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 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 1: Families on Archaeabacteria

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

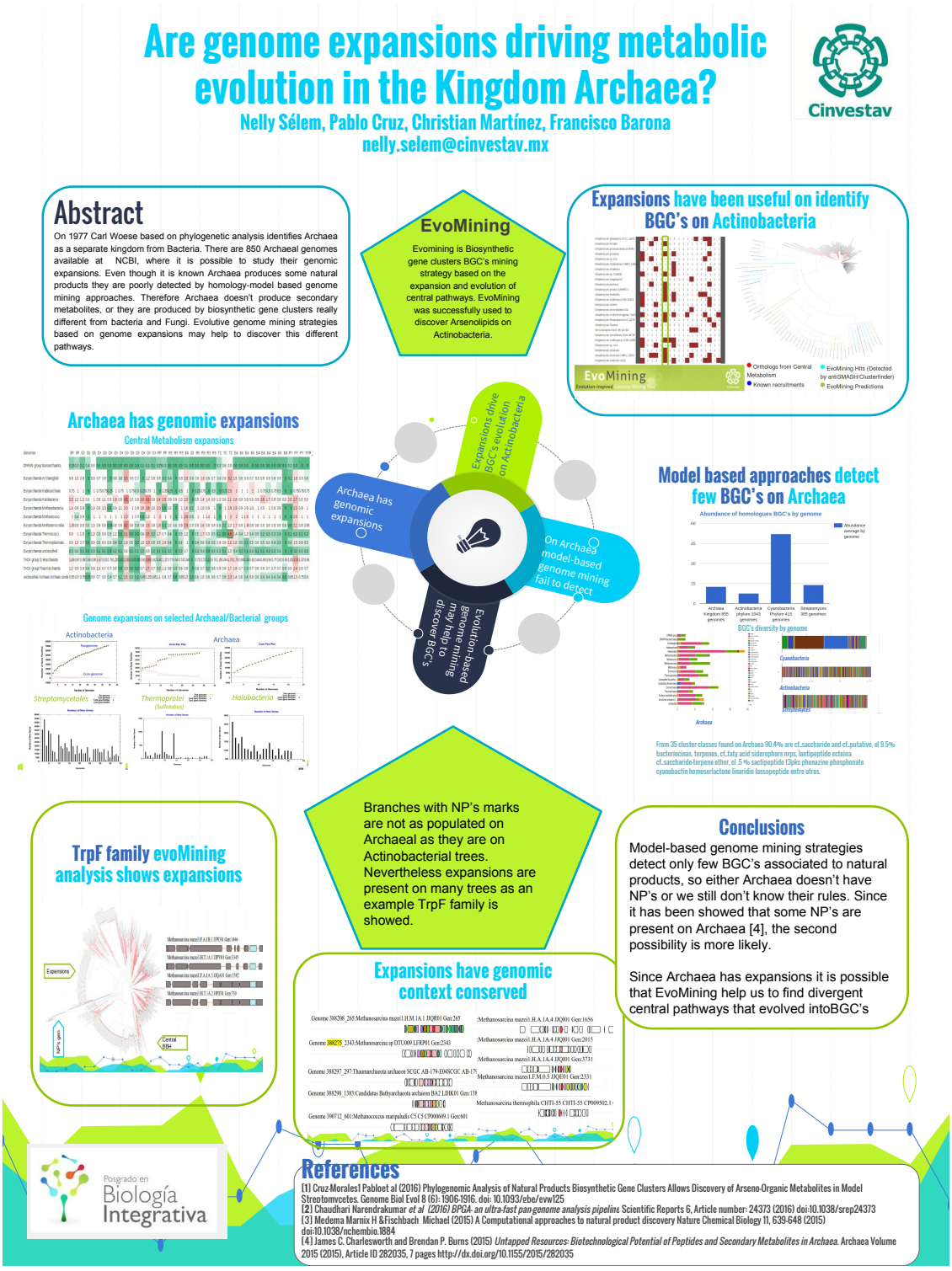


Figure 1: EvoMining Archaeas

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 Actinobacteria evoMining central pathways.

Expansions BoxPlot by metabolic family

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

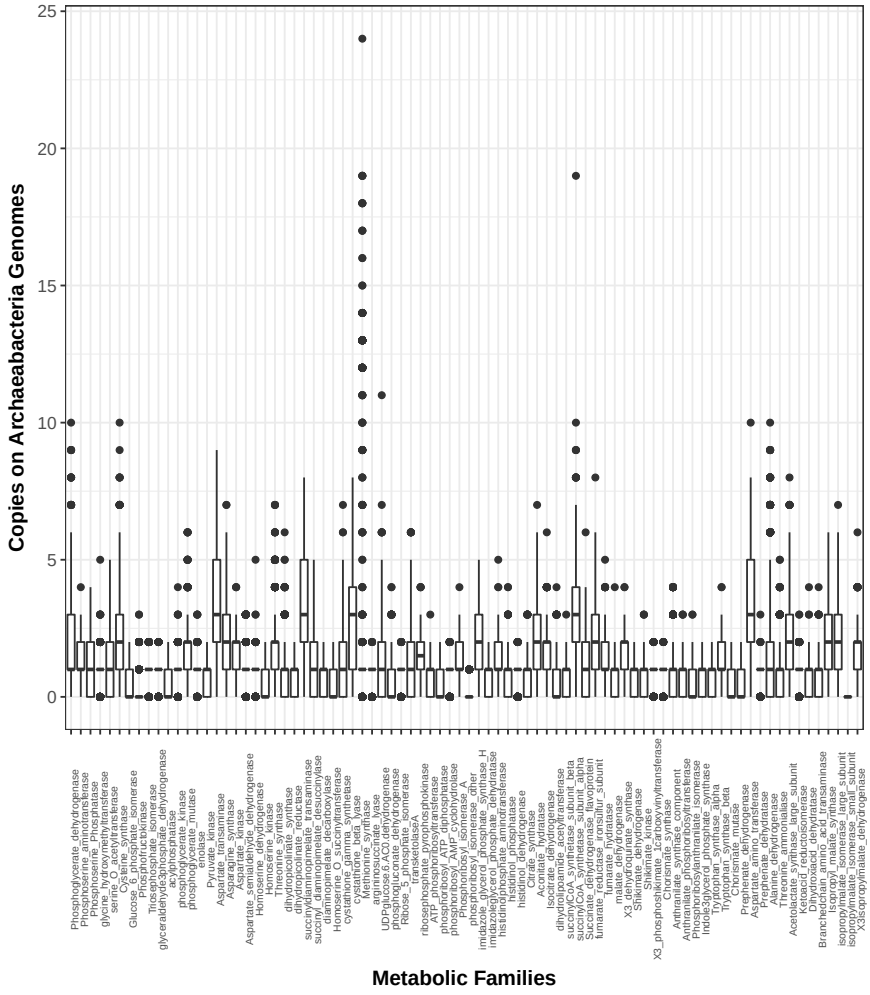
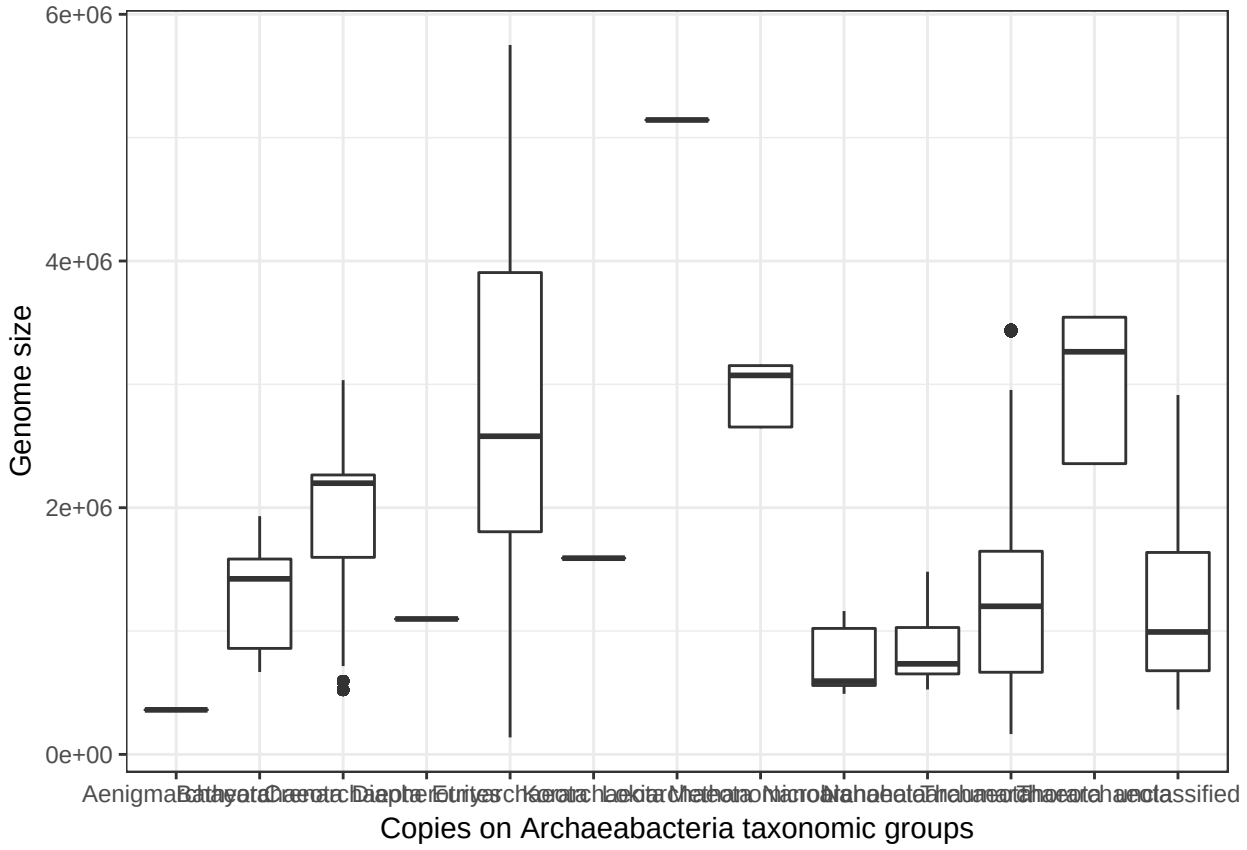


