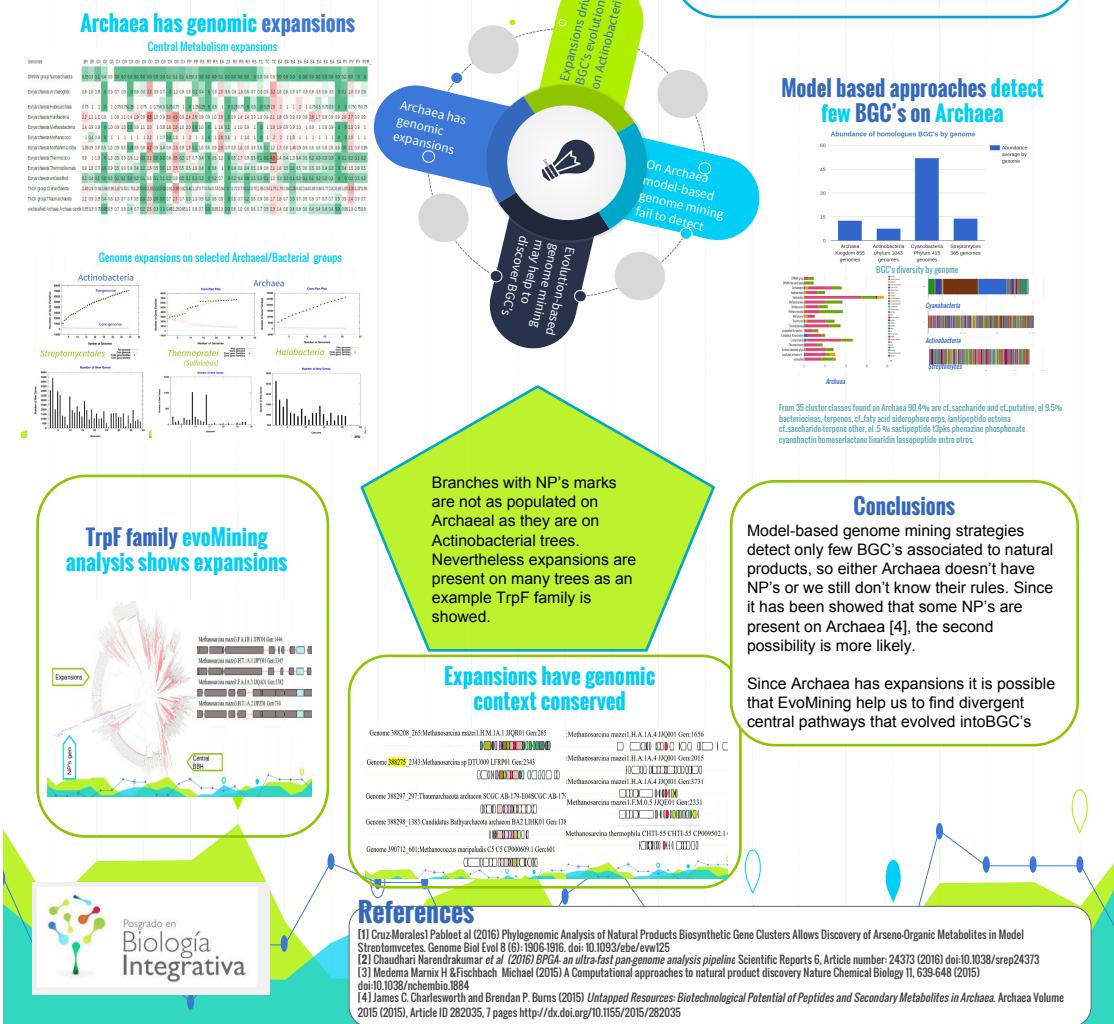


Figure 1: EvoMining Archaeas

First lets investigate if Archaea has expansions on families within central metabolic routes. Since main metabolic pathways are shared between Bacteria and Archaea makes sense to assemble Archeal EvoMining central database by using orthologous from Actinobacterial evoMining central pathways.

## Expansions BoxPlot by metabolic family

```
label(path = "chapter3/expansion_plotArchaeas.pdf", caption = "Expansions Boxplot", label = "Archaea_exp")
```

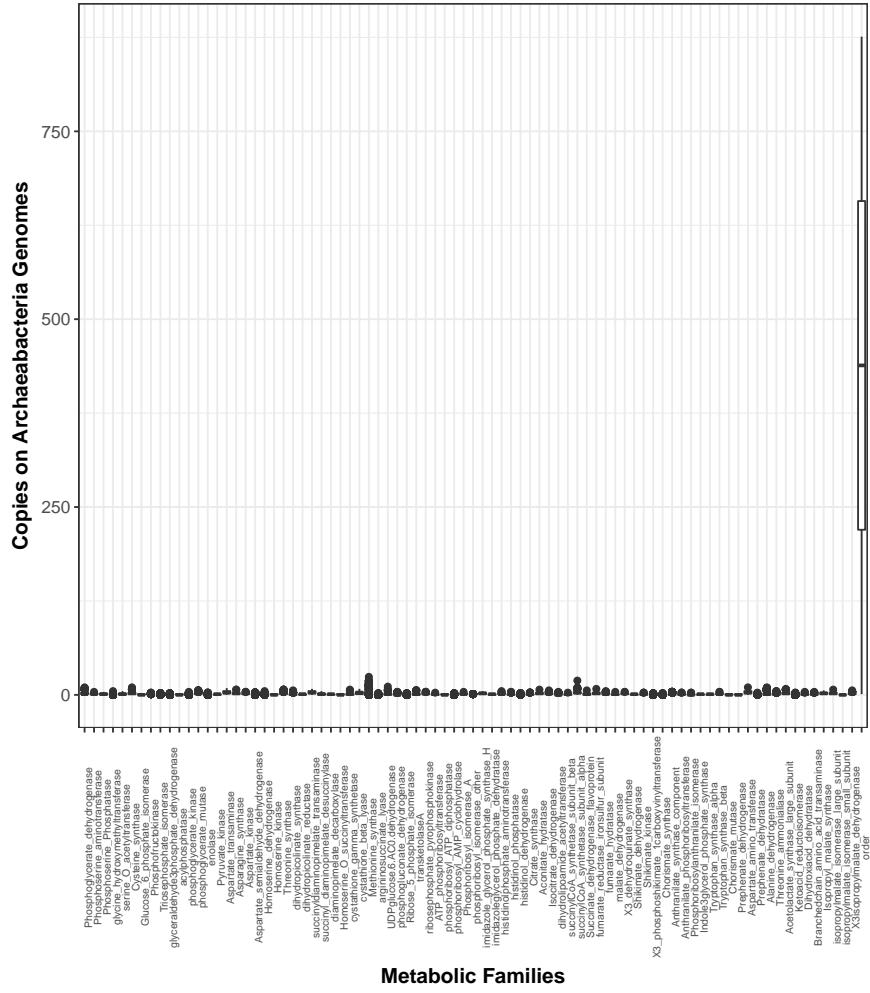


Figure 2: Expansions Boxplot

Here is a reference to the expansion boxplot: Figure 2.

## Expansions BoxPlot by metabolic family by phylum

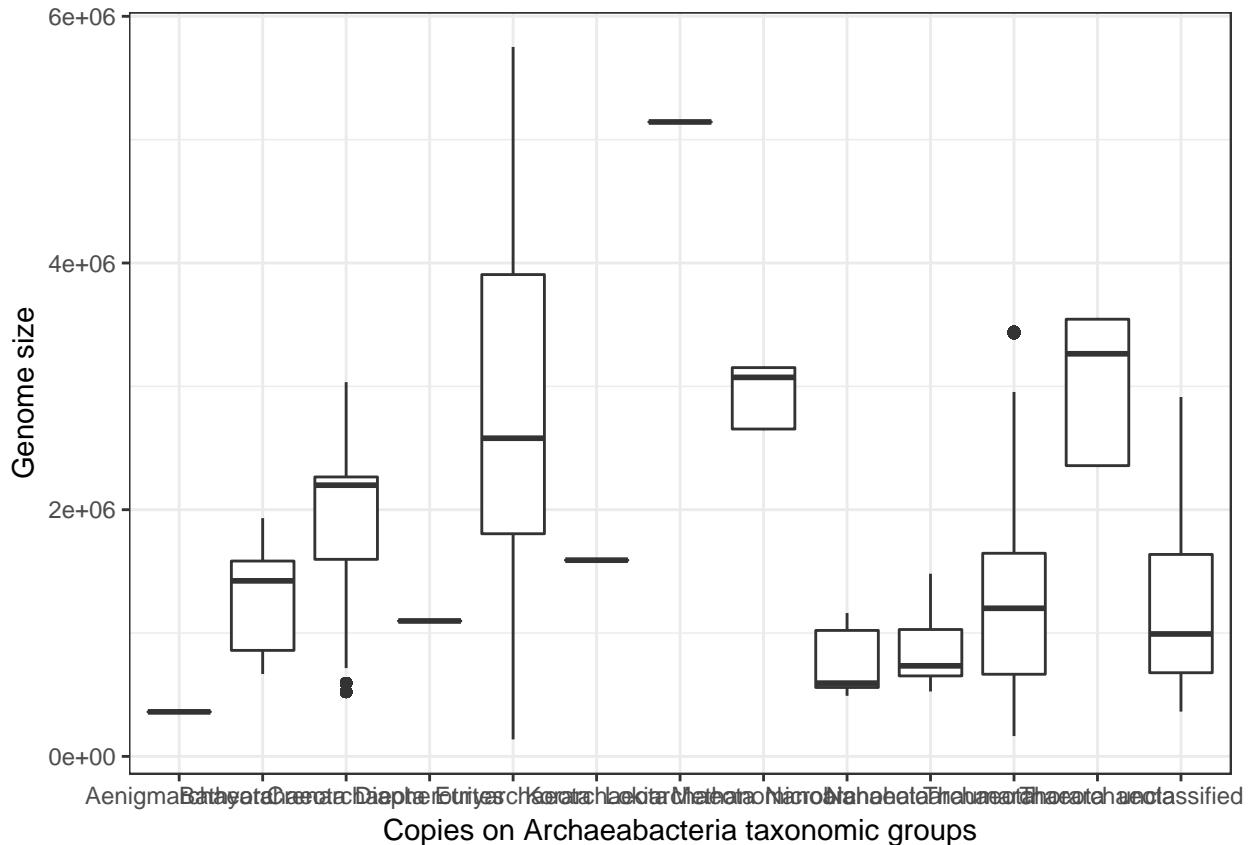
```
#+ geom_jitter()
#aes(fill = factor(vs))

ArchaeasTotalBP.m<-merge(ArchaeasHeatPlot,ArchaeasTaxa,by.x="RastId",by.y="RastId") ## works as expected
ArchaeasHeatPlotBP.m <- melt(ArchaeasTotalBP.m,id =c("RastId","Name","SuperPhylum","Phylum","Class","Order"))
ArchaeasHeatPlotBP.m<-subset(ArchaeasHeatPlotBP.m,variable!="TOTAL") ## works as expected
ArchaeasHeatPlotBP.m<-subset(ArchaeasHeatPlotBP.m,variable!="TOTAL") ## works as expected

## Each metabolic pathway se parte por phylum coloreado por order

#3PGA_AMINOACIDS
#Glycolysis
#OXALACETATE_AMINOACIDS
#R5P_AMINOACIDS
#TCA
#E4P_AMINO_ACIDS
#PYR_THR_AA

## Genome size
ggplot(ArchaeasHeatPlotBP.m, aes(x=ArchaeasHeatPlotBP.m$Phylum, y=ArchaeasHeatPlotBP.m$Size))+ geom_boxplot()
```

