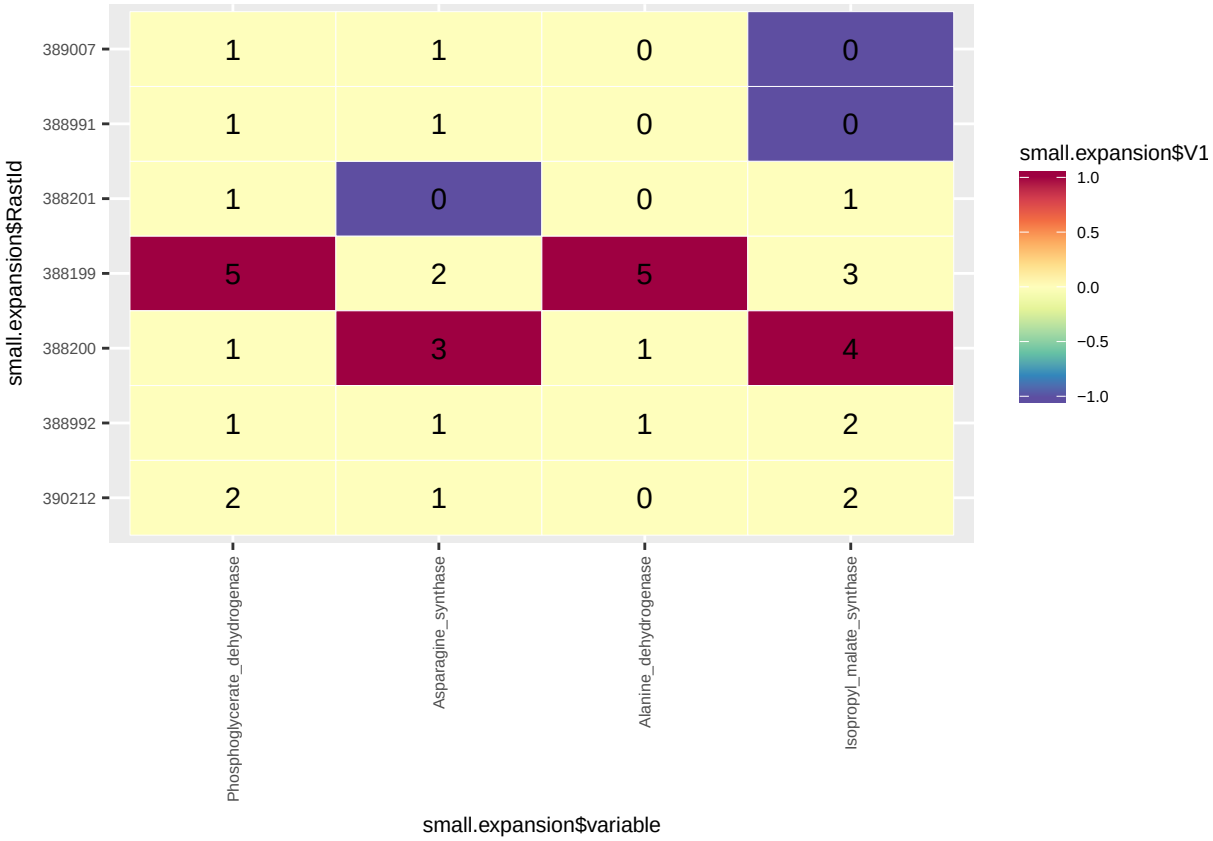
EvoMining




```
############################################################
## Trying to sort the heatplot
## Reading heatplot table and taxa information and saving it on data.frame data structure
ArchaeasHeatPlot <- read.table("chapter2/Archaeas/ArchaeasHeatPlot", header=TRUE, sep="\t")
ArchaeasTaxa <- read.table("chapter2/Archaeas/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$RastId, ArchaeasHeatPlot$orde
# 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 = r

rownames(HP_Archeas.calcs)<-HP_Archeas.calcs$variable
############################################################3
```

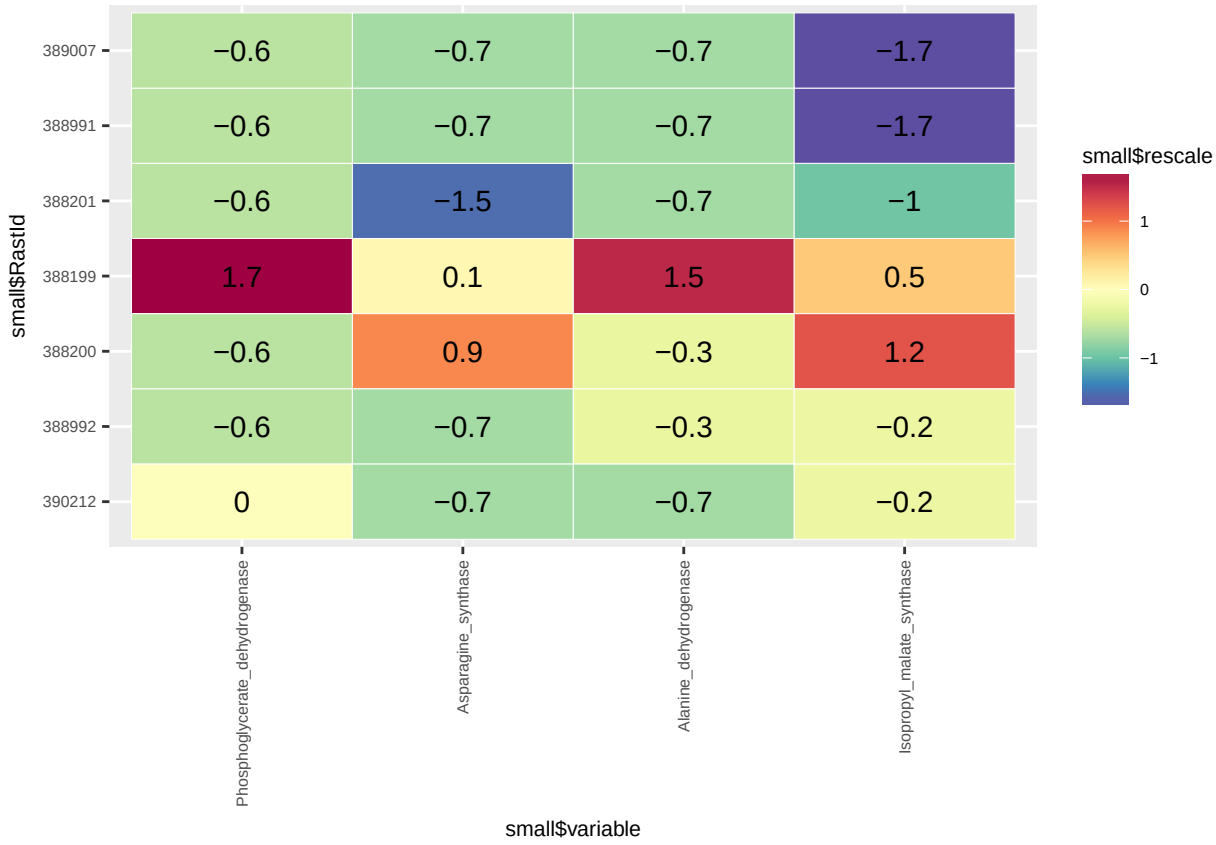
```
color_exp<-function(x){
  # Pendiente pasarle un dataframe en lugar de tener fijo small.m como variable global
  result=0
  expansion<-NULL
  reduction<-NULL
  local_value=x[1,"value"]
  met_family<-x[1,"variable"]
  Rast=x[1,1]
  #print ("rast",Rast)
  # print(paste(x,"family",met_family,"value",local_value,"Rast",Rast))
  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-AAA259O05 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-AAA259O05 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-AAA259O05 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-AAA259O05 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
```

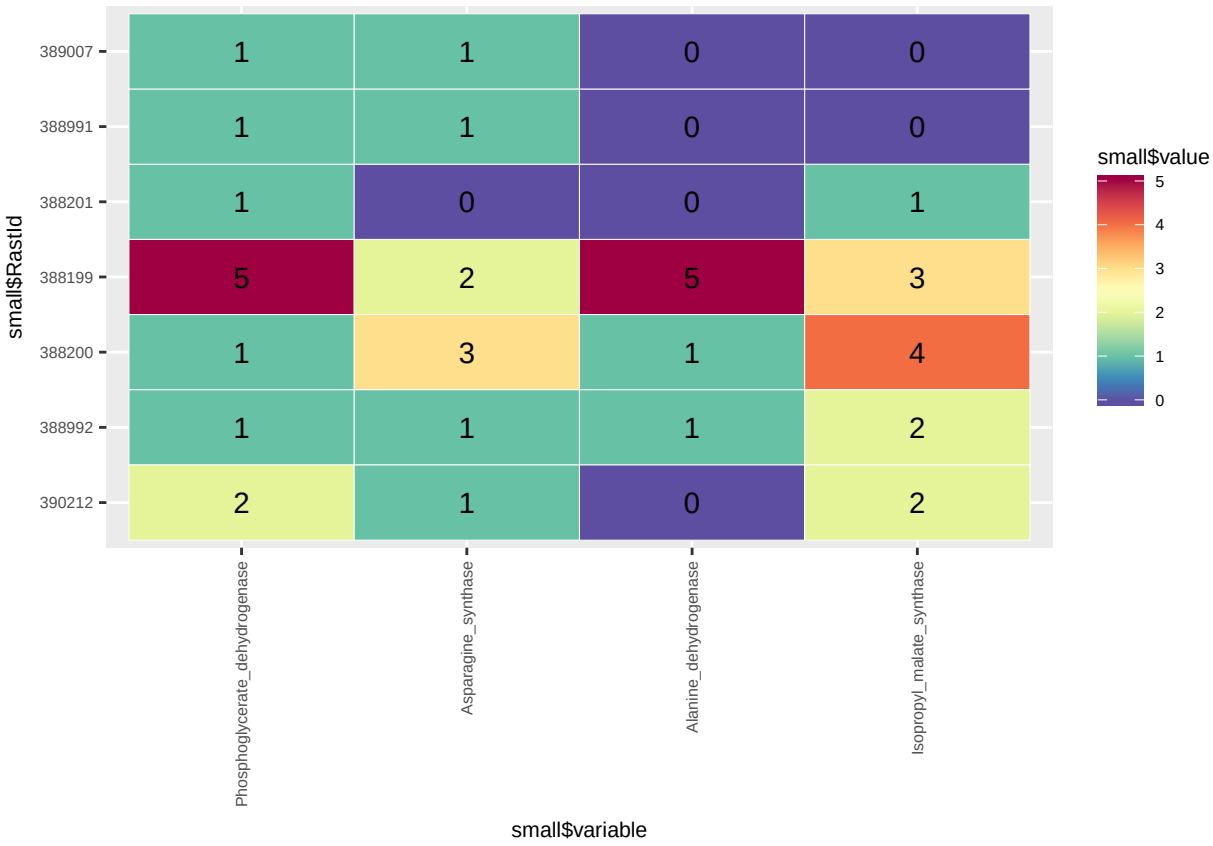
```
## 690   Euryarchaeota Euryarchaeota    Archaeoglobi   Archaeoglobales
## 14017 Euryarchaeota Euryarchaeota    Halobacteria     Haloferacales
## 14018 Euryarchaeota Euryarchaeota Methanomicrobia Methanosarcinales
## 14019 Euryarchaeota Euryarchaeota    unclassified      unclassified
## 14567    TACK 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_Archaeas_Taxa.m[!HP_Archaeas_Taxa.m$variable %in% c("Contigs", "Size","TOTAL
```

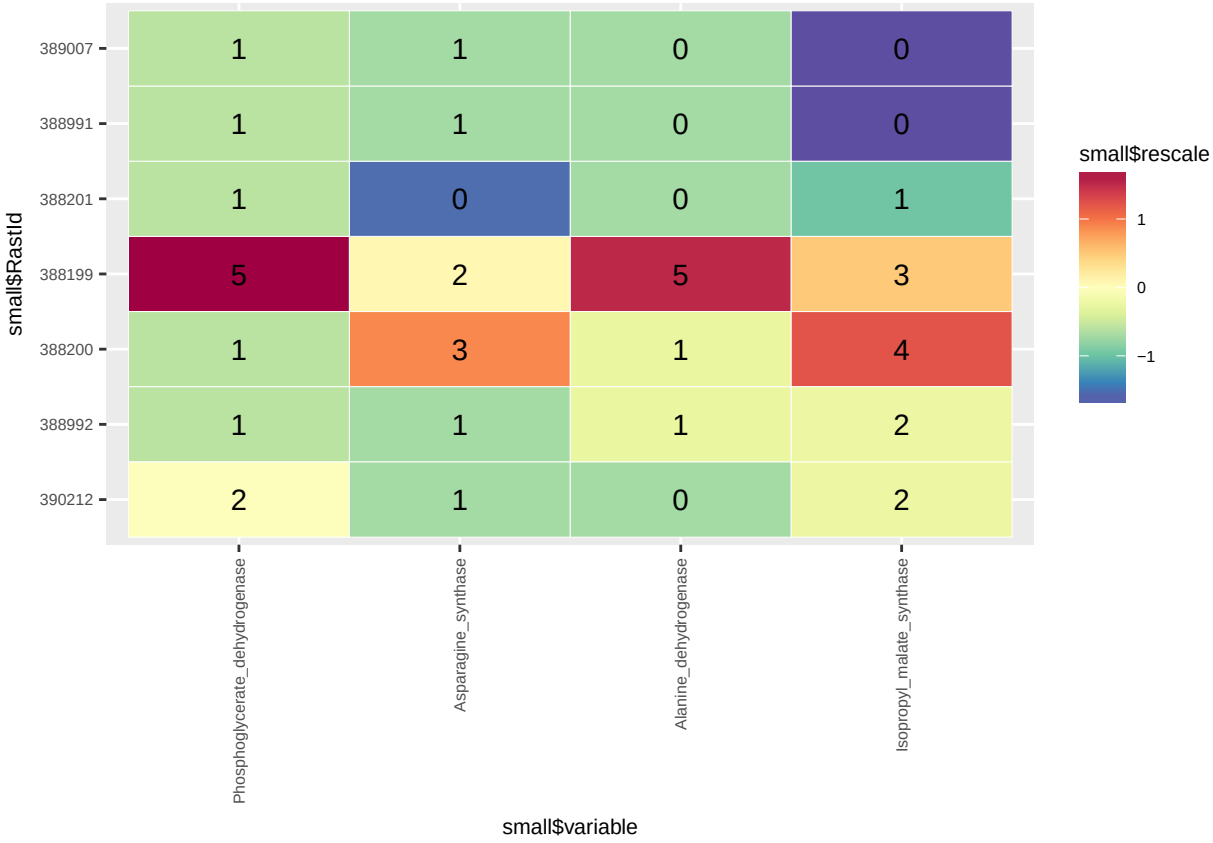
```
#####################################333
################################################


## 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(fil
```




```
## 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),expa
rownames(small.m)<-small.m$variable

#########################################################################
```

EvoMining

Introduction

La promiscuidad enzimática puede buscarse en familias envueltas en procesos de divergencia funcional. Uno de dichos procesos es la expansión y posterior reclutamiento de familias pertenecientes a rutas metabólicas conservadas hacia el metabolismo especializado. Los productos naturales o metabolitos especializados son sintetizados generalmente por clusters de genes distribuidos en un pequeño porcentaje de un linaje taxonómico. Estos clusters, conocidos como BGCs (Biosynthetical Gene Clusters), contienen reclutamientos de las copias extras de familias pertenecientes al metabolismo conservado. La similitud de secuencia de los genes que pertenecen a los BGCs así como su relativa sintenia en diversos organismos de un linaje hacen que genómica copmarativa sea de utilidad para intentar localizarlos. Finalmente, el auge en la cantidad de genomas disponibles públicamente así como la facilidad por sequenciar nuevos hace posible que los métodos bioinformáticos ayuden a encontrar nuevos BGCs. EvoMining es un método de para sugerir la formación de

nuevos BGCs y en consecuencia encontrar zonas donde puede estar ocurriendo la capcidad de adquirir un nuevo sustrato cambio en promiscuidad enzimátcica.

En este capítulo se explica el desarrollo de EvoMining como plataforma bioinformática dedicada a presentar una visualización del origen y destino de todas las copias de familia enzimátocas provenientes del emtabolismo conservado. Se discutirán también cuatro linajes genómicos Actinobacteria Cyanobacteria, Pseudomonas y Archaea y finalmente se analizará un BGCs scytonemin.

EvoMining es un método para encontrar BGCs no tradicionales

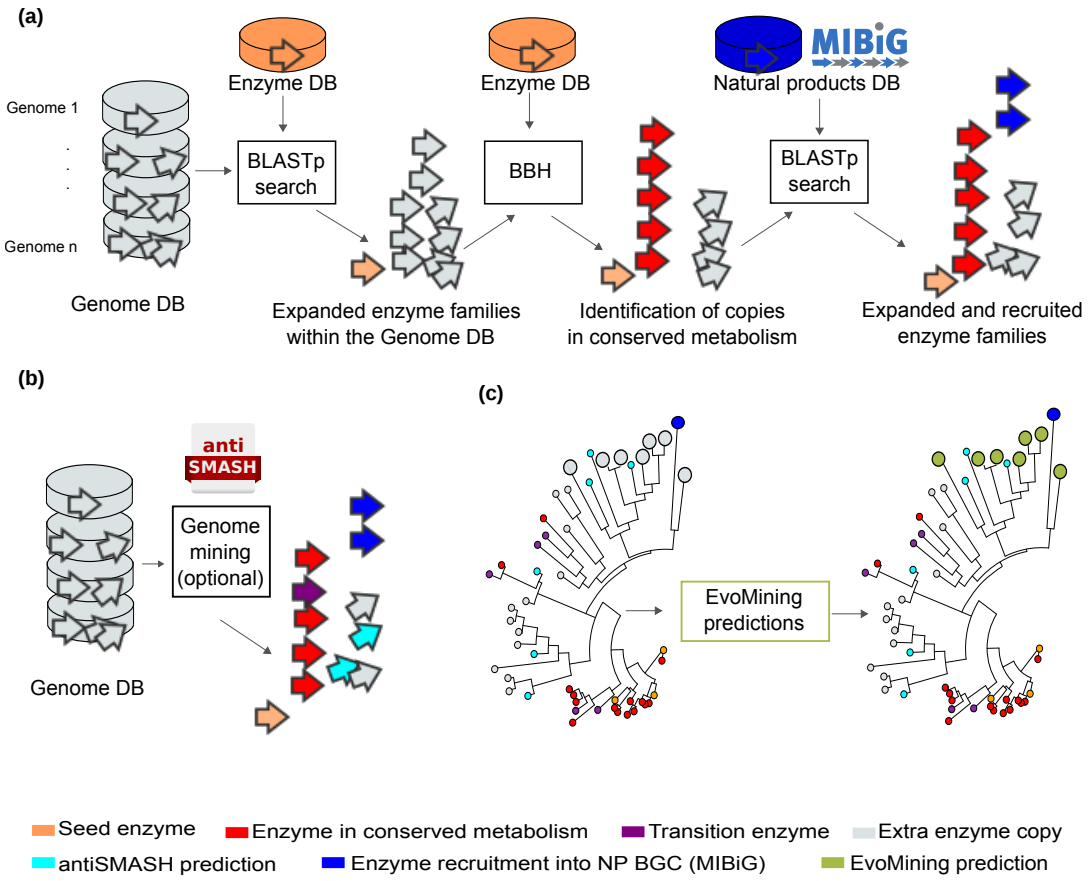


Figure 1: EvoMining Algorithm

Gen families expansions on genomes

Pangenomes

Las expansions están localizadas en el pangenoma, ya que si estuvieran en todo el genoma, Tools to analyse pangenome BPgA