Figure 2: Expansions Boxplot

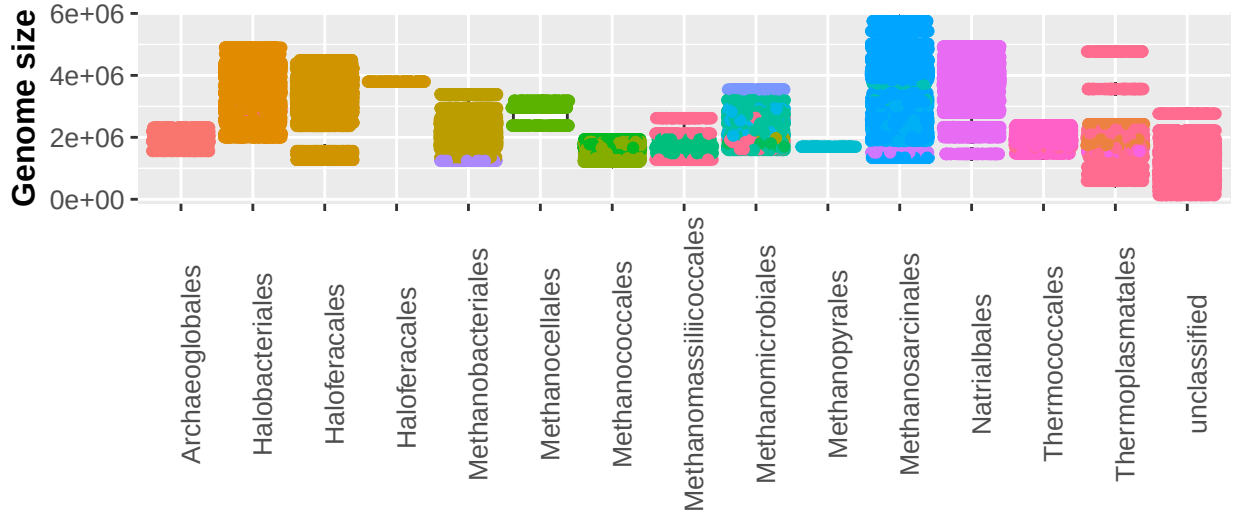
Here is a reference to the expansion boxplot: Figure 2.

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 expecte
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
```



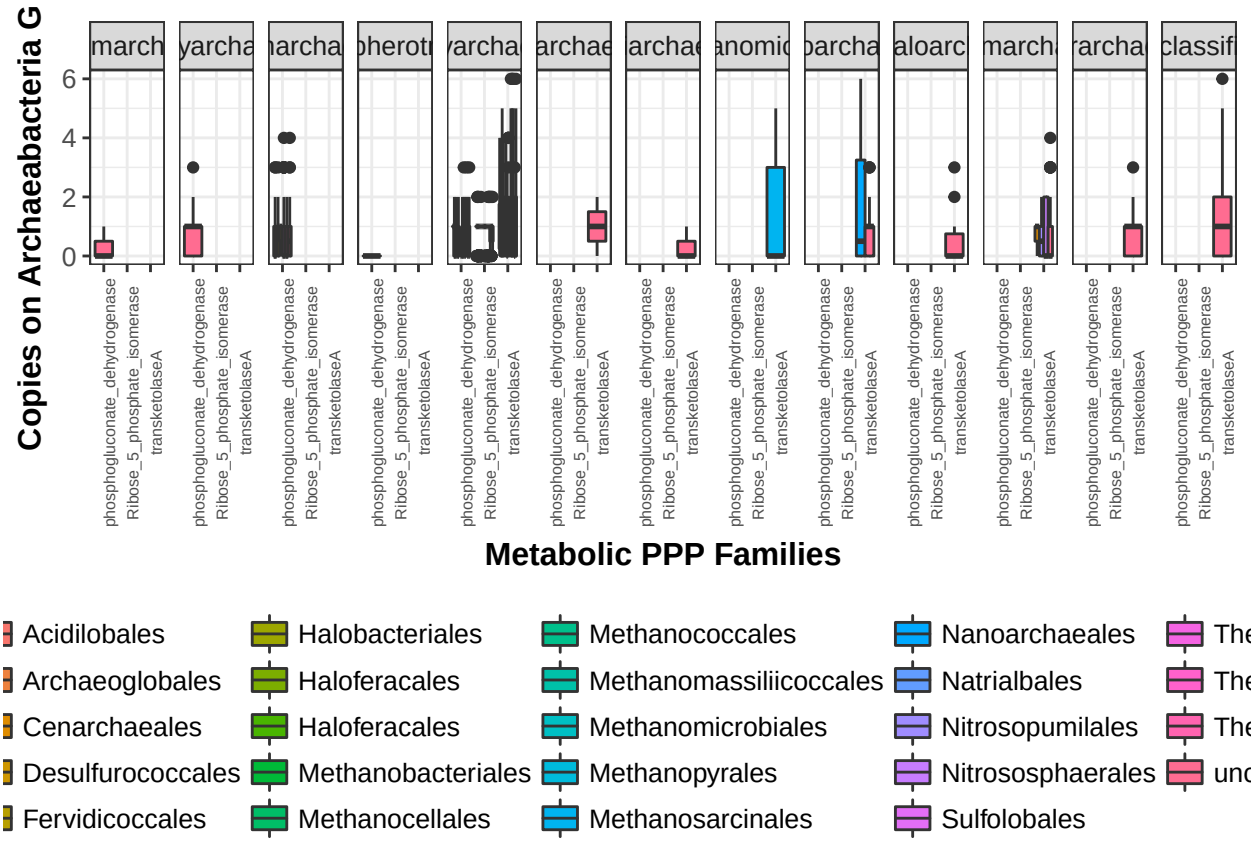
```
#+ geom_jitter(aes(color=ArchaeasHeatPlotBP.m$Phylum))