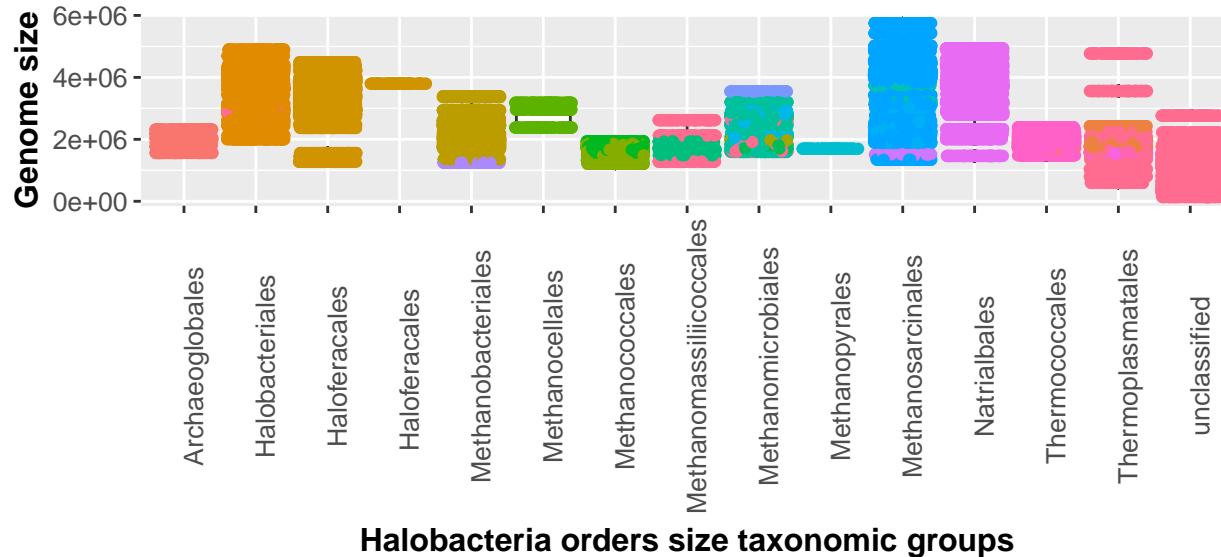


```
#+ geom_jitter(aes(color=ArchaeasHeatPlotBP.m$Phylum))
```

```
## Halobacteria
```

```
MetFam_BP.m=subset(ArchaeasHeatPlotBP.m, Phylum=="Euryarchaeota")
```

```
ggplot(MetFam_BP.m, aes(x=MetFam_BP.m$order, y=MetFam_BP.m$Size))+ geom_boxplot() +theme(plot.title = element_text(size = 14, face = "bold"), text = element_text(size = 12), axis.title = element_text(size = 12, face = "bold"))
```



haeoglobaceae	● Methanocalculaceae	● Methanomassiliicoccaceae	● Methanosaetaceae
roplasmaceae	● Methanocaldococcaceae	● Methanomicrobiaceae	● Methanosarcinaceae
obacteriaceae	● Methanocellaceae	● Methanoperedenaceae	● Methanospirillaceae
oferacaceae	● Methanococcaceae	● Methanopyraceae	● Methanothermaceae
thanobacteriaceae	● Methanocorpusculaceae	● Methanoregulaceae	● Methermicoccaceae

```
#MetFam_BP.m=subset(ArchaeasHeatPlotBP.m, Family=="Methanosaetaceae")
```

```
#ggplot(MetFam_BP.m, aes(x=MetFam_BP.m$Size, y=MetFam_BP.m$value))
```

```
+theme(plot.title = element_text(size = 14, face = "bold"), text = element_text(size = 12), axis.title = element_text(size = 12, face = "bold"))
```

```
#geom_jitter(aes(color=ArchaeasHeatPlotBP.m$Phylum))# + facet_grid(. ~ Phylum)+theme_bw()
```

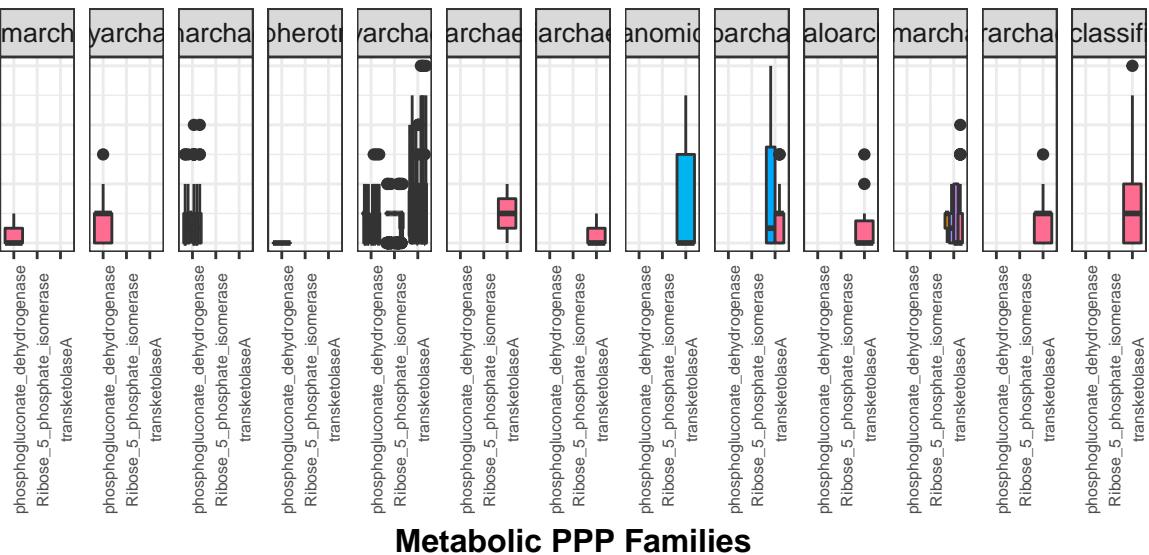
```
## Metabolic Pathways
```

```
MetFam=subset(ArchaeasCentral, Pathway=="PPP")
```

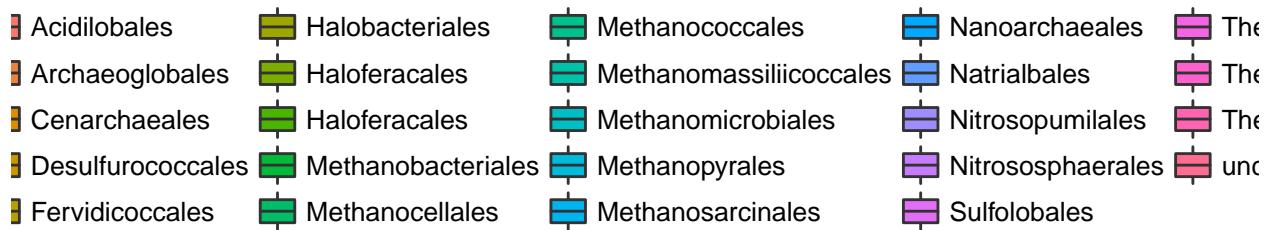
```
MetFam_BP.m=ArchaeasHeatPlotBP.m[ArchaeasHeatPlotBP.m$variable %in% MetFam$Enzyme,]
```

```
ggplot(MetFam_BP.m, aes(x=MetFam_BP.m$variable, y=MetFam_BP.m$value, fill=Order))+ labs(x = "Metabolic Pathways", y = "Value")
```