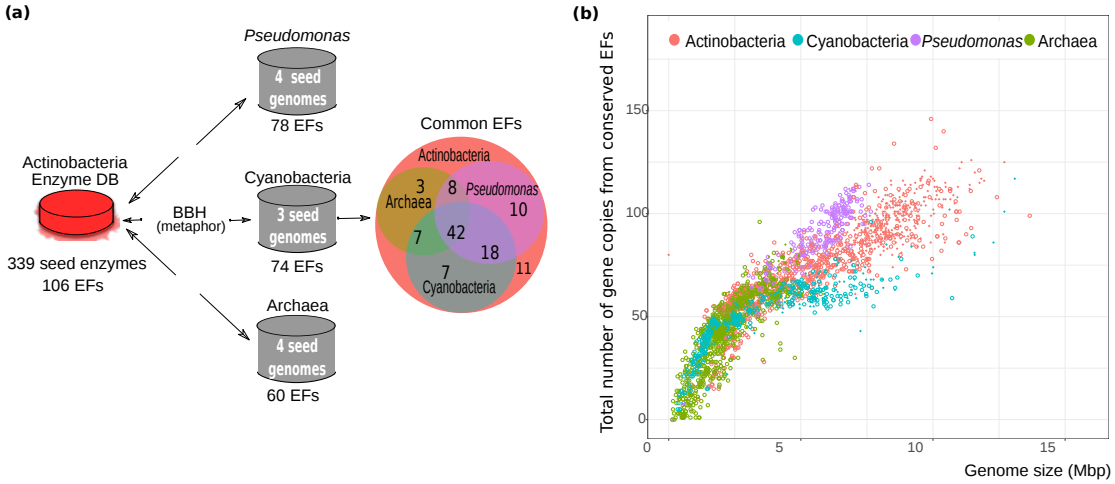


Figure 2: Seed genomes

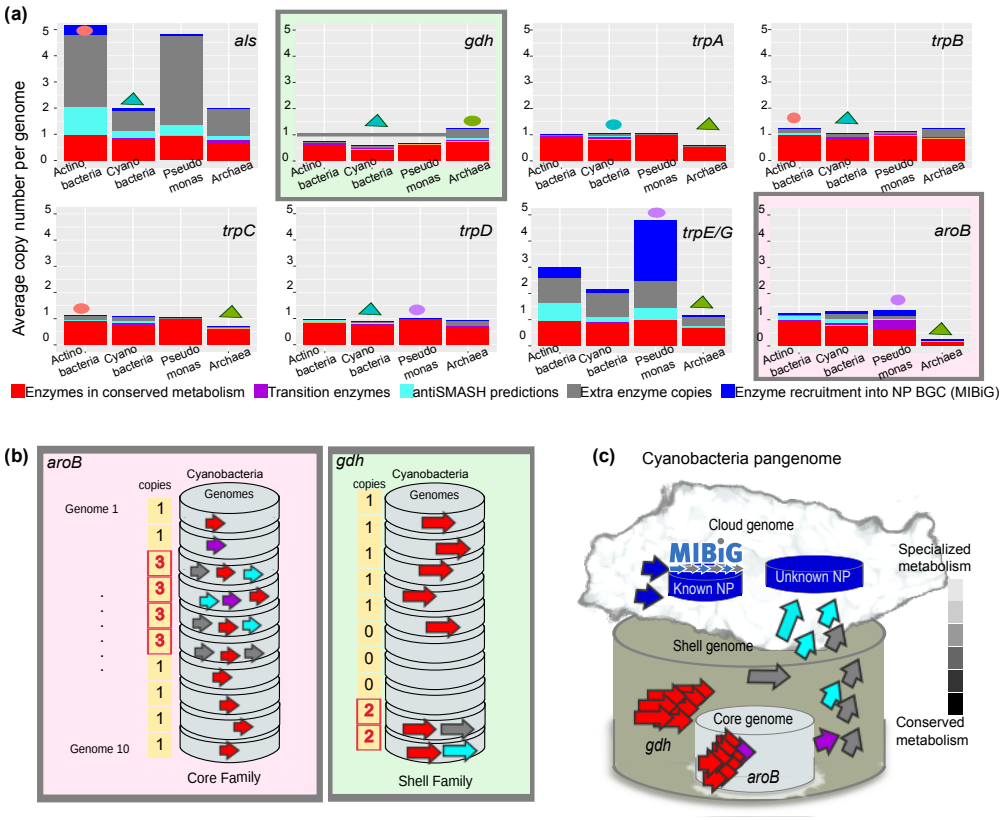


Figure 3: Expansion patterns in 42 conserved families

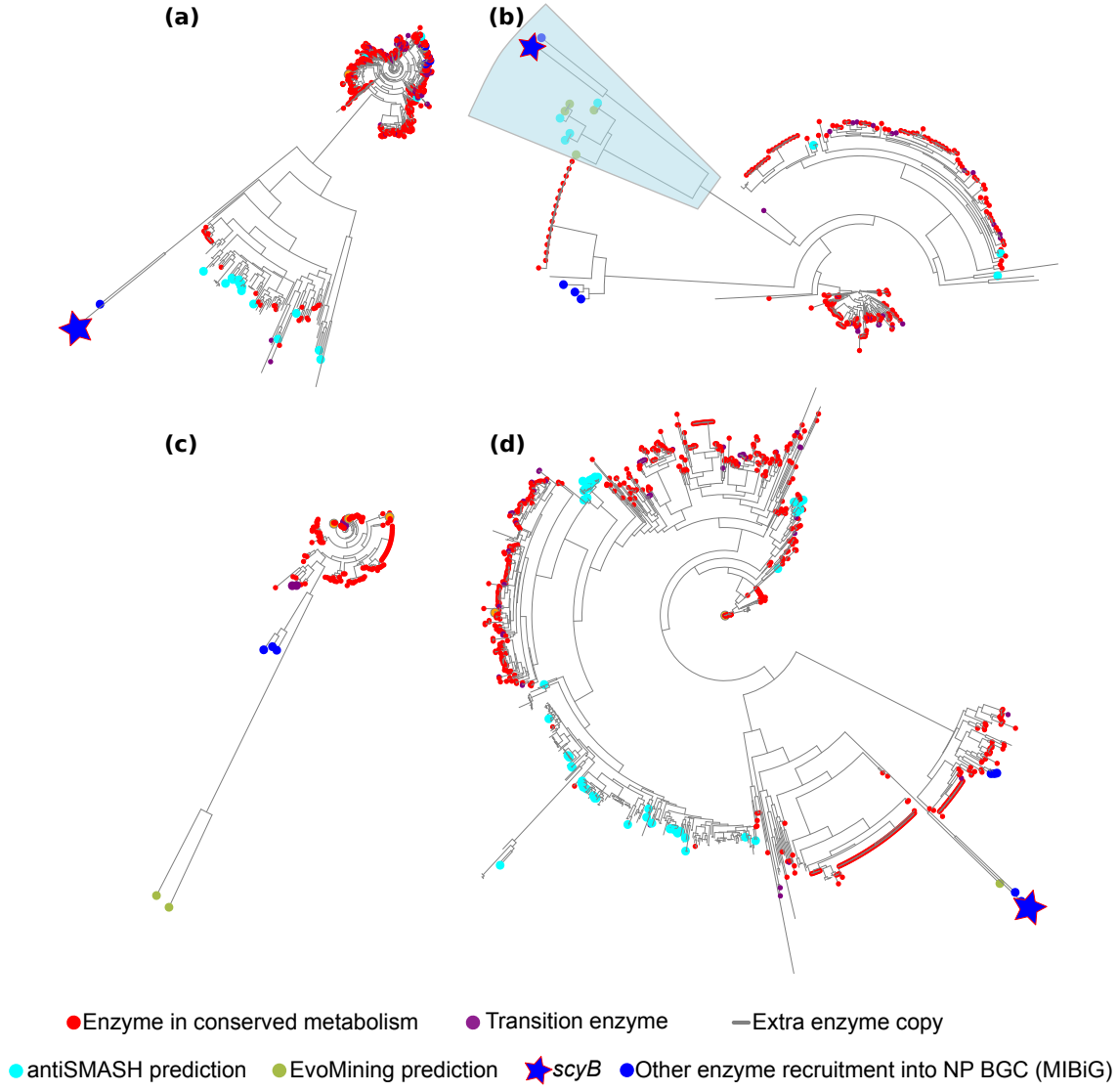


Figure 4: EvoMining Algorithm

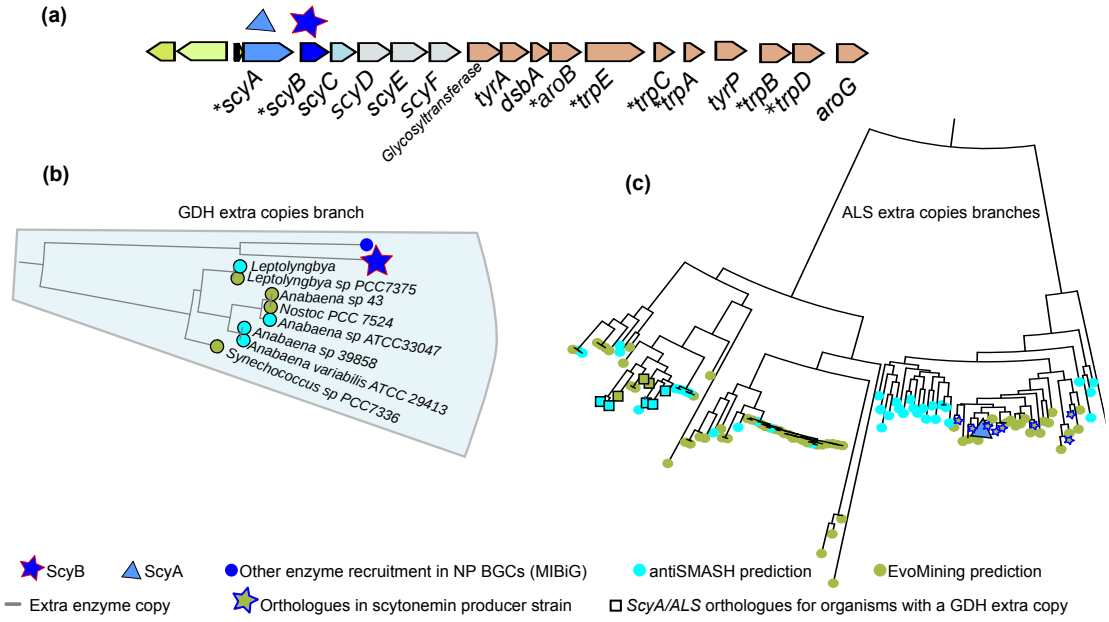


Figure 5: EvoMining Algorithm

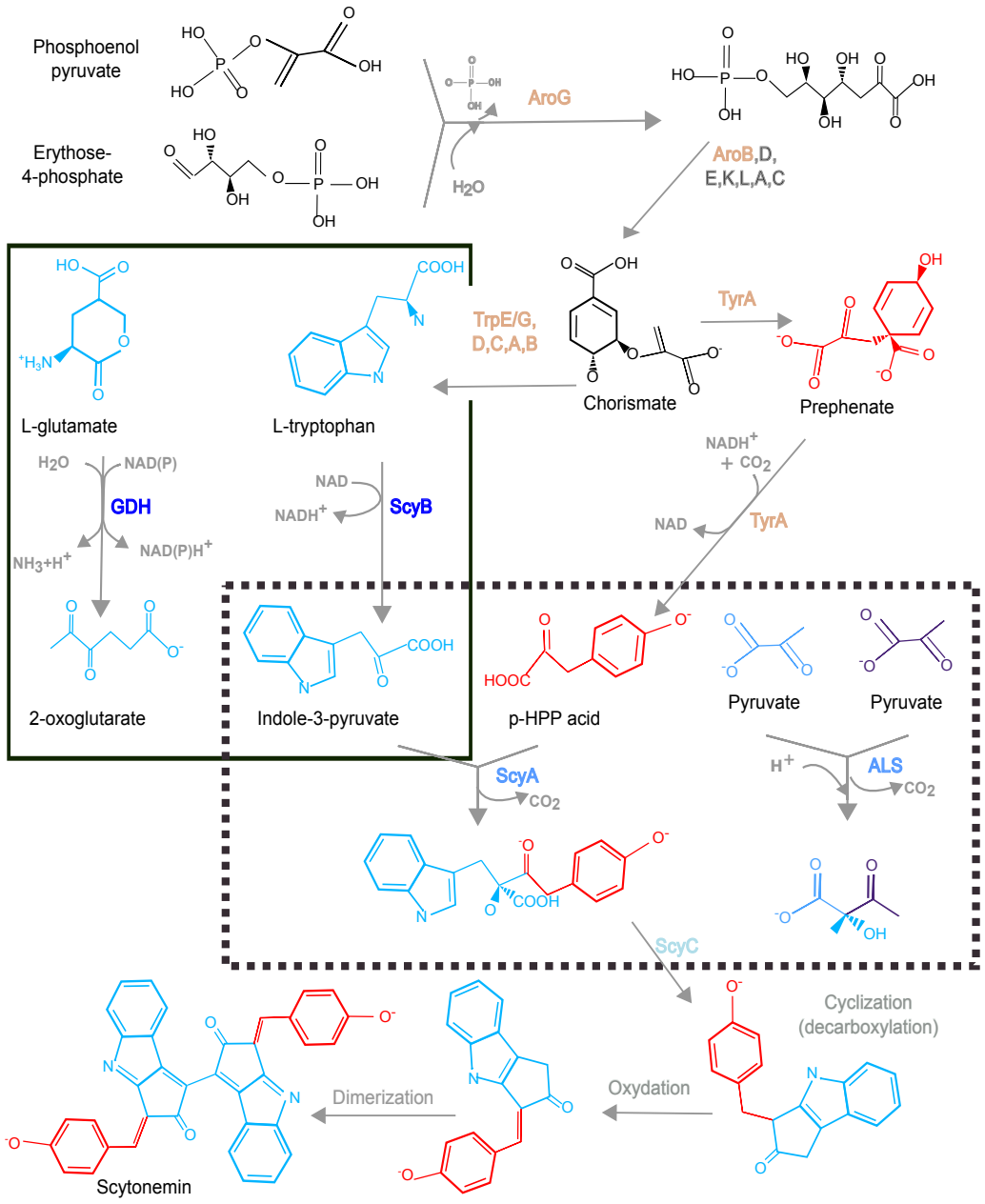


Figure 6: EvoMining Algorithm

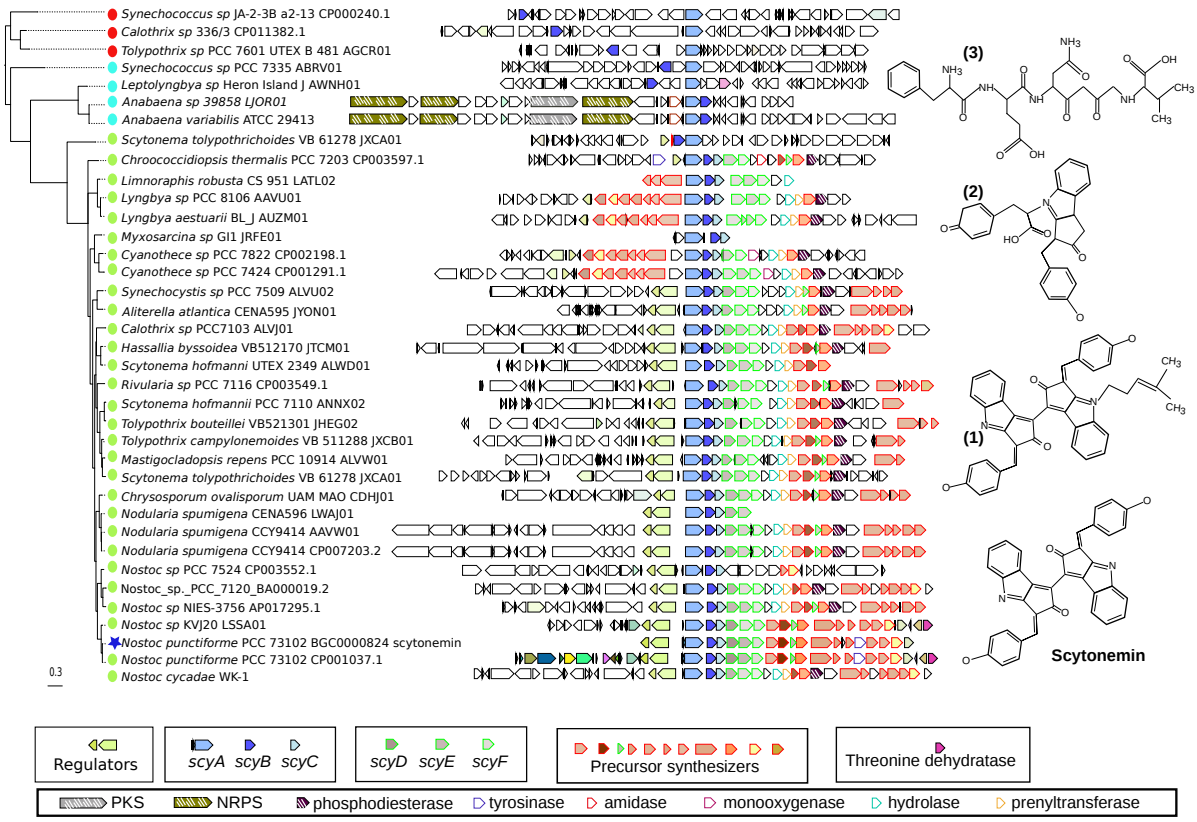


Figure 7: EvoMining Algorithm

EvoMining

EvoMining looks expansions on prokariotic pangenome.
Biological idea.

EvoMining was available as a consult website with 230 members of the Actinobacteria phylum as genomic data base, 226 unclassified nBGCs, and not interchangable central database 339 queries for nine pathways, including amino acid biosynthesis, glycolysis, pentose phosphate pathway, and tricarboxylic acids cycle. [@cruz-morales_phylogenomic_2016] EvoMining was proved on Actinobacteria Arseno-lipids

Pangenome

The sequenced genome of an individull in some species is just a partil print of the species geneticll repertoire Individualls can gain and loss genes.
[@koonin_turbulent_2015] Pangenome is the total sequenced gene pool in a taxonomically related group. Supergenome all the possible extant genes. About 10 times genomes. there are open, closed pangenomes.Most genomes has a core a shell and a unique genes.
Gene history its a tree history

HGT doubles mutation rate on prokarites.
Maybe HGT is an selected feature, if is the case, so could be np production.
Some archaeas has open pangenome. [@halachev_calculating_2011]

HGT doubles mutation rate on prokarites. [@koonin_turbulent_2015] Maybe HGT is an selected feature, if is the case, so could be np production.
Some archaeas has open pangenome. [@halachev_calculating_2011] Shell trees converge to core trees [@narechania_random_2012]

EvoMining Implementation

EvoMining was expanded from a website (http://evodivmet.langebio.cinvestav.mx/EvoMining/index.html) with limited datasets to an easy to install distribution that allows flexiblibilty on genomic, central and natural product databases. Evomining user distribution was developed on perl on Ubuntu-14.04 but wraped on Docker. Docker is a software containerization platform that allows repetibilty regardless of the environment. Docker engine is avilable for Linux, Cloud, macOS 10.10.3 Yosemite or newer and even 64bit Windows 10.

Dependencies that were packaged at EvoMining docker app are Apache2, muscle3.8.31, newick-utils-1.6,quicktree, blast-2.2.30, Gblocks_Linux64_0.91b perl and from cpan CGI, SVG and Statistics::Basic modules.

Github defines itself as an online project hosting using Git. Its free for open source-code hosting and facilitates team work. Includes source-code browser, in-line editing, and wikis.

Dockerhub is an apps project hosting.

Dockerhub nselem

EvoMining code is open source and it is available at a github repository github/EvoMining

Github and Dockerhub can be coneccted by the use of repositories automatically built. Among the advantages of automated builds are that the DockerHub repository is automatically kept up-to-date with code changes on GitHub and that its Dockerfile is available to anyone with access to the Docker Hub repository. EvoMining is stored on a DockerHub automated build repository linked to github EvoMining repository so that code is always actualized.

To download EvoMining image from docker Hub once Docker engine is installed its necessary to run the following command at a terminal:
`docker pull nselem/newevomining`

To run EvoMining container
`docker run-i -t  -v /home/nelly/docker-evomining:/var/www/html -p 80:80 evomining /bin/bash`

To start evoMining app `perl startEvomining`
“ Detailed tutorial, EvoMining description, pipeline and user guide are available at a wiki on github at EvoMining wiki.

Other genomic apps were containerized to docker images during this work.
- *myRAST* docker- https://github.com/nselem/myrast
RAST is a bacterial and Archaeal genome annotator [@aziz_rast_2008, @overbeek_seed_2014 , @brettin_rasttk:_2015] This app allows myRAST functionality to upload
It allows EvoMining genome database annotation.
-*Orthocores* docker-https://github.com/nselem/orthocore
Helps to obtain genomic core paralog free and construct genomic trees
-*CORASON* docker-https://github.com/nselem/EvoDivMet/wiki
-PseudoCore github- <>
Genomic Core with a reference genome has the advantage of more genomes, but it is not paralog free

-RadiCal docker image
To detect core diferrences on a set of genomes
-BPGA to analize pangenome

EvoMining Dockerization was chosen to avoid future compatibilty problems, for example dependencies unavailabilty, or incompatibility between future versions of its software components. As much as reproducible research was a concerned while developing EvoMining app, reproducibilty is also important on data analysis, for that reason this document was writen using R-markdown and latex template from Reed College [@chesterismay_updated_2016]. While R-markdown allows to write and run R code and interpolate text paragraph to explain scripts and analysis.

EvoMining Databases

Evomining containerized app is a user-interactive genomic tool dedicated to the study of protein function.

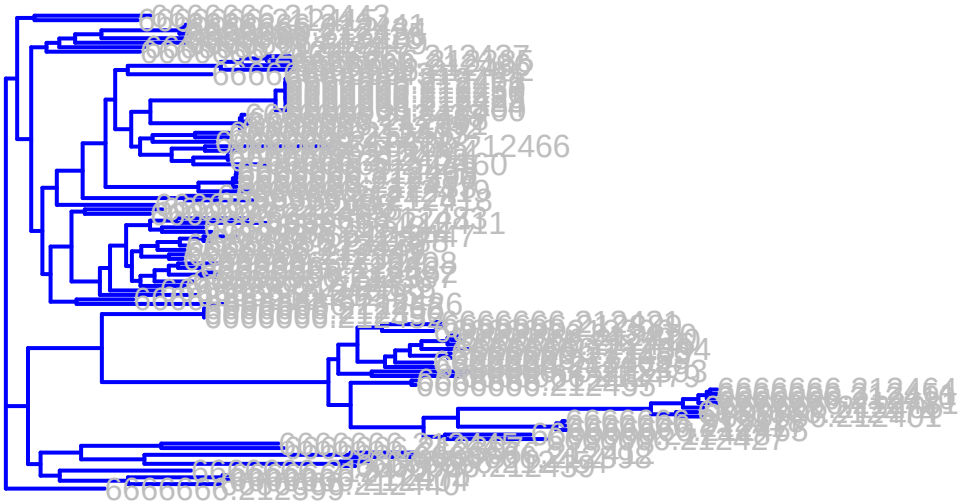
1. Genomes DB
2. Natural Products DB
3. Central Pathways DB

Archaea, *Actinobacteria*, *Cyanobacteria* were used as genome DB, MIBiG was used as Natural Product DB and different Central Pathways were used.

Genome DB

RAST annotation of genomes was done.

Phylogeny



To capture differences on genomes we sort them phylogenetically. Phylogenies can be constructed using different paradigms as Parsimony, Maximum Likelihood, and Bayesian inference. Short descriptions of the main phylogeny methods are included below.

Why is a tree useful {Book reference} why trees are useful for?
* Distance methods
* Parsimony * Maximum Likelihood * Mr bayes

General Trees
Actinobacteria Tree, ArchaeaTree, CyanobacteriaTree.

It's easy to create a list. It can be unordered like

To create a sublist, just indent the values a bit (at least four spaces or a tab). (Here's one case where indentation is key!)

1. Item 1
2. Item 2
3. Item 3
    - Item 3a
    - Item 3b

Central DB

We chose central pathways from [@barona-gomez_what_2012]
* BBH Best Bidirectional Hits with studied enzymes from Central Actinobacterial pathways were selected.