## 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
```



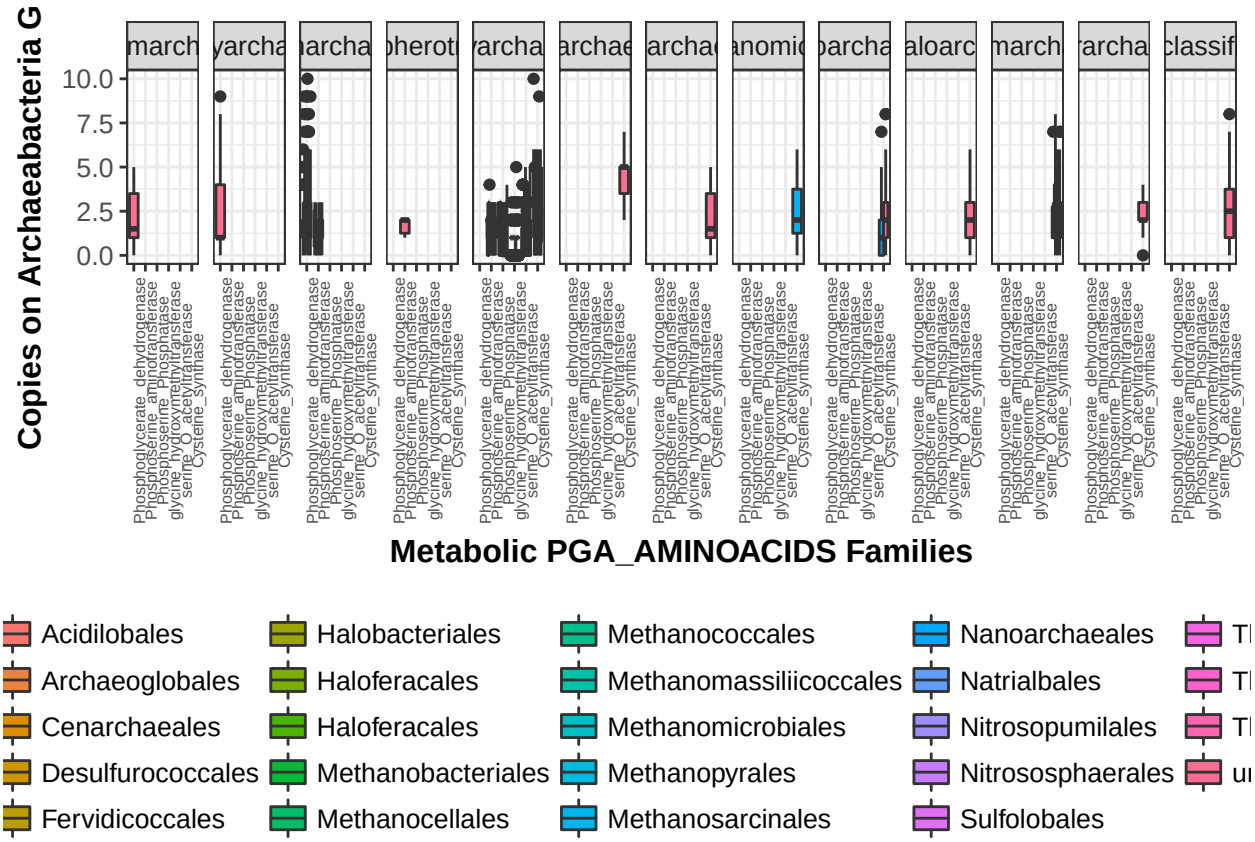
Halobacteria orders size taxonomic groups

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

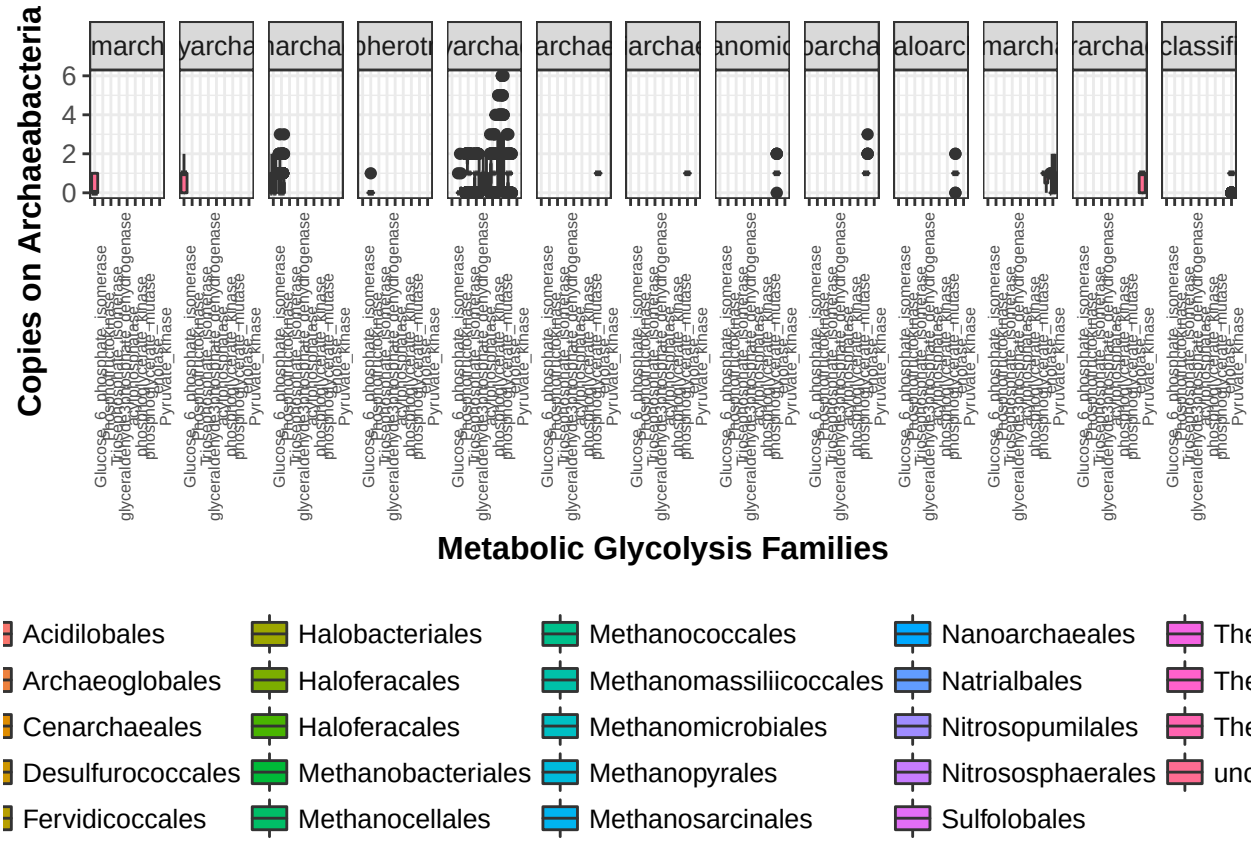
```
#MetFam_BP.m=subset(ArchaeasHeatPlotBP.m,Family=="Methanosarcinaceae")
#ggplot(MetFam_BP.m, aes(x=MetFam_BP.m$Size, y=MetFam_BP.m$value))
#+theme(plot.title = element_text(size = 14, face = "bold"), text = element_text(size = 12), axis.title