Copies on Archaeabacteria G



## Metabolic PPP Families



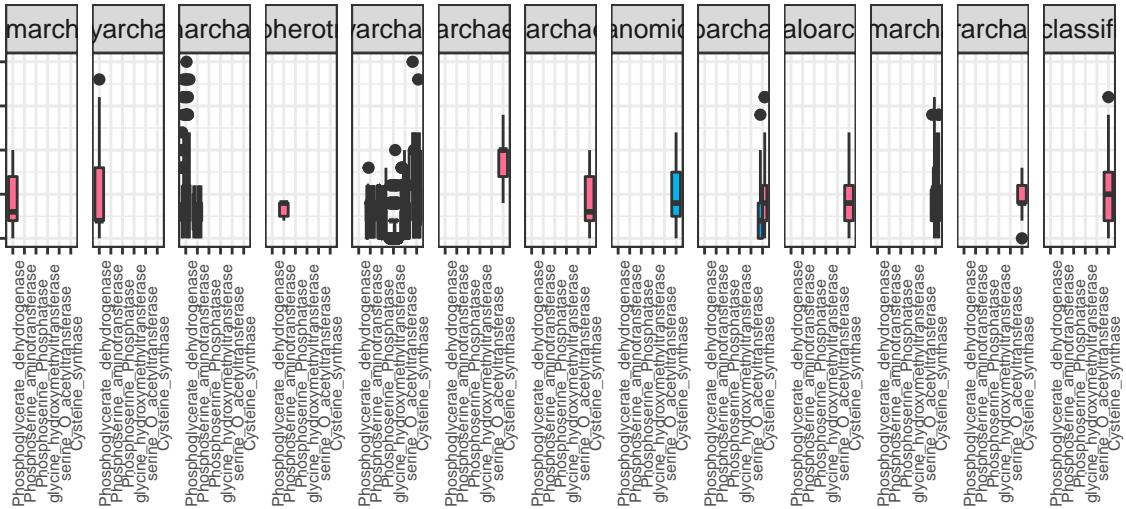
```
#+ geom_jitter(aes(color=MetFam_BP.m$Phylum))
```

```
MetFam=subset(ArchaeaCentral,Pathway=="3PGA_AMINOACIDS")
```

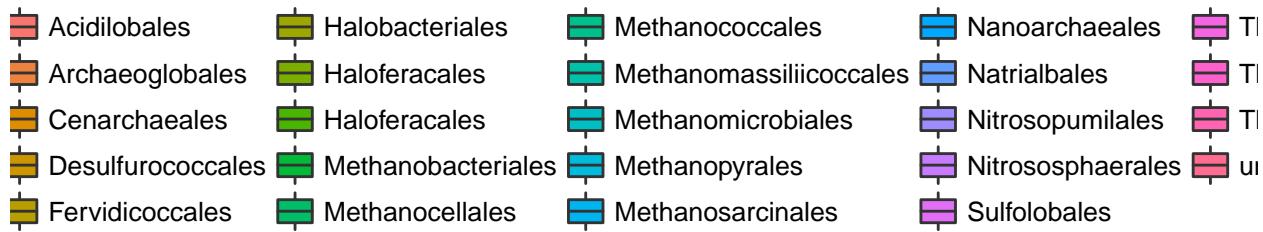
```
MetFam_BP.m=ArchaeasHeatPlotBP.m[ArchaeasHeatPlotBP.m$variable %in% MetFam$Enzyme,]
```

```
ggplot(MetFam_BP.m, aes(x=MetFam_BP.m$variable, y=MetFam_BP.m$value, fill=Order))+ labs(x = "Metabolic
```

Copies on Archaeabacteria G



## Metabolic PGA\_AMINOACIDS Families



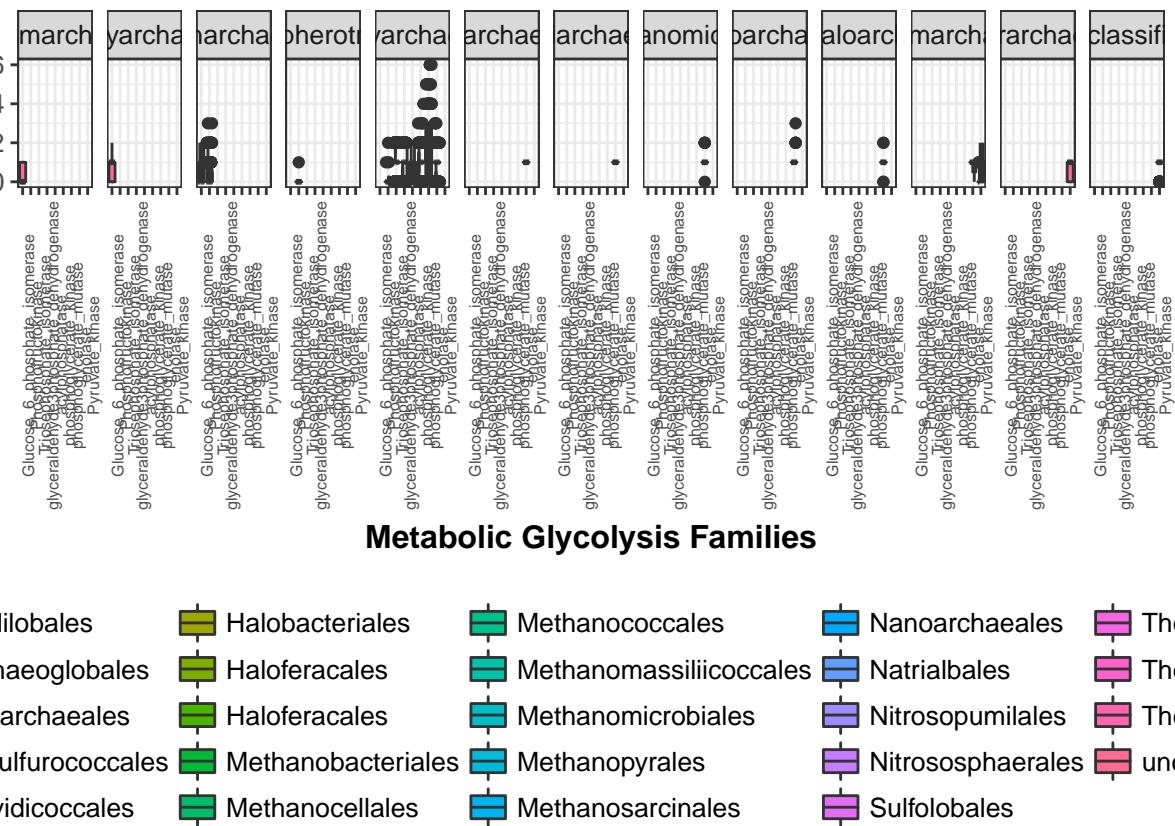
```
#+ geom_jitter(aes(color=MetFam_BP.m$Phylum))
```

```
MetFam=subset(ArchaeaCentral,Pathway=="Glycolysis")
```

```
MetFam_BP.m=ArchaeasHeatPlotBP.m[ArchaeasHeatPlotBP.m$variable %in% MetFam$Enzyme,]
```

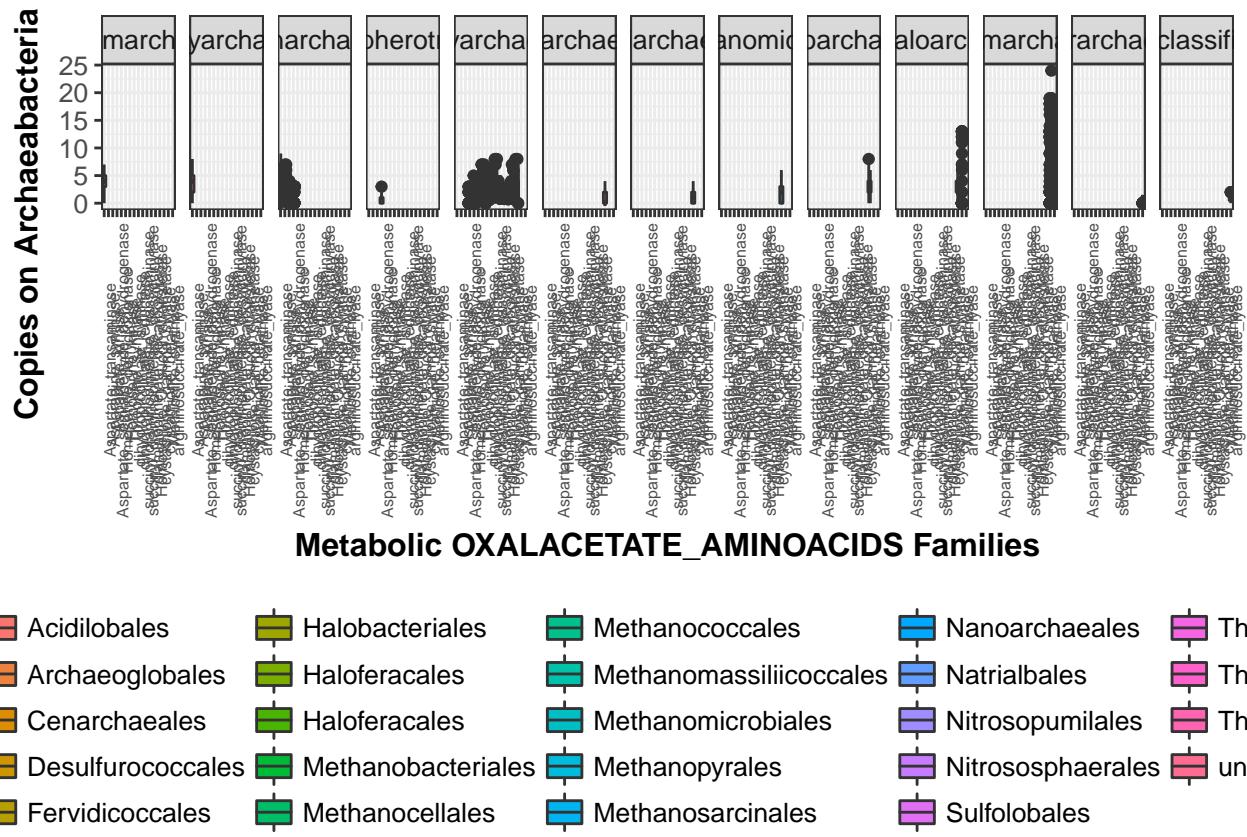
```
ggplot(MetFam_BP.m, aes(x=MetFam_BP.m$variable, y=MetFam_BP.m$value, fill=Order))+ labs(x = "Metabolic
```