- By abundance
- By expansions on genomes

[largefiles,https://help.github.com/articles/installing-git-large-file-storage/]

Data Bases

Central pathways

Central database were chosen by BBH from

```
table <- read.csv("chapter2/WC_Central/BBH_Organisms.txt", row.names = 1,sep="\t")
kable(table,  caption = "BBH_Organisms \\label{tab:BBH_Organisms}",caption.short = "BBH_Organisms ")
```

Table 1: BBH_Organisms

	RastId	Database	Taxa1	Taxa2
Corynebacterium glutamicum	6666666.112876	Actinobacteria		
Streptomyces coelicolor A3(2) NC_003888.3		Actinobacteria		
Mycobacterium tuberculosis H37Rv NC_000962.3	6666666.146923	Actinobacteria		
Methanosarcina acetivorans C2A AE010299.1	6666666.211599	Archaea	Euryarchaeota	Methanomicrob
Nanoarchaeum equitans Kin4-M - AE017199.1	6666666.211718	Archaea	DPANN group	Nanoarchaeota
Natronomonas pharaonis DSM 2160	CR936257.1	6666666.211909	Archaea	Euryarchaeota
Halobacteria				
Sulfolobus solfataricus P2 AE006641.1	6666666.211567	Archaea	TACK group	Crenarchaeota
Cyanothece sp. ATCC 51142 CP000806.1	6666666.212444	Cyanobacteria	Oscillatoriales	
Synechococcus sp. PCC 7002 CP000951.1	6666666.212477	Cyanobacteria	Synechococcales	
Arthrospira platensis C1	6666666.189647	Cyanobacteria	Cyanobacteria	

Genome Dynamics

Among BBH central databases, genomic dynamics was included.
Whats change site:WC Data

groups were formed with 100Cyanos, 100Archaea , 118 Actinos Closed, 43StreptosClosed
Selected organims were

```
table <- read.csv("chapter2/WC_Central/WC_Organisms.txt", row.names = 1,sep="\t")
kable(table,  caption = "WC_Organisms \\label{tab:WC_Organisms}",caption.short = "WC_Organisms ")
```

Table 2: WC_Organisms

	Rast.Id	Database
Arthrospira platensis NIES-39 AP011615.1	6666666.21	Cyanos
Synechococcus sp. PCC 7002	6666666.21	Cyanos
Cyanothece sp. ATCC 51142	6666666.21	Cyanos
Methanosarcina acetivorans	6666666.21	Archaea
Nanoarchaeum equitans Kin4-M	6666666.21	Archaea
Natronomonas pharaonis DSM 2160	6666666.21	Archaea
Sulfolobus solfataricus P2	6666666.21	Archaea
Mycobacterium tuberculosis H37Rv	83332.23	Actinos
Corynebacterium glutamicum ATCC 13032	196627.31	Actinos
Streptomyces coelicolor A3(2) NC_003888.3	6666666.11	Actinos and Streptomyces
Streptomyces sp. Mg1 NZ_CP011664.1	6666666.15	Streptomyces

Those families present on at least as much as genomes on the group
Cyanos 100 647
Abundant.Families.100Cyanos
Actinos 118 132
Abundant.Families.43Strepto
Archaea 100 35
Abundant.Families.Actinos
Streptomyces 43 1263
Abundant.Families.Archaeas

Those families expanded on at least two groups

```
cat *Abun* | cut -f3| sort | uniq -c | sort >Abundance.all
```

Those Families expanded on Archaea and not expanded on Actino
`comm -23 f3Archaeas f3Actinos >ArchaeasNoActinos`
Those Families expanded on Actino and not on Archaea
`comm -13 f3Archaeas f3Actinos >ActinosNoArchaea`

Those families expanded on Streptomyces but not in ActinoBacteria
`comm -13 f343Strepto f3Actinos >ActinosNoStrepto`
Those Families expanded on Actinobacteria and not in Streptomyces
`comm -23 f343Strepto f3Actinos >StreptoNoActinos`

Those Families expanded on Cyano and not in Actino
`comm -23 f3Cyanos f3Actinos >CyanosNoActinos`

Natural Products DB

Natural products was improved from previous version

AntisMASH optional DB

AntiSMASH is [@weber_antismash_2015,@medema_antismash:_2011]
Archaeas Results Archaea is a kingdom of recent discovery were not many natural products has been known. On Actinobacteria, evoMining has proved its value to find new kinds of natural products. The clue to this discovery was that Actinobacteria has genomic expanssions. Now Archaea has genomic expansions, even more has central pathways genomic expansions. Are this expansions derived from a genomic duplication? Has Archaea natural products detected by antismash, and if not, where are this NP's or may Archaea doesn't have NP's.

applying EvoMining to Archaea

Otras estrategias para los clusters Argon context Idea

Argonne

ssh nselem@login.mcs.anl.gov
phrase
ssh nselem@maple
password

cs close strain
wc whats chain

we source (edit bashrc)
link ln (create a link to ross directory)
run out of power:
screen

in Seqs (not mine)
cat
6666666.103569 6666666.112815 6666666.112823 6666666.112833 6666666.112841 6666666.112849 6666666.112857 > /home/nse/Concat_Full
to find paralogous sets
svr_representative_sequences -b -f Id_Clust -s 0.5 < Concat_Full > TempFull&

perl -p -i -e 's/̊//' readable.tree to clean the tree
To find contexts o pegs of paralogous sets

Context midle point 5000 bp (using text tables)
scp 6666666.112839.txt nselem@maple:/homes/nselem/Strepto_01/.

fig|6666666.112839.peg.26

copy families.all file
on the file we have column1 family name column 5 peg id

cluster_objects < elements_to_cluster > ClusteFile

write a file with pegs
1 peg1 adjacent1, adjacent2 ....
1 peg2
2
2

write a file similiar but with the family number

1 peg1 fn1, fn2 ....
1 peg2
2
2

compare each peg on this file from the same family

Write the conextions file
peg1 peg2
peg1 peg3
peg2 peg3

cluster this file and score the cluster

Define

```
1.  a "function set" is generated by the what's changed directory
```

as a "family"

```
2.  a "paralog set" is a set of function sets in which paralogous
```

members span the sets

```
3.  a PEG is in a paralog set if it is in one ofthe function sets
```

that make up the

```
4.  a "context" of a PEG is the set of close pegs
4.1 First cluster operation would give us: context sets  (CS)

5.  a "context set" is a set of PEGs with "similar contexts"
5.1 second clustering operation would give us:cluster  (Cl)

6.  a "cluster" is a set of context sets (each context set is a different
```

compute:
Compute the context sets that are made from PEGs that occur in PS.
Compute the contexts of PEGs in PS.

cluster these context using the "similar contexts" relation

This gives a set of clusters, and the members of the clusters are context sets
That is, a cluster is a set of context sets

```
  a. the number of contexts sets i
```

score the clusters
Take a paralog set PS.
Be the context sets: CS_1, CS_2,..., CS_k members of the paralogous set
k the number of contexts sets on the paralogous set
n_i the cardinality of CS_i

```
PS={CS1,CS2,...,CS3}
Cl={[CS_1,n_1],[CS_2,n_2],...,[CS_k,n_k]}

let be M=max(n_i)   i=1,2,..k (Maximum cardinality of Context sets)
    m=max(n_i)   i=1,2,..k, i!=M (second greatest cardinality of context sets)
    (We are intersted that a second copy is distributed)

We are interested on k,M,n to form a scoring function for the cluster set
S=f(k,m,M)=c_1*k+c_2*m+c_3*M
```

history

Para hacer un nuevo set de datos

591 cd Data/CS
592 mkdir Directorio
593 vi Directorio/rep.genomes
594 cd Directorio/
600 nohup svr_CS -d Directorio&

Contenido de rep.genomes
rast|390693 nselem35 q8Vf6ib
rast|390675 nselem35 q8Vf6ib
rast|388811 nselem35 q8Vf6ib

When you click the **Knit** button above a document will be generated that includes both content as well as the output of any embedded **R** code chunks within the document. You can embed an **R** code chunk like this (`cars` is a built-in **R** dataset):

```
summary(cars)
```

```
##      speed           dist
##  Min.   : 4.0   Min.   :  2.00
##  1st Qu.:12.0   1st Qu.: 26.00
##  Median :15.0   Median : 36.00
##  Mean   :15.4   Mean   : 42.98
##  3rd Qu.:19.0   3rd Qu.: 56.00
##  Max.   :25.0   Max.   :120.00
```

Inline code

If you'd like to put the results of your analysis directly into your discussion, add inline code like this:

> The `cos` of 2π is 1.

Another example would be the direct calculation of the standard deviation:

> The standard deviation of `speed` in `cars` is 5.2876444.

One last neat feature is the use of the `ifelse` conditional statement which can be used to output text depending on the result of an **R** calculation:

> The standard deviation is less than 6.

Note the use of `>` here, which signifies a quotation environment that will be indented.

As you see with `$2 \pi$` above, mathematics can be added by surrounding the mathematical text with dollar signs. More examples of this are in [Mathematics and Science] if you uncomment the code in [Math].

Recomendaciones de Luis

Para evoMining
Probar distintos métodos de filogenia y después hacer la coloración.
maximum likelihood, Protest phyml
Atracción de ramas largas.
raxml
trim all vs Gblocs (Tony Galvadon)

Comparar dos árboles
Para ver si la evolución de los genes concatenados ha sido simultánea
Robinson and foulds
Joe Felsestein
Phylip

2. dist tree
   quarter descomposition
   peter gogarten fendou Mao

Sets de experimentos.
Para el experimento de los streptomyces con ruta centrales el core, analizar el problema de dominios múltiples.
Dominios
Nan Song, Dannie durand
Después del blast

Para obtener
Pablo Vinuesa: Get Homologues

Burkhordelias y su toxina (Preguntar a Beto)
Cianobacterias y la ruta de fijación de nitrógeno.

Servidor Viernes a las 12:00

CORASON: Other genome Mining tools context-based

CORe Analysis of Syntenic Orthologs to prioritize Natural Product-Biosynthetic Gene Cluster

Bacterial biosynthetic gene clusters (BGCs) known are always increasing, almost all bacterial genome sequenced contributes with new genes and gene clusters to the known Bacterial Pangenome. In consequence of gene diversity and sequence technology advances researchers often have a large set of genomes to analize in search of a particular gene cluster variation. Answering BGCs analysis needs, CORASON allows users to find and visualice variations of a given gene cluster sorting them according to the conserved core cluster phylogeny.

The core genome on a taxonomical group is the set of coding sequences that are shared between all group members, this definition may be adapted to the cluster core by exploring a set of gene clusters instead of a set

of genomes. The cluster core attempts to identify a set of functions conserved on a particular BGC variations. A report about gene function using RAST technology will be provided whenever a cluster core exists and core sequences will be concatenated to construct a phylogenetic tree and sort variation clusters accordingly.

To find cluster variations, given a query protein sequence that belongs to a reference cluster, CORASON will search on a Bacterial genome database all gene clusters that contains orthologues of the query-protein and at least another sequence from the reference cluster. Orthologues on variation clusters are coloured within a gradient according to its identity percentage with the reference cluster sequences.

Finally, in order to provide an easy to install distribution, CORASON was packaged on docker containerization platform. Software dependencies such as BLAST 2.2.30, muscle3.8.3, GBlocksLinux64_0.91b, quicktree, newick-utils-1.6, and CORASON code were wrapped together on CORASON docker container. Tutorial and software are available at nselem/github.

CORASON inputs are a genomic database, a reference cluster and an enzyme inside this cluster, outputs are newick trees, core functional report and a cluster variation SVG file. SVG format among being high quality scalable graphics, also allow to display metadata such as gene function and genome coordinates just by mouse over figures on a browser facilitating genomic analysis.

In conclusion CORASON is an easy to install comparative genomic visual tool on a customizable genome database that allows users to visualice variations of a reference gene cluster identifing its core functions and finally sorting variations according to their evolutionary history helping to prioritize clusters that may be involved on chemical novelty.

Tree methods (from antiSMASH textual quotation)

Multiple methods exist to construct phylogenetic trees based on multiple sequence alignments. Depending on the desired output tree characteristics, the number of input sequences, and other constraints, the most appropriate method should be chosen. A popular algorithm among the distance-matrix based methods is the Neighbour-Joining algorithm that uses bottom-up clustering to create the tree. Neighbour-Joining is comparatively fast method, but the correctness of the tree depends on the accuracy and additivity of the underlying distance matrix. Maximum parsimony methods try to identify the tree that uses the smallest number of evolution events to explain the observed sequence data. While maximum parsimony algorithms build very accurate trees, their computation tends to be relatively slow compared to distance-matrix based methods. Maximum likelihood methods use probability distributions to assess the likelihood of a given 5 http://mc.manuscriptcentral.com/bibManuscripts submitted to Briefings in Bioinformatics phylogenetic tree according to a substitution model. This method unfortunately has a high complexity for computing the optimal tree. Many current tools use a combination of methods

EvoMining Results

Archaea

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 celular walls are not composed from peptidoglycans as in Bacteria. Archaeal proteins may be higlhy 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 mecanism, 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 3: Families on Archaeabacteria

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

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 Actinobacteria evoMining central pathways.

Expansions BoxPlot by metabolic family

```
label(path = "chapter2/Archaeas/expansion_plotArchaeas.pdf", caption = "Expansions Boxplot",label = "Ar
```

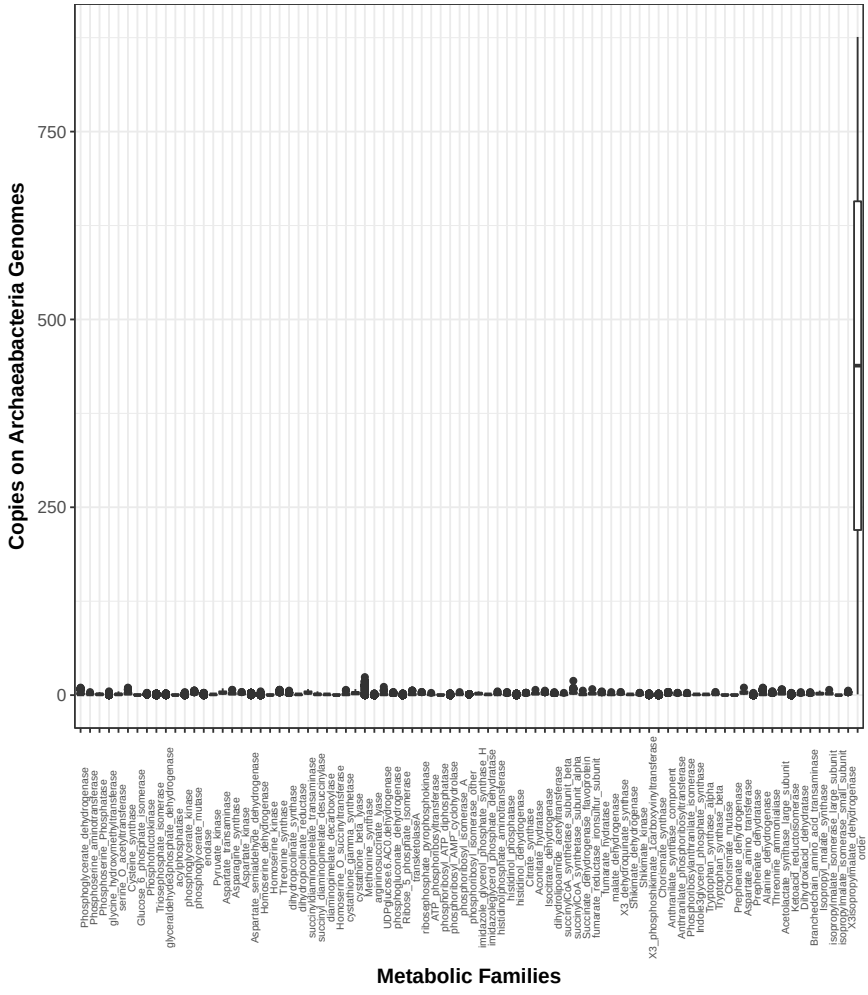
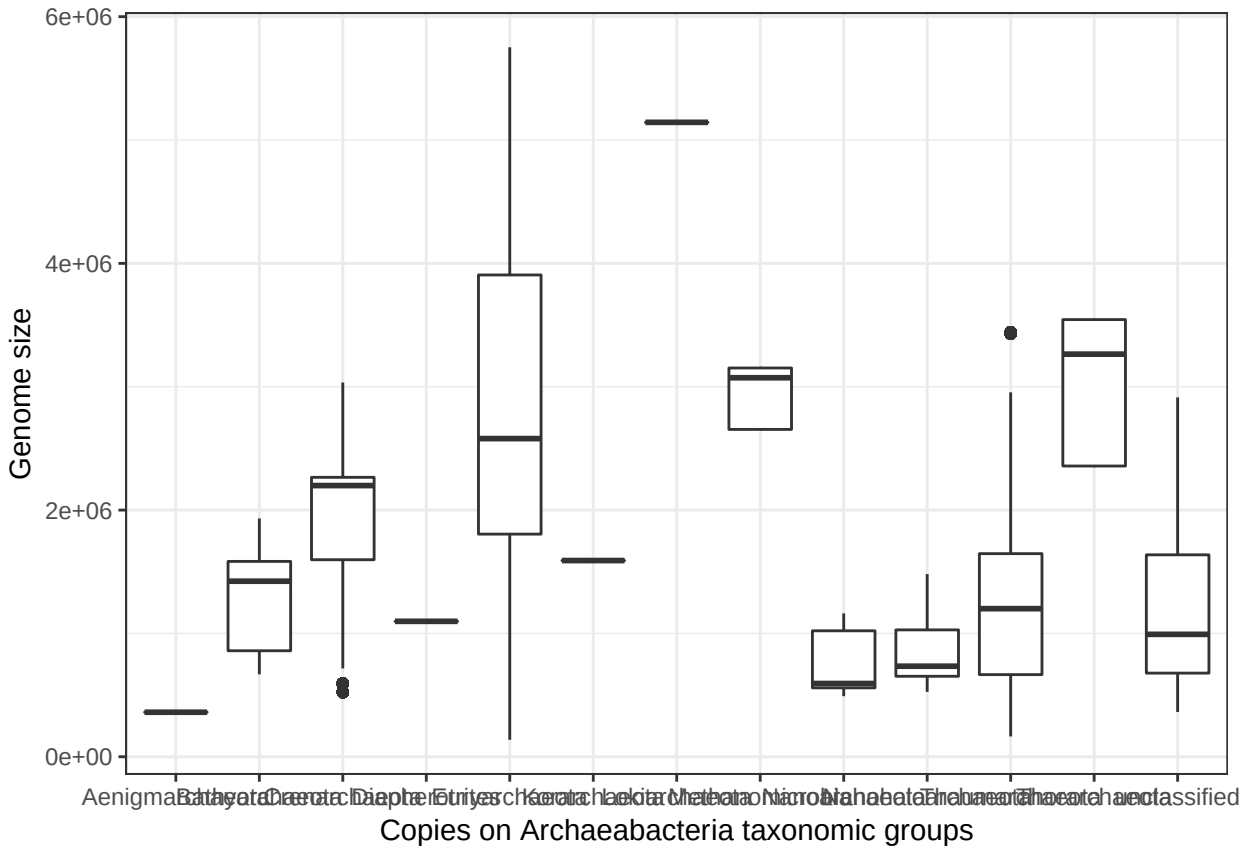


Figure 8: Expansions Boxplot

Here is a reference to the expansion boxplot: Figure 8.

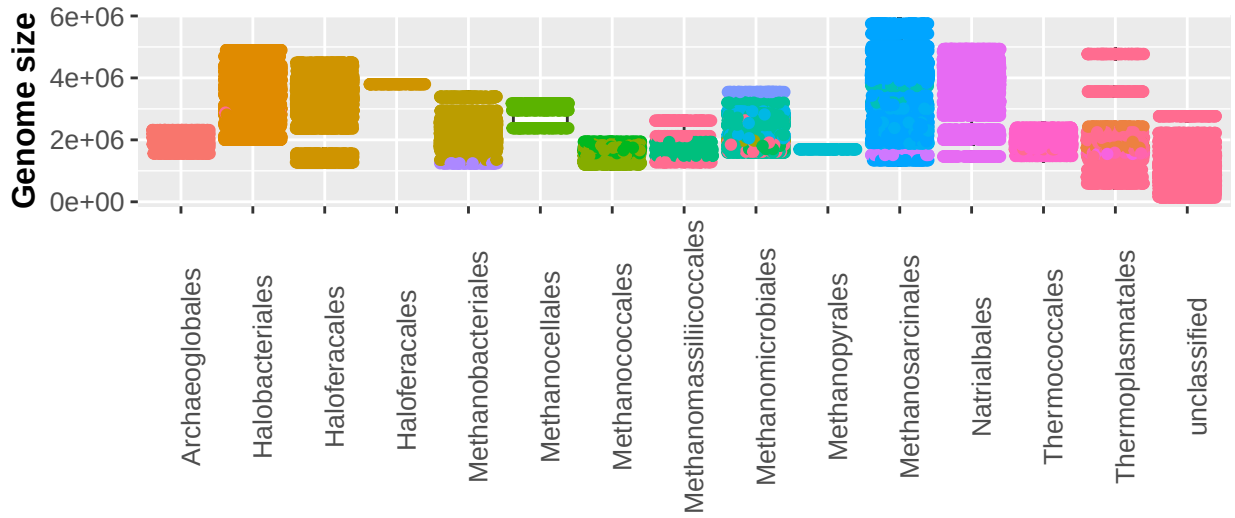
Expansions BoxPlot by metabolic family by phylum