  #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
```



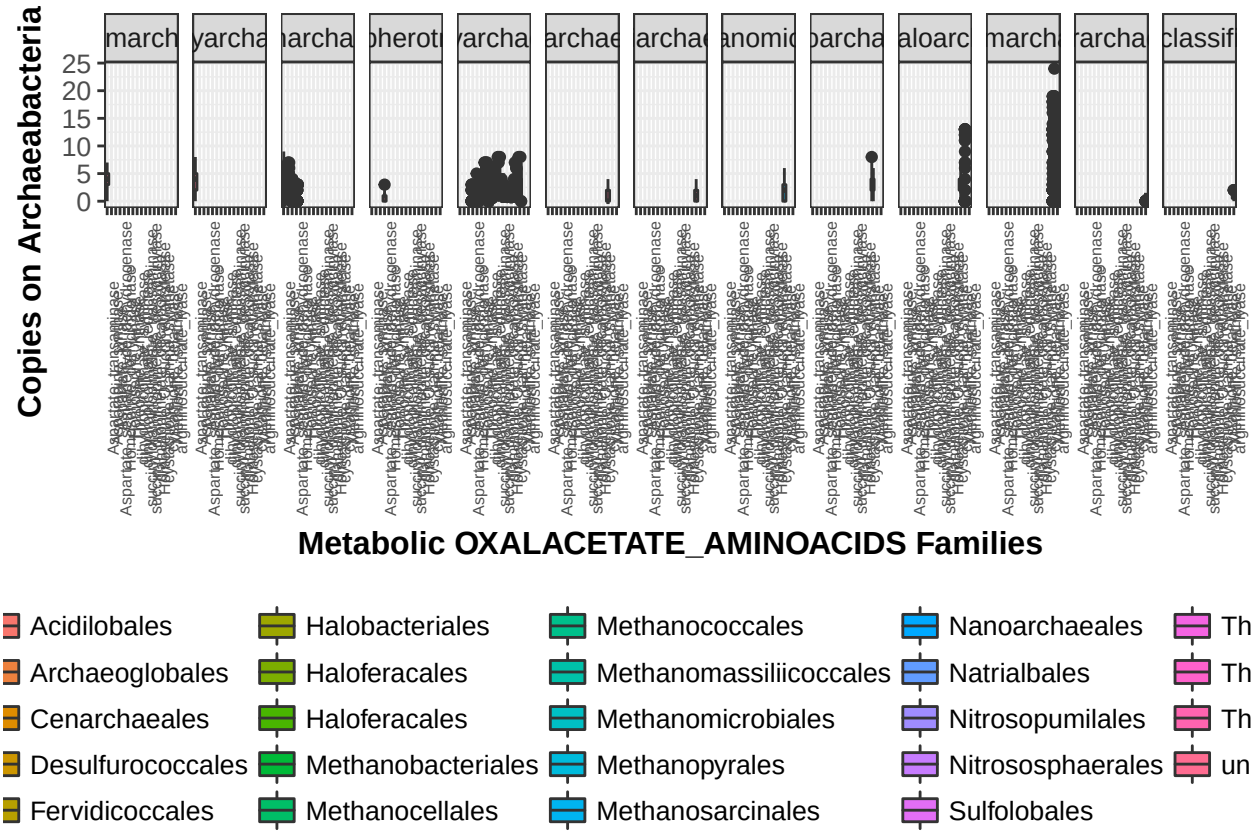

```
#+ geom_jitter(aes(color=MetFam_BP.m$Phylum))

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
```



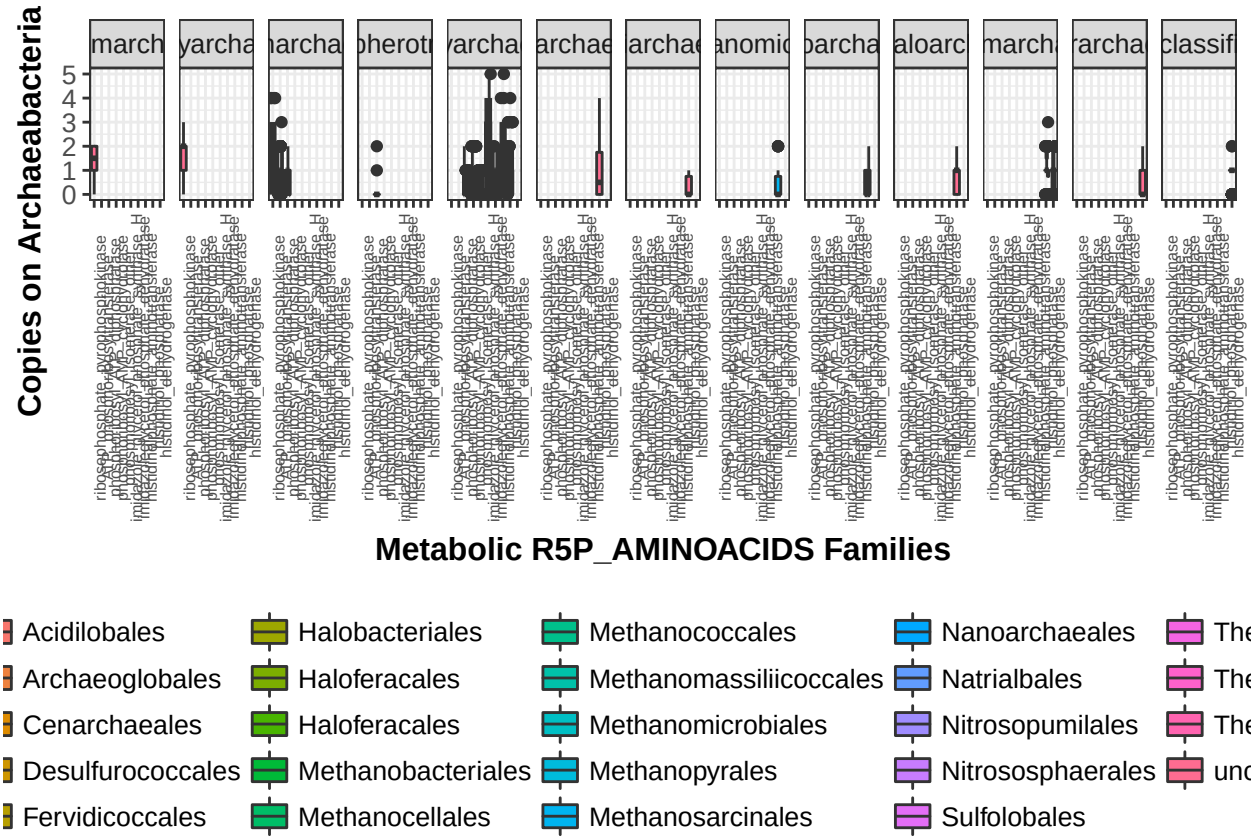

```
#+ geom_jitter(aes(color=MetFam_BP.m$Phylum))