Copies on Archaeabacteria



```
#+ geom_jitter(aes(color=MetFam_BP.m$Phylum))

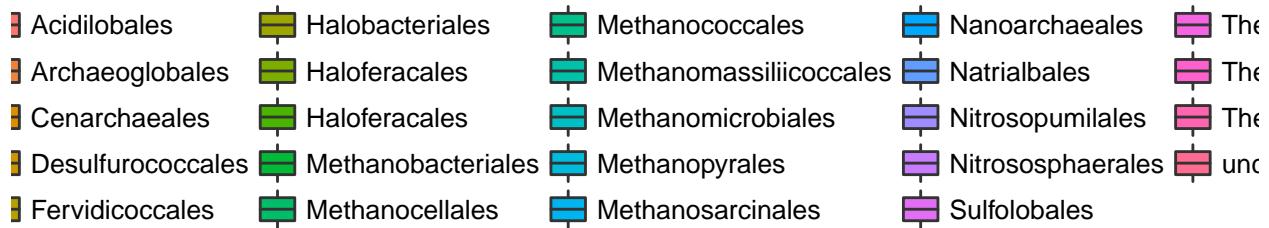
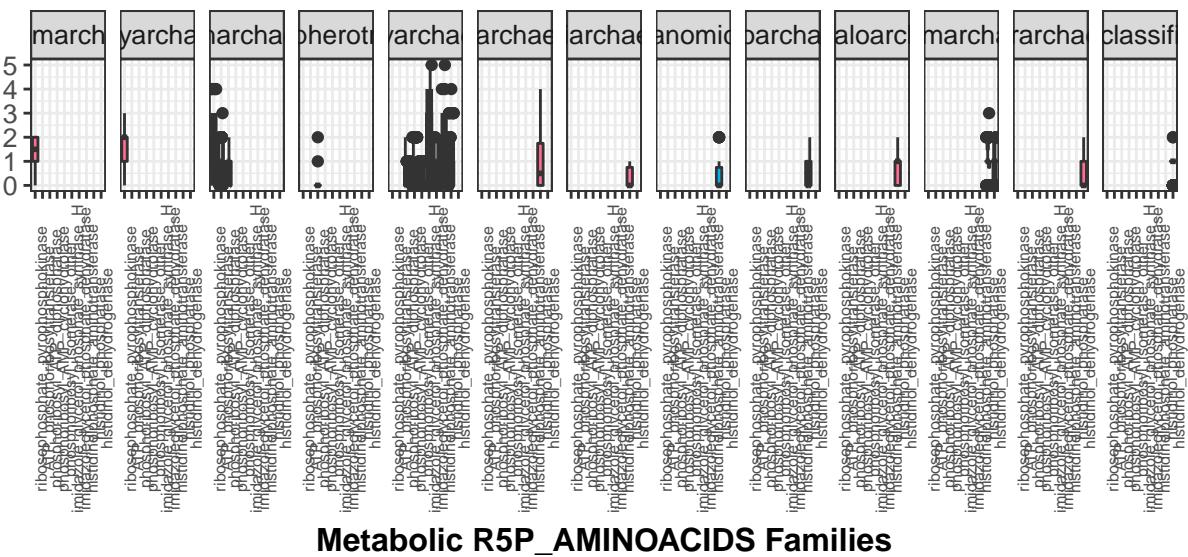
MetFam=subset(ArchaeaesCentral,Pathway=="OXALACETATE_AMINOACIDS")
MetFam_BP.m=ArchaeaesHeatPlotBP.m[ArchaeaesHeatPlotBP.m$variable %in% MetFam$Enzyme,]
ggplot(MetFam_BP.m, aes(x=MetFam_BP.m$variable, y=MetFam_BP.m$value, fill=Order))+ labs(x = "Metabolic
```



```
#+ geom_jitter(aes(color=MetFam_BP.m$Phylum))

MetFam=subset(ArchaeaHeatPlotBP.m,Pathway=="R5P_AMINOACIDS")
MetFam_BP.m=ArchaeaHeatPlotBP.m[ArchaeaHeatPlotBP.m$variable %in% MetFam$Enzyme,]
ggplot(MetFam_BP.m, aes(x=MetFam_BP.m$variable, y=MetFam_BP.m$value, fill=Order))+ labs(x = "Metabolic Pathways", y = "Relative abundance", color = "Phylum") +
```

## Copies on Archaeabacteria



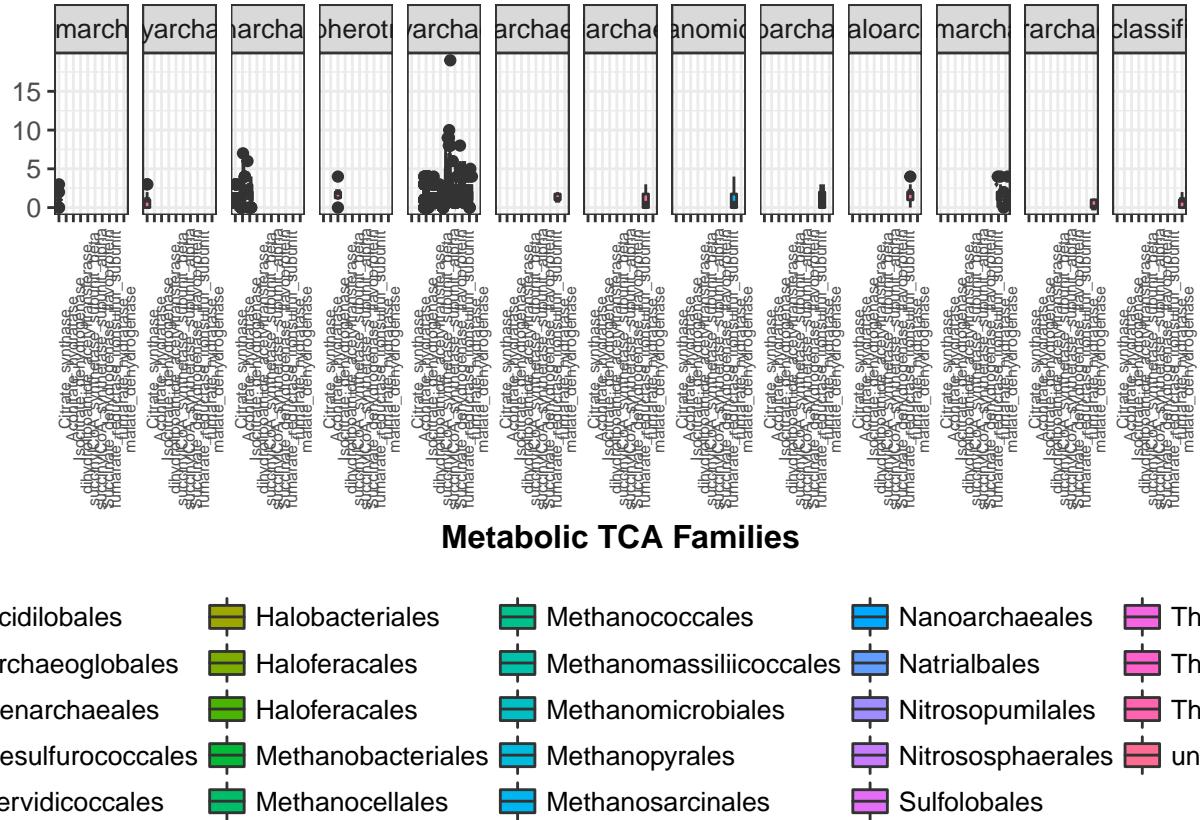
```

## geom_jitter(aes(color=MetFam_BP.m$Phylum))

#
MetFam=subset(ArchaeasCentral,Pathway=="TCA")
MetFam_BP.m=ArchaeasHeatPlotBP.m[ArchaeasHeatPlotBP.m$variable %in% MetFam$Enzyme,]
ggplot(MetFam_BP.m, aes(x=MetFam_BP.m$variable, y=MetFam_BP.m$value, fill=Order))+ labs(x = "Metabolic"

```

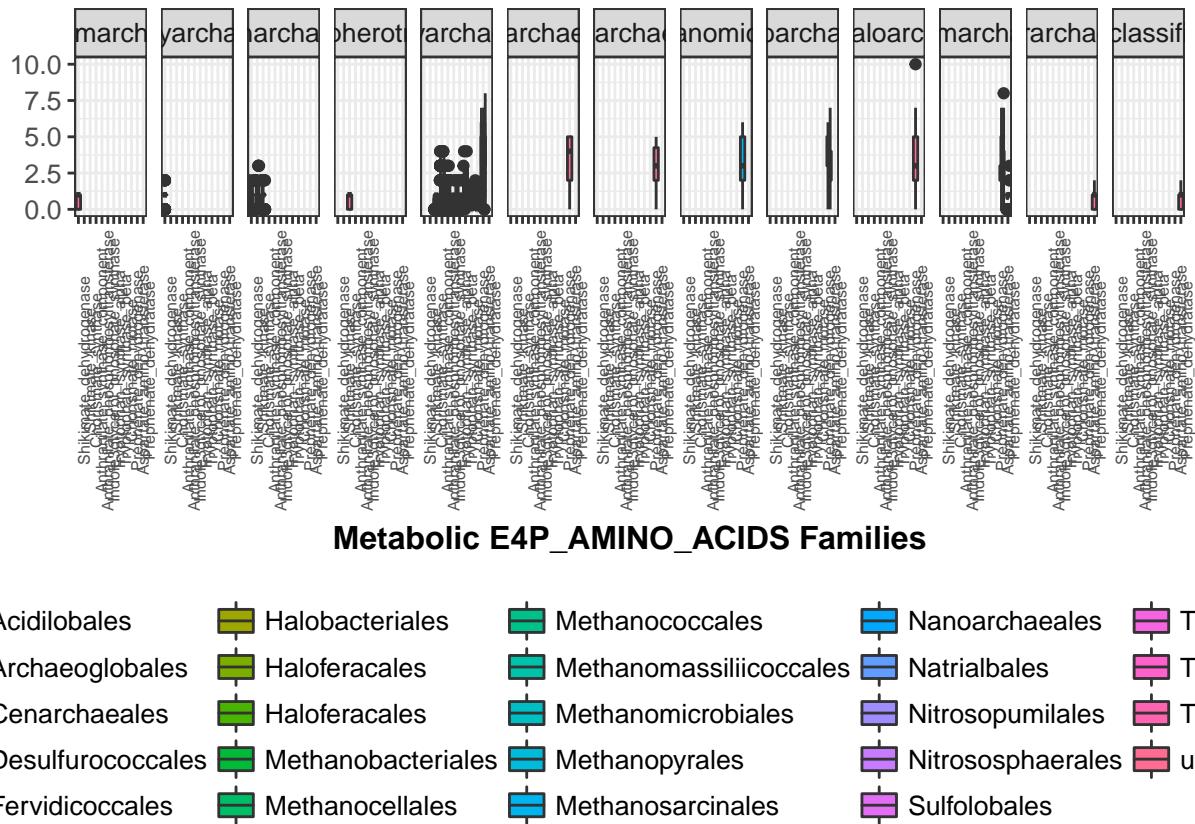
## Copies on Archaeabacteria (



```
#+ geom_jitter(aes(color=MetFam_BP.m$Phylum))

MetFam=subset(ArchaeasCentral,Pathway=="E4P_AMINO_ACIDS")
MetFam_BP.m=ArchaeasHeatPlotBP.m[ArchaeasHeatPlotBP.m$variable %in% MetFam$Enzyme,]
ggplot(MetFam_BP.m, aes(x=MetFam_BP.m$variable, y=MetFam_BP.m$value, fill=Order))+ labs(x = "Metabolic TCA Families")
```