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

ArchaeasTotalBP.m<-merge(ArchaeasHeatPlot,ArchaeasTaxa,by.x="RastId",by.y="RastId") ## works as expecte
##melt in r
ArchaeasHeatPlotBP.m <- melt(ArchaeasTotalBP.m,id =c("RastId","Name","SuperPhylum","Phylum","Class","Or
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_boxp
```




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

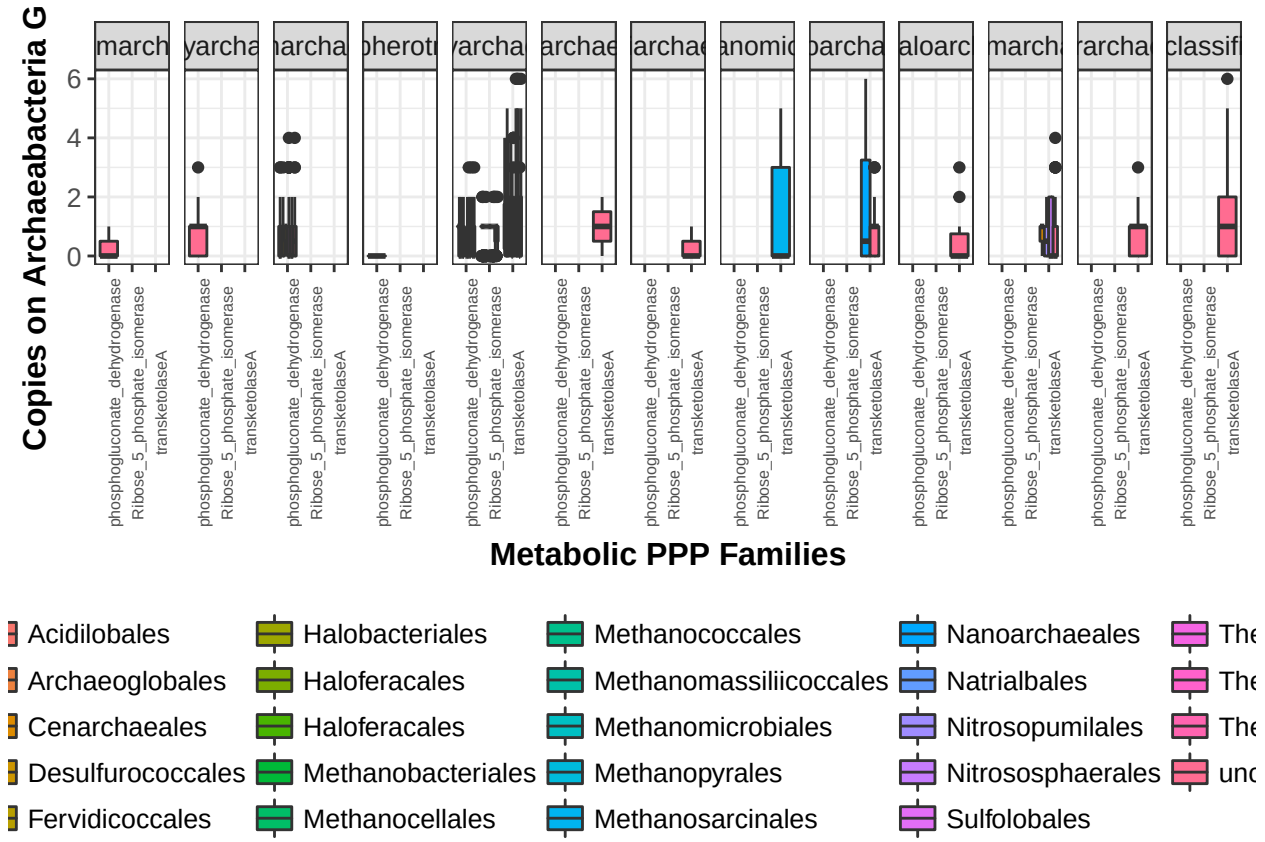
```
## Halobacteria
MetFam_BP.m=subset(ArchaeasHeatPlotBP.m,Phylum=="Euryarchaeota")
ggplot(MetFam_BP.m, aes(x=MetFam_BP.m$Order, y=MetFam_BP.m$Size))+ geom_boxplot() +theme(plot.title = el
```



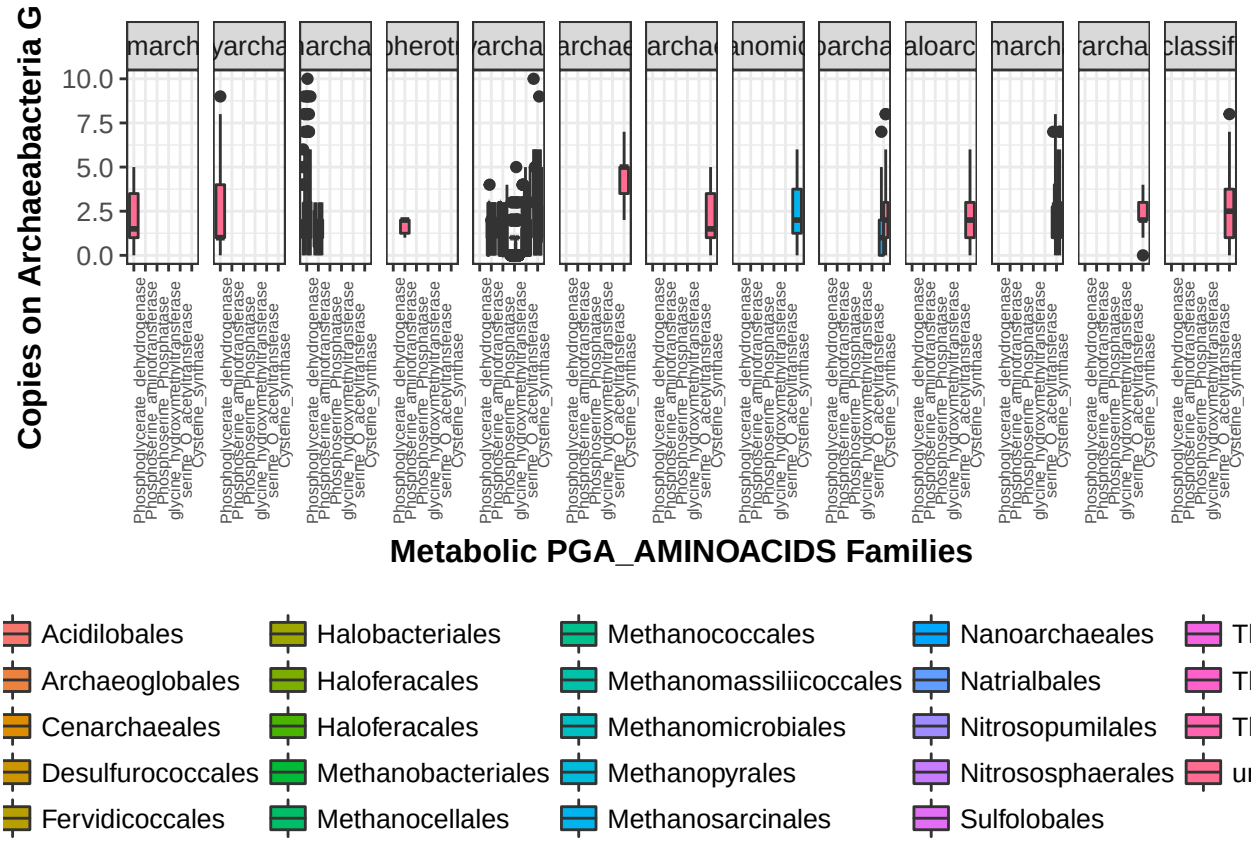
Halobacteria orders size taxonomic groups

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

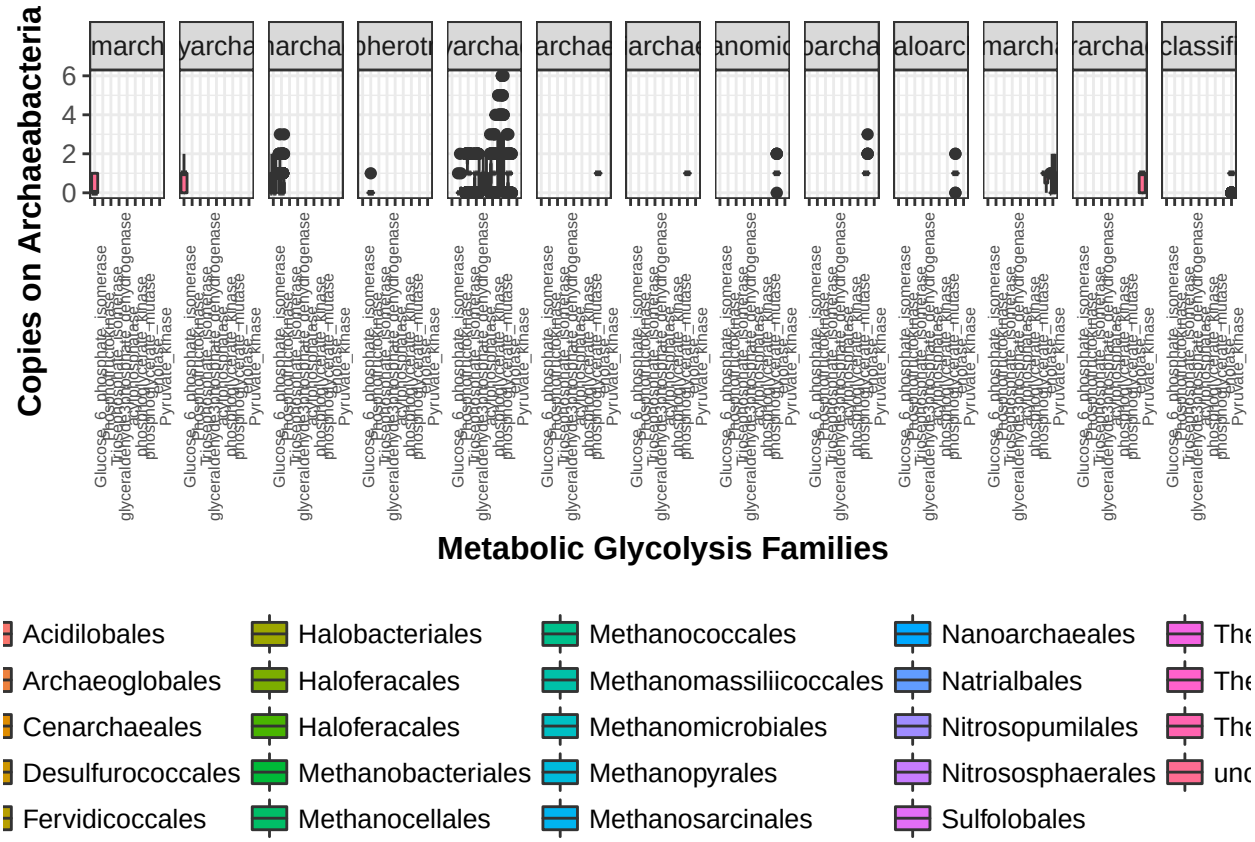
```
#MetFam_BP.m=subset(ArchaeasHeatPlotBP.m,Family=="Methanosarcinaceae")
#ggplot(MetFam_BP.m, aes(x=MetFam_BP.m$Size, y=MetFam_BP.m$value))
#+theme(plot.title = element_text(size = 14, face = "bold"), text = element_text(size = 12), axis.title


  #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
```



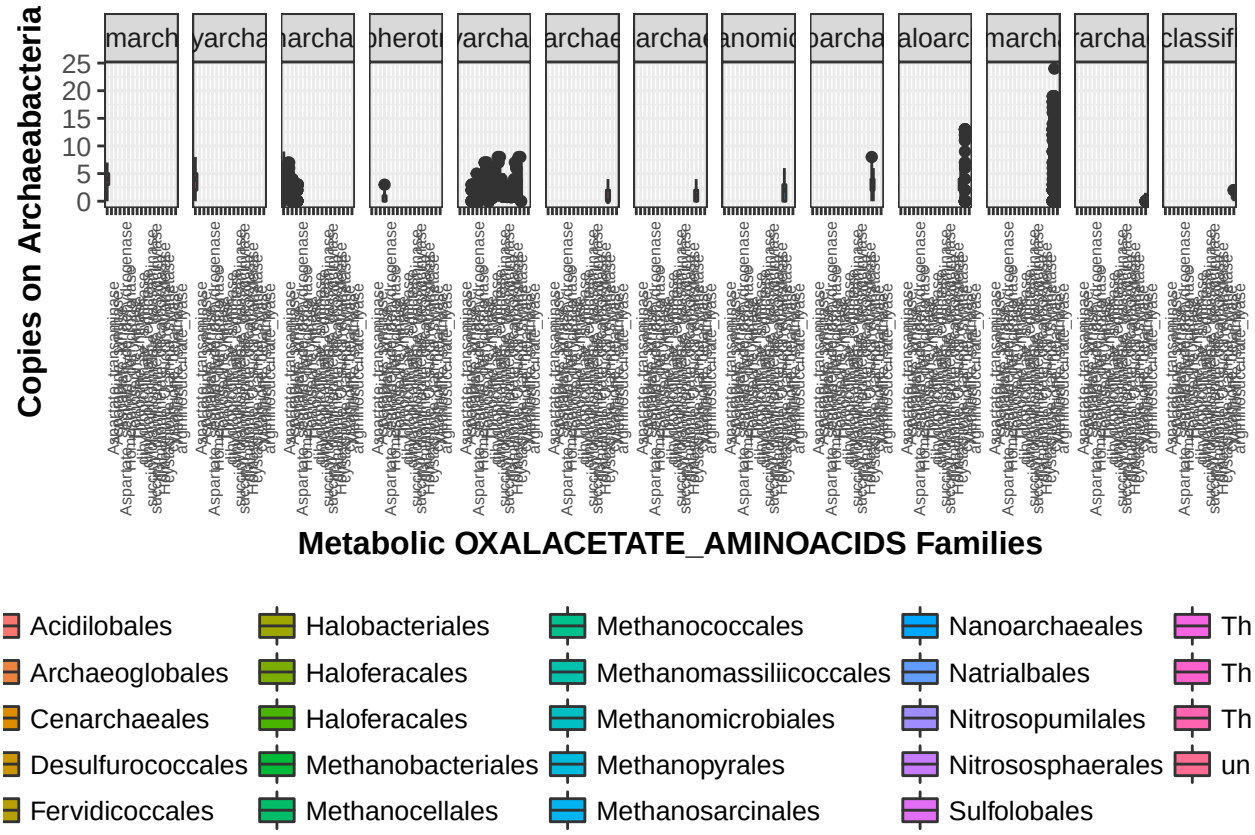

```
#+ geom_jitter(aes(color=MetFam_BP.m$Phylum))

MetFam=subset(ArchaeasCentral,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
```



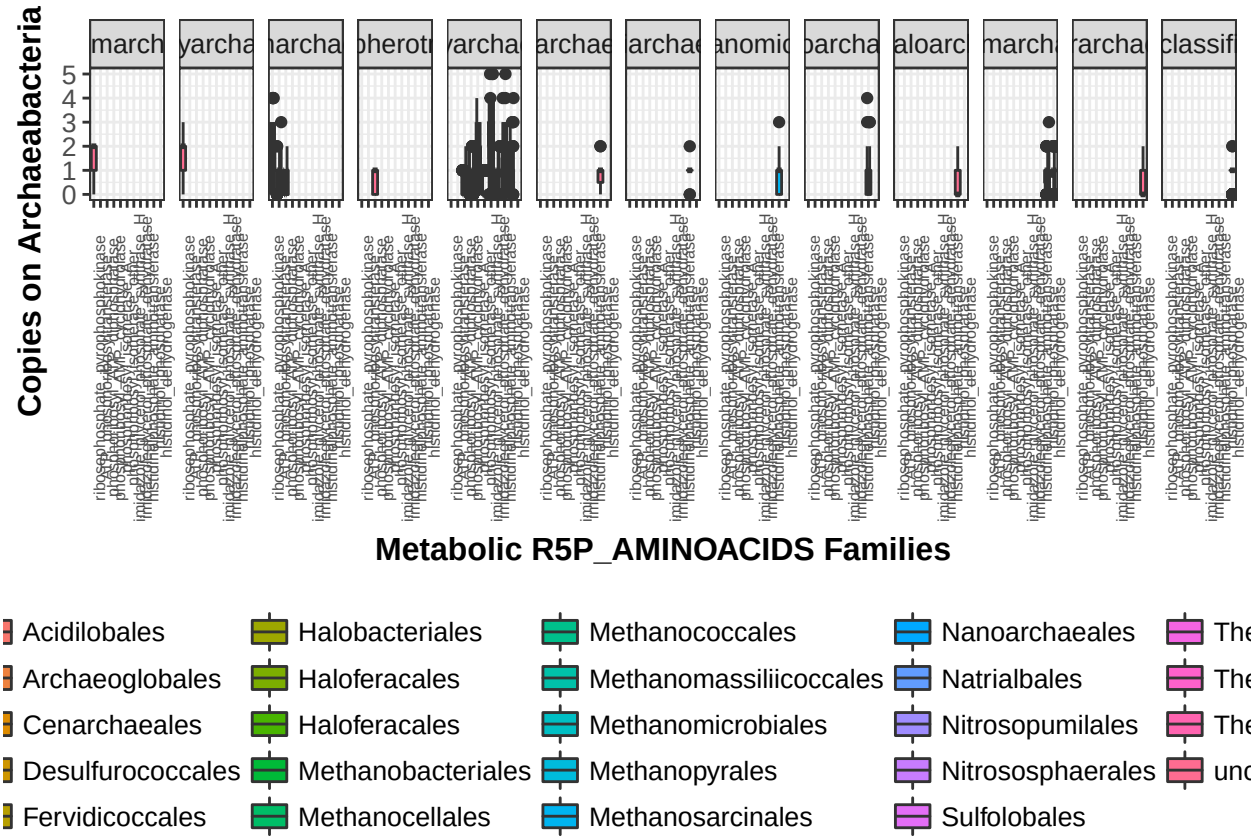

```
#+ geom_jitter(aes(color=MetFam_BP.m$Phylum))

MetFam=subset(ArchaeasCentral,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
```



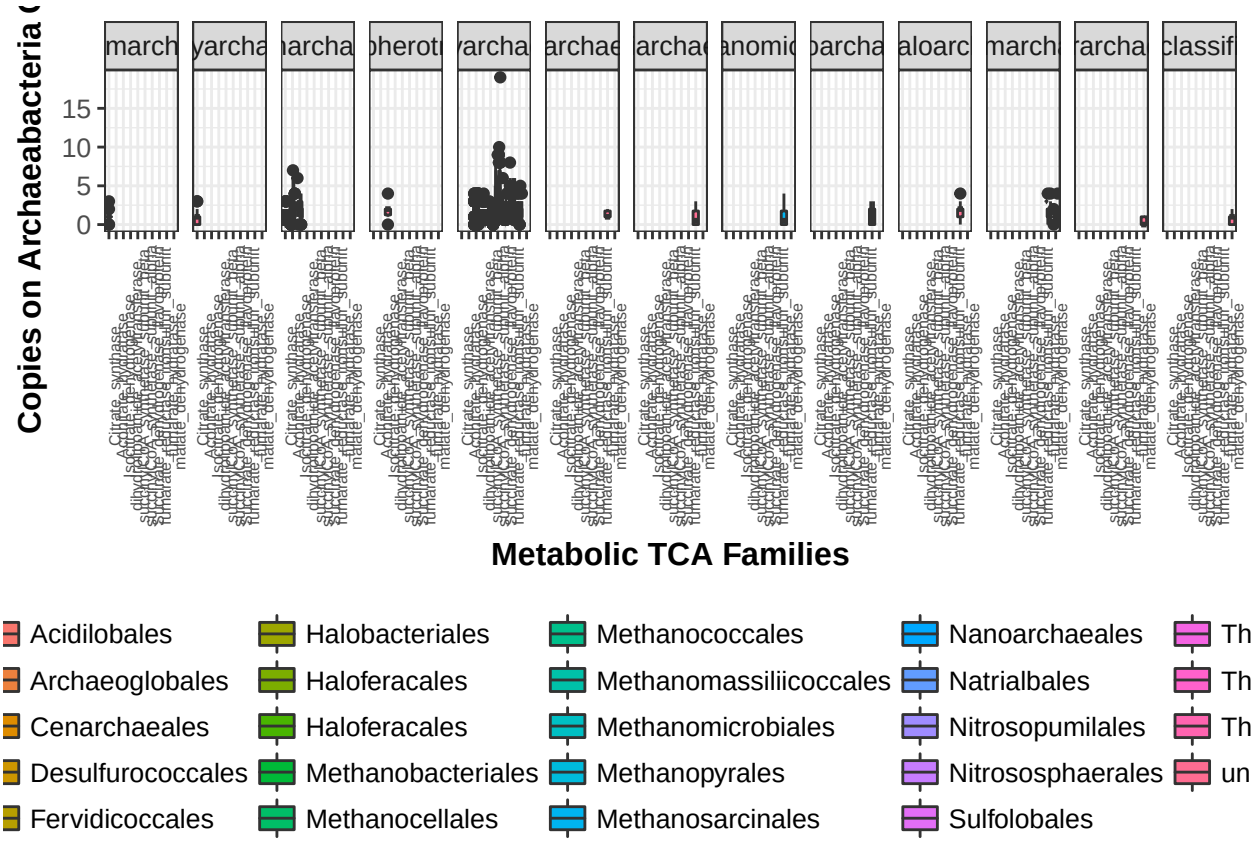

```
#+ geom_jitter(aes(color=MetFam_BP.m$Phylum))

MetFam=subset(ArchaeasCentral,Pathway=="OXALACETATE_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
```



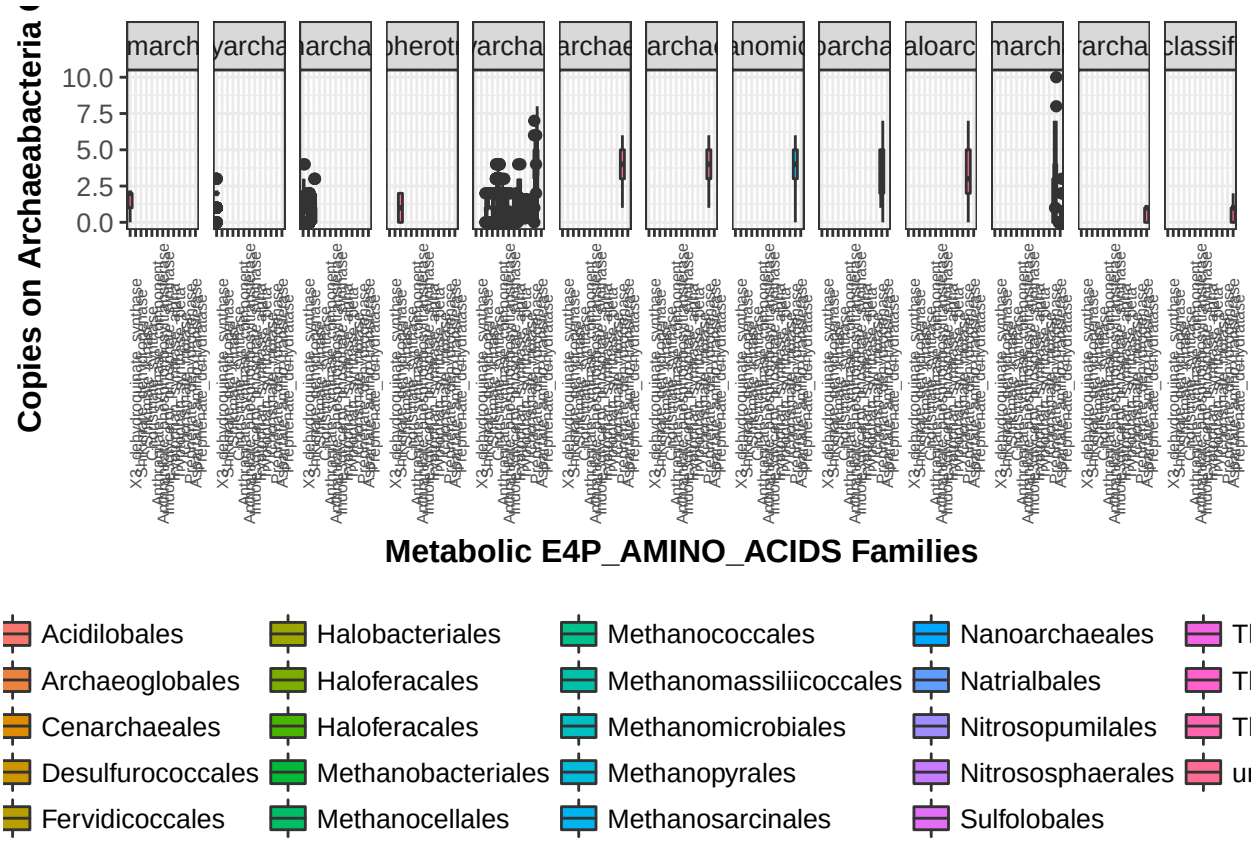

```
#+ geom_jitter(aes(color=MetFam_BP.m$Phylum))

MetFam=subset(ArchaeasCentral,Pathway=="R5P_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
```



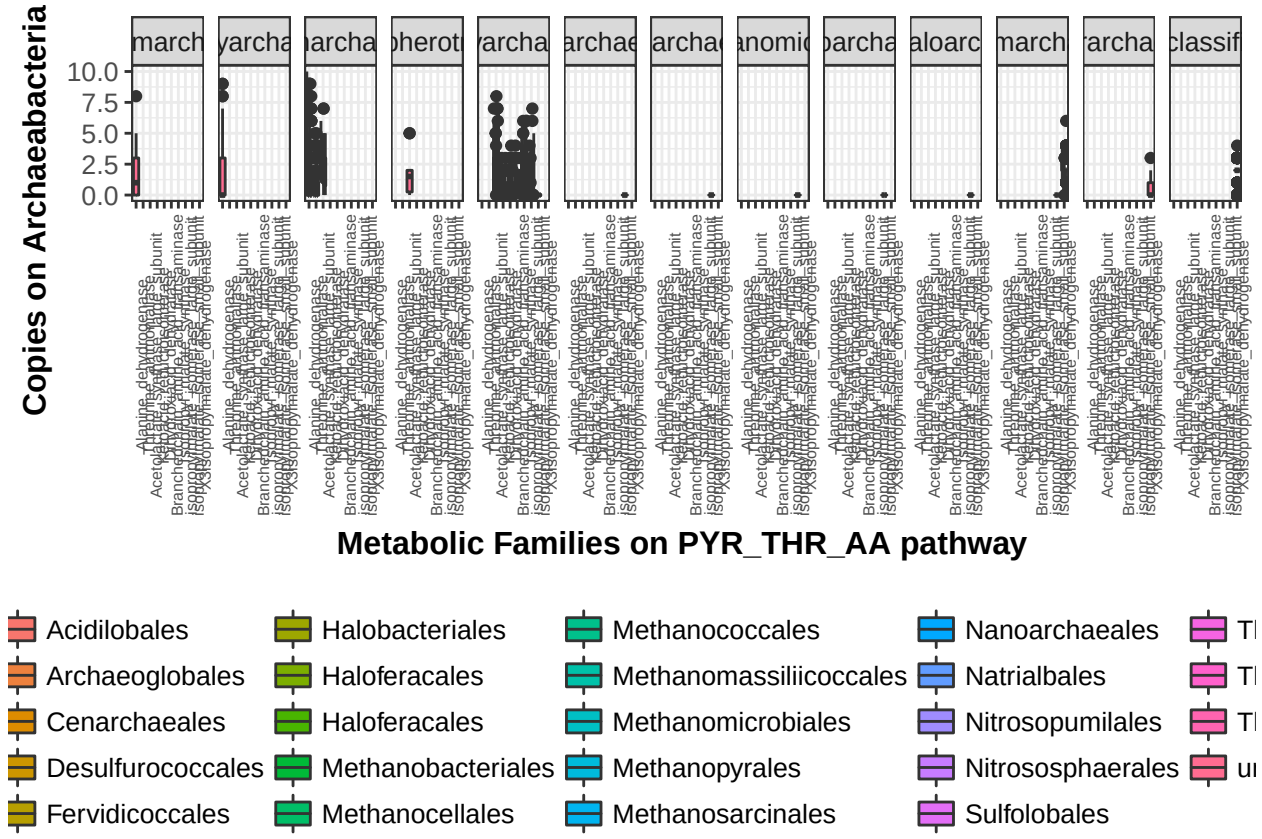