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
```



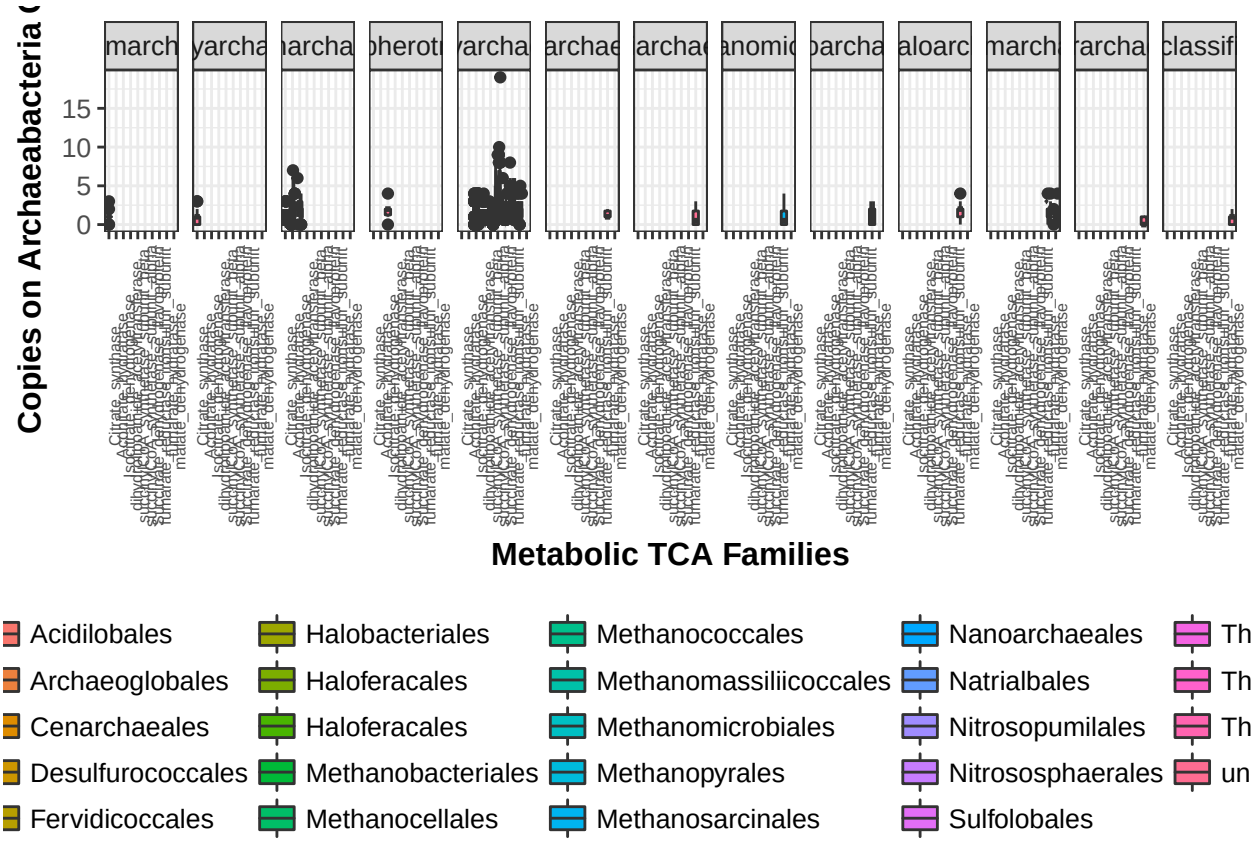

```
#+ geom_jitter(aes(color=MetFam_BP.m$Phylum))

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
```



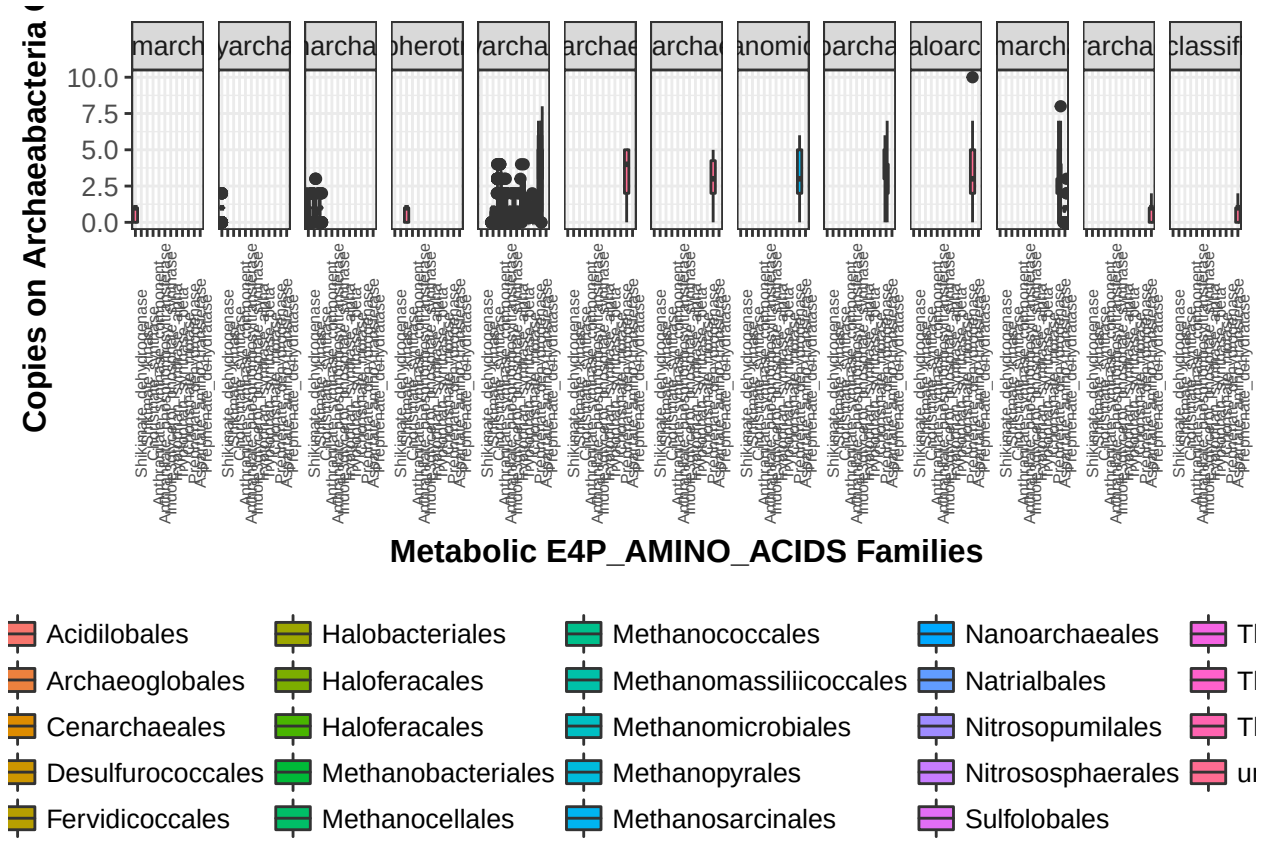

```
#+ geom_jitter(aes(color=MetFam_BP.m$Phylum))

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
```



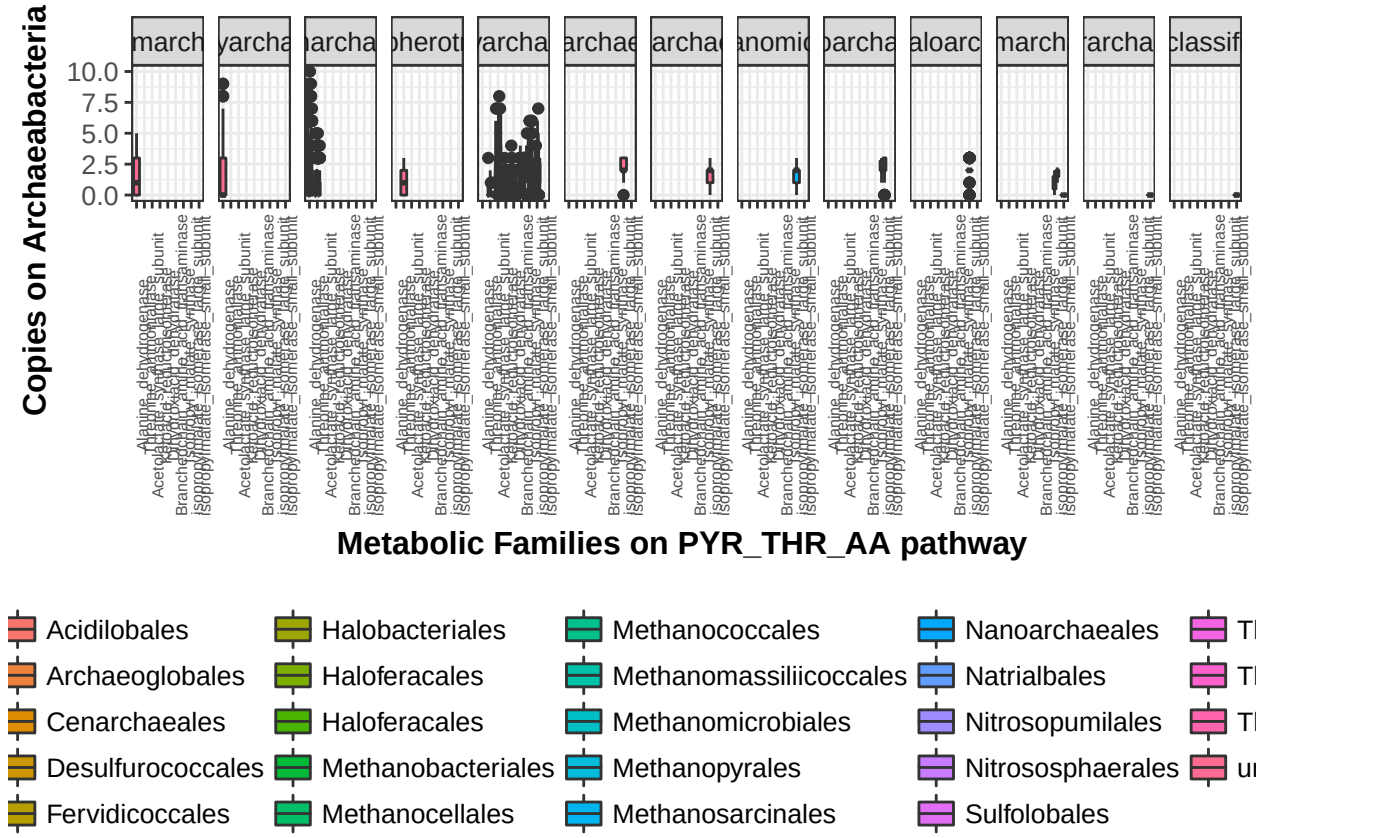