## Copies on Archaeabacteria (



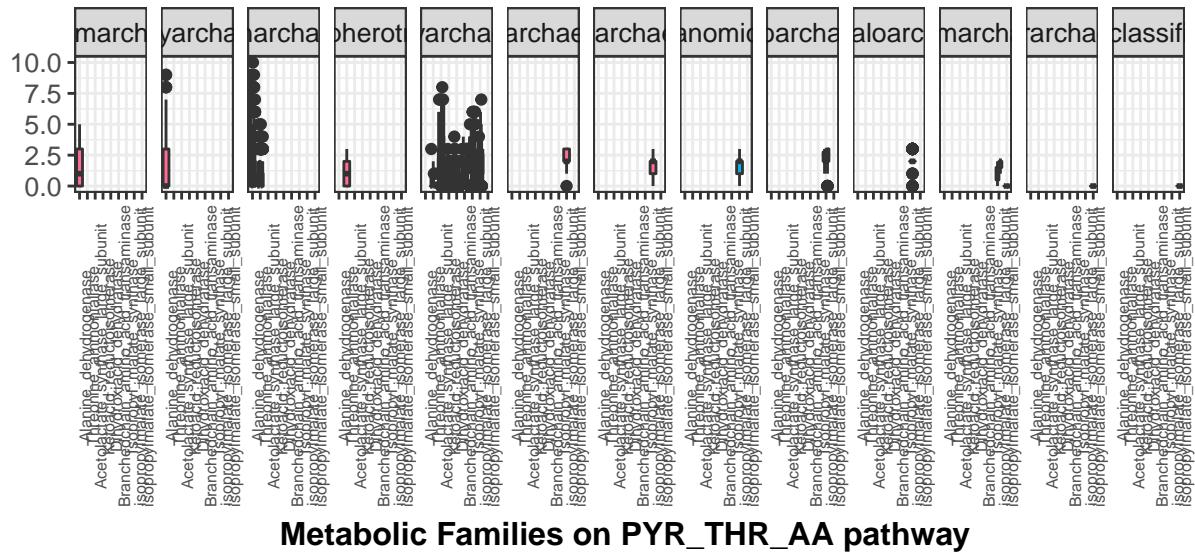
```
#+ geom_jitter(aes(color=MetFam_BP.m$Phylum))
```

```
MetFam=subset(ArchaeasCentral,Pathway=="PYR_THR_AA")
```

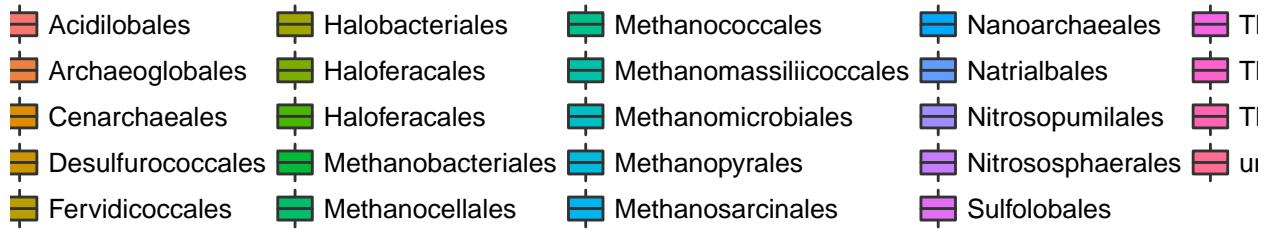
```
MetFam_BP.m=ArchaeasHeatPlotBP.m[ArchaeasHeatPlotBP.m$variable %in% MetFam$Enzyme,]
```

```
ggplot(MetFam_BP.m, aes(x=MetFam_BP.m$variable, y=MetFam_BP.m$value, fill=Order))+ labs(x = "Metabolic Enzyme")
```

## Copies on Archaeabacteria



Metabolic Families on PYR\_THR\_AA pathway

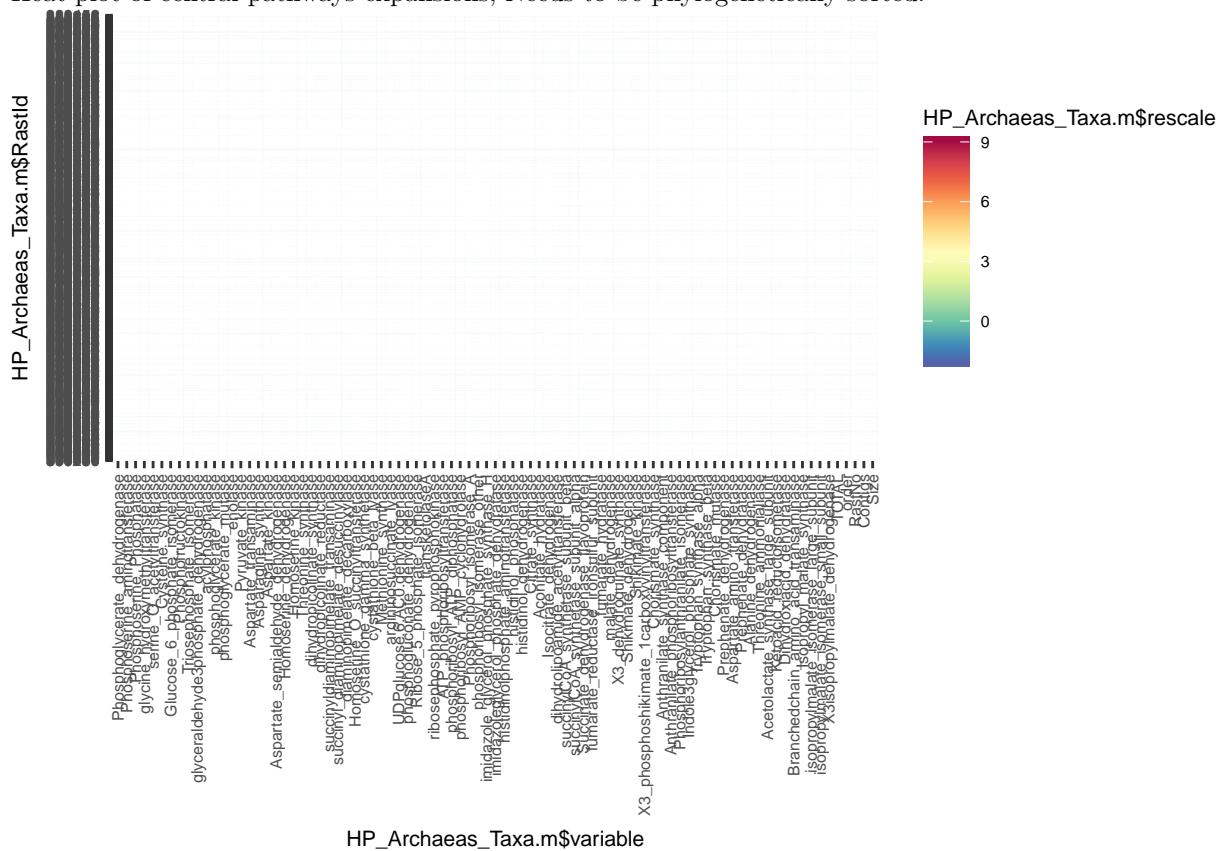


```
#+ geom_jitter(aes(color=MetFam_BP.m$Phylum))
```

```
#ggsave("chapter3/expansion_plotArchaea.pdf", plot = expansion_plotArchaea, height = 8, width = 7)
```

## Central pathway expansions

Heat plot of central pathways expansions, Needs to be phylogenetically sorted.



Here is a reference to the HeatPlot: Figure 3.