```
 #+ geom_jitter(aes(color=MetFam_BP.m$Phylum))

#
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
```




```
#+ 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
```



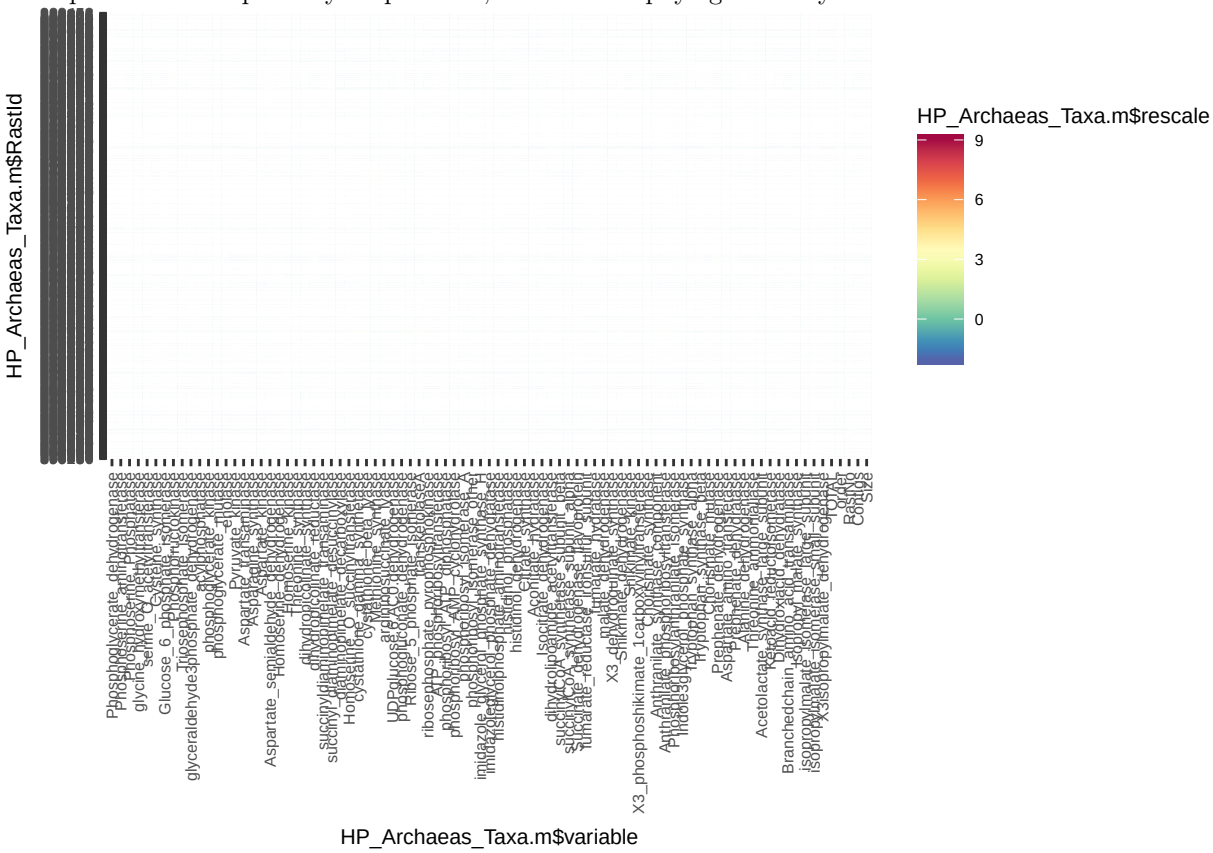

```
#+ geom_jitter(aes(color=MetFam_BP.m$Phylum))


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
```




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

#ggsave("chapter2/Archaeas/expansion_plotArchaeas.pdf", plot = expansion_plotArchaea,height = 8, width =
```

Central pathway expansions

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



Here is a reference to the HeatPlot: Figure 9.

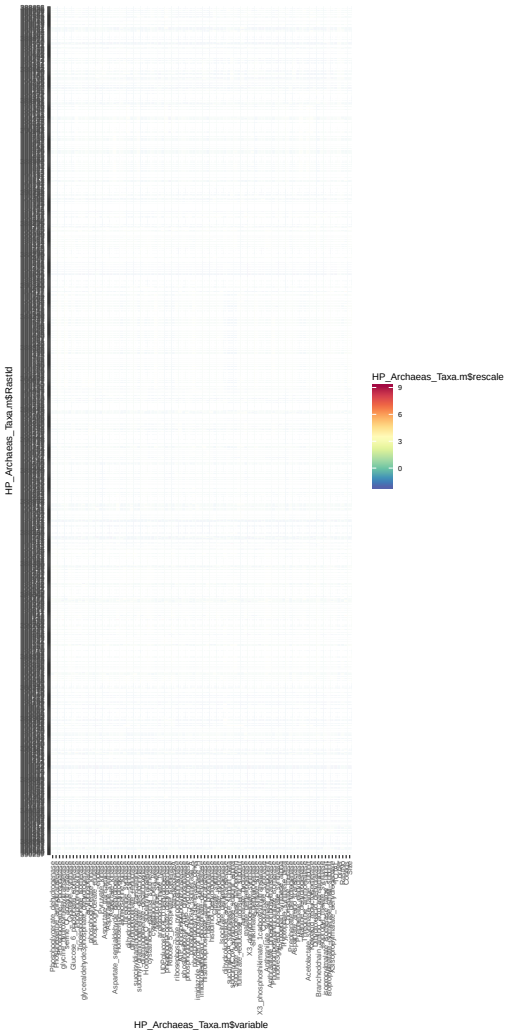


Figure 9: Archaeas Heatplot

Genome Size correlations

Correlation between genome size and AntiSMASH products

Genome size vs Total antismash cluster coloured by order

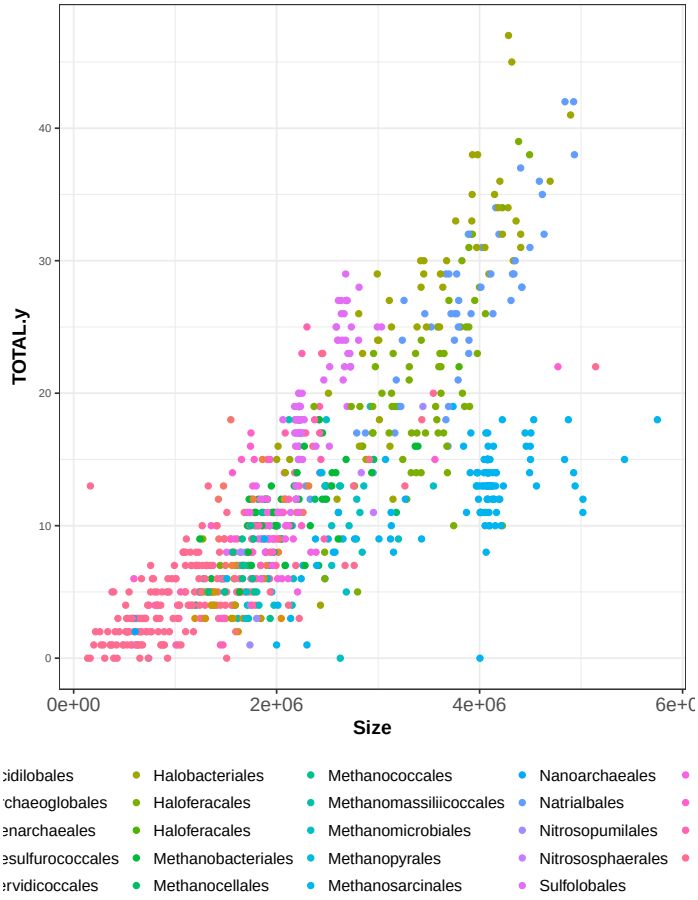


Figure 10: Correlation between Archaeas genome size and antismash Natural products detection colored by Order

Here is a reference to Genome size vs Total antismash cluster: Figure 10.

Genome size vs Total antismash cluster detected splitted by order

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

Correlation between genome size and Central pathway expansions

Genome size vs Total central pathway expansion coloured by order

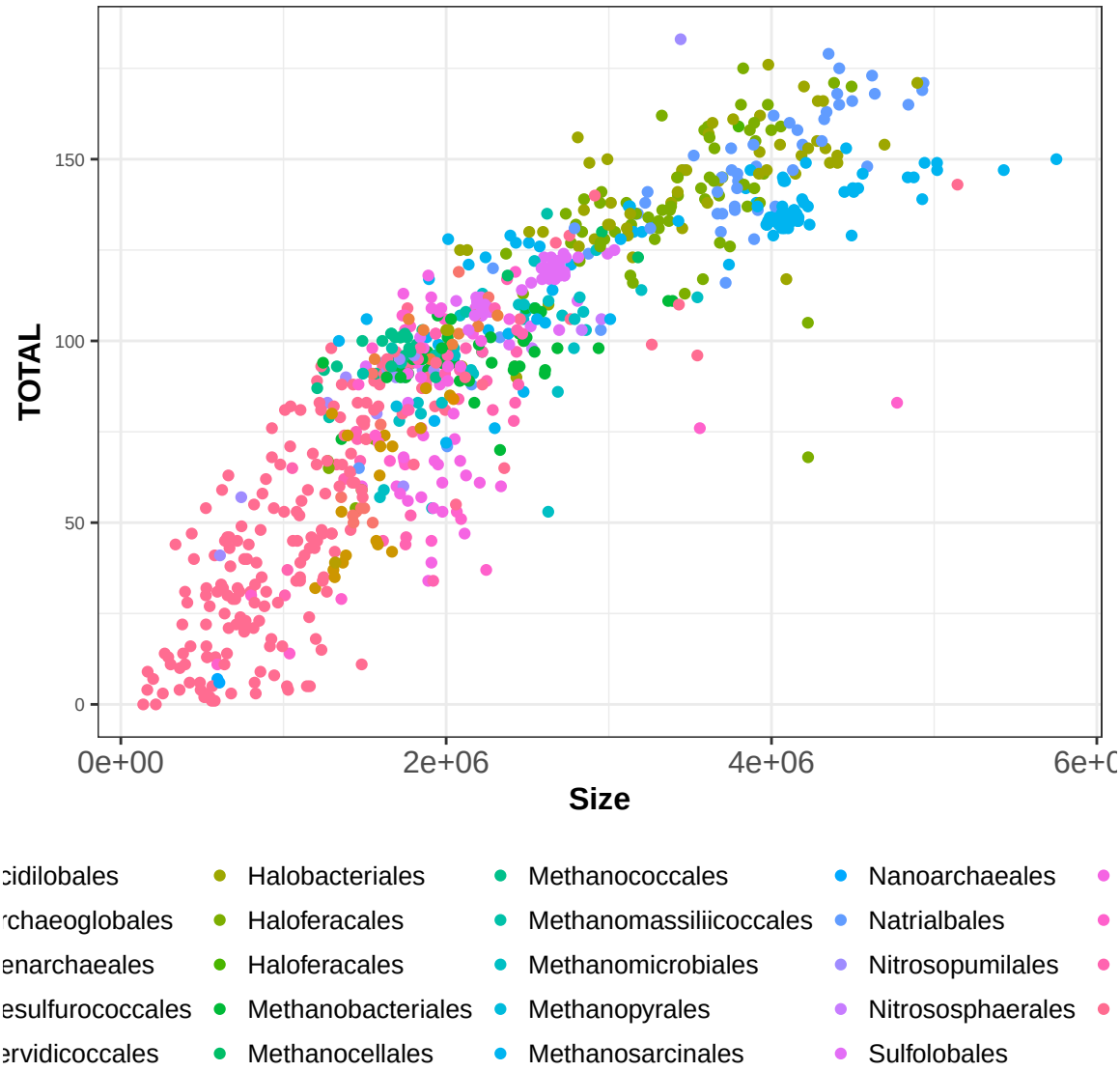


Figure 11: Correlation between Archaeas genome size and central pathway expansions

Here is a reference to the size vs Total central pathway expansion plot: Figure 11.

Genome size vs Total central pathway expansion grided by order

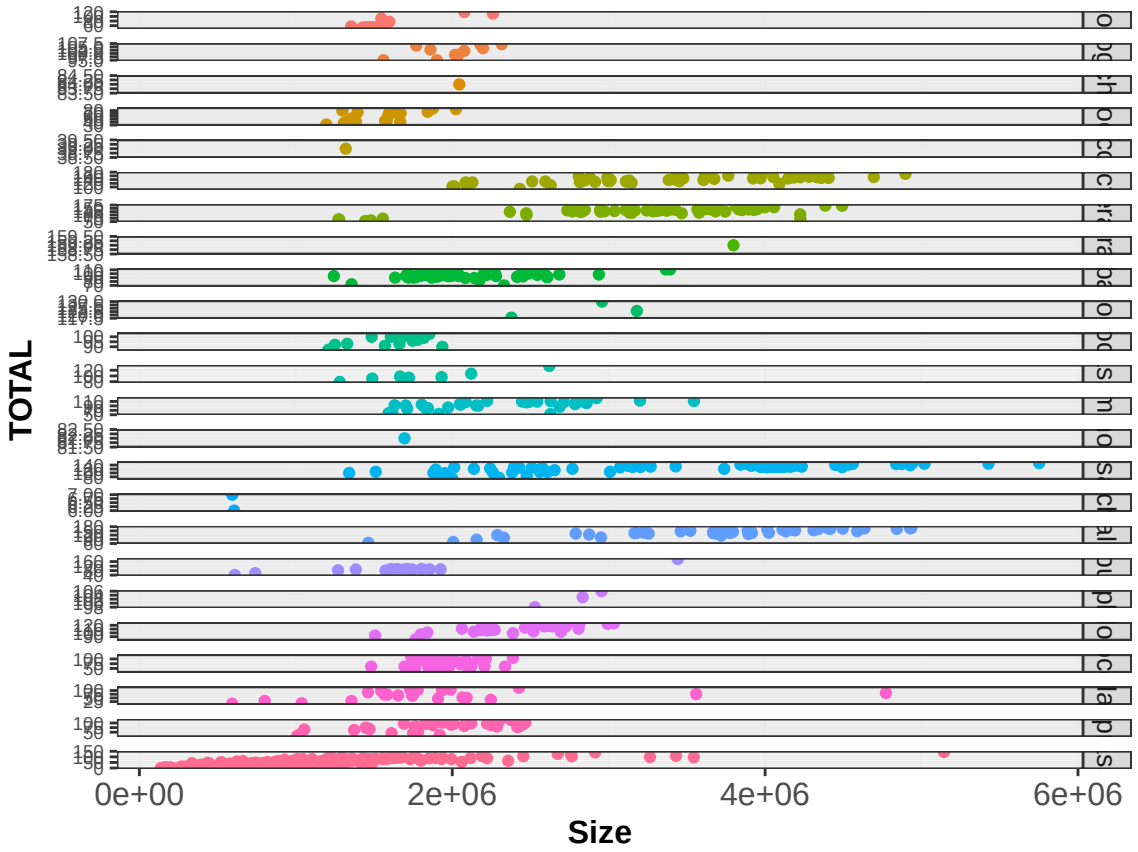


Figure 12: Correlation between Archaeas genome size and central pathway expansions grided by order

Here is a reference to the Genome size vs Total central pathway expansion grided by order plot: Figure 12.

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 differente 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

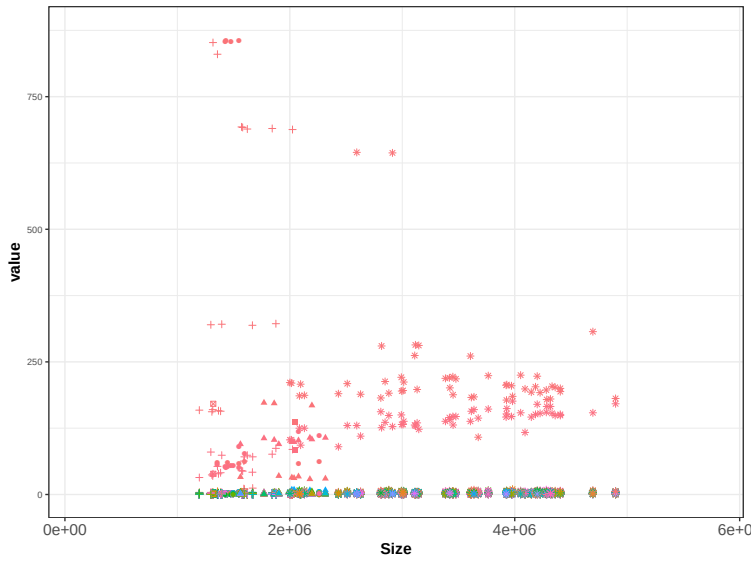


Figure 13: Correlation between Archaeas Genome size vs Total central pathway expansion coloured by metabolic Family

Here is a reference to the Genome size vs Total central pathway expansion coloured by metabolic Family plot: Figure 13.

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

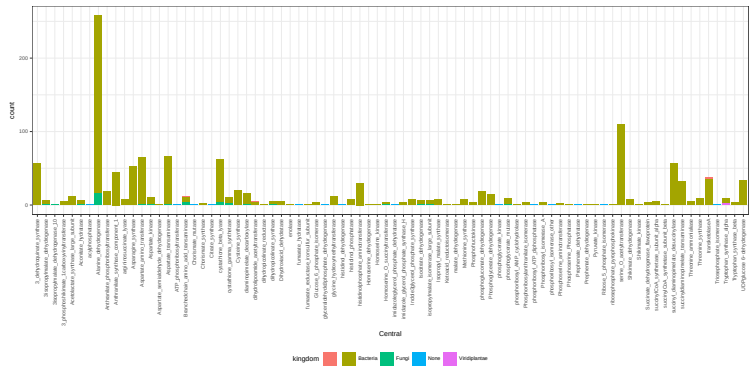


Figure 14: Archaeas Recruitmens on central families coloured by kingdom

Here is a reference to Recruitments after central pathways expansions colourd by Kingdom plot: Figure 14.

Recruitments after central pathways expansions colourd by taxonomy

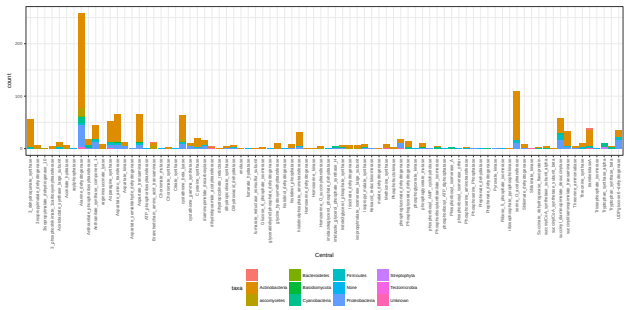


Figure 15: Archaeas Recruitmens on central families coloured by taxonomy

Here is a reference to Recruitments after central pathways expansions colourd by taxa plot: Figure 15.

Archaeas AntiSMASH

Taxonomical diversity on Archaeasbacteria Data

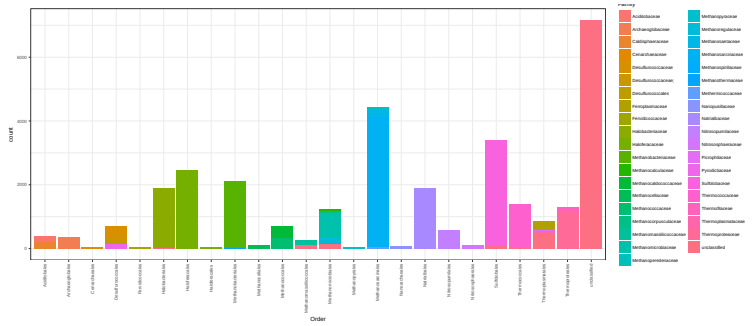


Figure 16: Archaeas Diversity

Here is a reference to Recruitments after central pathways expansions colourd by taxa plot: Figure 16.

Smash diversity

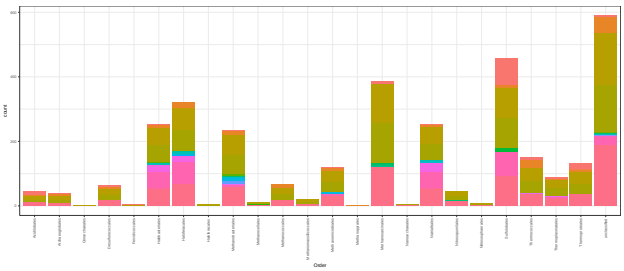


Figure 17: Archaeas Smash Taxonomical Diversity

Here is a reference to Recruitments after central pathways expansions colourd by taxa plot: Figure 17.

AntisSMASH 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 18: 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 18.

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

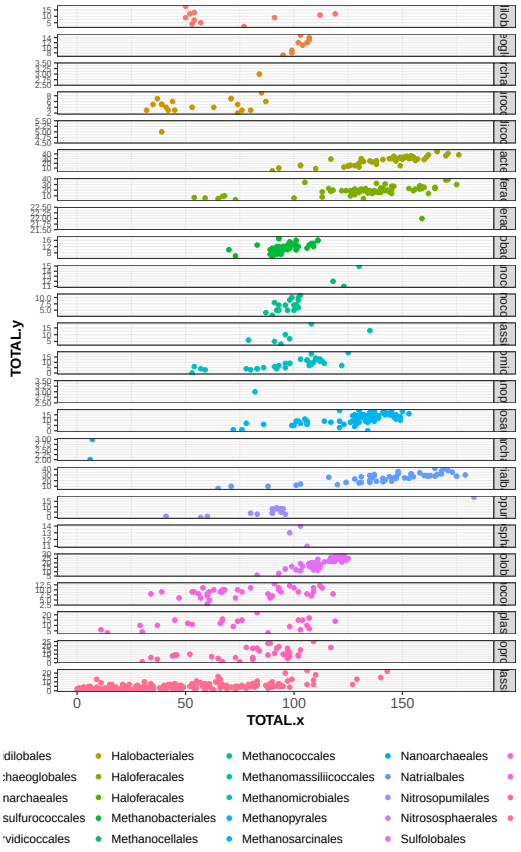


Figure 19: Correlation between Archaeas central pathway expnasions and antismash Natural products detection

Here is a reference to the expansions vs antismash NP's clusters splitted by order plot Figure 19.