```
 #+ geom_jitter(aes(color=MetFam_BP.m$Phylum))

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

Central pathway expansions

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

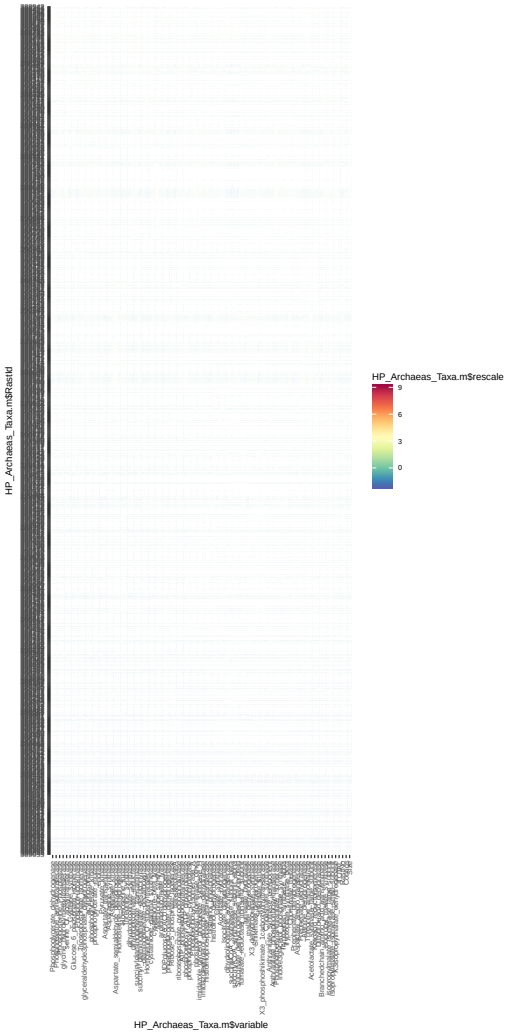


Figure 3: Archaeas Heatplot

Here is a reference to the HeatPlot: Figure 3.

Genome Size correlations

Correlation between genome size and AntiSMASH products

Genome size vs Total antismash cluster coloured by order

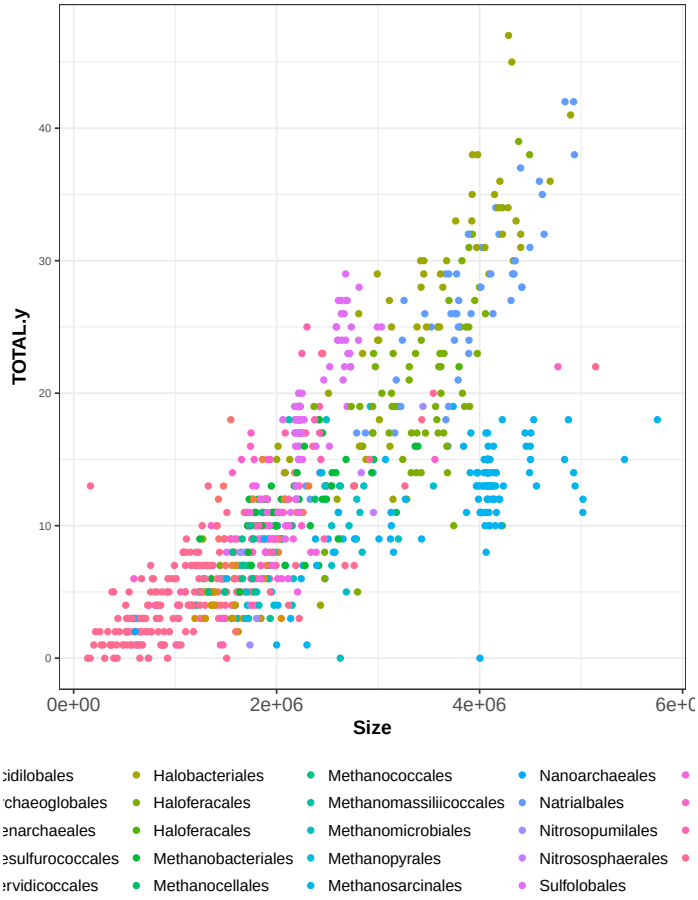


Figure 4: Correlation between Archaeas genome size and antismash Natural products detection colored by Order

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

Genome size vs Total antismash cluster detected splitted by order

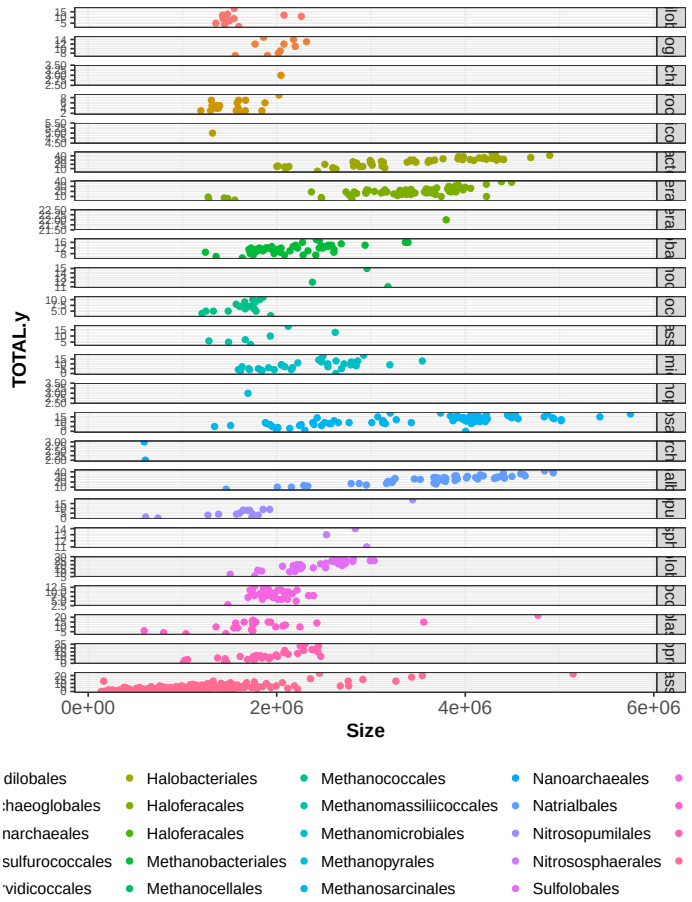


Figure 5: Correlation between Archaeas genome size and antismash Natural products detection grided by Order