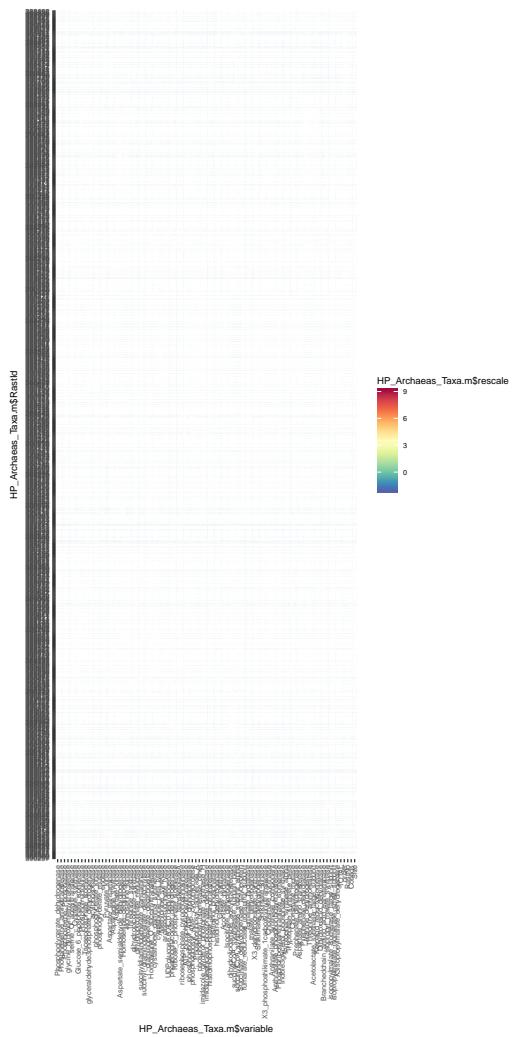


Figure 3: Archaeas Heatplot

## Genome Size correlations

### Correlation between genome size and AntiSMASH products

Genome size vs Total antismash cluster coloured by order

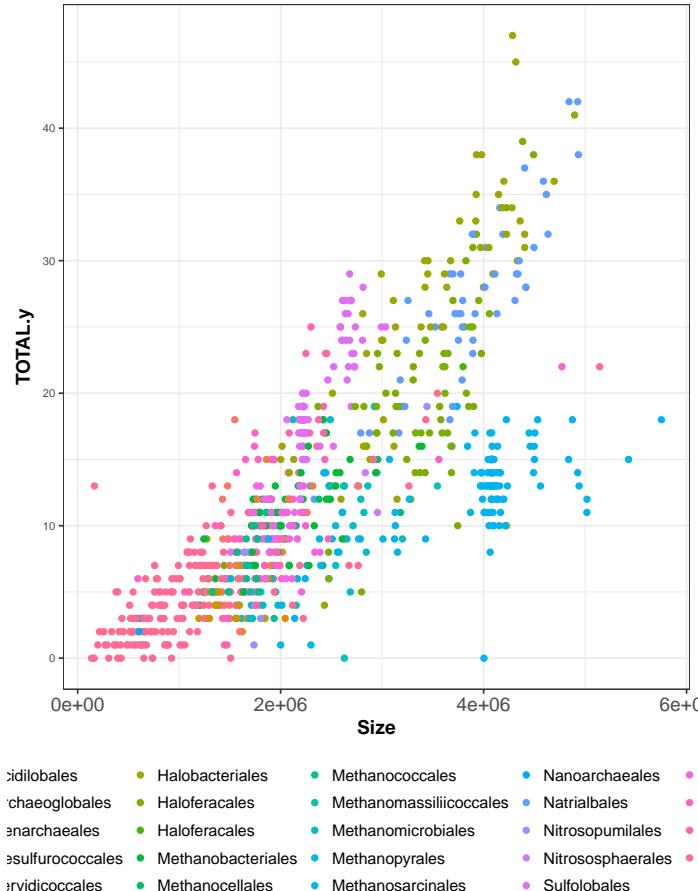


Figure 4: Correlation between Archaea genome size and antismash Natural products detection colored by Order

Here is a reference to Genome size vs Total antismash cluster: Figure 4.

Genome size vs Total antismash cluster detected splitted by order

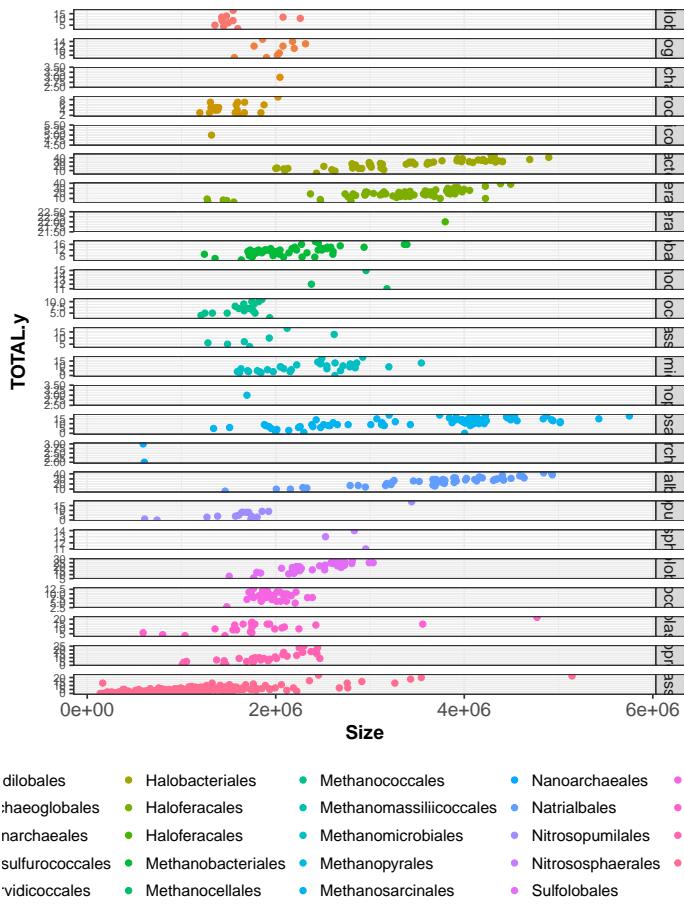


Figure 5: Correlation between Archaea genome size and antismash Natural products detection grided by Order

Here is a reference to Correlation between genome size and antismash Natural products detection grided by Order plot: Figure 5.

## Correlation between genome size and Central pathway expansions

Genome size vs Total central pathway expansion coloured by order

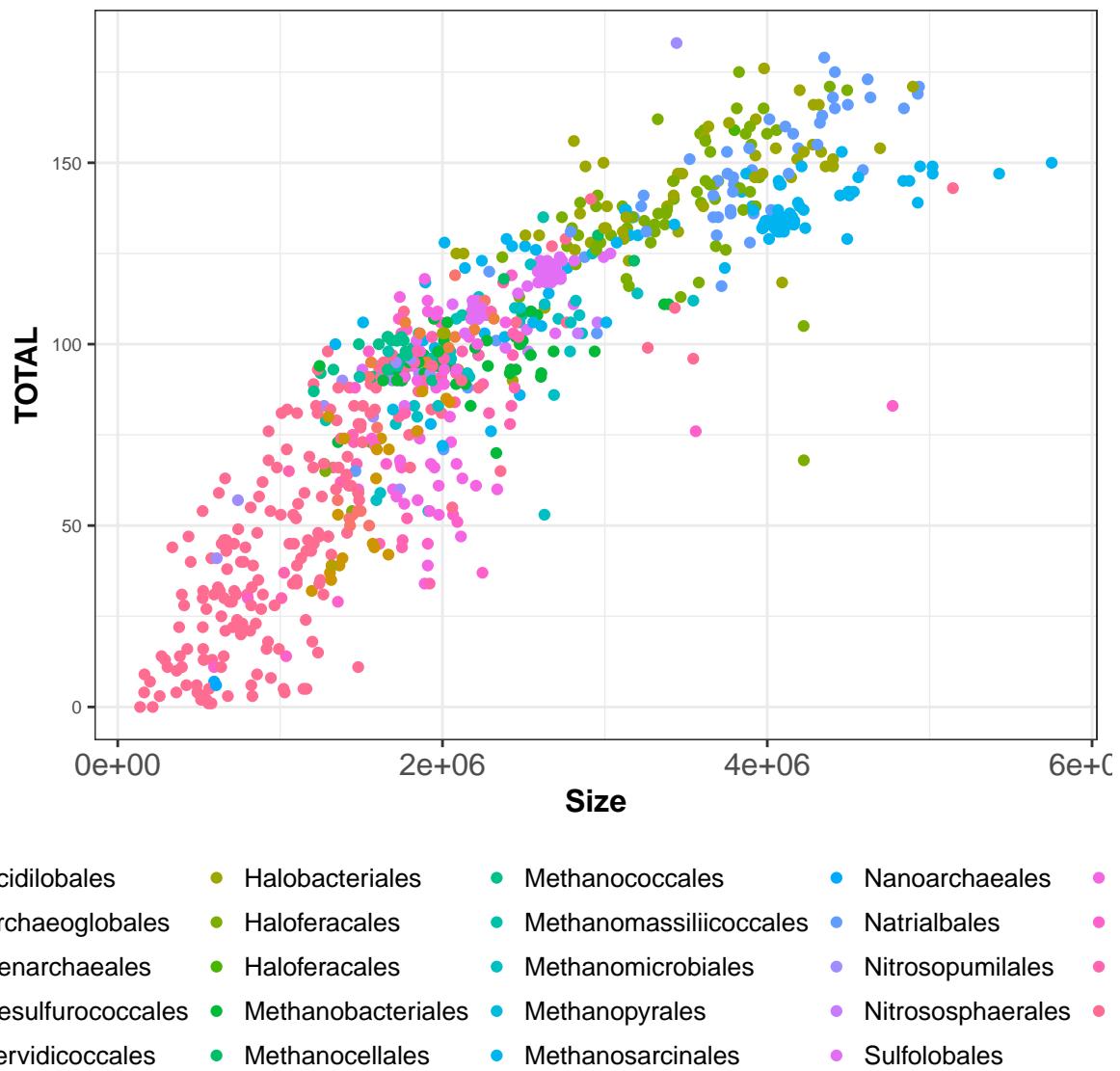


Figure 6: Correlation between Archaea genome size and central pathway expansions

Here is a reference to the size vs Total central pathway expansion plot: Figure 6.