AntisMAsh vs Expansions by taxonomic Family

Natural products colured by family

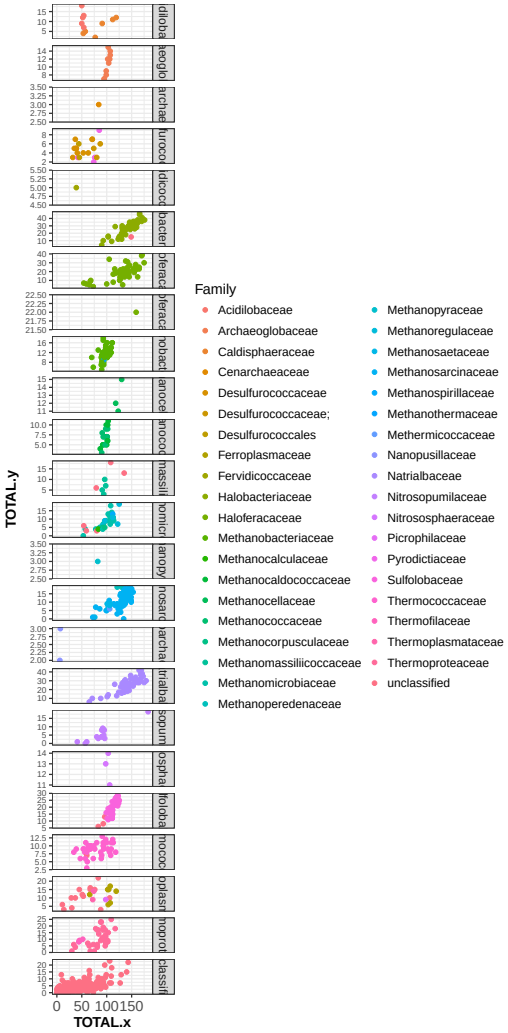


Figure 20: Archaeas Natural products by family

Here is a reference to the Natural products colured by family plot Figure 20.

Selected trees from EvoMining

Phosphoribosyl_isomerase_3 family
Figure from EvoMining

Figure 21: Phosphoribosyl isomerase A EvoMiningtree

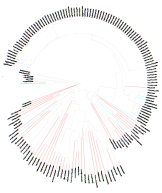


Figure 22: Phosphoribosyl isomerase other EvoMiningtree

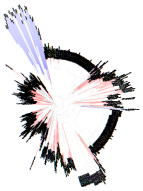


Figure 23: Phosphoribosyl anthranilate isomerase EvoMiningtree

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. [1] 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 LaTeX 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 to 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)[2]. 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 LaTeX 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[3] 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."

[1] footnote text
[2] @reedweb2007
[3] @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) with your suggestions. We love to see people using *R Markdown* for their theses, and are happy to help.

Actinobacteria

Actinobacteria is an ancient phylum {Referencia de luis}

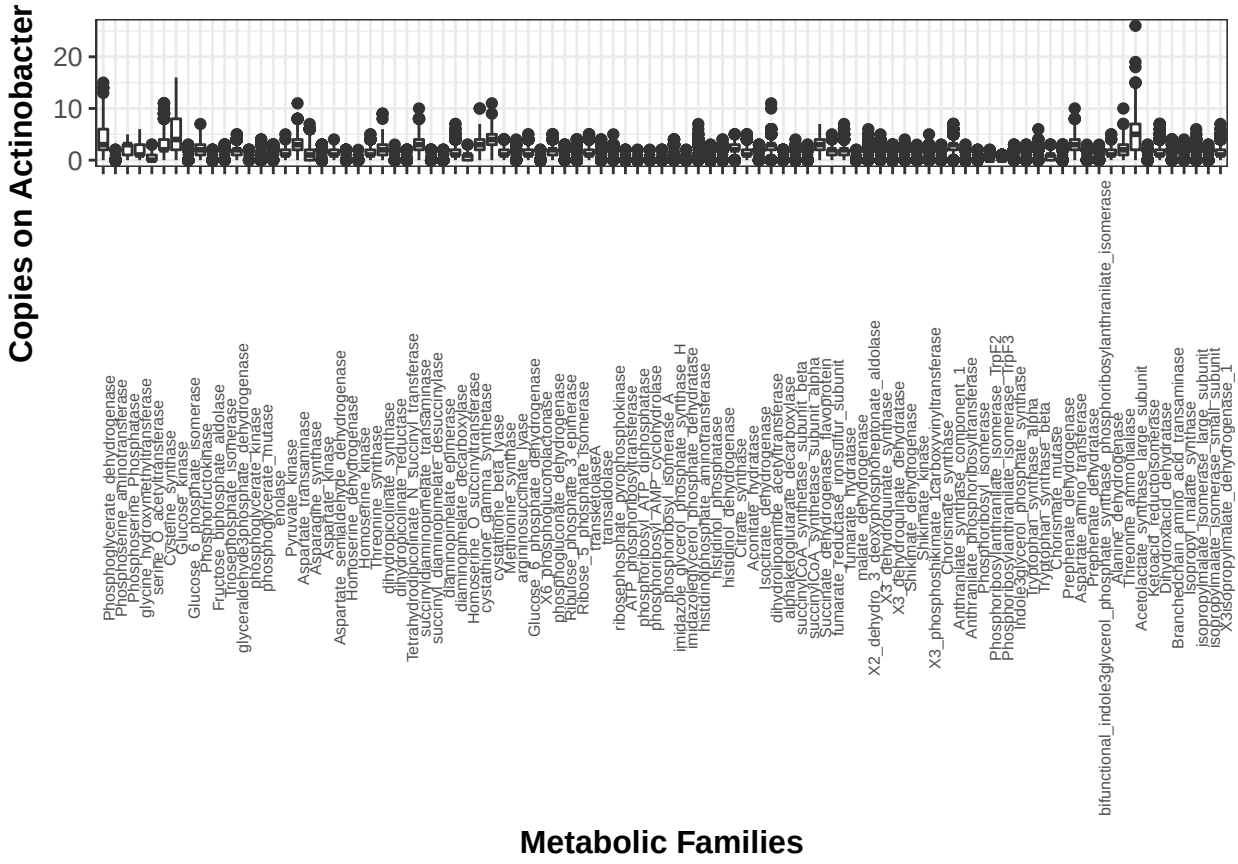
Tables

Table 4: Correlation of Inheritance Factors for Parents and Child

Factors	Correlation between Parents & Child
GenomeDB	1245
Families	65

<!– clearpage ends the page, and also dumps out all floats.Floats are things like tables and figures. –>

Expansions BoxPlot by metabolic family




```
label(path = "chapter2/Actinobacteria/expansion_plotActinos.pdf", caption = "Expansions Boxplot",label =
```

Here is a reference to the expansion boxplot: Figure 24.

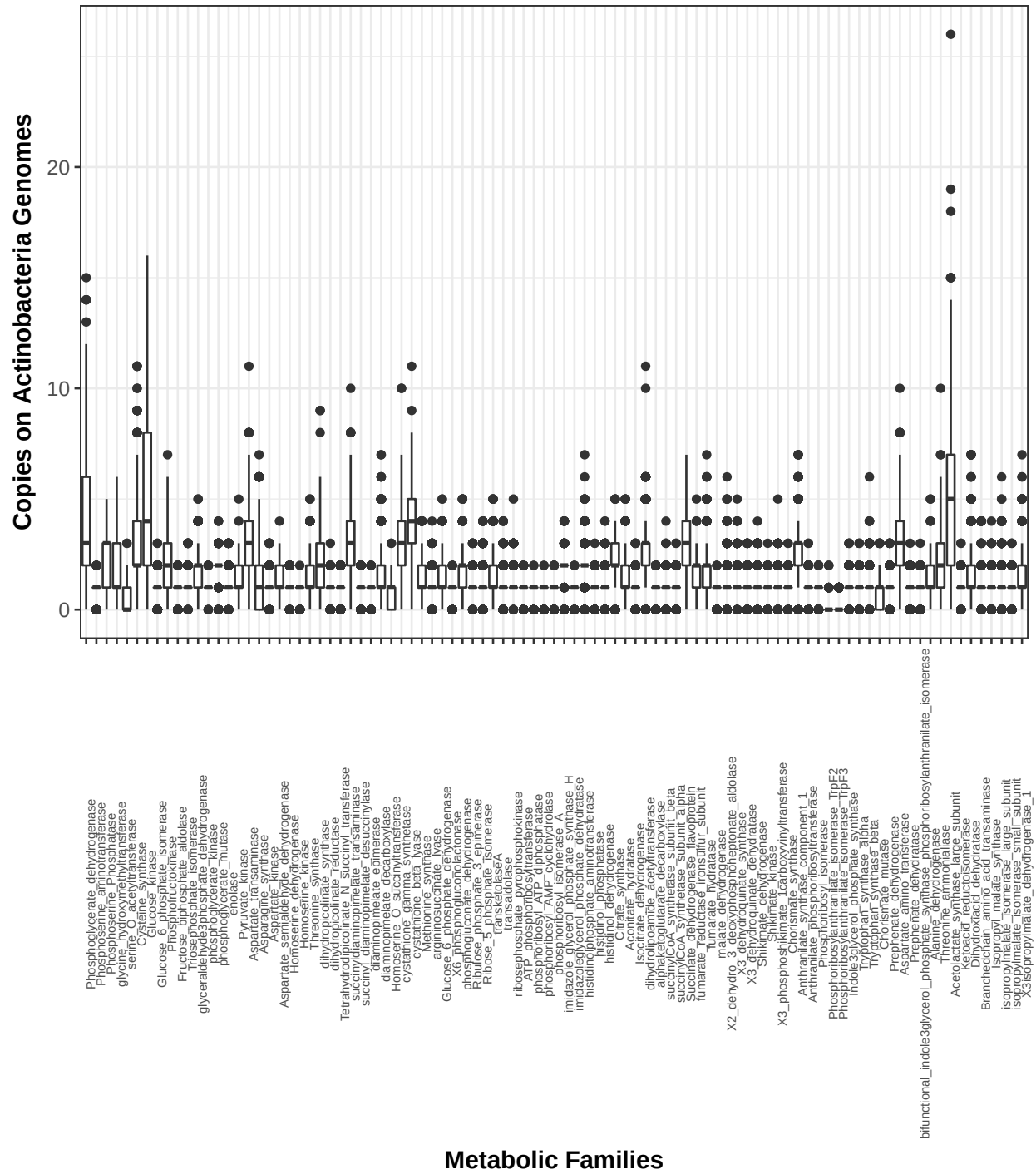


Figure 24: Expansions Boxplot

Central pathway expansions

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

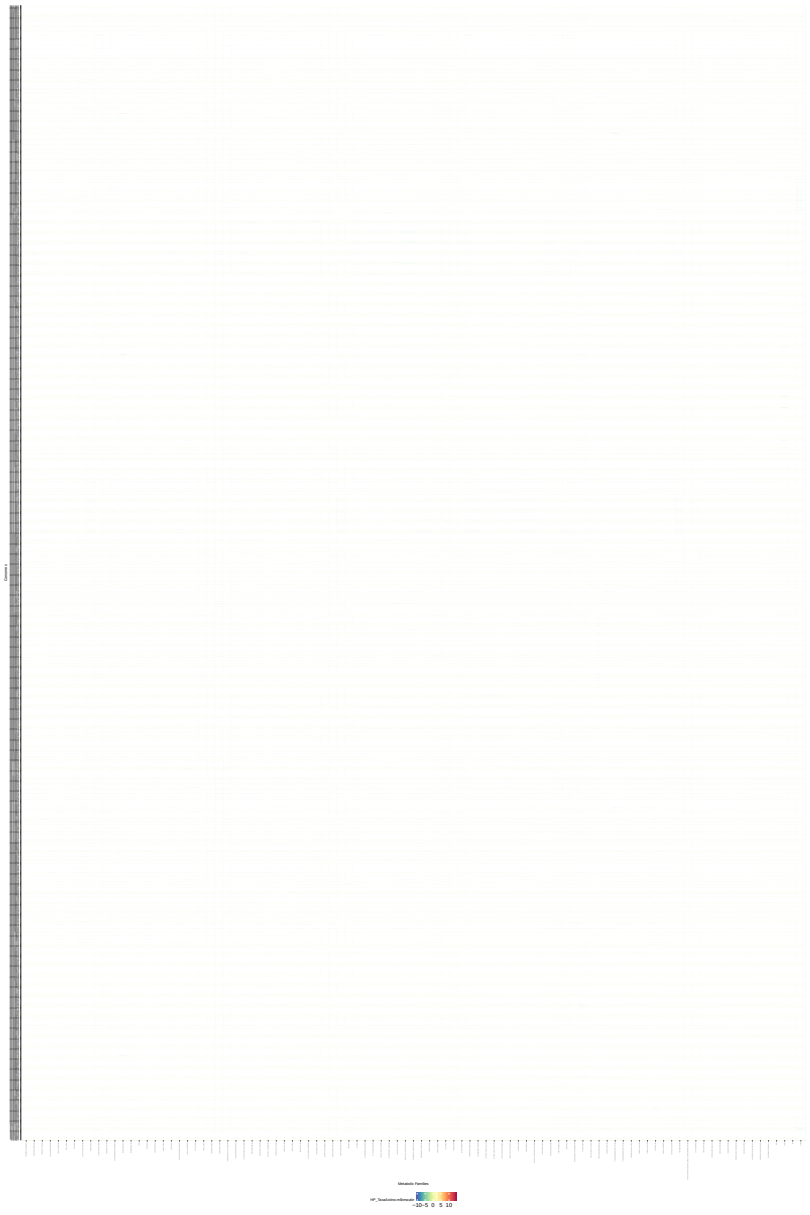


Figure 25: Actinobacterial Heatplot

Here is a reference to the HeatPlot: Figure 25.

PPP pahtway expansions restricted to *Streptomycetaceae* family HeatPlot: Figure 25.

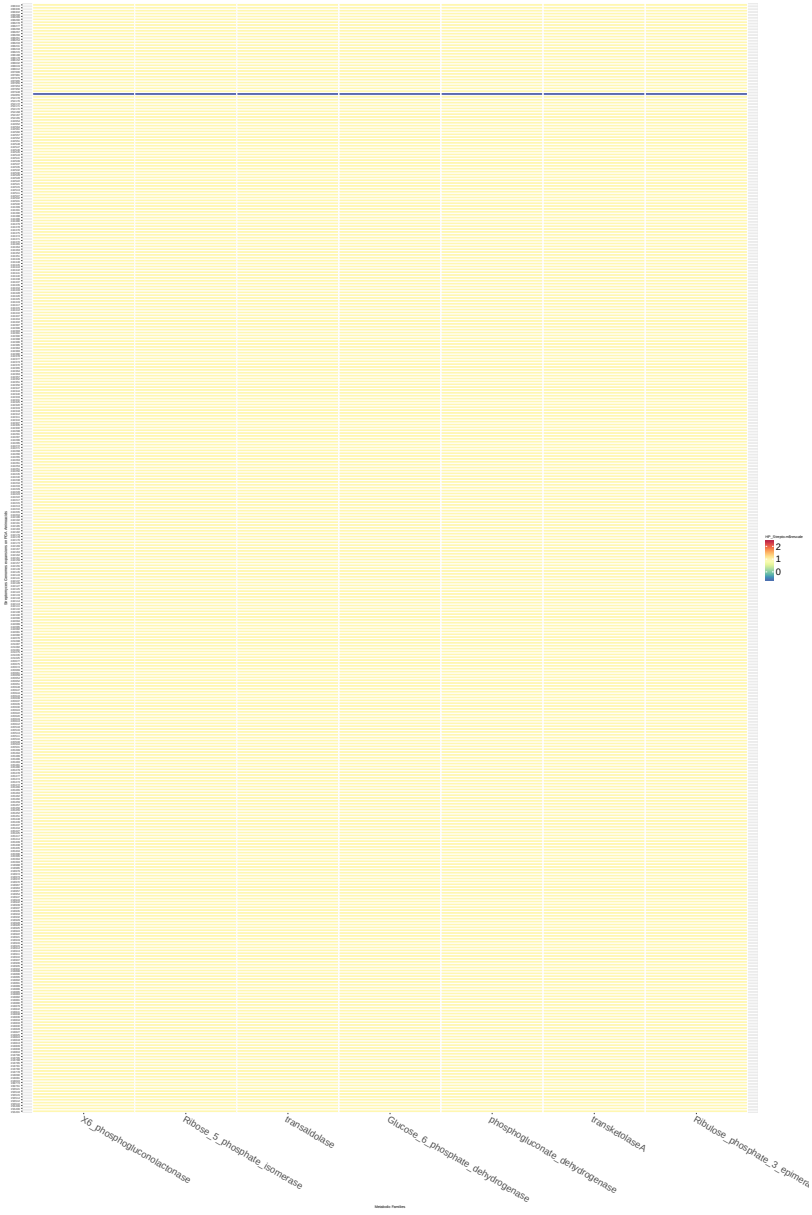


Figure 26: Streptomyces Genomes expansions on PGA Aminoacids HeatPlot

Here is a reference to the HeatPlot: Figure 26.

Genome Size correlations

Correlation between genome size and AntiSMASH products

```
## Warning: Removed 1 rows containing missing values (geom_point).
```

```
## Warning: Removed 1 rows containing missing values (geom_point).
```

Genome size vs Total antismash cluster coloured by order

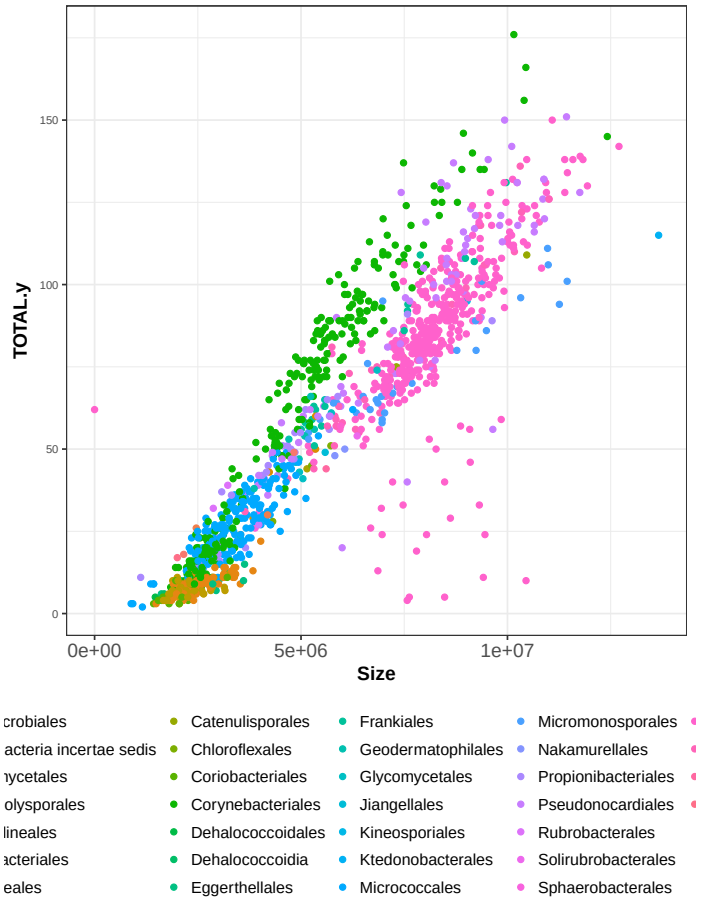


Figure 27: Correlation between Actinos genome size and antismash Natural products detection colored by Order

Here is a reference to Genome size vs Total antismash cluster: Figure 27.

Genome size vs Total antismash cluster detected splitted by order

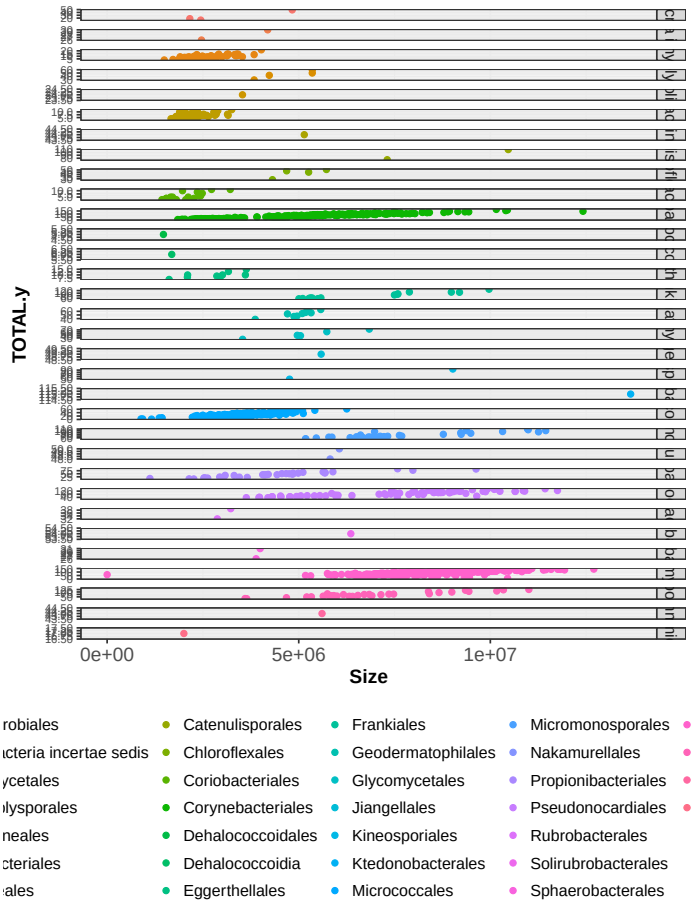


Figure 28: Correlation between Actinos genome size and antismash Natural products detection grided by Order

Here is a reference to Correlation between genome size and antismash Natural products detection grided by Order plot: Figure 28.

Correlation between genome size and Central pathway expansions

```
## Warning: Removed 1 rows containing missing values (geom_point).
```

```
## Warning: Removed 1 rows containing missing values (geom_point).
```

Genome size vs Total central pathway expansion coloured by order

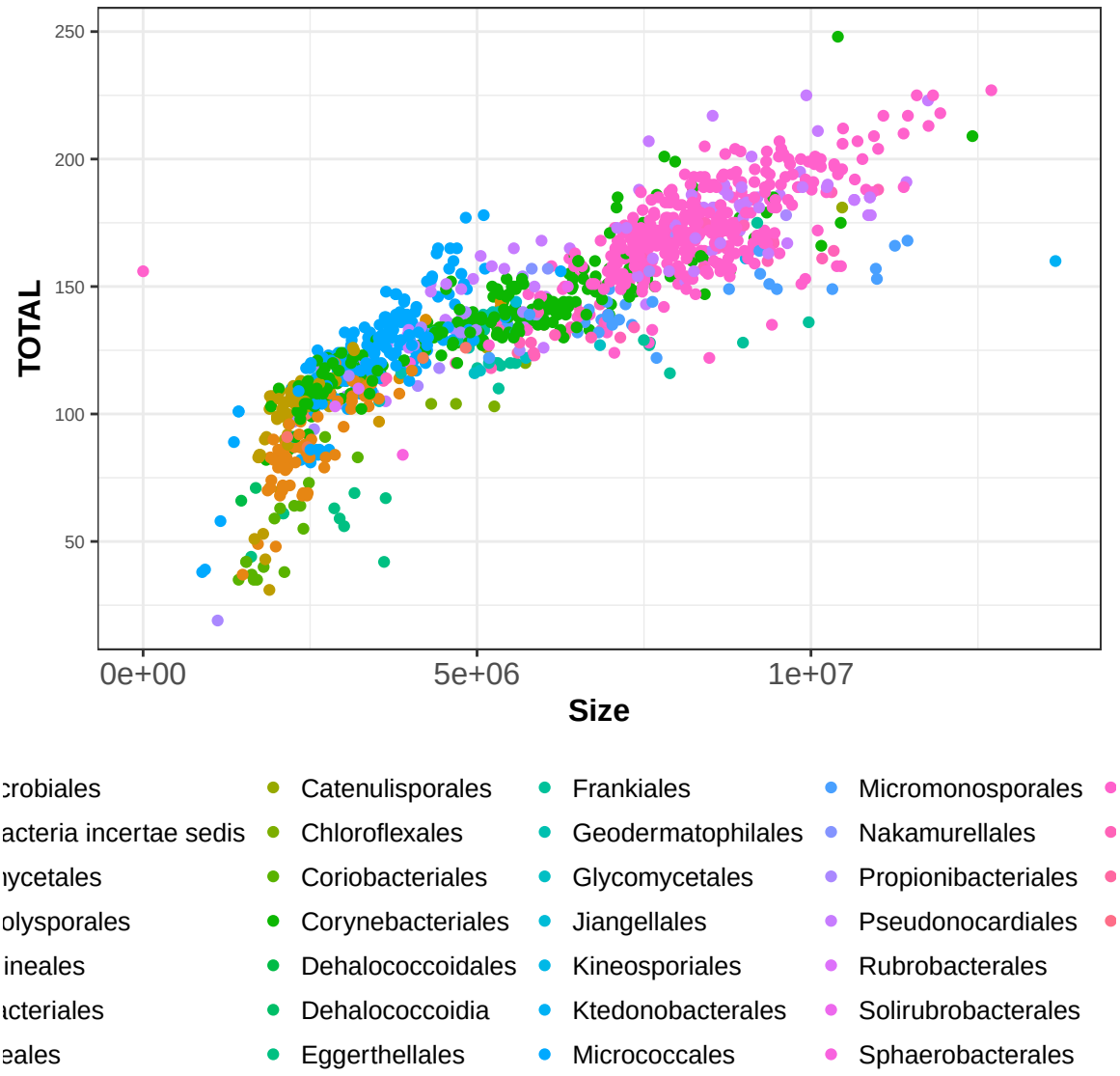


Figure 29: Correlation between Actinos genome size and central pathway expansions

Here is a reference to the size vs Total central pathway expansion plot: Figure 29.