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

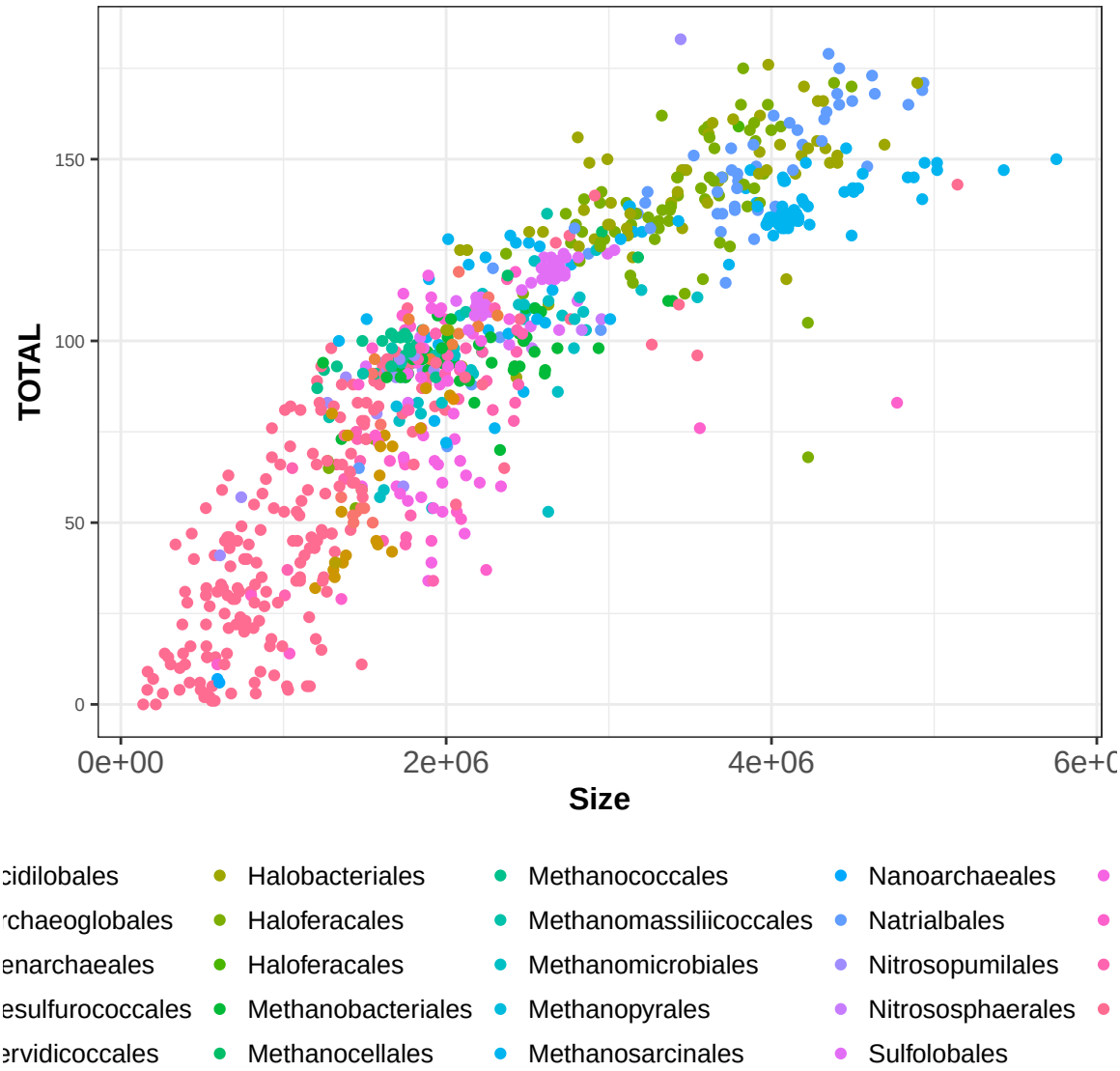


Figure 6: Correlation between Archaeas genome size and central pathway expansions

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

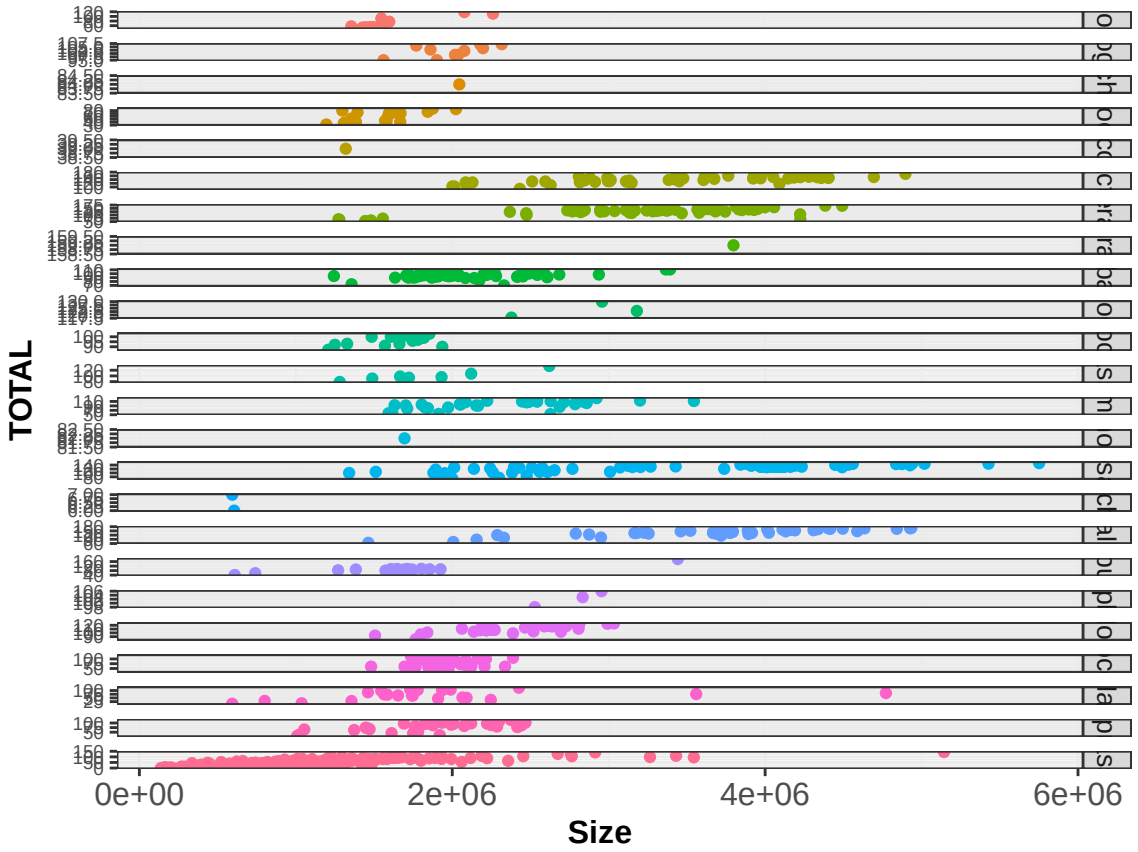


Figure 7: Correlation between Archaeas genome size and central pathway expansions grided by order