Genome size vs Total central pathway expansion grided by order

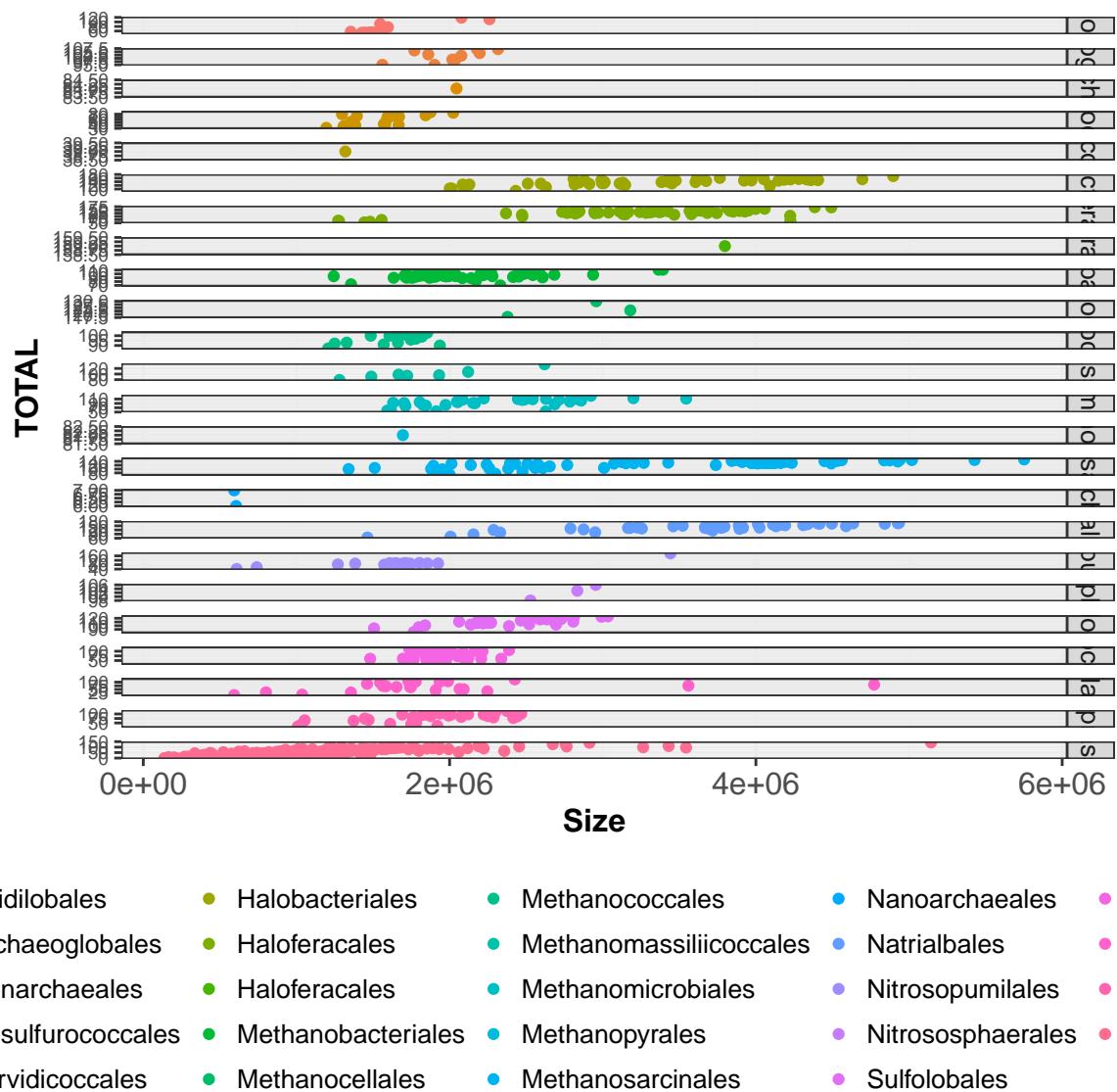


Figure 7: Correlation between Archaea genome size and central pathway expansion grided by order

Here is a reference to the Genome size vs Total central pathway expansion grided by order plot: Figure 7.

Correlation between genome size and each of the central pathway families. Data are coloured by metabolic family instead of coloured by taxonomical order. This treatment allows to answer how different metabolic families grows when genome size grow.

Also I want to add form given by taxonomical order.

```
## Warning: The shape palette can deal with a maximum of 6 discrete values
## because more than 6 becomes difficult to discriminate; you have
## 24. Consider specifying shapes manually if you must have them.

## Warning: Removed 65604 rows containing missing values (geom_point).
```

Genome size vs Total central pathway expansion coloured by metabolic Family

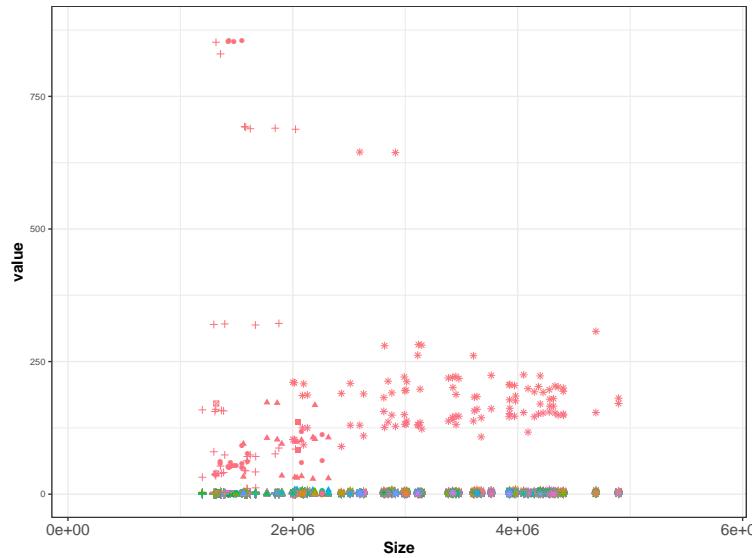


Figure 8: Correlation between Archaea Genome size vs Total central pathway expansion coloured by metabolic Family

Here is a reference to the Genome size vs Total central pathway expansion coloured by metabolic Family plot: Figure 8.

Future Work: Genome size vs Total central pathway expansion grided by metabolic Family For clarity I need to also grid and group by Metabolic Pathway

Here is a reference to Genome size vs Total central pathway expansion grided by metabolic Family plot: ??.

## Natural products

### Natural products recruitments from EvoMining heatplot

We can see natural products recruitment after central pathways expansions colored by their kingdom.  
Natural products recruited by metabolic family, colored by phylogenetic origin.

Recruitments after central pathways expansions coloured by Kingdom

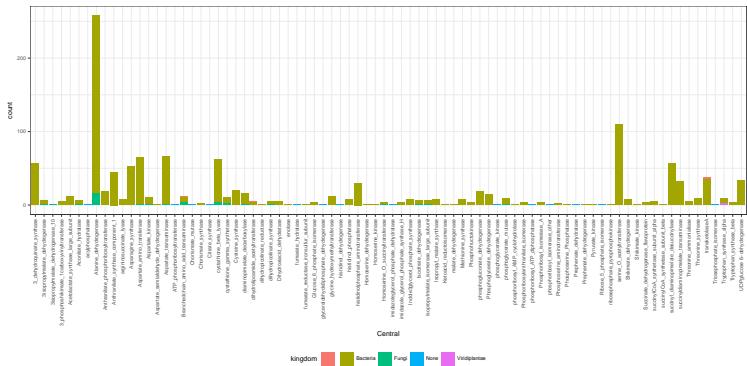


Figure 9: Archaeas Recruitmens on central families coloured by kingdom

Here is a reference to Recruitments after central pathways expansions colourd by Kingdom plot: Figure 9.

### Recruitments after central pathways expansions coloured by taxonomy

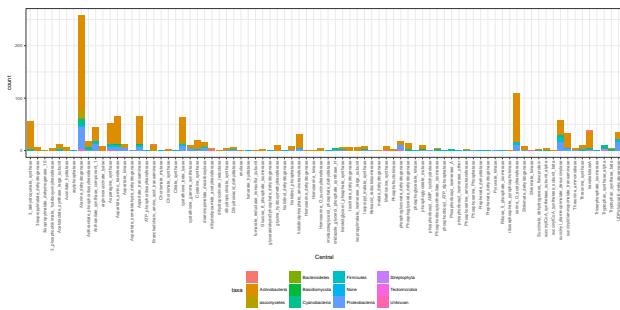


Figure 10: Archaeas Recruitmens on central families coloured by taxonomy

Here is a reference to Recruitments after central pathways expansions coloured by taxa plot: Figure 10.

Archaeas AntiSMASH

## Taxonomical diversity on Archaeasbacteria Data

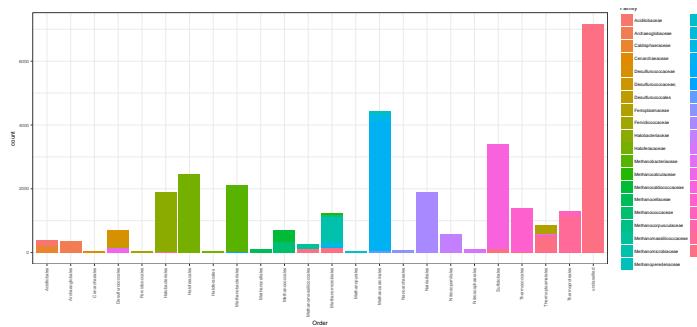


Figure 11: Archaeas Diversity

Here is a reference to Recruitments after central pathways expansions coloured by taxa plot: Figure 11.