Genome size vs Total central pathway expansion grided by order

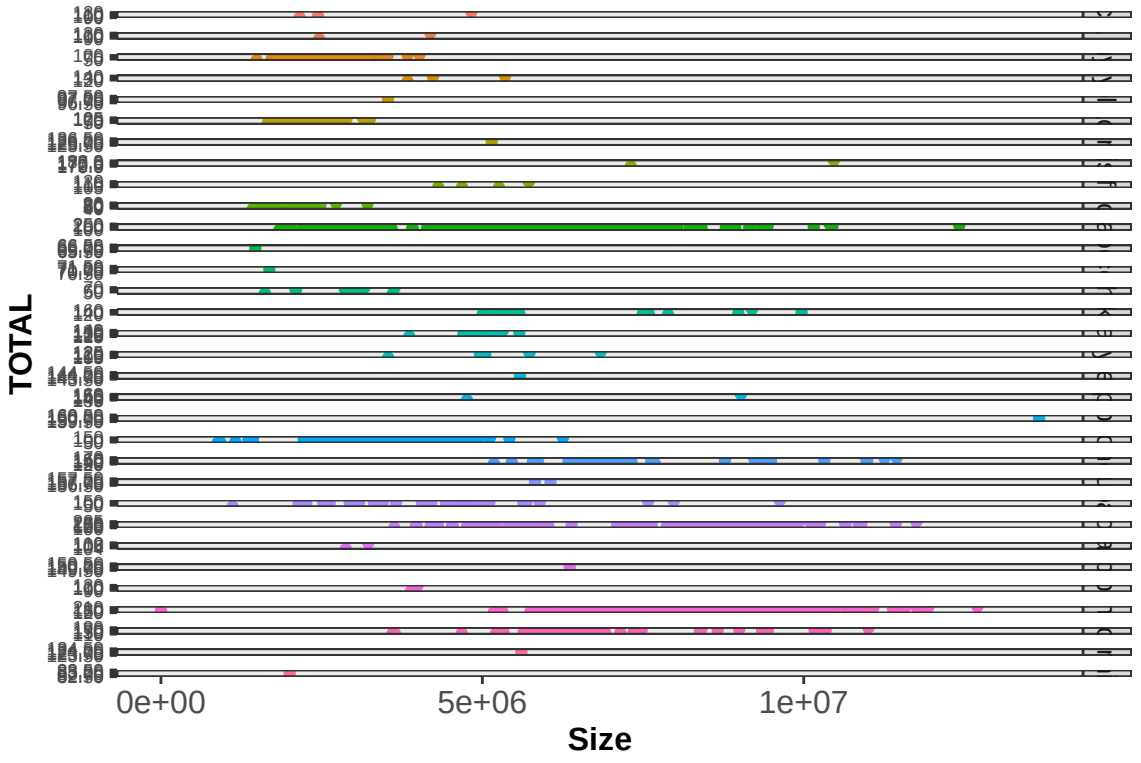


Figure 30: Correlation between Actinos genome size and central pathway expansions grided by order

Here is a reference to the Genome size vs Total central pathway expansion grided by order plot: Figure 30.

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 differente 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
## 32. Consider specifying shapes manually if you must have them.
```

```
## Warning: Removed 103306 rows containing missing values (geom_point).
```

```
## Warning: Removed 94 rows containing missing values (geom_point).
```

Genome size vs Total central pathway expansion coloured by metabolic Family

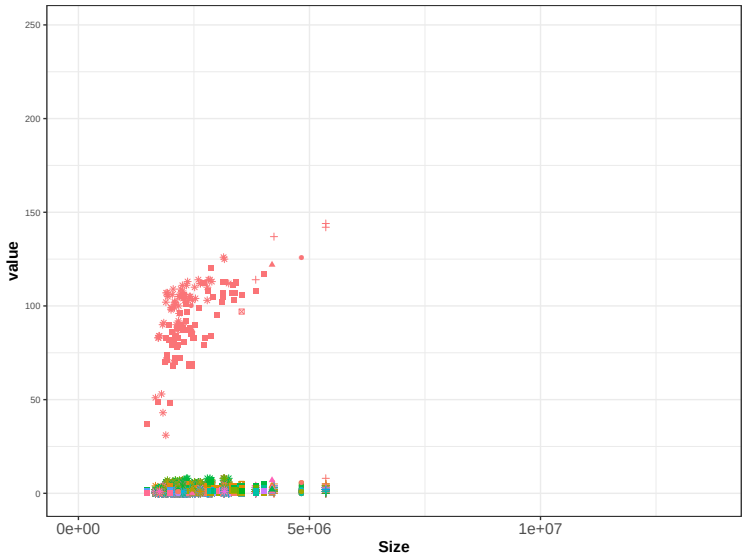


Figure 31: Correlation between Actinos Genome size vs Total central pathway expansion coloured by metabolic Family

Here is a reference to the Genome size vs Total central pathway expansion coloured by metabolic Family plot: Figure 31.

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

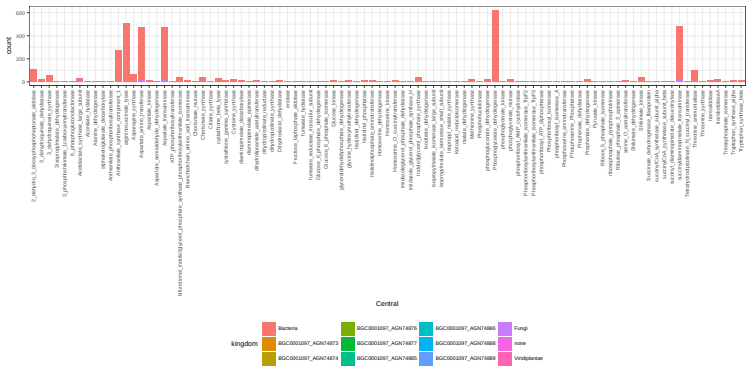


Figure 32: Actinos Recruitmens on central families coloured by kingdom

Here is a reference to Recruitments after central pathways expansions colourd by Kingdom plot: Figure 32.

Recruitments after central pathways expansions colourd by taxonomy

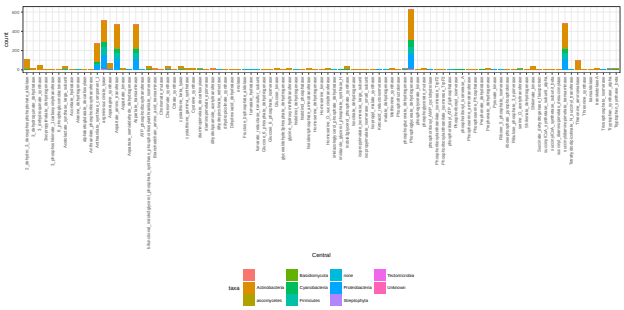


Figure 33: Actinos Recruitmens on central families coloured by taxonomy

Here is a reference to Recruitments after central pathways expansions colourd by taxa plot: Figure 33.

Actinos AntiSMASH

Taxonomical diversity on Actinosbacteria Data

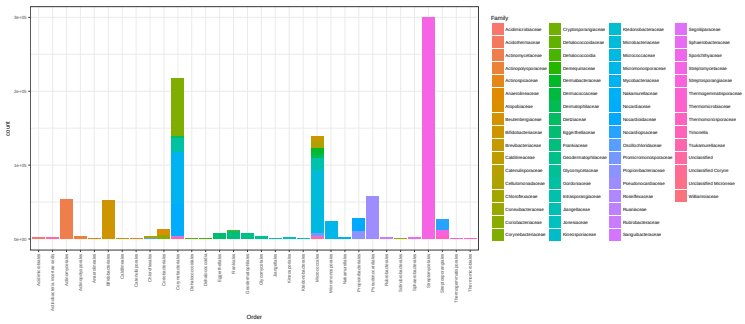


Figure 34: Actinos Diversity

Here is a reference to Recruitments after central pathways expansions colourd by taxa plot: Figure 34.

Smash diversity

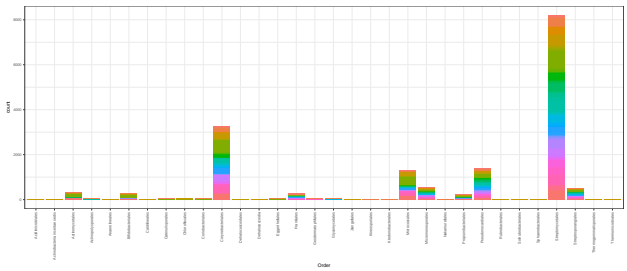


Figure 35: Actinos Smash Taxonomical Diversity

Here is a reference to Recruitments after central pathways expansions colourd by taxa plot: Figure 35.

AntisSMASH 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

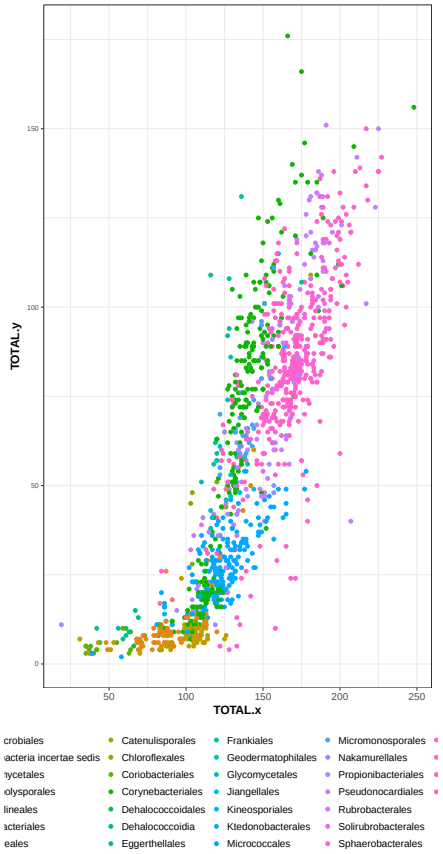


Figure 36: Correlation between Actinos central pathway expansions and antismash Natural products detection

Here is a reference to the expansions vs antismash NP's clusters plot: Figure 36.

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

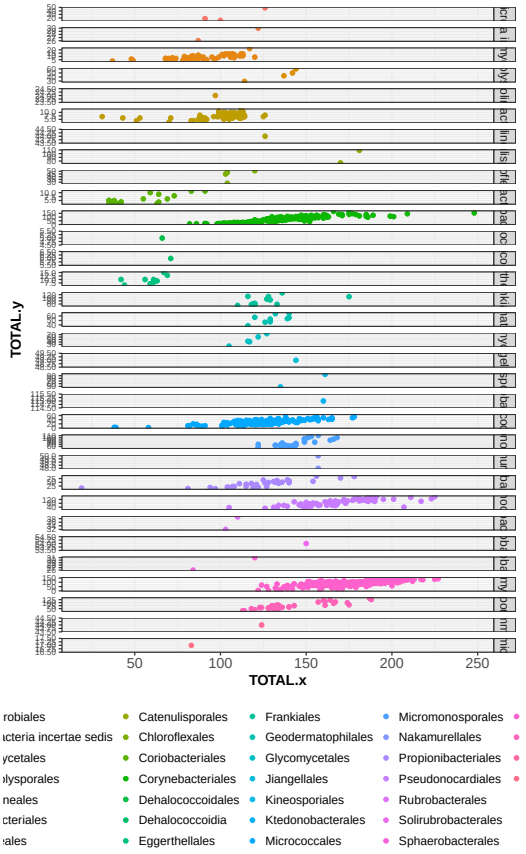


Figure 37: Correlation between Actinos central pathway expnasions and antismash Natural products detection

Here is a reference to the expansions vs antismash NP's clusters splitted by order plot Figure 37.

AntisMAsh vs Expansions by taxonomic Family

Natural products colured by family

Figure 38: Actinos Natural products by family

Here is a reference to the Natural products colured by family plot Figure 38.

Selected trees from EvoMining

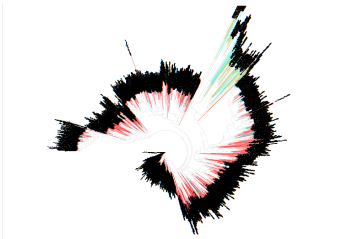


Figure 39: Enolase EvoMiningtree

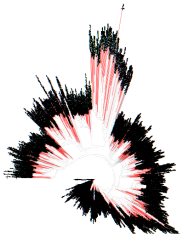


Figure 40: Phosphoribosyl isomerase EvoMiningtree

Figure 41: Phosphoribosyl isomerase A EvoMiningtree

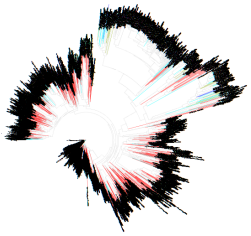


Figure 42: phosphoshikimate carboxyvinyltransferase EvoMiningtree

Cyanobacteria

Cyanobacteria phylum {Referencia}
Cyanobacteria is a photosynthetic phylum that inhabits a broad range of habitats. The broad adaptive potential is on part driven by gene-family enlargment [@larsson_genome_2011] by the analysis of 58 Cyanobacterial genomes concludes ancestor of cyanobacteria had a genome size of approx. 4.5 Mbp. Cyanobacteria produces natural products as pigments and toxins [@whitton_ecology_2012] Example of a PriA cluster toxins[@moustafa_origin_2009]

Fossil record situates Cyanobacteria [@whitton_ecology_2012] Molecular record and metabolic propoerties at [@battistuzzi_genomic_2004]

Tables

Table 5: Families on Cyanobacteria

Factors	Correlation between Parents & Child
GenomeDB	1245
Families	65

<!– clearpage ends the page, and also dumps out all floats.Floats are things like tables and figures. –>

Expansions BoxPlot by metabolic family