Here is a reference to the Genome size vs Total central pathway expansion grided by order plot: Figure 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 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 64823 rows containing missing values (geom_point).
```

Genome size vs Total central pathway expansion coloured by metabolic Family

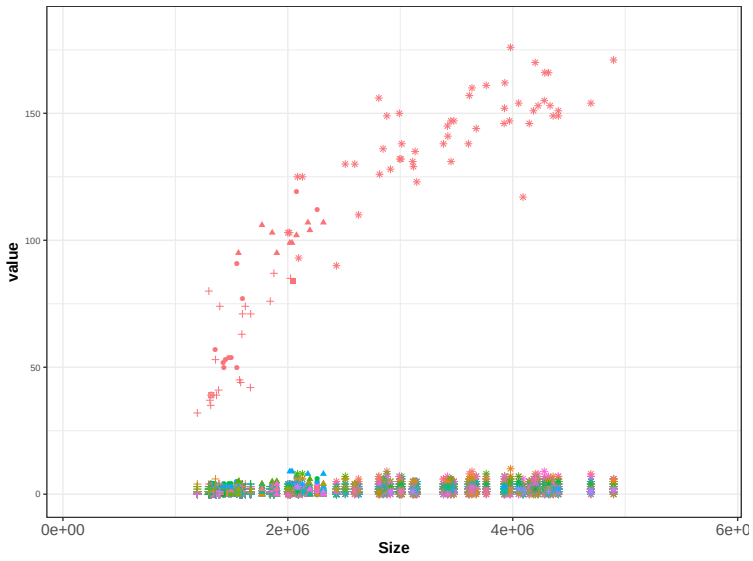


Figure 8: Correlation between Archaeas Genome size vs Total central pathway expansion coloured by metabolic Family

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

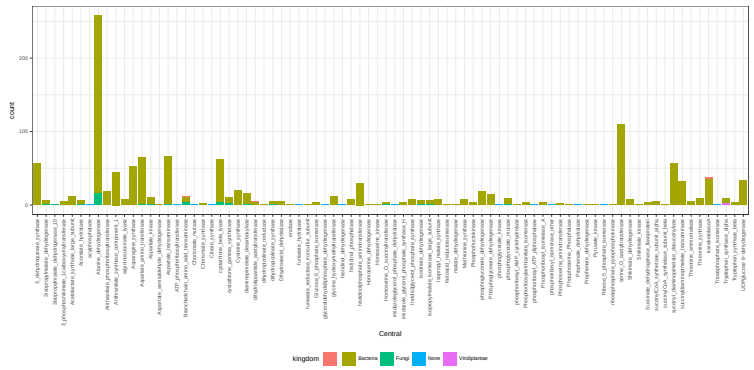


Figure 9: Archaeas Recruitmens on central families coloured by kingdom

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

Recruitments after central pathways expansions colourd by taxonomy

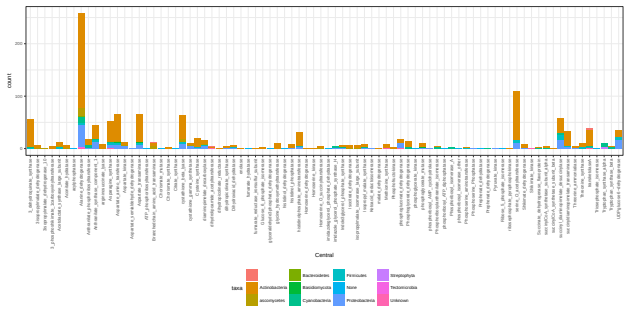


Figure 10: Archaeas Recruitmens on central families coloured by taxonomy

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

Archaeas AntiSMASH

Taxonomical diversity on Archaeasbacteria Data

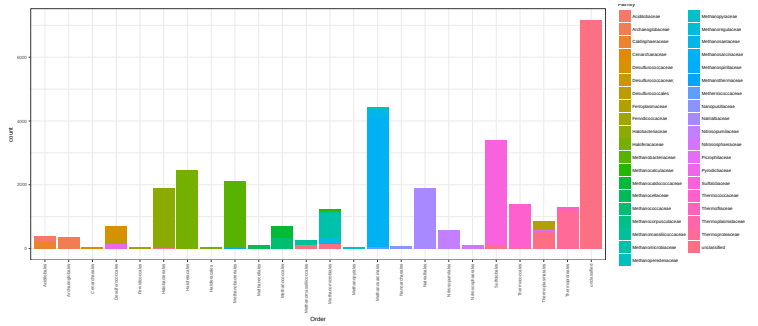


Figure 11: Archaeas Diversity

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

Smash diversity

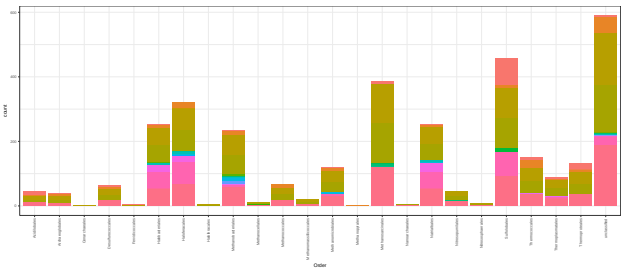


Figure 12: Archaeas Smash Taxonomical Diversity

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

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 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 antismash cluster detected splitted by order

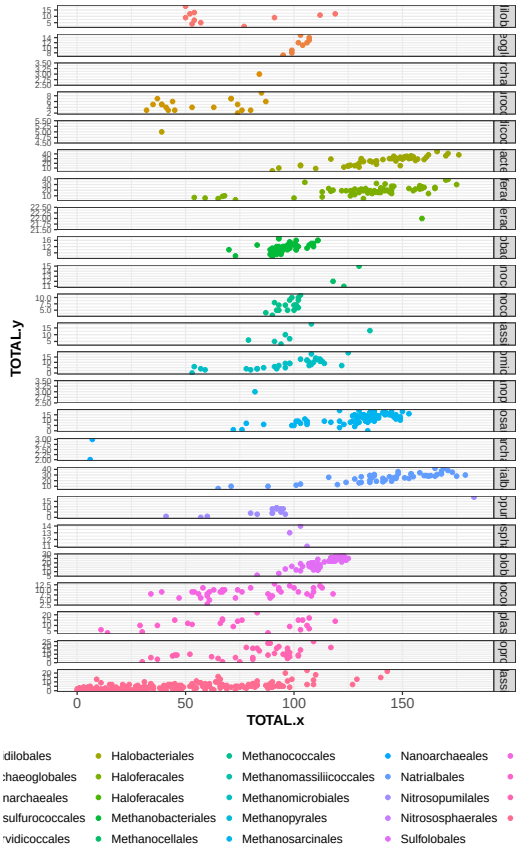


Figure 14: Correlation between Archaeas central pathway expnasions and antismash Natural products detection

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

AntisMAsh vs Expansions by taxonomic Family

Natural products colured by family

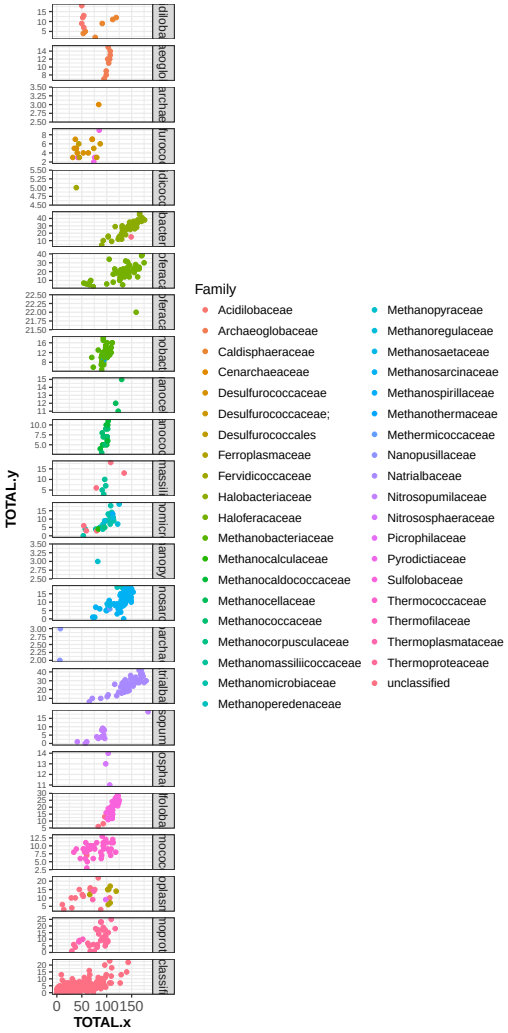


Figure 15: Archaeas Natural products by family

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

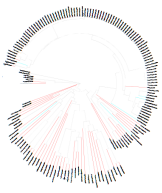


Figure 17: Phosphoribosyl isomerase other EvoMiningtree

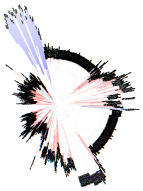


Figure 18: Phosphoribosyl anthranilate isomerase EvoMiningtree

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. [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.