## Smash diversity

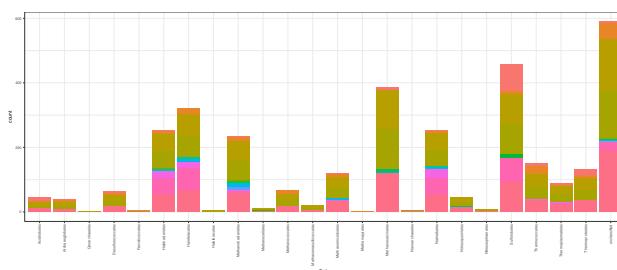


Figure 12: Archaeas Smash Taxonomical Diversity

Here is a reference to Recruitments after central pathways expansions coloured by taxa plot: Figure 12.

## AntisMASH vs Central Expansions

Is it a correlation between pangenome grow and central pathways expansions?

Total central pathway expansions by genome vs Total antismash cluster detected coloured by order

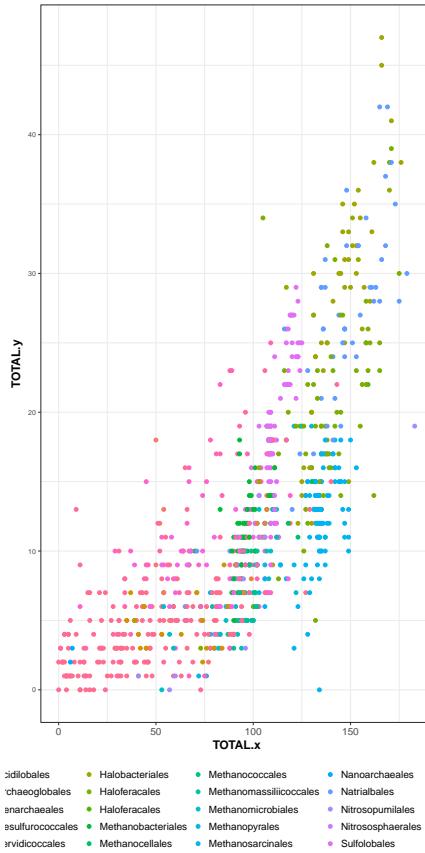


Figure 13: Correlation between Archaeas central pathway expansions and antismash Natural products detection

Here is a reference to the expansions vs antismash NP's clusters plot: Figure 13.

Total central pathway expansions by genome vs Total antimash cluster detected splitted by order

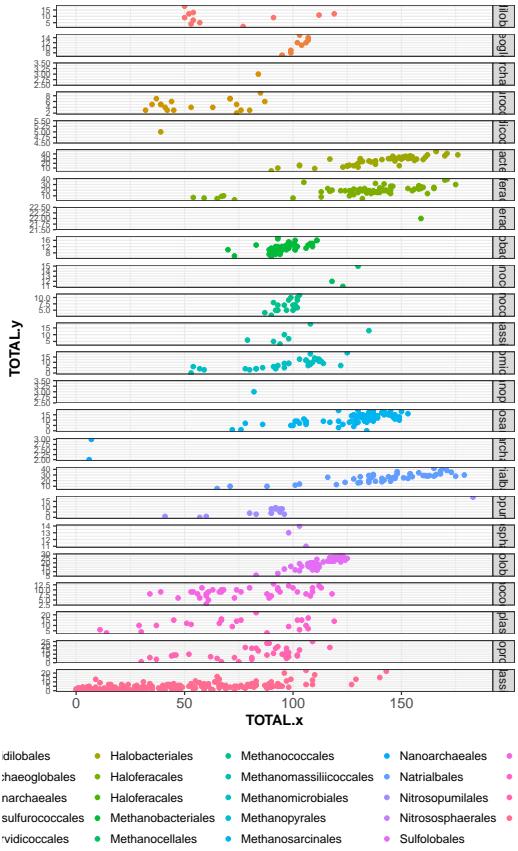


Figure 14: Correlation between Archaeas central pathway expansions and antimash Natural products detection

Here is a reference to the expansions vs antimash NP's clusters splitted by order plot Figure 14.

## AntisMAsh vs Expansions by taxonomic Family

Natural products colured by family

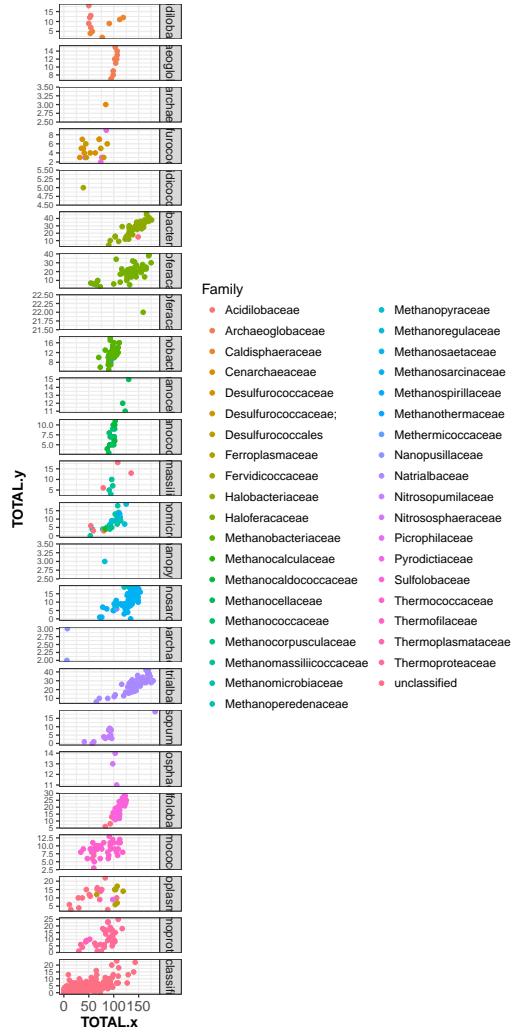


Figure 15: Archaeas Natural products by family

Here is a reference to the Natural products colured by family plot Figure 15.

## Selected trees from EvoMining

Phosphoribosyl\_isomerase\_3 family  
Figure from EvoMining



Figure 16: Phosphoribosyl isomerase A EvoMiningtree

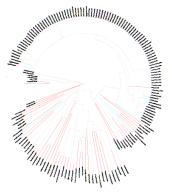


Figure 17: Phosphoribosyl isomerase other EvoMiningtree

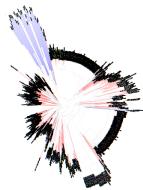


Figure 18: Phosphoribosyl anthranilate isomerase EvoMiningtree

Other possible databases Archaeal signatures *set of protein-encoding genes that function uniquely within the Archaea; most signature proteins have no recognizable bacterial or eukaryal homologs* [@graham\_archaeal\_2000]  
## Footnotes and Endnotes

You might want to footnote something.<sup>1</sup> The footnote will be in a smaller font and placed appropriately. Endnotes work in much the same way. More information can be found about both on the CUS site or feel free to reach out to data@reed.edu.

## Bibliographies

Of course you will need to cite things, and you will probably accumulate an armful of sources. There are a variety of tools available for creating a bibliography database (stored with the .bib extension). In addition to BibTeX suggested below, you may want to consider using the free and easy-to-use tool called Zotero. The Reed librarians have created Zotero documentation at <http://libguides.reed.edu/citation/zotero>. In addition, a tutorial is available from Middlebury College at <http://sites.middlebury.edu/zoteromiddlebury/>.

R Markdown uses *pandoc* (<http://pandoc.org/>) to build its bibliographies. One nice caveat of this is that you won't have to do a second compile to load in references as standard L<sup>A</sup>T<sub>E</sub>X requires. To cite references in your thesis (after creating your bibliography database), place the reference name inside square brackets and precede it by the "at" symbol. For example, here's a reference to a book about worrying: [@Molina1994]. This Molina1994 entry appears in a file called **thesis.bib** in the **bib** folder. This bibliography database file was created by a program called BibTeX. You can call this file something else if you like (look at the YAML header in the main .Rmd file) and, by default, is placed in the **bib** folder.

For more information about BibTeX and bibliographies, see our CUS site (<http://web.reed.edu/cis/help/latex/index.html>)<sup>2</sup>. There are three pages on this topic: *bibtex* (which talks about using BibTeX, at <http://web.reed.edu/cis/help/latex/bibtex.html>), *bibtexstyles* (about how to find and use the bibliography style that best suits your needs, at <http://web.reed.edu/cis/help/latex/bibtexstyles.html>) and *bibman* (which covers how to make and maintain a bibliography by hand, without BibTeX, at <http://web.reed.edu/cis/help/latex/bibman.html>). The last page will not be useful unless you have only a few sources.

If you look at the YAML header at the top of the main .Rmd file you can see that we can specify the style of the bibliography by referencing the appropriate csl file. You can download a variety of different style files at <https://www.zotero.org/styles>. Make sure to download the file into the csl folder.

## Tips for Bibliographies

- Like with thesis formatting, the sooner you start compiling your bibliography for something as large as thesis, the better. Typing in source after source is mind-numbing enough; do you really want to do it for hours on end in late April? Think of it as procrastination.
- The cite key (a citation's label) needs to be unique from the other entries.
- When you have more than one author or editor, you need to separate each author's name by the word "and" e.g. Author = {Noble, Sam and Youngberg, Jessica},.
- Bibliographies made using BibTeX (whether manually or using a manager) accept L<sup>A</sup>T<sub>E</sub>X markup, so you can italicize and add symbols as necessary.
- To force capitalization in an article title or where all lowercase is generally used, bracket the capital letter in curly braces.
- You can add a Reed Thesis citation<sup>3</sup> option. The best way to do this is to use the phdthesis type of citation, and use the optional "type" field to enter "Reed thesis" or "Undergraduate thesis."

---

<sup>1</sup>footnote text

<sup>2</sup>@reedweb2007

<sup>3</sup>@noble2002

## **Anything else?**

If you'd like to see examples of other things in this template, please contact the Data @ Reed team (email [data@reed.edu](mailto:data@reed.edu)) with your suggestions. We love to see people using *R Markdown* for their theses, and are happy to help.