```
label(path = "chapter2/Cyanobacteria/expansion_plotCyanos.pdf", caption = "Expansions Boxplot",label =
```

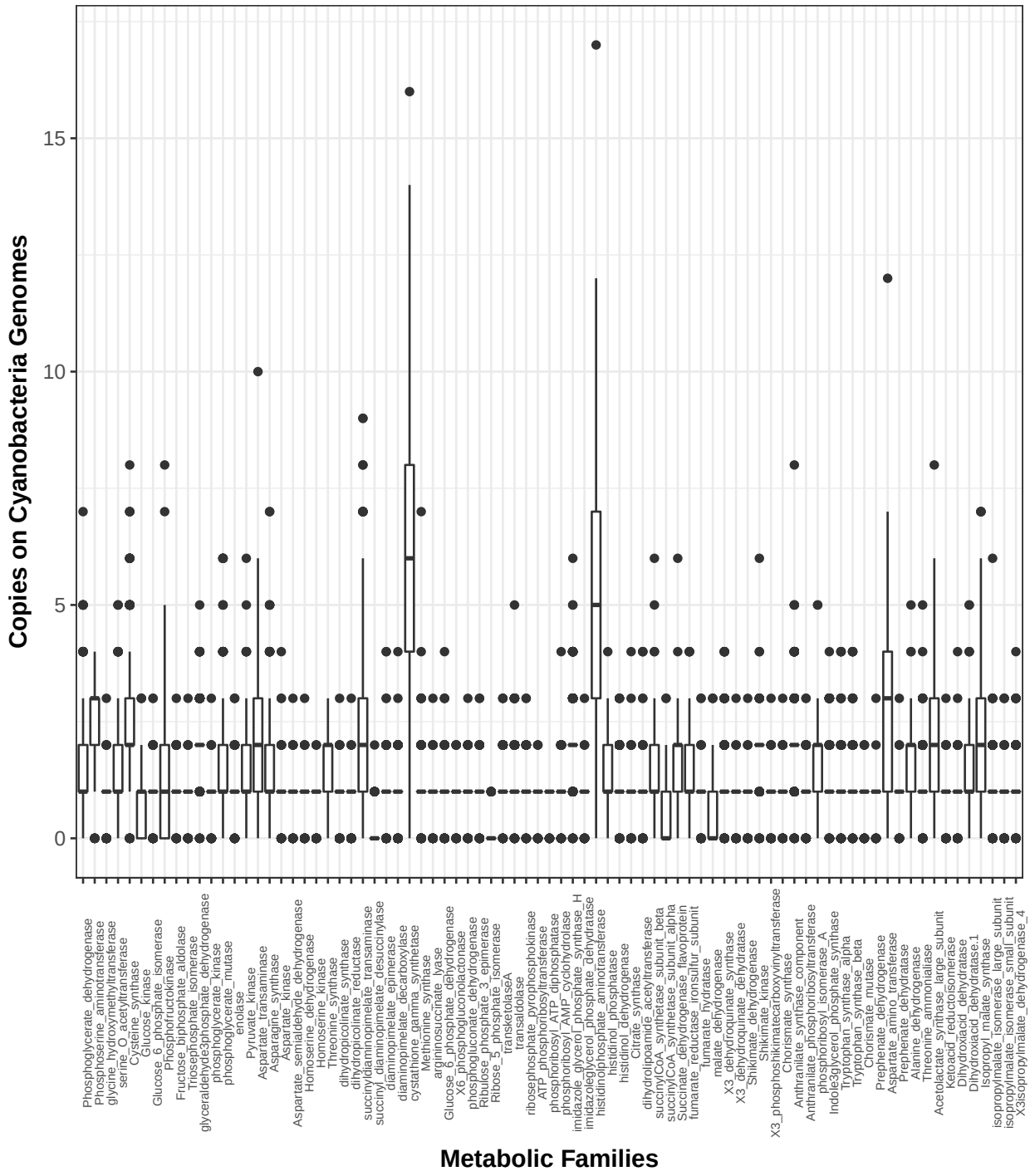


Figure 43: Expansions Boxplot

Here is a reference to the expansion boxplot: Figure 43.

Central pathway expansions

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

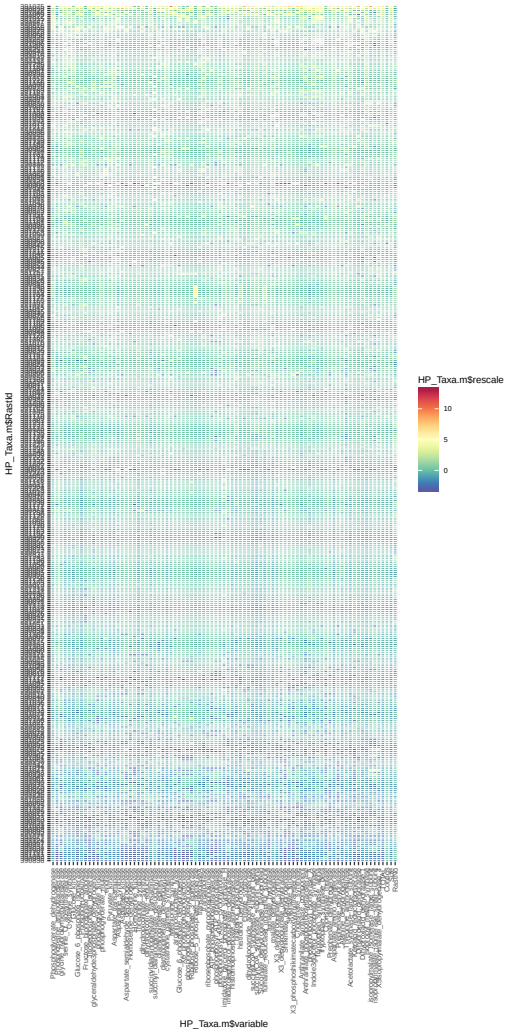


Figure 44: Cyanobacterial Heatplot

Here is a reference to the HeatPlot: Figure 44.

Genome Size correlations

Correlation between genome size and AntiSMASH products

Genome size vs Total antismash cluster coloured by order

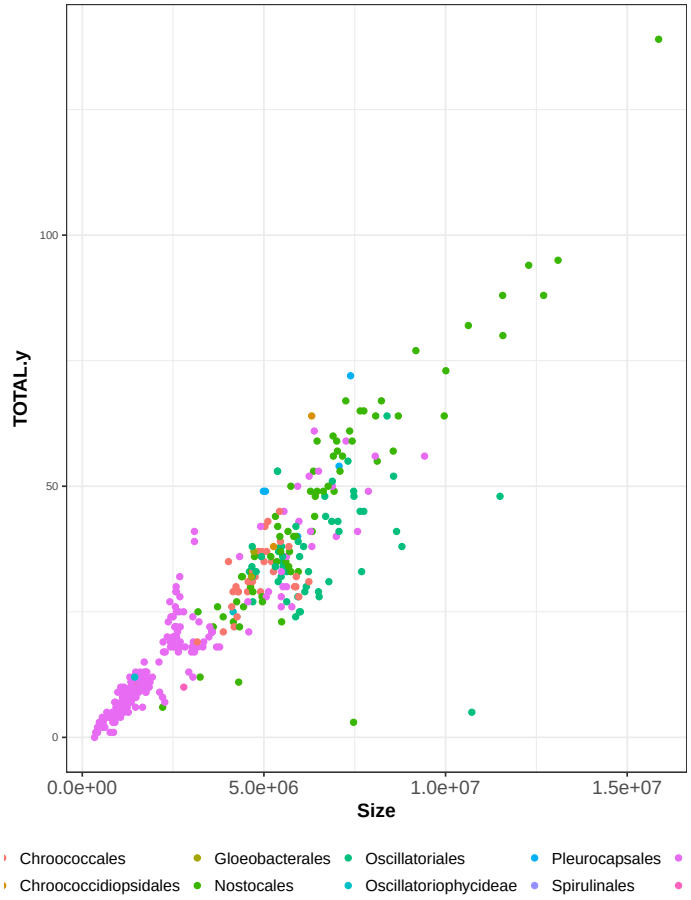


Figure 45: Correlation between genome size and antismash Natural products detection colored by Order

Here is a reference to Genome size vs Total antismash cluster: Figure 45.

Genome size vs Total antismash cluster detected splitted by order

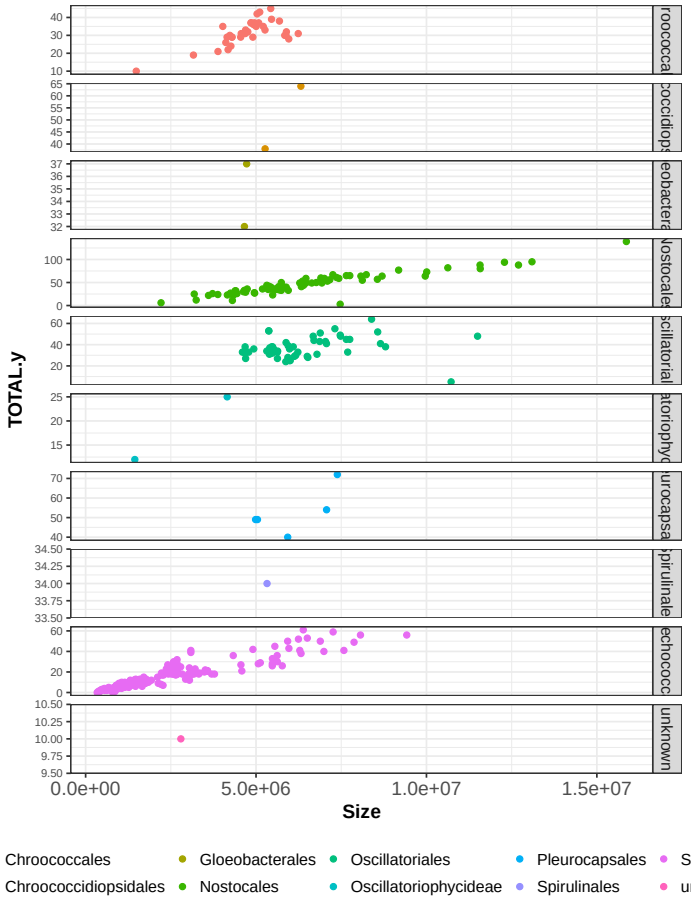


Figure 46: Correlation between genome size and antismash Natural products detection grided by Order

Here is a reference to Correlation between genome size and antismash Natural products detection grided by Order plot: Figure 46.

Correlation between genome size and Central pathway expansions

Genome size vs Total central pathway expansion coloured by order

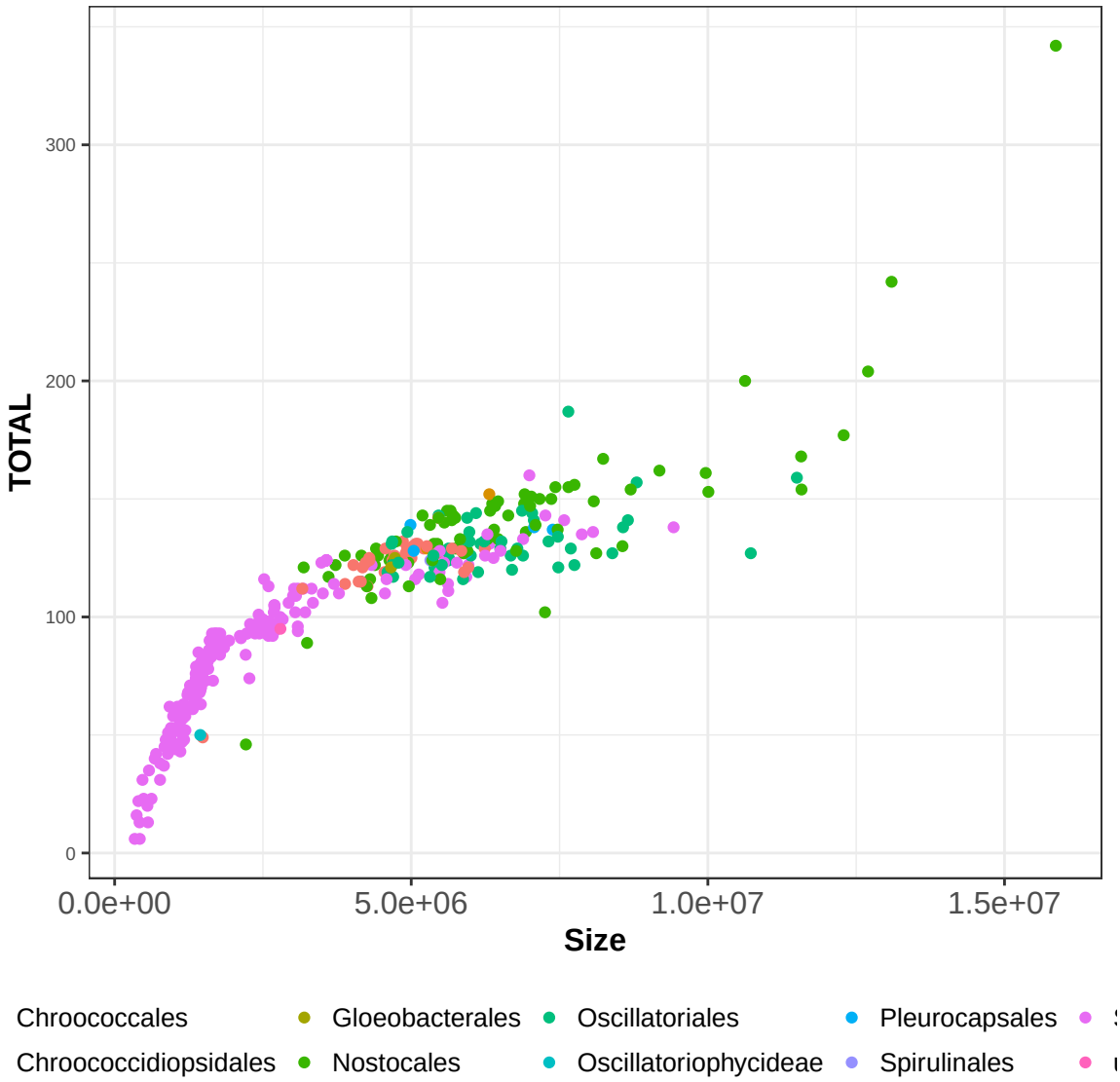


Figure 47: Correlation between genome size and central pathway expansions

Here is a reference to the size vs Total central pathway expansion plot: Figure 47.

Genome size vs Total central pathway expansion grided by order

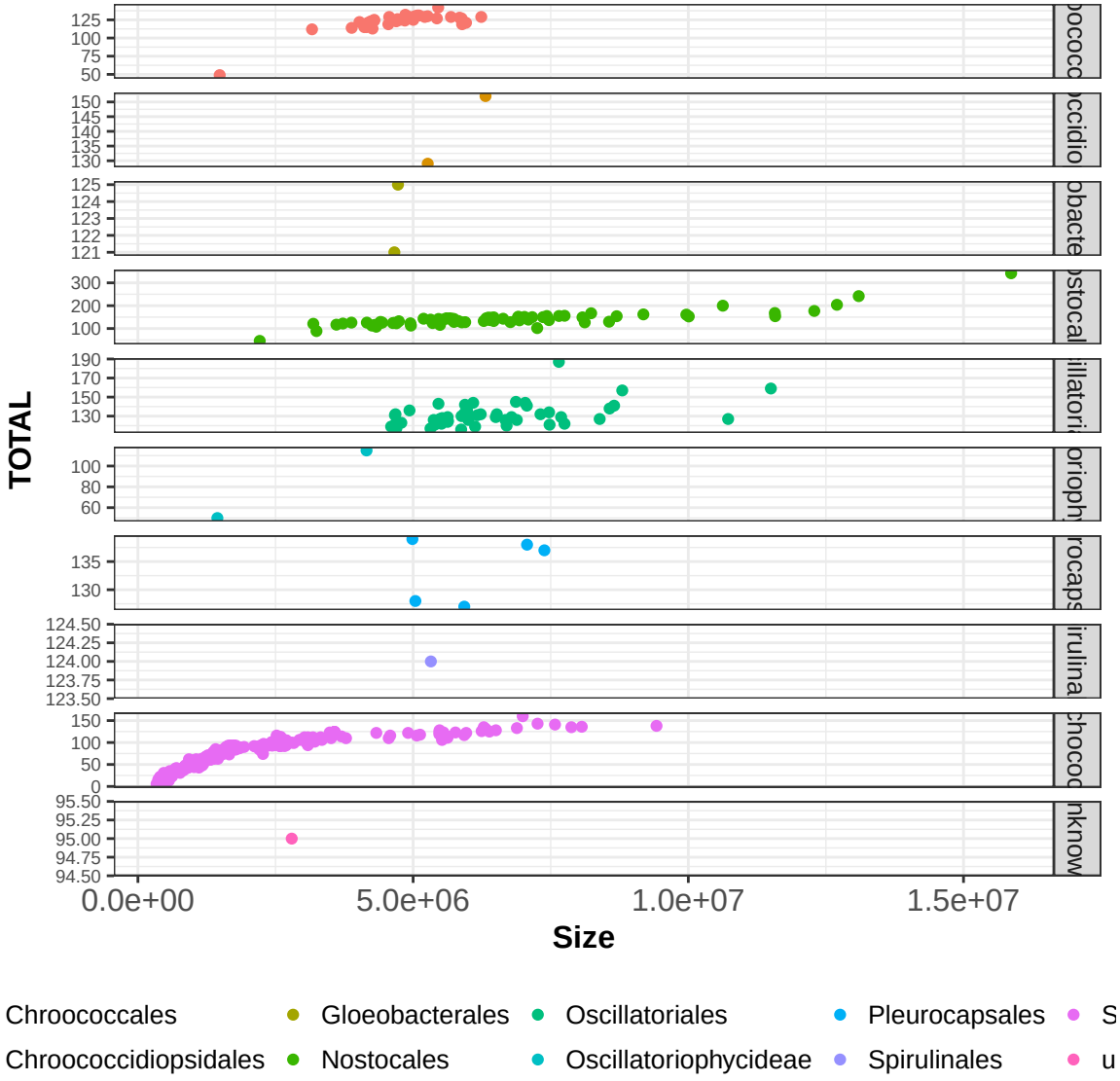


Figure 48: Correlation between genome size and central pathway expansions grided by order

Here is a reference to the Genome size vs Total central pathway expansion grided by order plot: Figure 48.

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 differente 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
## 10. Consider specifying shapes manually if you must have them.
```

```
## Warning: Removed 20418 rows containing missing values (geom_point).
```

Genome size vs Total central pathway expansion coloured by metabolic Family

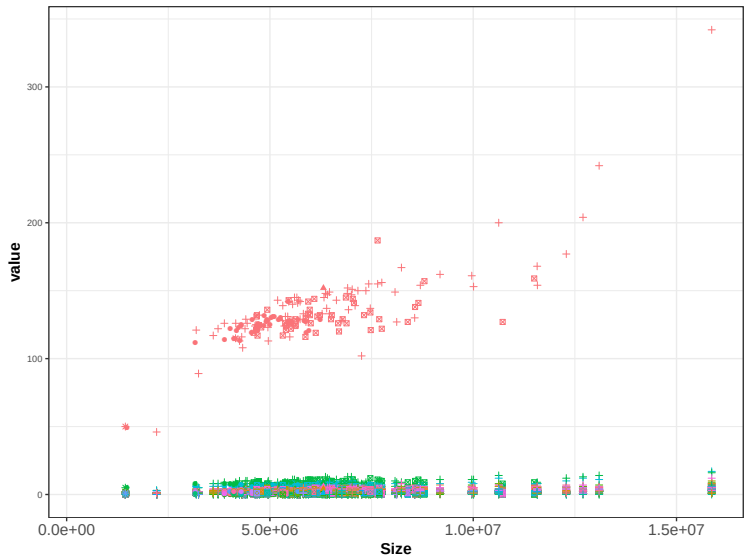


Figure 49: Correlation between Genome size vs Total central pathway expansion coloured by metabolic Family

Here is a reference to the Genome size vs Total central pathway expansion coloured by metabolic Family plot: Figure 49.

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

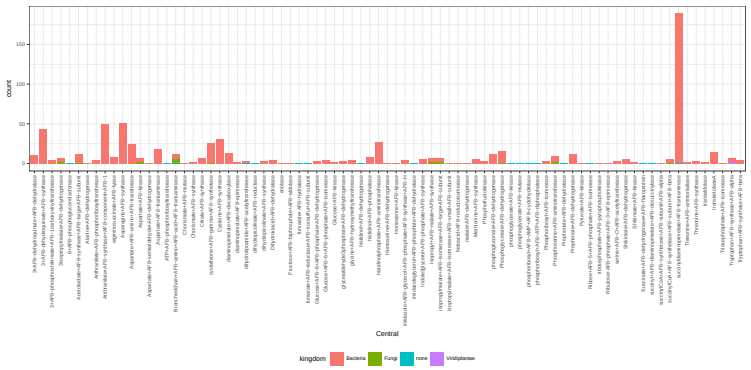


Figure 50: Recruitmens on central families coloured by kingdom

Here is a reference to Recruitments after central pathways expansions colourd by Kingdom plot: Figure 50.

Recruitments after central pathways expansions colourd by taxonomy

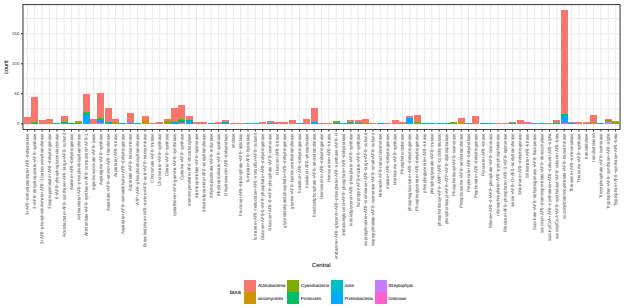


Figure 51: Recruitmens on central families coloured by taxonomy

Here is a reference to Recruitments after central pathways expansions colourd by taxa plot: Figure 51.

Cyanobacterias AntiSMASH

Taxonomical diversity on Cyanobacteria Data

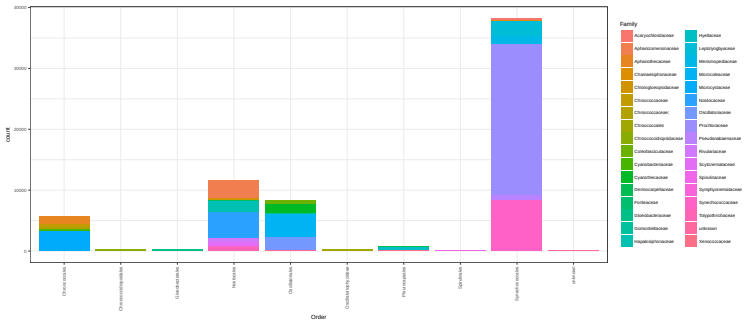


Figure 52: Diversity

Here is a reference to Recruitments after central pathways expansions colourd by taxa plot: Figure 52.

Smash diversity

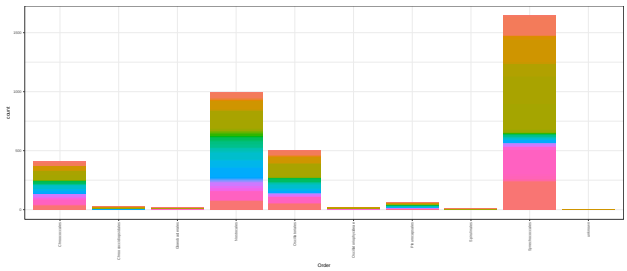


Figure 53: Smash

Here is a reference to Recruitments after central pathways expansions colourd by taxa plot: **??**.

AntisSMASH 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

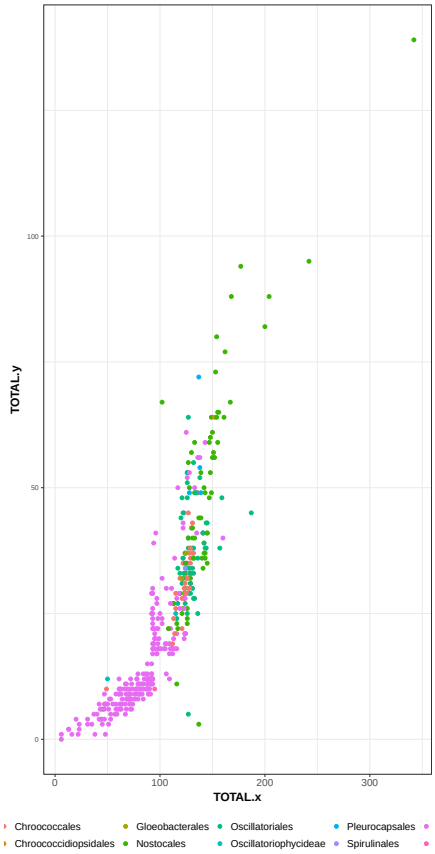


Figure 54: Correlation between central pathway axpnasions and antismash Natural products detection

Here is a reference to the expansions vs antismash NP's clusters plot: Figure 54.

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

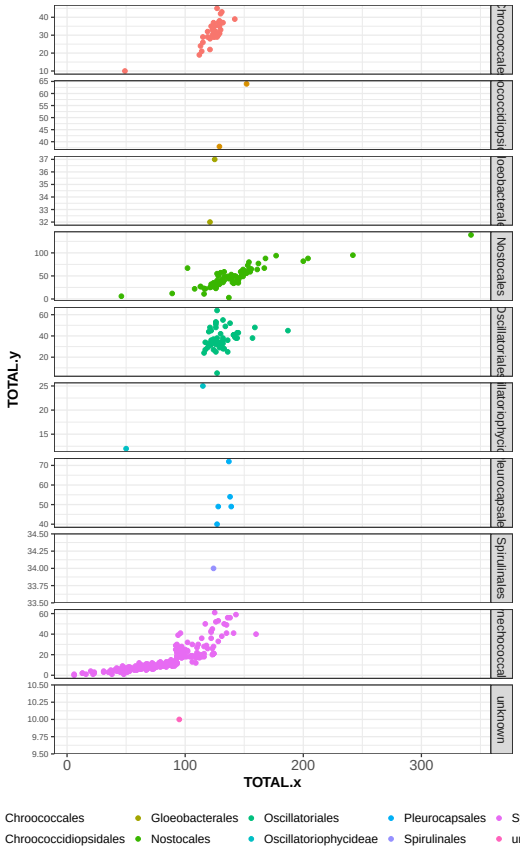


Figure 55: Correlation between central pathway axpnasions and antismash Natural products detection

Here is a reference to the expansions vs antismash NP's clusters splitted by order plot **??**.

AntisMAsh vs Expansions by taxonomic Family

Natural products colured by family

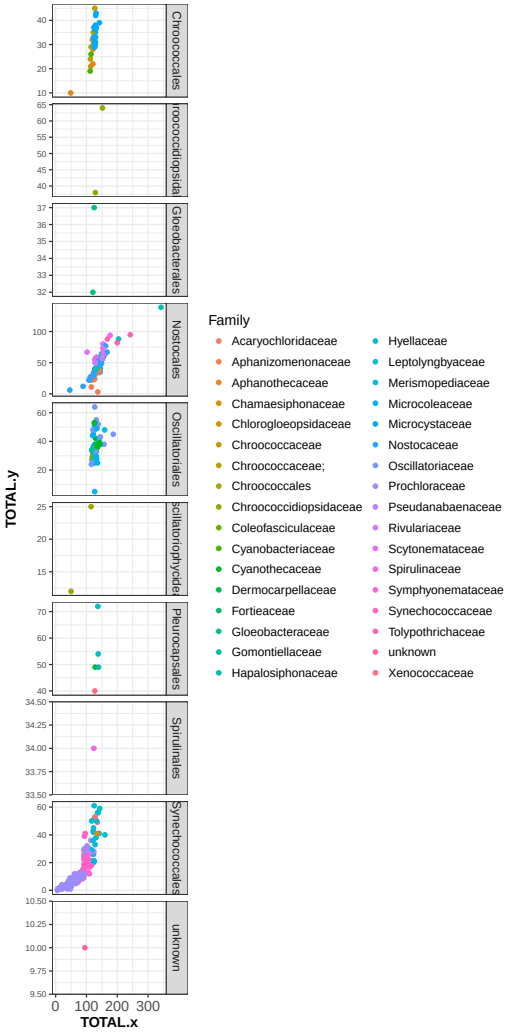


Figure 56: Natural products by family

Here is a reference to the Natural products colured by family plot Figure 56.

Selected trees from EvoMining

Phosphoribosyl_isomerase_3 family
Figure from EvoMining

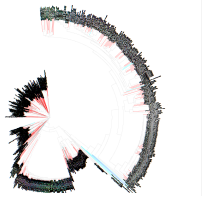


Figure 57: Phosphoribosyl isomerase EvoMiningtree

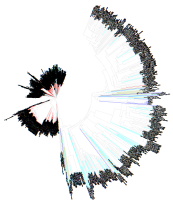


Figure 58: Phosphoglycerate dehydrogenase EvoMiningtree

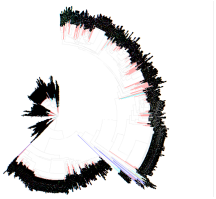


Figure 59: Phosphoserine aminotransferase EvoMiningtree

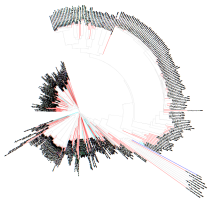


Figure 60: Triosephosphate isomerase EvoMiningtree

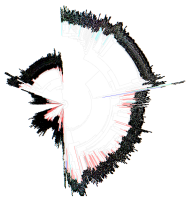


Figure 61: glyceraldehyde3phosphate dehydrogenase EvoMiningtree

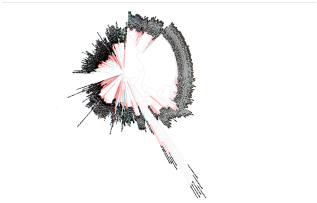


Figure 62: phosphoglycerate kinase EvoMiningtree

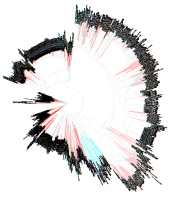


Figure 63: phosphoglycerate mutaseEvoMiningtree

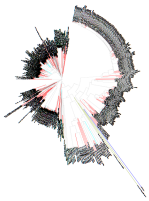


Figure 64: enolase EvoMiningtree

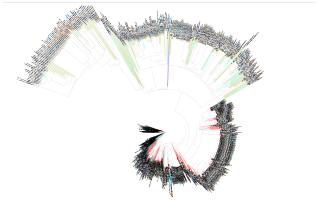


Figure 65: Pyruvate kinase EvoMiningtree

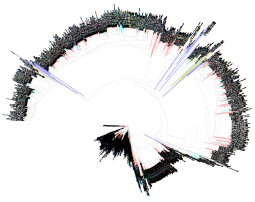


Figure 66: Aspartate transaminase EvoMiningtree

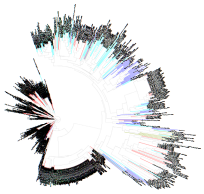


Figure 67: Asparagine synthase EvoMiningtree

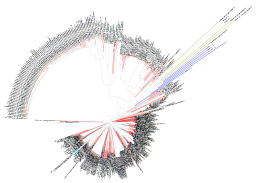


Figure 68: Aspartate kinase EvoMiningtree

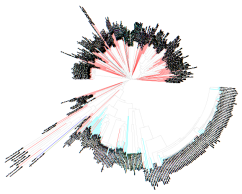


Figure 69: Aspartate semialdehyde dehydrogenase EvoMiningtree

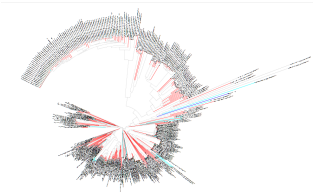


Figure 70: Homoserine dehydrogenase EvoMiningtree